Spatial data infrastructures for environmental sciences

GIULIANI, Gregory

Abstract
Today we are living in a globalized world with rapidly evolving processes including climate and land cover change, population growth that are impacting the environment. In parallel, means of communication have expanded to take on a remarkable place in our society, allowing us to access an enormous and continuous flow of information. The underlying hypothesis of this thesis is that making data interoperable and providing access to computing resources can potentially allow data users to spend more time in data analysis than in data discovery, and enabling more people to benefit from using geospatial data on the environment. Our research has shown that SDIs and related concepts, methods and technologies are suitable and can bring major benefits to support and facilitate environmental data discovery, accessibility, visualization, dissemination and analysis. Moreover, facilitating access, integration and use of geospatial data can answer the requirements of a specific community as well as making these data available to the widest possible audience. The challenges that humankind is facing require acting now and we need to [...]
Spatial Data Infrastructures for Environmental Sciences

THOSE

présentée à la Faculté des Sciences de l'Université de Genève
pour obtenir le grade de Docteur ès Sciences, mention Sciences de l'Environnement

par

Grégory GIULIANI
De
Vernier (Genève)

Thèse N° 4348

GENEVE
Atelier d'impression ReproMail
2011
This work is dedicated to my family

“Information is the seed for an idea, and only grows when it’s watered.”
(Heinz V. Bergen)
Table of contents
# Table of Contents

## Abstract

ABSTRACT  9  RÉSUMÉ  10  RIASSUNTO  13

## Acknowledgments

ACKNOWLEDGMENTS  20  REMERCIEMENTS  23  RINGRAZIAMENTI  25

## Chapter 1: Introduction

1.1 Structure of the Thesis  28  1.2 Projects  29  1.3 List of Contributing Research Papers  30  1.4 Background: Responding to Our Changing Environment  31  1.5 Research Problem and Questions

1.5.1 Research Problem  33  1.5.2 Research Questions  34

## Chapter 2: Theoretical Framework

2.1 Spatial Data Infrastructure

2.1.1 Definition, Concepts and Rationale  37  2.1.2 Objectives  39  2.1.3 Components  40  2.1.4 SDI Hierarchy  42  2.1.5 SDI Evolution and (Emerging) Trends  43  2.1.6 Benefits  46

2.2 Interoperability and Standards

2.2.1 Definition and Concepts  47  2.2.2 Types of Interoperability  49  2.2.3 Interoperability Enablers  50  2.2.4 Standards  50  2.2.5 Benefits  51

2.3 Initiatives

2.3.1 Infrastructure for Spatial Information in the European Community (INSPIRE)  52  2.3.2 Global Earth Observation System of Systems (GEOSS)  55  2.3.3 United Nations Spatial Data Infrastructure (UNSDI)  56  2.3.4 Global Monitoring for the Environment and Security (GMES)  57  2.3.5 Global Spatial Data Infrastructure (GSDI)  58

2.4 Standards Organizations Relevant for GIS/SDI

2.4.1 Open Geospatial Consortium (OGC)  59  2.4.2 International Organization for Standardization (ISO)  60  2.4.3 The World Wide Web Consortium (W3C)  60  2.4.4 Organization for the Advancement of Structured Information Standards (OASIS)  61

2.5 Standards Description

2.5.1 Catalogue Service for the Web (CSW)  63  2.5.2 Web Map Service (WMS)  64  2.5.3 Web Feature Service (WFS)  67  2.5.4 Web Coverage Service (WCS)  68
Table of contents

2.5.5 **WEB PROCESSING SERVICE (WPS)** 70
2.5.6 **SENSOR OBSERVATION SERVICE (SOS)** 71
2.5.7 **ISO 19115/19139** 72
2.5.8 **ISO 19119** 72
2.5.9 **KEYHOLE MARKUP LANGUAGE (KML)** 73
2.5.10 **GEOGRAPHIC MARKUP LANGUAGE (GML)** 73

2.6 **TOOLS** 74
2.6.1 **OGC WEB SERVICES** 74
2.6.2 **METADATA EDITOR AND CATALOG SYSTEM** 77
2.6.3 **DATA STORAGE** 78
2.6.4 **WEB MAPPING** 79

CHAPTER 3: **IS THERE A NEED TO ACCESS AND PROCESS ENVIRONMENTAL DATA IN A BETTER AND EFFICIENT WAY?** 81

3.1 **PREAMBLE** 82
3.2 **SHARING ENVIRONMENTAL DATA THROUGH GEOSS** 85
3.2.1 **ABSTRACT** 85
3.2.2 **INTRODUCTION** 85
3.2.3 **THE NEED FOR DATA SHARING AND INTEGRATION** 87
3.2.4 **SERVING DATA INTO GEOSS** 89
3.2.5 **TECHNICAL COMPARISON AND COMMON GROUNDS** 92
3.2.6 **CHALLENGES AND PROMISES** 96
3.2.7 **CONCLUSIONS** 99

3.3 **SUMMARY AND LESSONS LEARNED** 101

CHAPTER 4: **HOW CAN SDI IMPROVE OUR CAPACITY TO DISCOVER, SHARE, RETRIEVE AND INTEGRATE ENVIRONMENTAL DATA?** 102

4.1 **PREAMBLE** 103
4.2 **THE PREVIEW GLOBAL RISK DATA PLATFORM: A GEOPORTAL TO SERVE AND SHARE GLOBAL DATA ON RISK TO NATURAL HAZARDS** 106
4.2.1 **ABSTRACT** 106
4.2.2 **INTRODUCTION** 106
4.2.3 **SDI AND ITS ROLE FOR DISASTER RISK REDUCTION COMMUNITY** 108
4.2.4 **THE PREVIEW GLOBAL RISK DATA PLATFORM** 109
4.2.5 **PREVIEW SDI CONCEPTUAL MODEL** 111
4.2.6 **PREVIEW GLOBAL RISK DATA PLATFORM, THE GATEWAY TO GLOBAL NATURAL DISASTER DATA** 116
4.2.7 **USES OF THE PREVIEW AND LESSONS LEARNT** 120
4.2.8 **CONCLUSIONS** 123

4.3 **OGC WEB FEATURE AND WEB COVERAGE SERVICES PERFORMANCE TESTING: TOWARDS AN EFFICIENT ACCESS TO GEOSPATIAL DATA** 125
4.3.1 **ABSTRACT** 125
4.3.2 **INTRODUCTION** 125
4.3.3 **GEOSPATIAL DATA INTEROPERABILITY** 127
4.3.4 **INSPIRE NETWORK SERVICES** 129
4.3.5 **METHODOLOGY OF TESTING** 131
4.3.6 **TECHNICAL ARCHITECTURE & SOFTWARE** 134
4.3.7 **RESULTS** 135
4.3.8 **DISCUSSION** 141
4.3.9 **CONCLUSIONS** 143

4.4 **SUMMARY AND LESSONS LEARNED** 145
CHAPTER 5: CAN SDI TAKE ADVANTAGE OF DISTRIBUTED COMPUTING POWER TO PROCESS THE INCREASING AMOUNT OF HIGH-RESOLUTION DATA? 147

5.1 PREAMBLE 148
5.2 GRID-ENABLED SPATIAL DATA INFRASTRUCTURE FOR ENVIRONMENTAL SCIENCES: CHALLENGES AND OPPORTUNITIES 151
  5.2.1 ABSTRACT 151
  5.2.2 INTRODUCTION 151
  5.2.3 BACKGROUND 156
  5.2.4 DESCRIBING AND CATALOGUING GEOSPATIAL DATA 157
  5.2.5 ACCESSING AND SHARING GEOSPATIAL DATA 159
  5.2.6 PROCESSING GEOSPATIAL DATA 160
  5.2.7 BENEFITS AND CHALLENGES TO USE GRIDS WITHIN SDIs 163
  5.2.8 EnviroGRIDS APPROACHES TO INTEROPERABILITY BETWEEN SDIs AND GRIDS 167
  5.2.9 CONCLUSIONS AND OUTLOOK 172
5.3 WPS MEDIATION: AN APPROACH TO PROCESS GEOSPATIAL DATA ON DIFFERENT COMPUTING BACKENDS 174
  5.3.1 ABSTRACT 174
  5.3.2 INTRODUCTION 174
  5.3.3 WEB PROCESSING SERVICE AND DISTRIBUTED COMPUTING 176
  5.3.4 GRIDIFICATION APPROACHES 180
  5.3.5 IMPLEMENTATION AND ARCHITECTURE 182
  5.3.6 USE CASE: NDVI COMPUTATION 192
  5.3.7 DISCUSSION & PERSPECTIVES 195
  5.3.8 CONCLUSIONS 199
5.4 SUMMARY AND LESSONS LEARNED 200

CHAPTER 6: CONCLUSIONS & RECOMMENDATIONS 202

6.1 CONCLUSIONS 203
6.2 LIMITATIONS/CONSTRAINTS TO SDI DIFFUSION AND UTILIZATION 217
6.3 RECOMMENDATIONS/PERSPECTIVES 220
6.4 CONCLUDING REMARKS 227

REFERENCES 230

ANNEXES 244

A.1 INSPIRE THEMES 245
A.2 BRINGING GEOSS SERVICES INTO PRACTICE WORKSHOP 249
A.3 WITH OR WITHOUT SDI 250

LISTS OF FIGURES, TABLES & WEBSITES 251

FIGURES 252
TABLES 253
WEBSITES 253

ABBREVIATIONS & ACRONYMS 256
Abstract
Abstract

Today we are living in a globalized world with rapidly evolving processes including climate and land cover change, population growth that are impacting the environment. In parallel, means of communication have expanded to take on a remarkable place in our society, allowing us to access an enormous and continuous flow of information.

Our planet is a multi-dimensional system made of complex interactions highly interconnected and continuously evolving at many spatial and temporal scales. To understand these interactions, we need to gather and integrate different sets of data about physical, chemical and biological systems, as well as socio-economical ones. Altogether, these sets of data constitute environmental data sets, or data related to the environment. These data are often georeferenced, describing a geographical location through a set of attributes thus are part of geospatial data. An environmental data set is seldom interesting in itself, but rather displays its full information potential when used in conjunction with other data sets, allowing one to monitor and assess the actual status of the global, regional or local environments, to discover complex relationships between them, to model future changes, and to potentially support sound and reliable decisions-making processes at all scales (from local to global), and in many disciplines.

However, it has been reported that data accessibility, availability and compatibility are among the most frequent difficulties evidenced while preparing various environmental assessments in Europe. Additionally, it is estimated that up to 50% of users’ time is spent in data discovery and transformation in order to make them compatible. This is mainly caused because geospatial data are voluminous, geographically distributed, heterogeneous in format, complex, and bound to institutional arrangements and policies. All these factors influence the way that data providers store, publish and deliver geospatial data. Moreover, users are often lacking computing resources to analyze data. Current environmental research projects regularly need to handle several terabytes of data, and accessing high-performance hardware and specialized software is expensive. This explains why data sources are often fragmented, integrating geospatial data to answer a scientific problem is difficult and expensive, and diffusion of geospatial information may be problematic and not achieved as widely as possible.

With this in mind, the underlying hypothesis of this thesis is that making data interoperable and providing access to computing resources can potentially improve the above mentioned situations allowing data users to spend more time in data analysis than in data discovery, and enabling more people to benefit from using geospatial data on the environment.

In this thesis, we first show that there is a growing need to better organize and share environmental data in order to understand the complexity of earth-system processes and to convey improved information on the environment to decision-makers and the general public. Addressing this need by sharing environmental data is challenging because it requires a common agreed framework that allows easy and seamless integration of data from different sources giving access to services that could be linked together to process and generate new understandable knowledge and information. The establishment
and implementation of initiatives such as the Global Earth Observation System of Systems (GEOSS) and the Infrastructure for Spatial Information in the European Community (INSPIRE) reflect a growing commitment to better and efficiently manage environmental data, and to share them more openly using interoperability arrangements. This indicates increasing recognition of the potential benefits of informed decision-making from evaluation, access, integration and processing of various environmental, economical, statistical and other data sources within a common framework. We also highlight a unique added value of interoperable data and processing services that allows users to perform functions that cannot be made with any single component. By integrating/composing different services, new properties are emerging that offer possibilities to better understand the complex relationships between the different components of the Earth system.

We then demonstrate the benefits of Spatial Data Infrastructure (SDI) in a real use case targeting the Disaster Risk Reduction community that needs timely access and easy integration of geospatial data. The development of the PREVIEW Global Risk Data Platform has exemplified that having geospatial data in digital form allows easy storage into databases and file systems, facilitates data exchange/sharing, enables faster updates, gives the ability to integrate data from multiple sources, and finally favors the development of customized products and services. SDI concepts, methods and technologies provide a solid ground to facilitate and coordinate the exchange and sharing of geospatial data. The development of such a platform has highlighted benefits and raised various issues. In particular, computational needs to process large data sets and efficient access to geospatial data through OGC services are two factors that will strongly influence the future success of SDIs. Ensuring user satisfaction through sufficiently responsive services giving access to vector and raster data sets requires to measure and monitor them to track latencies, bottlenecks and errors. Consequently, we developed an approach to measure performances of different data services and provided some guidance to data providers to improve the quality of their services. Our tests showed that overall performances of the tested implementations are globally satisfactory, even without tuning different parameters. However, to achieve reliable services, tuning memory on the server side is an essential and critical factor. Additionally, optimizing data and storage are factors that can easily increase efficiency of services. Some differences were highlighted regarding the various implementations of Web Feature Service (WFS) and Web Coverage Service (WCS) specifications. This can potentially limit data integration if clients do not implement the different flavors of these specifications. Finally, by their nature these specifications are not well suited to transfer large volume of data, and the current specifications are more appropriate to share local medium-resolution data than global high-resolution data. This can be a potential issue, especially given the ever-increasing volume of available high-resolution data.

The thesis finally discusses actual processing limitations of SDIs and presents a possible approach to extend their capabilities using the OGC Web Processing Service (WPS) specification on distributed computing infrastructures. Transforming raw data into understandable information is an essential task that SDIs cannot currently fully satisfy. Environmental sciences are data and computing-intensive domains where data are in general processed on desktop
Abstract

computers. This clearly limits the types of analyses that can be conducted due to their reduced power given the ever-growing size of data that need to be analyzed. Therefore, Environmental sciences and Geographical Information communities are trying to benefit from the superior storage and computing capabilities offered by distributed computing related methods and technologies. Our research showed that grid-enabled SDIs have the potential to become a powerful tool within the multi-disciplinary field of environmental sciences, empowering researchers to explore new venues to better understand the vast complexity of the interactions between anthropic and natural systems. In our view, connecting Grids and SDIs could potentially mark the advent of a new generation of SDIs extending their capacities to, and benefiting from, Grid infrastructures both in term of data processing and data management.

We conclude that SDIs and related concepts, methods and technologies are suitable and can bring major benefits to support and facilitate environmental data discovery, accessibility, visualization, dissemination and analysis. On a technical level, all the building blocks are available, supported by OGC and ISO standards, allowing data providers to start sharing and disseminating their data and metadata in an interoperable way. Our work has highlighted that this is feasible, it is not difficult to develop, and software implementation are reliable. Moreover, facilitating access, integration and use of geospatial data can answer the requirements of a specific community as well as making these data available to the widest possible audience. We argued through 13 recommendations that it is time to make all these web-based geospatial information components operational, otherwise SDI will remain only an innovative concept. Let the dream come true because we cannot wait! The challenges that humankind is facing require acting now and we need to provide decision-makers with tools that allow them accessing rapidly and efficiently good and reliable environmental information. SDIs have clearly the potential to be a part of the answer to bridge the gap between science and policy-making. It is obvious that to achieve this objective in the shortest term possible it will mostly depend on political, social, economical constraints. In this sense, the human component is probably the most influencing one. It is on this component that future endorsements depend because, at the technological level it is no more a problem to share data and metadata. At this stage it is only a matter of human/political will to make it happen or not. In our view, capacity building (at human, institutional and technical levels) will certainly help to reach endorsement on the use of such technologies, raising and increasing awareness on the benefits of sharing geospatial data, and finally creating new commitments. Scattering efforts and energies in discussions that often concern only details can block entire process, resulting in lost of (precious) time and motivation, and finally leading to the risk of disappointment and disengagement.... So it is time to SHARE!
Abstract

Résumé

Actuellement, nous vivons dans un monde globalisé, où tous les processus évoluent rapidement, dont entre autres les changements climatiques et de couverture du sol ainsi que la croissance démographique qui impactent l’environnement. Parallèlement, les moyens de communication ont évolué et pris une place remarquable dans notre société, nous permettant d’accéder à un flux énorme et continu d’informations.

Notre planète est un système multi-dimensionnel fait d’interactions complexes fortement interconnectées et en constante évolution à de nombreuses échelles spatiales et temporelles. Pour comprendre ces interactions, nous avons besoin de recueillir et d’intégrer des données différentes sur les propriétés physiques, chimiques et biologiques ainsi que socio-économiques. Ensemble, ces jeux de données constituent les données environnementales ou les données liées à l’environnement. Ces données sont souvent géoréférencées, décrivant une localisation géographique à travers un ensemble d’attributs et peuvent donc être entendues comme faisant partie des données géospatiales. Un ensemble de données environnementales est rarement intéressant en soi, mais montre plutôt son réel potentiel informatif lorsqu’il est utilisé en conjonction avec d’autres ensembles de données, permettant de surveiller et d’évaluer l’état de l’environnement mondial, régional ou local, de découvrir les relations complexes entre elles, de modéliser les changements futurs, et potentiellement de soutenir la prise de décisions pertinentes et fiables à toutes les échelles (du local au global) et dans de nombreuses disciplines.

Toutefois, il a été rapporté que l’accessibilité, la disponibilité et la compatibilité des données sont parmi les difficultés les plus fréquentes mises en évidence lors de la préparation de diverses évaluations environnementales en Europe. En outre, on estime que 50% du temps des utilisateurs est passé à découvrir et transformer des données afin de les rendre compatibles. Ceci est principalement dû au fait que les données géospatiales sont volumineuses, géographiquement distribuées, hétérogènes en terme de format, complexes, et liées à des arrangements institutionnels et politiques. Tous ces facteurs peuvent influencer la façon dont les fournisseurs de données stockent, publient et fournissent des données géospatiales. En outre, les utilisateurs manquent souvent de ressources de calculs pour analyser les données. Les projets de recherche actuels sur l’environnement ont régulièrement besoin de gérer plusieurs téraoctets de données et l’accès au matériel de haute performances ainsi qu’aux logiciels spécialisés est cher. Ceci explique pourquoi actuellement les sources de données sont souvent fragmentées, l’intégration de données géospatiales pour répondre à un problème scientifique est difficile et coûteux, et la diffusion de l’information géospatiale peut être problématique et n’est pas utilisée aussi largement que possible.

En gardant ce problème à l’esprit, l’hypothèse de base de cette thèse est que rendre les données interopérables et donner accès à des ressources de calcul peut potentiellement améliorer la situation mentionnée précédemment en permettant aux utilisateurs de passer plus de temps dans l’analyse qu’à la découverte de ces données et de permettre à davantage de personnes de bénéficier de l’accès aux données géospatiales sur l’environnement.
Abstract

Dans cette thèse, nous montrons d’abord qu’il existe un besoin croissant de mieux organiser et partager des données environnementales afin de comprendre la complexité des processus du système terrestre et de transmettre une meilleure information sur l’environnement aux décideurs et au grand public. Répondre à ce besoin par le partage des données sur l’environnement est difficile car elle nécessite un cadre commun qui permet une intégration facilitée et transparente des données provenant de différentes sources donnant accès à des services qui pourraient être reliés entre eux pour traiter et générer de nouvelles connaissances et informations. L’établissement et la mise en œuvre d’initiatives comme le Global Earth Observation System of Systems (GEOSS) et l’Infrastructure d’information Spatiale dans la Communauté Européenne (INSPIRE) démontrent un engagement et un besoin croissant de gérer efficacement les données environnementales, et de les partager plus ouvertement à l’aide de mécanismes permettant de les rendre interopérables. Cela indique aussi une reconnaissance croissante des avantages potentiels pour la prise de décision ainsi que pour l’évaluation, l’accès, l’intégration et le traitement des diverses sources de données environnementales, économiques, statistiques dans un cadre de référence commun. Nous soulignons également, une valeur ajoutée unique de données et de services interopérables: elles permettent aux utilisateurs d’exécuter des fonctions qui ne peuvent être faites avec un seul composant. En intégrant différents services, de nouvelles propriétés apparaissent et offrent des possibilités pour mieux comprendre les relations complexes entre les différentes composantes du système terrestre.

Nous démontrons ces avantages du SDI dans un cas d’utilisation réel ciblant la communauté de réduction des risques et catastrophes qui nécessite d’un accès rapide et une intégration facile des données géospatiales. Le développement de la plateforme PREVIEW a illustré qu’avoir des données géospatiales sous forme numérique permet: un stockage simplifié dans des bases de données et des systèmes de fichiers, facilite l’échange de données et le partage des données, et permet une mise à jour plus rapide, donne la possibilité d’intégrer des données provenant de sources multiples, et favorise enfin le développement de produits et services personnalisés. Les SDIs constituent donc une base solide pour faciliter et coordonner l’échange et le partage des données géospatiales. Le développement d’une telle plate-forme a mis en évidence des avantages et des inconvénients. En particulier, les besoins de calcul pour traiter de grands jeux de données et un accès efficace aux données géospatiales au moyen de web services OGC sont deux facteurs qui influencent fortement le succès futur des SDIs. Garantir la satisfaction des utilisateurs grâce à des services suffisamment efficaces donnant accès à des données vecteurs et rasters requièrent de pouvoir mesurer et surveiller ces services afin d’évaluer les temps de latence, les goulets d’étranglement et les erreurs. Par conséquent, nous avons développé une approche permettant de mesurer les performances de différents services de données et de fournir des orientations aux fournisseurs de données afin d’améliorer la qualité de leurs services. Nos tests ont montré que les performances globales des implémentations testées sont globalement satisfaisants déjà sans réglage de paramètres particuliers. Toutefois, pour atteindre une plus grande fiabilité des services, la mémoire est un facteur essentiel et critique à mettre au point. En outre, optimiser les données et leur stockage sont des facteurs qui peuvent facilement augmenter l’efficacité des
Abstract

services. Certaines différences ont été mises en évidence en ce qui concerne les diverses implémentations des spécifications WFS et WCS. Cela peut potentiellement réduire l'intégration de données si les clients n'implémentent pas ces différences de spécifications. Enfin, de par leur nature ces spécifications ne semblent pas bien adaptées à transférer de grands volumes de données. Les spécifications actuelles sont plus appropriées pour partager des données locales de moyenne résolution que des données globales à haute résolution. Cela peut être un problème potentiel compte tenu du volume croissant de données à haute résolution disponible.

Cette thèse aborde enfin les limites de capacité de calculs actuelles des SDIs et présente une approche pour étendre leurs capacités à l'aide de la spécification de l'OGC Web Processing Service (WPS) sur les infrastructures de calculs distribués. Transformer des données brutes en informations compréhensibles est une tâche essentielle que les SDIs ne peuvent pas actuellement satisfaire pleinement. Les sciences de l'environnement sont des domaines à forte intensité de calcul où les données sont généralement traitées sur des ordinateurs de bureau. Cela limite clairement les types d'analyses qui peuvent être réalisées en raison de leur puissance réduite, étant donné une taille sans cesse croissante des données qui doivent être analysées. Par conséquent les sciences de l'environnement et la communauté de l'information géographique tentent de bénéficier de la capacité de stockage accrue et des capacités de calculs offertes par les méthodes et technologies de calculs distribués. Nos recherches ont montré que les SDIs supportés par une grille de calculs ont le potentiel pour devenir un outil puissant dans le domaine multidisciplinaire des sciences de l'environnement, permettant aux chercheurs d'explorer de nouveaux domaines et technologies afin de mieux comprendre l'immense complexité des interactions entre les systèmes anthropiques et naturels. A notre avis, les grilles de calculs et les SDIs peuvent marquer l'avènement d'une nouvelle génération d'SDIs avec des capacités étendues et bénéficiant des grilles de calculs tant en terme d'analyse que de gestion des données.

Nous concluons que les SDIs et les concepts, méthodes et technologies associées sont appropriés et peuvent apporter des avantages importants pour soutenir et faciliter la découverte, l'accessibilité, la visualisation, la diffusion et l'analyse de données environnementales. Sur le plan technique, tous les éléments sont disponibles, supportés par les standards de l'OGC et les normes ISO, pour permettre aux fournisseurs de données de commencer à partager et diffuser leurs données et métadonnées d'une manière interopérable. Notre travail a souligné que cela est faisable et n'est pas difficile à mettre au point, les logiciels étant fiables et que faciliter l'accès, l'intégration et l'utilisation de données géospatiales permet de répondre aux exigences d'une communauté spécifique, et de rendre ces données accessibles à un public aussi large que possible. Nous soutenons à travers 13 recommandations, qu'il est temps de rendre tous ces éléments de l'information géospatiale opérationnels sinon les SDIs resteront uniquement un concept novateur. Il est temps que le rêve devienne réalité parce que nous ne pouvons pas attendre! Les défis auxquels l'humanité est confrontée exigent que l'on agisse maintenant et nous avons donc besoin de fournir aux décideurs des outils qui leur permettent d'accéder rapidement et efficacement à de l'information environnementale fiable et de qualité. Les SDIs ont clairement le potentiel d'être une partie de la réponse qui permettra de combler le fossé entre
Abstract

la science et les décideurs. Il est évident que pour atteindre cet objectif dans les plus brefs délais, tout dépendra en grande partie des contraintes politiques, sociales et économiques. En ce sens, la composante humaine est sans aucun doute la plus influente. C’est de cette composante que dépendra les futurs soutiens car au niveau technologique, le partage des données et métadonnées de manière interopérable n’est plus un problème. A ce stade, il s’agit simplement d’une question de volonté politique/humaine qui permettra d’y arriver ou non. A notre avis, le renforcement des capacités (au niveau humain, institutionnel et technique) aidera certainement à obtenir le soutien à l’utilisation de ces technologies, d’éléver et d’accroître la sensibilité des utilisateurs sur les avantages liés au partage des données géospatiales, et enfin de permettre la création de nouveaux engagements. La dispersion des efforts et des énergies dans des discussions qui concernent souvent des détails peuvent bloquer un processus entier, résultant en des pertes de temps (précieux) et de motivation, et enfin mener au risque de déception et de désengagement .... Alors il est temps de PARTAGER!
Abstract

Riassunto

Oggi viviamo in un mondo globalizzato con una rapida evoluzione dei processi tali i cambiamenti climatici e di copertura del suolo, la crescita della popolazione e impatando l’ambiente. In parallelo, i mezzi di comunicazione hanno acquisito un posto notevole nella nostra società, che ci permette di accedere a un enorme flusso continuo di informazioni.

Il nostro pianeta è un sistema multi-dimensionale di interazioni complesse altamente interconnesse e in continua evoluzione in molte scale spaziali e temporali. Questo significa che per capire queste interazioni, abbiamo bisogno di raccogliere e integrare diversi dati sui sistemi fisici, chimici e biologici, ed anche socio-economici. Nel complesso, questi insiemi di dati ambientali costituiscono insiemi di dati o di dati relativi all’ambiente. Questi dati sono spesso georeferenziati, che descrivono una posizione geografica attraverso una serie di attributi, e quindi potrebbero essere intesi come dati geospaziali. Un set di dati ambientali raramente è interessante in sé, ma piuttosto mostra il suo pieno potenziale informativo quando usato in combinazione con altri insiemi di dati, permettendo di monitorare e valutare lo stato attuale dell’ambiente globale, regionale o locale, per scoprire le complesse relazioni tra di loro, per modellizzare i cambiamenti futuri e, potenzialmente, sostenere decisioni valide e affidabili su tutte le scale (dalla locale alla globale) e discipline.

Tuttavia, è stato riferito che l’accessibilità, la disponibilità e la compatibilità dei dati sono tra le difficoltà più frequenti evidenziate durante la preparazione di diverse valutazioni ambientali in Europa. Inoltre, si stima che il 50% del tempo degli utenti viene speso nella scoperta e trasformazione dei dati, al fine di renderli compatibili. Ciò è dovuto principalmente perché i dati geospaziali sono voluminosi, geograficamente distribuiti, eterogenei in formati, complessi, e dipendono da accordi istituzionali e politici. Tutti questi fattori possono influenzare il modo in cui i provider di dati pubblicano e distribuiscono i dati geospaziali. Inoltre, gli utenti spesso mancano di risorse di calcolo per analizzare i dati. I progetti attuali di ricerca ambientale hanno regolarmente la necessità di gestire diversi terabyte di dati ed accedere ad alte prestazioni hardware e software specifici e sono costosi. Questo spiega perché attualmente le fonti di dati sono spesso frammentate, l'integrazione dei dati geospaziali per rispondere a un problema scientifico è difficile e costoso, e la diffusione d’informazioni geospaziali può essere problematico e non viene utilizzato nel modo più ampio possibile.

Tenendo in conto tutto ciò, l'ipotesi di questa tesi è di rendere i dati interoperabili e fornire un accesso a risorse di calcolo che possono migliorare la situazione sopra-citata consentendo agli utenti di dedicare più tempo all’analisi dei dati piuttosto che al rilevamento e consentire a più persone di trarre beneficio dall’uso dei dati geospaziali per l’ambiente.

In questa tesi, abbiamo dimostrato che c’è una crescente necessità di organizzare e condividere meglio i dati ambientali al fine di capire la complessità dei processi del sistema Terra e trasmettere una migliore informazione in materia ambientale all’opinione pubblica. Affrontare questa esigenza di condivisione di dati ambientali è impegnativo, perché richiede un quadro di comune accordo che permette una facile integrazione e ininterrotta di dati da fonti diverse che danno accesso a servizi che potrebbero essere collegati tra loro.
Abstract

per elaborare e generare nuove conoscenze e informazioni comprensibili. L'istituzione e l'attuazione di iniziative come il Global Earth Observation System of Systems (GE OSS) e l'Infrastruttura per l'Informazione Territoriale nella Comunità Europea (INSPIRE) riflettono un crescente impegno per una gestione dei dati ambientali migliore ed efficiente, con il fine di condividerli in modo più aperto con accordi di interoperabilità. Questo indica una crescente consapevolezza circa i potenziali vantaggi per un processo decisionale informato, e un accesso, un'integrazione ed un'elaborazione di varie fonti ambientali, economiche, statistiche e di altri dati all'interno di un quadro comune. Si evidenzia inoltre un valore aggiunto unico di dati e servizi interoperabili di elaborazione che permette agli utenti di eseguire funzioni che non possono essere fatte con ogni singolo componente. Integrando diversi servizi, nuove proprietà emergono e offrono la possibilità di comprendere al meglio i complessi rapporti tra le diverse componenti del sistema Terra.

Abbiamo dimostrato i benefici degli SDI in un caso d'uso reale della comunità nella riduzione dei rischi di catastrofi che hanno bisogno di accedere tempestivamente ed integrare facilmente i dati geospaziali. Lo sviluppo della piattaforma PREVIEW ha dimostrato che avere dati geospaziali in forma digitale consente un facile stoccaggio in banche dati e sistemi di file, facilita lo scambio di dati e la condivisione, consente aggiornamenti più veloci, dà la capacità di integrare dati provenienti da più fonti, e favorisce prodotti e servizi personalizzati in via di sviluppo. I concetti, metodi e tecnologie SDI possono fornire una solida base per facilitare e coordinare lo scambio e la condivisione di dati geospaziali. Lo sviluppo di tale piattaforma ha messo in evidenza i vantaggi e ha sollevato diverse questioni. In particolare, le esigenze di calcolo per elaborare grandi quantità di dati e un accesso efficiente ai dati geospaziali mediante servizi OGC sono due fattori che influenzano fortemente il futuro successo delle SDI. Garantire la soddisfazione degli utenti attraverso servizi efficienti che danno accesso ai dati vettoriali e raster richiede misurare e monitorare le latenze, le strozzature e gli errori. Di conseguenza, abbiamo sviluppato un approccio che permette di misurare le prestazioni dei diversi servizi di dati e di fornire alcune indicazioni ai provider per migliorare la qualità dei loro servizi. I nostri test hanno dimostrato che le prestazioni complessive delle implementazioni testate sono globalmente soddisfacenti anche senza modificare i parametri di ottimizzazione. Tuttavia, per realizzare servizi affidabili la memoria è un fattore essenziale e critico. Inoltre, l'ottimizzazione dei dati e del loro stoccaggio sono fattori che possono facilmente aumentare l'efficienza dei servizi. Alcune differenze sono state evidenziate per quanto riguarda le varie implementazioni di WFS e WCS. Ciò può potenzialmente limitare l'integrazione dei dati se i clienti non implementano queste diversi variazioni. Infine, dovuto alla loro natura, queste specificazioni non sono adatte per trasferire grandi quantità di dati e attualmente sono più adatte per condividere dati locali a media risoluzione che globali ad alta risoluzione. Questo può essere un potenziale problema, dato il volume sempre crescente di dati ad alta risoluzione disponibile.

Questa tesi infine discute i limiti analitici delle SDI e presenta un possibile approccio per estendere le loro capacità utilizzando l'OGC Web Processing Service (WPS) specificazione usando delle infrastrutture di calcolo distribuito. Trasformare i dati grezzi in informazioni comprensibili è un compito fondamentale che i SDI al momento non possono soddisfare pienamente. Le
Abstract

scienze ambientali sono domini di calcolo ad alta intensità in cui i dati sono elaborati in generale sui desktop computer. Questo limita chiaramente i tipi di analisi che possono essere effettuate a causa della loro potenza ridotta rispetto a una dimensione sempre crescente di dati che devono essere analizzati. Quindi le scienze ambientali e la comunità dell’informazione geografica stanno cercando di beneficiare di questa struttura superiore e della capacità di calcolo offerte dai modelli di calcolo distribuito e dalle tecnologie. La nostra ricerca ha dimostrato che i grid-SDI hanno il potenziale per diventare un strumento potente nel settore multi-disciplinare delle scienze ambientali, consentendo ai ricercatori di esplorare nuove strade per capire meglio la grande complessità delle interazioni tra sistemi antropici e naturali. A nostro avviso, le griglie di calcolo e i SDIs potrebbero segnare l’avvento di una nuova generazione di SDI che estende le proprie capacità e beneficiano delle infrastrutture Grid sia in termini di elaborazione dei dati che di gestione dei dati.

Concludendo, SDI e relativi concetti, metodi e tecnologie sono adattati e possono portare benefici importanti per sostenere e facilitare la scoperta, l’accessibilità, la visualizzazione, la diffusione e l’analisi dei dati ambientali. A livello tecnico, tutti i componenti sono disponibili, sostenuti da OGC e ISO, e consentono ai fornitori di dati di iniziare a condividere e diffondere i loro (meta) dati in modo interoperabile. Il nostro lavoro ha evidenziato che questo è fattibile e non è difficile da sviluppare: l’implementazione di software sono affidabili e facilitano l’accesso, l’integrazione e l’uso di dati geospaziali in grado di rispondere alle esigenze di una specifica comunità oltre a rendere tali dati accessibili al più vasto pubblico possibile. Abbiamo sostenuto con 13 raccomandazioni, che è il momento di rendere operativi tutti questi componenti altrimenti l’SDI rimarrà solo un concetto innovativo. Bisogna che il sogno diventi realtà, perché non possiamo aspettare! Le sfide che l’umanità si trova ad affrontare richiedono di agire ora e quindi abbiamo bisogno di fornire alle persone di rilievo la possibilità di accedere rapidamente ed efficacemente ad informazioni attendibili sull’ambiente. I SDI hanno chiaramente il potenziale per essere una parte della risposta a colmare il divario tra scienza e politica. E’ ovvio che per raggiungere questo obiettivo nel più breve termine possibile, dipenderanno da considerazioni politiche, sociali, e vincoli economici. In questo senso la componente umana è probabilmente quella più influente e il fatto di discutere e di raggiungere accordi è importante. Quindi è su questa componente che il futuro dipenderà perché a livello tecnologico non è più un problema condividere dati e metadati. In questa fase è solo una questione di volontà politica/umana di realizzarlo o meno. A nostro avviso, lo sviluppo di capacità (a livello umano, istituzionale e tecnico) sarà certamente di aiuto per raggiungere l’approvazione per l’uso di tali tecnologie, la sensibilizzazione e la presa di coscienza crescente sui vantaggi di utilizzare e condividere dati geospaziali e, infine, creare nuovi impegni. No bisogna spendere energie in discussioni che spesso riguardano solo dettagli in grado di bloccare un intero processo, con conseguente perdita di tempo prezioso e di motivazione, e, infine, portando a delusione e disimpegno... E temps di CONDIVIDERE!
Acknowledgments
Acknowledgments

At the end of this PhD work, I’m pleased to thank various people that are very important to me.

First of all, I’d like to warmly thank Dr. Anthony Lehmann, director of this thesis, for his scientific analysis and discussions, his pedagogy and encouragements, his perpetual enthusiasm, passion and humanity. They were for me a source of motivation and pleasure during these years of work. A very big thank also for giving me the chance and opportunity to work with him on projects like enviroGRIDS and ACQWA and integrating the enviroSPACE laboratory. All these moments of exchange, sharing, discussions and friendship have enriched and broadened my scientific and human horizons.

I’d like to thank very much Prof. Stefano Nativi from the Institute of Methodologies for Environmental Analysis of Prato, co-supervisor of this thesis, for its valuable advices and exchanges on Spatial Data Infrastructures, data analysis and distributed systems. His comments, suggestions and advices have been invaluable to me.

I am very pleased to thank Dr. Nicolas Ray, with whom I really appreciated to collaborate on many projects related to SDIs and distributed computing. Thank you for guiding my first steps into the world of Grid computing. And thank you also for all those moments of sharing and friendship so rewarding.

I would also like to thank Prof. Martin Beniston for his warm welcome in the Institute for Environmental Sciences (ISE) of the University of Geneva and for allowing me to work in his research group. I am also grateful to Prof. Walter Wildi for having welcoming me at the F.-A. Forel Institute. It is a unique opportunity for a young researcher to be able to work in contact with such great scientists.

A very big thank to Pascal Peduzzi from UNEP/GRID-Geneva for giving me the chance to develop the PreView platform. It was a great opportunity to work and develop my research in the field of SDIs.

Finally I would like to thank Pierre Lacroix and Alain Dubois from the University of Geneva for their valuable advices and techno-scientific exchanges during our fruitful collaboration around our joint research on geospatial data access.

I realize how lucky I was working with scientists like you and thank you for allowing me to enrich myself on human and scientific levels.

My work would not exist without the help and good humor of my colleagues at ISE: Kazi Rahman, Ana Silva, Christophe Etienne, Walter Silverio, Nicole Gallina, Margot Hill, Bastienne Uhlmann, Maura Brunetti, Stéphane Goyette, Helder Santiago, Regina Gama, Audrey Reverdin, Violeta Djambazova, Yann Pittet and UNEP / GRID-Geneva: Andrea De Bono, Diana Rizzolio, Karin Allenbach, Yaniss Guigouz, Géraldine Boezio, Stéphane Kluser, Stefan Schwarzer Bénédicte Boudol, Jean-Philippe Richard, Bruno Chatenoux, Christian Herold, Hy Dao, and Ron Witt. Thank you all for your friendship, these moments of sharing, laughter and all that gave me on human and professional perspectives.

A special thank to Dr. Jean-Michel Jaquet for giving me (a few years ago) the taste and passion for GIS and remote sensing. Thank you for his advices, his eternal encouragements and enthusiasm. He is for me an example to follow.
Acknowledgments

My thanks also go the “Système d’Information du Territoire Genevois” (SITG) for giving access to orthophotos, and GAR, ACQWA and enviroGRIDS partners. In particular, I wish to thank Prof. Dorian Gorgan and Denisa Rodila from the Technical University of Cluj-Napoca (Romania) for their valuable assistance on Grid computing. A special thank to Jay and Françoise Pearlman from IEEE and all the members of the GEO STC committee for the valuable discussions and exchange I had with them.

I would also like to thank all administrative and technical staff of the F.-A. Forel Institute, ISE and GRID-Geneva for giving me a framework particularly suited to my research.

Of course, I deeply thank my friends for their support, encouragements, and their presence throughout these years of research. I’m very lucky to have you!

Finally, I thank my family: my parents and my nonni. This thesis would never have been possible without your support, your presence, your patience, your encouragements and for giving me the sense of values, work ethic and effort, respect and willingness to excel. Thank you for trusting me. If I’m happy and I am what I am now, is to you that I owe. Thank you from the deepest of my heart!

******************************************************************************

This PhD thesis has benefited from various European and United Nations projects. I would like to acknowledge the European Commission “seventh framework programme” that funded the enviroGRIDS (Grant Agreement n° 226740) and ACQWA (Grant Agreement n° 212250) projects and also the United Nations International Strategy for Disaster Reduction (UNISDR), United Nations Development Programme (UNDP), United Nations Environment Programme (UNEP), the World Bank and the Swiss Federal Environment Agency for their financial and in kind support in the development of the PREVIEW Global Risk Data Platform and associated research. Finally, I am grateful to the Free and Open Source – GIS community because without its open and sharing spirit, I would not have been able to learn and develop the tools used in my research.
Remerciements

Au terme de ce travail de thèse, il m’est très agréable de remercier de nombreuses personnes qui toutes et tous comptent beaucoup pour moi.

En premier lieu je souhaiterais chaleureusement remercier le Dr. Anthony Lehmann, directeur de cette thèse, pour son analyse scientifique, ses discussions, sa pédagogie, ses encouragements, son perpétuel enthousiasme, sa passion et son humanité. Ils ont été pour moi une source de motivation et de plaisir durant ces années de travail. Un très grand merci aussi pour m’avoir donné la chance et l’opportunité de travailler avec lui sur de nombreux projets dont enivroGRIDs et ACQWA ainsi que d’intégrer le laboratoire enivroSPACE. Tous ces moments d’échange, de partage, de discussions et d’amitié ont enrichi et élargi mon horizons scientifique et humain.

Je souhaiterais vivement remercier le Prof. Stefano Nativi de l’Institut des Méthodologies d’Analyses de l’Environnement de Prato, co-superviseur de cette thèse, pour ces précieux conseils et échanges concernant les infrastructures des données géospatiales, l’analyse de données et les systèmes distribués. Ces commentaires, suggestions et conseils ont été très précieux pour moi.

Il m’est très agréable de remercier le Dr. Nicolas Ray avec qui j’ai eu beaucoup de plaisir à collaborer autour de nombreux projets en relation aux SDIs et du calculs distribués. Merci pour avoir guidé mes premiers pas dans le monde du Grid computing. Et merci aussi pour tout ces moments de partage et d’amitié tellement enrichissants.

Je souhaite aussi remercier le Prof. Martin Beniston pour son chaleureux accueil au sein de l’Institut des Sciences de l’Environnement (ISE) de l’Université de Genève et pour m’avoir permis de travailler au sein de son groupe de recherche. Je suis aussi très reconnaissant au Prof. Walter Wildi pour son accueil au sein de l’Institu F.-A. Forel. C’est une chance unique pour un jeune chercheur que de pouvoir travailler au contact de tels scientifiques.

Un très grand merci aussi à Pascal Peduzzi du PNUE/GRID-Genève pour m’avoir donné la chance de développer la plateforme PreView. Cela a été une grande opportunité de pouvoir ainsi travailler et valoriser mes recherches dans le domaine des SDIs.

Finalement je souhaiterais remercier Pierre Lacroix et Alain Dubois de l’Université de Genève pour leur précieux conseils et échanges technico-scientifiques lors de notre fructueuse et agréable collaboration autour de notre recherche commune sur l’accès aux données géospatiales.

Je mesure la chance que j’ai de pouvoir travailler avec des scientifiques comme vous et merci pour m’avoir permis de m’enrichir tant d’un point de vue humain que scientifique.

Mon travail n’aurait pas été ce qu’il est sans l’aide et la bonne humeur de mes collègues de l’ISE: Kazi Rahman, Ana Silva, Christophe Etienne, Walter Silverio, Nicole Gallina, Margot Hill, Bastienne Uhlmann, Maura Brunetti, Stéphane Goyette, Helder Santiago, Regina Gama, Audrey Reverdin, Violeta Djambazova, Yann Pittet et du PNUE/GRID-Genève: Andrea De Bono, Diana Rizzolio, Karin Allenbach, Yaniss Guigoz, Géraldine Boezio, Stéphane Kluser, Stefan Schwarzer, Bénédicte Boudol, Jean-Philippe Richard, Bruno Chatenoux, Christian Herold, Hy Dao et Ron Witt.Merci à toutes et à tous pour votre amitié,
Acknowledgments

les moments de partage, de rires et pour tout ce que vous m’avez apporté sur le plan humain et professional.

Un merci tout particulier au Dr. Jean-Michel Jaquet pour m’avoir donné (il y a quelques années maintenant) le goût et la passion de SIG et de la télédétection. Merci pour ces conseils, ces encouragements et son éternel enthousiasme. Il est pour moi un exemple à suivre.

Mes remerciements vont aussi au Système d’Information du Territoire Genevois (SITG) pour m’avoir donné accès à des orthophotos ainsi qu’aux partenaires des projets GAR, ACQWA et enviroGRIDS. En particulier, je souhaite remercier le Prof. Dorian Gorgan et Denisa Rodila de l’Université Technique de Cluj-Napoca (Roumanie) pour leur aide précieuse sur le Grid computing. Un merci spécial à Jay and Françoise Pearlman de IEEE ainsi que tous les membres du GEO STC pour les intéressantes et enrichissantes discussions que j’ai pu avoir avec eux.

Je tiens également à remercier tout le personnel administratif et technique de l’Institut F.-A. Forel, de l’ISE et du GRID pour m’avoir fourni un cadre de travail particulièrement propice à mes recherches.

Bien entendu, j’aimerais vivement remercier mes amis et amies pour leur soutien, leurs encouragements, leur présence tout au long de ces années de recherche. J’ai beaucoup de chance de vous avoir!

Finalement, j’aimerais remercier ma famille: mes parents et mes nonni. Ce travail de thèse n’aurait jamais vu le jour sans votre soutien, votre présence, votre patience, vos encouragements ainsi que pour m’avoir donné le sens des valeurs, le goût du travail et de l’effort, le respect et la volonté de se dépasser. Merci de m’avoir fait confiance. Si je suis heureux et je suis ce que je suis devenu, c’est à eux que je le dois. Merci du fond du cœur!

***************************

Ce travail de thèse a bénéficié de différents projets européens et onusiens. Je souhaite remercier le 7ème programme cadre de la Commission Européenne qui a financé les projets enviroGRIDS (Grant Agreement n° 226740) et ACQWA (Grant Agreement n° 212250) ainsi que la Stratégie Internationale pour la Réduction des Désastres (UNISDR), le Programme des Nations-Unies pour le Développement (PNUD), le Programme des Nations-Unies pour l’Environnement (PNUE) et l’Office Fédéral de l’Environnement qui ont financé et soutenu le développement et la recherche associée de la plateforme PREVIEW des données sur les risques globaux. Finalement je remercie vivement la communauté libre et open source – GIS pour leur esprit de partage car sans eux je n’aurais pu apprendre et développer les outils présentés dans cette thèse.
Acknowledgments

Ringraziamenti

Al termine di questa tesi, mi fa molto piacere ringraziare molte persone che sono molto importanti per me.

Prima di tutto vorrei ringraziare di cuore il Dr. Anthony Lehmann, direttore di questa tesi, per le sue analisi scientifiche, discussioni, il suo insegnamento, l'incoraggiamento, il suo entusiasmo perpetuo, passione e umanità. Essi sono stati per me una fonte di motivazione e di piacere nel corso di questi anni di lavoro. Un grande grazie anche per avermi dato la possibilità e l'opportunità di lavorare con lui su progetti come enviroGRIDS ed ACQWA e integrare il laboratorio enviroSPACE. Tutti questi momenti di scambio, condivisione, discussioni e amicizia hanno arricchito ed ingrandito il mio horizonte scientifico ed umano.

Mi piacerebbe anche ringraziare il Professore Stefano Nativi del Istituto di Metodologie per l'Analisi Ambientale di Prato, co-supervisore di questa tesi, per i suoi preziosi consigli e gli scambi sulle infrastrutture di dati geospaziali, analisi dei dati e dei sistemi distribuiti. I suoi commenti, suggerimenti e consigli sono stati preziosi per me.

Sono molto lieto di ringraziare il Dr. Nicolas Ray, con il quale ho avuto tanto piacere di collaborare su numerosi progetti in materia SDIs e di calcolo distribuito. Grazie per avermi guidato i miei primi passi nel mondo del Grid computing. E grazie anche per tutti quei momenti di condivisione e di amicizia così gratificante.

Vorrei anche ringraziare il Professore Martin Beniston per la sua calda accoglienza nel Istituto di Scienze Ambientali (ISE) dell'Università di Ginevra e per avermi consentito di lavorare nel suo gruppo di ricerca. Sono anche molto riconoscente al Prof. Walter Wildi per il suo sostegno all'Istituto F.-A. Forel. Si tratta di un'occasione unica per un giovane ricercatore di essere in grado di lavorare a contatto di scienziati del loro livello.

Un grande ringraziamento a Pascal Peduzzi del UNEP/GRID-Ginevra per avermi dato la possibilità di sviluppare la piattaforma PreView. E 'stata una grande occasione per poter lavorare e sviluppare la mia ricerca nel campo degli SDIs.

Infine vorrei ringraziare Pierre Lacroix e Alain Dubois del Università di Ginevra per i loro preziosi consigli e gli scambi scientifici e tecnici durante la nostra proficua collaborazione durante le nostre ricerche comuni sull'accesso dei dati geospaziali.

Mi rendo conto di quanto sono stato fortunato di poter lavorare con degli scienziati come voi e vi ringrazio per avermi permesso di arricchire me stesso sia di un punto di vista umano che scientifico.

Il mio lavoro non esisterebbe senza l'aiuto e il buon umore dei miei colleghi del ISE: Kazi Rahman, Ana Silva, Christophe Etienne, Walter Silverio, Nicole Gallina, Margot Hill, Bastienne Uhmann, Maura Brunetti, Stéphane Goyette, Helder Santiago, Regina Gama, Reverbín Audrey, Violeta Djambazova, Pittet Yann e UNEP/GRID-Ginevra: Andrea De Bono, Diana Rizzolio, Karin Allenbach, Yaniss Guigoz, Géraldine Boezio, Stéphane Kluser, Stefan Schwarzer, Bénédicte Boudol, Jean-Philippe Richard, Bruno Chatenoux, Christian Herold, Hy Dao, Ron Witt. Grazie a tutti voi per la vostra amicizia, i momenti di condivisione, di risate e tutto quello che mi avete dato al livello umano e professionale.
Un ringraziamento particolare al Dr. Jean-Michel Jaquet per avermi dato (alcuni anni fa), il gusto e la passione per il GIS e la teledetezione. Grazie per i suoi consigli, eterni incoraggiamenti ed entusiasmo. Per me è un esempio da seguire.

I miei ringraziamenti vanno anche al “Système d'Information du Territoire Genevois” (SITG) per avermi concesso l’accesso a ortofoto ed ai colleghi dei progetti GAR, ACQWA ed enviroGRIDS. In particolare, desidero ringraziare il Professore Dorian Gorgan e Denisa Rodila del Università Tecnica di Cluj-Napoca (Romania) per il loro prezioso aiuto sul Grid computing. Un ringraziamento speciale per Jay e Françoise Pearlman del IEEE e tutti i membri del comitato GEO STC per le interessante discussione che ho potuto avere con loro

Vorrei inoltre ringraziare tutto lo staff amministrativo e tecnico del Istituto F.-A. Forel, del ISE e del GRID per avermi dato un quadro particolarmente adatto alla mia ricerca

Naturalmente, un grande ringraziamento a miei amici per il loro supporto, il loro incoraggiamento, la loro presenza nel corso di questi anni di ricerca. Sono molto fortunate di avervi!

Infine, ringrazio la mia famiglia: i miei genitori e miei nonni. Questa tesi non sarebbe mai stato possibile senza il vostro sostegno, la vostra presenza, la vostra pazienza, il vostro incoraggiamento e per avermi dato il senso dei valori, etica del lavoro e lo sforzo, il rispetto e la volontà per eccellere. Grazie per la fiducia me. Se sono felice e sono quello che sono ora, è a vuoi che lo devo. Grazie di cuore!

***************************

Questa tesi di dottrato ha beneficiate di diversi progetti europei e delle Nazioni Unite. Vorrei ringraziare la Commissione europea "Settimo programma quadro" che ha finanziato I progetti enviroGRIDS (Grant Agreement N ° 226740) e ACQWA (Grant Agreement n ° 212250) ed anche i la Strategia Internazionale per la Riduzione dei Disastri (UNISDR Programma delle Nazioni Unite per lo Sviluppo (UNDP), Programma delle Nazioni Unite per l’Ambiente (UNEP) e l’Ufficio federale dell’ambiente per il loro sostegno finanziario nello sviluppo del piattaforma PREVIEW. Infine vorrei ringraziare la comunità di open source - GIS per il loro spirito di condivisione, perché senza di loro non avrei potuto imparare e sviluppare gli strumenti presentati in questa tesi.
Chapter 1: 
Introduction
Chapter 1: Introduction

1.1 Structure of the thesis

This thesis is structured in six chapters contributing to fulfill the research questions that will be defined later on (under section 1.5.2).

Chapter 1 introduces the thesis by setting the scene and explaining the necessity of sharing data. It then focuses on environmental data and their specificities, and defines three research questions to be addressed. It also briefly presents the projects that helped in conducting the associated research.

Chapter 2 gives an overview of the basic concepts underlying Spatial Data Infrastructures and related topics, with a special focus on interoperability and standards. The aim of this chapter is to give the necessary knowledge to readers allowing them to understand the associated research developed in the three following chapters.

Chapter 3 shows that there is a growing need to better organize and share environmental data in order to understand the complexity of the Earth-system, and to convey improved information on the environment to decision-makers and the general public. Addressing this need by sharing environmental data is challenging because it requires a common agreed framework that allows easy and seamless integration of data from different sources giving access to services that could be linked together to process and generate new understandable knowledge and information.

Based on the recognition that there is a clear commitment in sharing data, Chapter 4 aims at demonstrating the applicability of SDI concepts and methods targeting specifically the Disaster Risk Reduction community. Indeed, timely access and easy integration of geospatial data are essential to support efforts in reducing disaster risk and promoting a culture of disaster resilience. The development of such a platform has highlighted benefits and raised different issues. In particular, computational needs to process large data sets and efficient access to geospatial data through OGC services are two factors that will strongly influence the future success of SDIs. Ensuring user satisfaction through sufficiently responsive services giving access to vector and raster data sets require to measure and monitor them to track latencies, bottlenecks and errors. In consequence, this chapter also adds some insights into performance measurements of different WFS and WCS services. It provides also some guidance to data provider to improve the quality of their services.

Chapter 5 discusses actual analytical limitations of SDIs and presents a possible approach to use the OGC Web Processing Service (WPS) specification on distributed computing infrastructures. Transforming raw data into understandable information is an essential task that SDIs cannot currently fully satisfy. Environmental sciences are a computing-intensive domain where data are in general processed on desktop computers. This can clearly limit the types of analyses that can be conducted in light of the ever-growing size of data that need to be analyzed. Therefore, Environmental sciences and Geographical Information communities are trying to benefit from the superior storage and computing capabilities offered by distributed computing related methods and technologies.

Finally, Chapter 6 concludes this research by answering the three research questions, and by making recommendations for future directions of research.
Chapter 1: Introduction

1.2 Projects

This research was conducted and has benefited from three international projects at the global and regional scales supported by the United Nations and the 7th Framework Program for EU Research (FP7).

EnviroGRIDS (http://www.envirogrids.net) is a FP7 project that will last from 2009 until 2013. The Black Sea Catchment is largely following an ecologically unsustainable pathway based on inadequate resource management that could lead to severe environmental, social and economical problems, especially in a changing climate (WWF 2008). The aim of the project is to build capacities in the Black Sea region to use new international standards to gather, store, distribute, analyze, visualize and disseminate crucial information on past, present and future states of this region, in order to assess its sustainability and vulnerability. EnviroGRIDS objective is to federate and strengthen existing Observation Systems to address several GEOSS Societal Benefit Areas within a changing climate framework. The expected result will be a shared information system that operates on the boundary of scientific/technical partners, stakeholders and the public. It will contain early warning systems able to inform in advance decision-makers and the public about risks to human health, biodiversity and ecosystems integrity, agriculture production or energy supply caused by climatic, demographic and land cover changes on a 50-year time horizon. To achieve and support the enviroGRIDS vision and objectives, a grid-enabled Spatial Data Infrastructure (gSDI) is under construction. The aim of the gSDI is to host and analyze the data for the assessment of GEOSS Societal Benefit Areas, as well as the data produced within the project. These data must be gathered and stored in an organized form and accessible in an interoperable way on the grid infrastructure in order to provide a high performance and reliable access through standardized interfaces.

The PREVIEW (Project of Risk Evaluation, Vulnerability, Information, and Early Warning) Global Risk Data Platform (http://preview.grid.unep.ch) is a collaborative effort of United Nations Environment Programme (UNEP), United Nations Development Programme (UNDP/BCPR), and United Nations International Strategy for Disaster Reduction (UNISDR) and the World Bank to share geospatial data on global risk from natural hazards. Users can freely visualize, download or extract data on past hazardous events, human and economical hazard exposure and risk from natural hazards. The platform covers nine types of natural hazard: tropical cyclones and related storm surges, drought, earthquakes, biomass fires, floods, landslides, tsunamis and volcanic eruptions. The collection of data is made via a wide range of partners. This geoportal was developed as a support to the 2009 Global Assessment Report on Disaster Risk Reduction (United Nations International Strategy for Disaster Reduction Secretariat, 2009), replacing the previous PREVIEW platform initially designed by UNEP/GRID-Europe and already available since 2000. The new PREVIEW platform is fully compliant with the OGC Web Services (OWS) to access data using Web Map Service (WMS), Web Feature Service (WFS), Web Coverage Service (WCS), geo-enabled Really Simple Syndication (GeoRSS) or Keyhole Markup Language (KML), as well as metadata using Catalogue Service for the Web (CS-W).
Chapter 1: Introduction

ACQWA ([http://www.acqwa.ch](http://www.acqwa.ch)) stands for Assessing Climate impacts on the Quantity and quality of Water. It is also a FP7 European research project lasting from 2008 until 2013. As the evidence for human induced climate change becomes clearer, so does the realization that its effects have impacts on natural environment and socio-economic systems. Some regions are more vulnerable than others, both to physical changes and to the consequences for ways of life. The project assess the impacts of a changing climate on the quantity and quality of water in mountain regions which are particularly affected by rapidly rising temperatures, prolonged droughts and extreme precipitation. Modeling techniques are used to project the influence of climatic change on the major determinants of river discharge at various time and space scales. Regional climate models provide the essential information on shifting precipitation and temperature patterns. Snow, ice, and biosphere models feed into hydrological models in order to assess the changes in seasonality, amount, and incidence of extreme events in various catchment areas. Environmental and socio-economic responses to changes in hydrological regimes are analyzed in terms of hazards, aquatic ecosystems, hydropower, tourism, agriculture, and the health implications of changing water quality. Attention is also devoted to the interactions between land use/cover changes, and changing or conflicting water resource demands. Adaptation and policy options will be elaborated on the basis of the results. The chain of processes involved in climatic, cryospheric and hydrologic models is complex because each process impacts on different compartments of human and natural systems. Different types of data covering various geographical regions are therefore necessary to build different sets of scenarios, which translates into substantial amount of data. In ACQWA, SDI is used to extend the outputs of the project to potential end-users.

1.3 List of contributing research papers


1.4 Background: Responding to our changing environment

Today we live in a globalized world where everything is changing rapidly (e.g., population, climate, land cover, biodiversity), and where communication means have taken a remarkable place in our life. Every day scientists as well as the general public access an enormous and continuous flow of information and much of it refers to a position or a specific place on the surface of our planet. This information is therefore, and by definition, georeferenced (or geospatial).

In the last 30 years, the amount of geospatial data available has grown dramatically following the evolution of the communication means and the rapid development of spatial data capture technologies such as Global Positioning System (GPS), remote sensing images, sensors, etc... (Philips, Williamson et al. 1999). Over the last ten years, with the advent of applications like Google Earth, we have seen that geographical information has been incorporated and routinely embedded into business and workflows of agencies at all levels of government, as well as in the private sector (Booz, Allen et al. 2005).

Despite the fact that administrations and governments are recognizing that geospatial information is important and must be part of the basic information infrastructure that need to be efficiently coordinated and managed for the interest of all citizens (Ryttersgaard 2001), this huge amount of geospatial data is stored in different places, managed by different organizations, and the vast majority of the data are not being used as effectively as they should.

Moreover, at the United Nations Conference on Environment and Development in Rio de Janeiro in 1992, the so-called Agenda 21 resolution underscored the importance of geospatial information to support decision-making and management on the degradation and threats that are affecting the environment (Nebert 2005). This means that there is a strong need for availability and access to appropriate information. The development of databases and exchange of information are the conditions for creating the basis for a sustainable development and to support the information management needs for implementing and monitoring sustainable development policies and goals such as the UN Millennium Development Goals (Henricksen 2007).

Thus, geospatial information is a critical element underpinning decision making for many disciplines (Rajabifard and Williamson 2001) at all scales, from local to global. Experiences from developed countries show that more than two-thirds of human decision-making is driven by geospatial information (Ryttersgaard 2001).

However, geospatial information is an expensive resource, that is time consuming to produce, and so it is of high importance to improve the access and availability of data, and to promote its reuse. Many of the decisions that different organizations need to make depend on good and consistent geospatial data, that are both available and readily accessible (Rajabifard and Williamson 2001).

In 1998, the former vice-president of the United States, Al Gore, presented his visionary concept of a Digital Earth (Gore 1998), “a multi-resolution, three-dimensional representation of the planet, into which we can embed vast quantities of georeferenced data”. As of today this vision is clearly not fully realized, but it is still relevant and it gives us an interesting support to our purpose.

Talking about geospatial data, Al Gore (1998) mentioned that the difficult part in taking advantage of this vast amount of information will be “making sense
Chapter 1: Introduction

of it, turning raw data into understandable information” because at the moment we have more information than we are able to handle and it is stored in “electronic silos of data”, remaining mostly unused. He envisioned applications where “information can be seamlessly fused with the digital map or terrain data” allowing the user to move through space and time. To achieve this vision, a collaborative effort (from government, industry, academia and citizens) is needed.

All the technologies and capabilities required to transform this vision into reality and to build a Digital Earth have become available:

- **computational science**: even a simple desktop computer can process complex models and simulations. With the potential of technologies such as the Grid new insights into the data are possible, giving us the ability to simulate phenomena that are impossible to observe.
- **mass storage**: storing Tera-bytes of data on a desktop computer is not a problem anymore.
- **remote sensing imagery**: many satellites and airborne sensors are continuously observing the Earth offering high spatial and temporal multispectral observations.
- **sensors**: it is now possible to access real time data from sensors that are continuously monitoring and measuring various environmental variables.
- **broadband networks**: are already a reality giving the ability to connect different databases together.
- **interoperability**: this is a key point to allow communication and integration of distributed data, allowing the geospatial data generated by one software to be read by another. Still if technology is available, data interoperability is not yet achieved.
- **metadata**: are important as they describe the data, allowing a user to evaluate and discover the data before using them.

Even if all technologies are ready, organizations and agencies around the world are still spending billions of dollars every year to produce, manage and use geospatial data without getting the information they need to answer the challenges our world is facing (Rajabifard and Williamson 2001). These authors also highlight the facts that most organizations and/or agencies need more data than they can afford, and these data are often outside their jurisdictions, and the data collected by different organizations are often incompatible. This inevitably leads to inefficiencies and duplication of effort, and thus it is evident that countries can benefit both economically and environmentally from a better management of their data (Nebert 2005; Henricksen 2007).

In consequence, it is now essential to make these data easily available and accessible in order to give the opportunity to the users to turn them into understandable information with clear and broad benefits for the society and the economy, because “working together, we can help and solve many of the most pressing problems facing our society...” (Gore 1998).

It is clear that there are many challenges to face, both tangible and intangible, when we start sharing data but we have to overcome them in order to improve our knowledge, share our experiences and try to build a better-informed society. Achieving the goal of a sustainable development requires the integration of a large number of different types of data from different sources. Through agreed common standards and a clear political/institutional will, these data can be
interchanged and integrated in an interoperable way, leading to a new collaborative approach to decision-making.

For Arzberger et al. (2004), ensuring that data are easily accessible, so that they can be used as often and as widely as possible, is a matter of sound stewardship of public resources. Availability should be restricted only in certain specific cases like national security. These authors argue that “publicly funded research data should be openly available to the maximum extent possible”, because publicly funded data are a public good, produced in the public interest.

1.5 Research problem and questions

1.5.1 Research problem

Our planet is a multi-dimensional system made of complex interactions highly interconnected and continuously evolving at many spatial and temporal scales (GEO secretariat 2005). To understand these interactions, we need to gather and integrate different sets of data about physical, chemical and biological, as well as social and economical systems. Altogether, these sets of data constitute environmental data sets or data related to the environment. These data are often georeferenced, describing a geographical location through a set of attributes and thus are part of so-called geospatial data. An environmental data set is seldom interesting in itself, but rather displays its full information potential when used in conjunction with other data sets, allowing one to monitor and assess the actual status of the global, regional or local environments, to discover complex relationships between them, or to model future changes.

Currently, data accessibility, availability, compatibility, and lack of sufficient resources to analyze these data are among the most frequent difficulties that are negatively influencing the way that scientists, researchers, decision-makers and the general public are accessing and using these data (Bernard and Craglia 2005; Vandenbroucke 2010). These authors estimate that up to 50% of users’ time is spent in data discovery and transformation in order to make them compatible and integrable. This is mainly due to the fact that geospatial data are voluminous, complex (e.g., geometries, relationships, attributes), geographically distributed, and heterogeneous in term of format. Additionally, institutional arrangement and policies (e.g., copyrights, intellectual property rights) can impede the diffusion of geospatial data. All these factors influence the way that data providers store, publish and deliver environmental data. Moreover, users are often lacking the appropriate computational resources to analyze these data. Current environmental research projects regularly need to handle several terabytes of data and accessing high-performance hardware and specialized software is expensive (Di 2004; Di, Chen et al. 2008). This explains why currently data sources are often fragmented, integrating geospatial data to answer a scientific problem is difficult and expensive, and diffusion of geospatial information is problematic and not applied efficiently. Therefore, we assume that making data interoperable and providing access to computing resources can potentially improve the above mentioned situation allowing data users to spend more time in data analysis than in data discovery, and enabling more people to benefit from using geospatial data.
Chapter 1: Introduction

1.5.2 Research questions

Recognizing the problem of environmental data accessibility, availability and compatibility together with the need to enable high-performance computing to analyze ever-increasing amounts of high-resolution data sets leads us to:

(a) define the aim of this research:

*Examine how well Spatial Data Infrastructure concepts, methods and related technologies are useful to support and facilitate environmental data discovery, accessibility, visualization, dissemination and analysis.*

(b) formulate associated research questions:

1. *Is there a need to access and process environmental data in a better and efficient way?*

   One of the challenges we are facing today is to make sense of the vast amount of data and information we continuously receive and access in order to turn them into understandable information (Gore 1998). This will participate to building a better-informed society because achieving the goal of a sustainable development involves the integration of a large number of different types of data from different sources (Henricksen 2007). Through agreed common standards and a clear will to share data and information, these data can be interchanged and integrated in an interoperable way, and can lead to a collaborative approach to decision-making.

2. *How can SDI improve our capacity to discover, share, retrieve and integrate environmental data?*

   Geospatial data can be a shared resource that can be maintained continuously. Having geospatial data in digital form allows easy storage into databases and file systems, facilitates data exchange/sharing, faster updates, gives the ability to integrate data from multiple sources, and finally favors developing customized products and services (Nebert 2005). Therefore, a collaborative environment based on the concept of partnership in data production, management, and integration would bring major benefits (Mansourian, Rajabifard et al. 2006; Alinia and Delavara 2009). As a result, the concept of SDI appears an interesting framework to facilitate and coordinate the exchange and sharing of geospatial data (Rajabifard and Williamson 2001; Masser 2007) encompassing data sources, systems, network linkages, standards and institutional issues in delivering geospatial data and information, from many different sources to the widest possible group of potential users (Coleman, McLaughlin et al. 1997). SDIs intend to avoid duplication of efforts and expenses by enabling users to save resources, and time when trying to acquire or maintain data sets (Rajabifard and Williamson 2001; Mansourian, Rajabifard et al. 2004). Finally SDIs can be seen as an
Chapter 1: Introduction

integrated information highway which links together environmental, socio-economic and institutional geospatial data resources providing a movement of data from local to national and eventually to global levels (Rajabifard and Williamson 2001; Masser 2005).

3. Can SDI take advantage of distributed computing power to process the increasing amount of high-resolution data?

Current SDIs are essentially supporting data discovery, visualization and retrieval, but have typically limited analysis capabilities (Kiehle, Greve et al. 2006). This means that the processing of geospatial data is done in general on the client's desktop computer, which is an inhibiting factor when processing large and high-resolution data sets. With the recently introduced Open Geospatial Consortium Web Processing Service specification and the promises of high storage and computing capacities offered by Grid and Cloud infrastructures, new opportunities are emerging within environmental sciences (Padberg and Greve 2009).
Chapter 2: Theoretical framework

Based on:
Chapter 2: Theoretical framework

2.1 Spatial Data Infrastructure

2.1.1 Definition, concepts and rationale

The term Spatial Data Infrastructure (SDI) is often used to describe the mechanisms or the enabling environment that support easy access to, and utilization of, geographical data and information (Ezigbalike 2004). This definition is quite reductive as it gives the idea that SDIs are essentially technical. The primary objective of SDIs is to provide a basis for geospatial data discovery, evaluation, and application for users and providers within all levels of government, commercial and the non-profit sectors, academia and citizens (Nebert 2005).

This means that SDIs are more than just data repositories. SDIs store data, their attributes, and their related documentation (metadata), offering a means to discover, visualize, and evaluate their fitness to different purposes, and finally provide access to data themselves. In addition to these basic services, there are often additional services or software supporting the use of the data. Finally, to make an SDI working efficiently, it is necessary to include all the organizational agreements needed to coordinate and administer it. In consequence, following Masser (2005) and Nebert (2005), we can give a more complete definition of what are SDIs:

“A spatial data infrastructure supports ready access to geographic information. This is achieved through the coordinated actions of nations and organizations that promote awareness and implementation of complementary policies, common standards and effective mechanisms for the development and availability of interoperable digital geographic data and technologies to support decision making at all scales for multiple purposes. These actions encompass the policies, organizational remits, data, technologies, standards, delivery mechanisms, and financial and human resources necessary to ensure that those working at the national and regional scale are not impeded in meeting their objectives”.

Before going further into details, we have to explain the concepts underlying the rational of SDI, in particular geospatial data and information (also named geodata or georeferenced data). A geospatial data describes a location on Earth, giving through its attributes a comprehensive picture of the physical, biological, chemical world both in term of spatial and/or temporal extent. Geospatial data are extremely valuable as users can build spatial relationships between features and data. For example, just after a flood event, one can overlay remote sensing images with existing georeferenced data of settlements to evaluate the extent of the damage and then focus humanitarian assistance. In consequence, geospatial data have a key role to play in our knowledge-based economy affecting directly or indirectly different sectors such as forestry, urban planning, security, telecommunication, environmental protection, etc... (figure 1).
Chapter 2: Theoretical framework

If, previously, geographical information was mostly presented in the form of paper maps, with increasing means to capture information in digital formats, geospatial data are nowadays used and viewed within a Geographical Information System (GIS). This computer system is capable of assembling, storing, manipulating, and displaying geographically referenced information (Ezigbalike 2004).

A GIS gives the ability to merge different existing information from different sources facilitating collaboration in creating and analyzing data. Due to these new possibilities of reusing existing data and working on collaboratively greater scale, new challenges arise. When someone wishes to create a new information layer based on different data sets or different formats, with different terminology, and perhaps different projection, it is quite difficult to bring them together. Harmonizing geospatial data is a complex, costly and time-consuming task, but could be in general achieved by agreeing among data capturers before the work begins.

Figure 1: GIS and economy (Source: Geoconnections).
Chapter 2: Theoretical framework

The growing recognition that once a geospatial data set has been created it can be used within public and private sectors (Ryttersgaard 2001) reinforces the need to store data into databases that are made accessible for different purposes (Philips, Williamson et al. 1999). This leads to the concept that geospatial data can be a shared resource, which will be maintained continuously.

The advantage of having geospatial data in digital form (Ezigbalike 2004; Henricksen 2007) are:

- easy storage,
- easy dissemination,
- facilitation of data exchange/sharing,
- faster and easier updates and corrections,
- ability to integrate data from multiple sources,
- customization of products and services.

As a result of the previous considerations, the concept of SDI was developed in order to facilitate and coordinate the exchange and sharing of geospatial data (Rajabifard and Williamson 2001), encompassing data sources, systems, network linkages, standards and institutional issues involved in delivering geodata and information from many different sources to the widest possible group of potential users (Coleman, McLaughlin et al. 1997).

The vision of an SDI incorporates different databases, ranging from the local to the national, into an integrated information highway and constitutes a framework, needed by a community, in order to make effective use of geospatial data (Ezigbalike 2004).

2.1.2 Objectives

Following Masser’s definition (2005) and the different considerations highlighted previously we can list different objectives underpinning SDIs:

- The overall objective of an SDI is to maximize the reuse of geospatial data and information.
- SDIs cannot be realized without coordination (especially by governments).
- SDIs must be user-driven, supporting decision-making for many different purposes.
- SDIs implementation involves a wide range of activities, including not only technical topics such as data, standards, interoperability, and delivery mechanisms, but also institutional arrangements, policies, financial and human resources.
- The term infrastructure is used to promote the idea of a reliable and supporting environment, analogous to a road or a telecommunication network, facilitating the access to geoinformation using a minimum set of common practices, protocols, and specifications (Nebert 2005). This allows the movement of spatial information instead of goods.
- SDIs are all about (Henricksen 2007):
  - re-use: of data, technical capabilities, skills developed, invested effort and capital.
  - sharing: “sharing-not-wearing” the costs of data, people, technology,... helping to realize more rapid returns on investment.
  - learning from others: avoiding the pitfalls experienced by others.
Chapter 2: Theoretical framework

- Avoid duplication efforts and expenses and enables users to save resources, time and effort when trying to acquire or maintain datasets (Rajabifard and Williamson 2001).
- SDIs are “about working smarter, not harder” (Henricksen 2007).
- Implies to scale from specific and monolithic (data-centric) towards independent and modular (service-oriented) information systems.
- Integrate these systems together into an information highway, which both links together environmental, socio-economic and institutional databases and provides a movement of information from local to national and global levels.
- Encompass the sources, systems, network linkages, standards and institutional issues involved in delivering spatially-related information from many different sources to the widest possible group of potential users.

Altogether these objectives intend to create an environment that foster activities (figure 2) for using, managing and producing geospatial data and in which all stakeholders can co-operate with each other and interact with technology, to better achieve their objectives at different political/institutional levels (Rajabifard and Williamson 2004).

![Diagram of EnviroGRIDS vision of improved data access and geoprocessing](image)

**Figure 2: EnviroGRIDS vision of improved data access and geoprocessing**

2.1.3 Components

Masser (2005) identifies the most important stakeholders with special interests in geoinformation/SDIs matters and shows their diversity both in terms of size and resource of the large numbers of players involved:
Chapter 2: Theoretical framework

- Central government organizations,
- Local government organizations,
- Commercial sector,
- Non-for-profit/non-governmental organizations,
- Academics,
- Individuals.

Therefore the temptation can be creating a centralized “one-size-fits-all” spatial database, in order to provide all the information needed by a country or a specific community of common interest. But as reported by Ezigbalike (2005), Henricksen (2007) and Nebert (2005) the existence of geospatial data and information does not alone ensure that it will be used for decision-making. Different other factors are important to consider in order ensuring that information will be effectively used and reused:

- To be used, people need to know that the data exist, and where to obtain them.
- They need to be authorized to access and use the data.
- They need to know the history of the data capture, in order to interpret it correctly, trust it and be able to integrate it meaningfully with data coming from other sources.
- They need to know if the data depends on other data sets, in order to make sense of data.

Consequently, to leverage the full potential of geospatial data, an SDI must be made of different components allowing users to find, discover, evaluate, access and use these data, namely:

- A clearly defined core of geospatial data.
- The adherence to known and accepted standards and procedures.
- Databases to store data and accessible documentation about the data, the so-called metadata.
- Policies and practices that promote the exchange and reuse of information.
- Adequate human and technical resources to collect, maintain, manipulate and distribute geospatial data.
- Good communication channels between people/organizations concerned with geospatial data, allowing the establishments of partnerships and share knowledge.
- The technology for acquiring and disseminating data through networks.
- Institutional arrangements to collaborate co-operate and coordinate actions.

But as stated by Rajabifard and Williamson (2001), there is an important additional component represented by people. This includes not only users of geospatial data but also data providers and any other data custodians. For these authors, people are the key to transaction processing and decision-making. Facilitating the role of people and data in governance that appropriately supports decision-making and sustainable development objectives is central to the concept of SDI.

In order to meet the requirements of all stakeholders involved, an SDI must (Coleman, McLaughlin et al. 1997):

- be widely available,
Chapter 2: Theoretical framework

- be easy to use,
- be flexible,
- form the foundation for other activities.

In summary, Rajabifard and Williamson (2001) suggest that an SDI cannot be seen only as composed of geospatial data, services and users but instead involves other issues regarding interoperability, policies and networks.

![Diagram of SDI components]

Figure 3: Nature and relations between SDI components (Source: GISCache).

Figure 3 shows that an SDI is by nature really dynamic, as people who want to access data must interact with technological components.

2.1.4 SDI hierarchy

As a result of the fact that SDI initiatives range from local to national and regional levels (Crompvoets and Bregt 2003; Masser 2007) and they all aim to promote economic development, to stimulate better government and to foster environmental sustainability (Masser 2005), Rajabifard (2002) proposed a model of SDI hierarchy that is made of inter-connected SDIs developed at different levels (from local to global). Each SDI of a higher level is primarily formed by the integration of geospatial datasets developed and made available by the lower level (figure 4).
Chapter 2: Theoretical framework

Such a hierarchy has two views: in one hand it is an umbrella in which SDI at a higher level encompasses all the components of SDIs at levels below. On the other hand, it can be seen as the building block supporting the access of geospatial data needed by SDIs at higher levels. The SDI hierarchy allows to create an environment in which users working at any level can rely on data from other levels and integrate geospatial data from different sources (Mohammadi, Rajabifard et al. 2008). Such a hierarchy is envisioned by regional and global initiatives such as INSPIRE and GEOSS that will be further discussed.

For Masser (2006), the SDI hierarchy poses the challenge of a multistakeholder participation in SDI implementation because the bottom-up vision differs a lot from the top-down approach that is implicit in most of the SDI literature. The top-down approach emphasizes the need for standardization and uniformity while the bottom-up stresses the importance of diversity and heterogeneity due to different aspirations of the various stakeholders. In consequence, it is necessary to find a consensus to ensure some measure of standardization and uniformity while recognizing the diversity and heterogeneity of the different stakeholders performing different tasks at different levels.

2.1.5 SDI evolution and (emerging) trends

Different authors (Crompvoets and Bregt 2003; Masser 2007) have studied the diffusion and evolution of SDI around the world and show that driving forces behind SDI initiatives are generally similar:

- promoting economic development,
- stimulating better government,
- fostering environmental sustainability,
- modernization of government,
- environmental management.

They all agree on the fact that, as of today, a critical mass of SDI users has been reached as a result of the diffusion of SDI concepts during the last ten to fifteen years. This provides a basic network of people and organizations that is essential for future development of SDIs.
Chapter 2: Theoretical framework

Rajabifard et al. (2001) find that the first generation of SDIs, based on a product model, gave way to a second generation at the beginning of years 2000, the latest being characterized by a process model. Indeed, the first generation of SDIs were product-based, aiming to link existing and future databases while the second generation aims to define a framework to facilitate the management of information assets allowing reuse of collected data by a wide range of people and/or organizations for a great diversity of purposes at various times and scales. For Masser (2005) this evolution emphasizes the shift from concerns of data producers to those of data users and the shift from centralized structures to decentralized and distributed networks like the Web.

The process-based model emphasizes the communication channel of knowledge infrastructure and capacity building towards the creation of an SDI facilitating cooperation and exchange of data and knowledge (Rajabifard and Williamson 2001). They also highlight the fact that characteristics of social systems strongly influence the approach taken to implement and develop a Spatial Data Infrastructure. They propose key issues and strategies to be considered for the design process:

- development of a strategic vision and associated implementation strategy,
- recognition that SDI is not an end in itself,
- key institutional strategy is to have all coordinating processes administered by one group.

Today's effort on the technical development of SDI components clearly focus on the exchange of geospatial data in an interoperable way (Bernard and Craglia 2005) through services that allow efficient access to spatially distributed data. The shift towards an infrastructure offering services to answer requests rather than a “simple” network allowing to find, view and exchange geospatial data is highlighted by the concept of web services and the related Service Oriented Architecture (SOA).

Web services are a “new paradigm” (Cömert 2004) where different systems or providers offer some services for certain user groups, allowing an easy access to distributed geographic data and geoprocessing applications. The web services emphasize the necessity that systems involved could talk to each other and the provision of this talk should be easy and cost-effective for businesses to profit. In other words, web services rely on interoperability. Web services enable the possibility to construct web-based application using any platform, object model and programming language. A service is no more than a collection of operations that allows users to invoke a service, which could be as simple as requesting to create a map or complicated as processing a remote sensing image. In summary, web services are for application-to-application communication over Internet and are based, in general, on open standards like XML (Cömert 2004). SOA is the basic principle concept supporting web services development. It promotes loose coupling between software components so that they can be reused (Sahin and Gumusay 2008). In a SOA, the key component is services. They are well-defined set of actions, self-contained, and stateless (i.e., do not depend on the state of other services).

There are three components on the web services architecture: service provider, service requester and service broker and three operations: publish, find and bind. A SOA relates the three components to the three operations to allow automatic discovery and use of services.
Chapter 2: Theoretical framework

In a traditional scenario, a service provider hosts a web service and “publishes” a service description to a service broker. The service requester uses a “find” operation to retrieve the service description and uses it to “bind” with the service provider and invoke the web service itself (figure 5).

![Diagram of SOA operations](image)

**Figure 5: Basic operations in SOA (Source: IBM).**

SOA is the underlying concept for an interoperable environment based on reusability and standardized components and thus is of high importance for SDIs allowing applications and related components to exchange data, share tasks, and automate processes over the Internet (Open Geospatial Consortium 2004). The OGC web services are, by far, the most important and relevant web services in the GI community and they will be further discussed in details. With the advent of web services into the SDI community new trends/opportunities could be foreseen:

- Actual SDIs are lacking analysis capabilities, an essential task to turn data into understandable information. This means that processing geospatial data is done in general on desktop computers and thus limit the analytical capacities caused by the huge execution time that geoprocessing tasks require processing a vast amount of data. With the recently introduced Web Processing Service and the promises of high storage and computing capacities offered by Grid infrastructures, new opportunities are emerging within geosciences and environmental communities (Padberg and Greve 2009).

- Semantic web-developed vocabularies (called ontologies) for geospatial data with the goal to increase understanding of such data by machines, allowing automated process through web services (Boes and Pavlova 2008; Van Oosterom and Zlatanova 2008).

- Web services are one aspect of the Web 2.0 revolution. The Web 2.0 refers to a second generation of Internet based services that allow people to collaborate and share information (Boes and Pavlova 2007). GIS is also taking advantage of the Web 2.0 revolution, with a good example being Google Maps that opened some of the more straightforward capabilities of GIS to the general public (Goodchild 2007) and allowing, with other tools, the general public to create and generate news sources of data and information. This phenomenon is also know as Volunteered Geographic Information (VGI) (Boes and Pavlova 2007; Craglia, Goodchild et al. 2008; Coleman, Georgiadou et al. 2009). VGI offers new opportunities and
perhaps will influence the development of SDIs and the production of data in a near future.

Finally, it is necessary to mention that building an efficient SDI is almost impossible without partnership because a single agency is unlikely to have all resources, skills or knowledge to undertake the development of all aspects of a SDI (Ezigbalike 2004; Henricksen 2007). This is why different authors (Williamson, Rajabifard et al. 2003; Rajabifard and Williamson 2004) stress the importance of capacity building component in the SDI implementation process.

2.1.6 Benefits

To conclude, we can highlight some of the (expected) benefits that SDIs can offer:

- universal (anywhere and anytime) access to geospatial data and related information,
- services and applications to discover and access distributed data sources,
- integration of different geospatial information to provide seamless visualization,
- seamless combination (chaining) of data, services and related applications,
- geospatial data update and maintenance made easy,
- sharing and reuse capabilities,
- collaborative activities,
- wide-scale interoperability, agreeing on open and common standards,
- development of partnerships, collaboration between different stakeholders.

The Canadian Geospatial Data Infrastructure (CGDI) claimed that developing applications using such an infrastructure allows to:

- Reduce costs: Applications can be built by reusing existing services.
- Reduce complexity: Service interfaces hide the underlying complexity.
- Permits less costly integration and interoperability: Standard interfaces simplify interconnection and integration.
- Allow direct access to current, authoritative source data.

Finally an effective and working SDI leads to:

- Informed decision-making: easy access to current information, knowledge and expertise.
- Increased efficiency: standards and specifications, as well as access to services, reduce duplication of effort.
- Enhanced usability: providing reliable access to geospatial information to all levels, from the citizens to governments.
- Push for economic growth.

We would not depict SDIs as a “perfect tool” that can solve all problems. It is evident that SDIs represent a great opportunity and a framework with great perspective but as Masser reminds us (2005), SDIs can facilitate access to data to a wide range of users only if profound changes in “sharing spirit” take place. He also mentions the fact that building an SDI is a long-term process. In order to be fully operational such a process depends on sustainability and commitment.
Chapter 2: Theoretical framework

In addition, there are several others issues that could limit the implementation of SDI concepts such as: collaboration, funding, political stability, legislation, priorities, awareness, copyrights, privacy, licensing, capacity building and cultural issues (table 1).

<table>
<thead>
<tr>
<th>Technical</th>
<th>Institutional</th>
<th>Policy</th>
<th>Legal</th>
<th>Social</th>
</tr>
</thead>
<tbody>
<tr>
<td>Computational heterogeneity</td>
<td>Collaboration models</td>
<td>Political stability</td>
<td>Rights, Restrictions</td>
<td>Cultural</td>
</tr>
<tr>
<td>Semantic</td>
<td>Funding model</td>
<td>Legislation</td>
<td>Copyright,</td>
<td>Capacity building</td>
</tr>
<tr>
<td>Reference sys.</td>
<td>Linkage between data units</td>
<td>Priorities/ Sustainable dev.</td>
<td>Intellectual Property rights</td>
<td>Equity</td>
</tr>
<tr>
<td>Data quality</td>
<td>Awareness of data existence</td>
<td>Data access and prices</td>
<td>Privacy</td>
<td>Licensing</td>
</tr>
<tr>
<td>Metadata</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Format</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Table 1: Integration issues (Williamson, Rajabifard et al. 2003)*

To conclude this section, it is important to keep in mind that geospatial data sharing and related SDI developments rely primarily on individuals (Craig 2005) that have many interests in common. First, their idealism, their sense that better data will lead to better decisions. Second, their self-interest: by sharing, they receive something in return even if it is intangible, they are viewed as cooperative partners, and finally they are involved in a professional culture that honors serving society and cooperating with others.

2.2 Interoperability and standards

2.2.1 Definition and concepts

Previously, we have seen that we are living in a world that is changing rapidly with communication means that are taking an increasing place in our everyday life. This communication revolution is mostly based on the Internet, whose successes are due to interoperability. Interoperability is “the ability of a system or a product to work with other systems or products without special effort on the part of the customer.” (Open Geospatial Consortium 2004). This means that two or more systems or components are able to transmit or exchange information through a common system and to use the information that has been exchanged.

When systems are interoperable, it gives users the ability to:

- find what they need,
- access it,
- understand and employ it,
Chapter 2: Theoretical framework

- have goods and services responsive to their needs.

As of today, in a climate of economic constraint, interoperability and standardization have never been so important because a non-interoperable system impedes sharing of data, information and computing resources (Open Geospatial Consortium 2004), leading organizations to spend much more than necessary on data, software and hardware. Moreover being non-interoperable increases the risk for a system or an infrastructure to not deliver its expected benefit and in consequence to lead users to disappointment and system failure.

In order to achieve interoperability, there are two approaches:

- adhering to standards
- making use of a "broker" of services that can convert one product's interface into another product's interface, "on the fly".

One good example of the first approach is the Web, where standards like HTTP, TCP/IP or HTML have been developed by organizations that wish to create standards to "meet everyone's needs without favoring any single company or organization" (Open Geospatial Consortium 2004; Open Geospatial Consortium 2005).

The great advantage of interoperability, and that is why it is an essential building block for the GIS and SDI industry, is that it describes the ability of locally managed and distributed heterogeneous systems to exchange data and information in real time to provide a service (Open Geospatial Consortium 2004). This allows users to maximize the value of past and future investments in geoprocessing systems and data.

As a response to the need of GIS standards to support interoperability, the OGC aims to tackle the non-interoperability caused by the diversity of systems creating, storing, retrieving, processing and displaying geospatial data in different formats. In addition to this, software vendors often did not communicate among themselves to agree on how data should be structured and stored and how systems must exchange information, leading inevitably to a non-interoperable environment, isolating geospatial data in "electronic silos" and resulting in expensive duplication of data and difficulty in sharing and integrating information (Open Geospatial Consortium 2004).

The OGC (2005) has pointed out general user needs:

- Need to share and reuse data in order to decrease costs (avoid redundancy collection), obtain additional or better information, and increase the value of data holdings.
- Need to choose the best tool for the job and the related need to reduce technology and procurement risks (avoid being locked in to one vendor).
- Need to leverage investment in software and data, enabling more people to benefit from using geospatial data across applications without the need for additional training.

The OGC believes that responding to user needs of interoperability will have a profound and positive impact in public and private sectors, opening doors of new business opportunities and new human activities.

In summary, interoperability enhances: communication, efficiency and quality for the benefit of all citizens allowing them to access data in a good, consistent and transparent way.
2.2.2 Types of interoperability

There are two types of interoperability (Open Geospatial Consortium 2004):

- **syntactic** (or technical): when two or more systems are capable of communicating and exchanging data, they are exhibiting syntactic interoperability. Specified data formats and communication protocols are fundamental. In general, XML or SQL standards provide syntactic interoperability. Syntactical interoperability is required for any attempts of further interoperability.

- **semantic**: Beyond the ability of two or more computer systems to exchange information, semantic interoperability is the ability to automatically interpret the information exchanged meaningfully and accurately in order to produce useful results as defined by the end users of both systems. To achieve semantic interoperability, both sides must defer to a common information exchange reference model. The content of the information exchange requests are unambiguously defined: what is sent is the same as what is understood (e.g., explaining why INSPIRE is producing data specifications).

Different types of geoprocessing systems (vector, raster, CAD, etc.) producing different types of data, different vendors of systems using internal data formats and producing proprietary formats, different vendors systems using proprietary libraries and interfaces and reducing the possibilities of communication between systems... are all causes of syntactic non-interoperability while different data producers using different metadata schemas and/or different naming conventions lead to semantic non-interoperability.

The World Wide Web and its associated technologies offer a great opportunity to overcome both syntactic and semantic non-interoperability because it is an almost universal platform for distributed computing and it provides facilities to semantically process structured text. The web is thus a key enabler for interoperability, by increasing access to geospatial data and processing resources, which in consequence increases the value of those resources (Open Geospatial Consortium 2004). To ensure effective interoperability, it is not only a matter of technology but also and often it requires a change of philosophy, of spirit to go “open”. This is classified under human or legal/policy on the following table summarizing the different types of interoperability (table 2).

<table>
<thead>
<tr>
<th>Technical</th>
<th>Semantic</th>
<th>Human</th>
<th>Legal/Policy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Machine to machine connections</td>
<td>Common understanding concepts, terms, ...</td>
<td>Cooperation</td>
<td>Digital rights, ownerships</td>
</tr>
<tr>
<td>Software interaction</td>
<td>Inter-disciplinary vocabularies</td>
<td>Training</td>
<td>Responsibility</td>
</tr>
<tr>
<td>APIs</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Formats</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Table 2: Different types of interoperability.*
Chapter 2: Theoretical framework

As expressed by the OGC (2004) “If an organization does not fully embrace the tenets of interoperability and interoperable architectures, then long-term success in integrating geospatial processes into an organization’s overall business processes may be problematic”. In consequence, organizations will need:

- commitment to interoperability and geospatial standards,
- commitment to collaboration,
- commitment to define a geospatial interoperability and information framework that meets the requirements of the organization,
- commitment to the collection and maintenance of geospatial metadata,
- commitment to training and education

Through all these commitments, an organization will be truly interoperable, maximizing the value and reuse of data and information under its control and will be able to exchange these data and information with other interoperable systems, allowing new knowledge to emerge from relationships that were not envisioned previously.

2.2.3 Interoperability enablers

To enable effective interoperability, we have already seen that the Internet and standards are probably the most important components at a technical level but there are a lot of other possible enablers, both human and technical, that could help an organization to promote its commitment to interoperability:

- web and networks,
- standards,
- infrastructure,
- metadata,
- support for multiple: languages, views, data formats, projections, datums,...
- sharing of best practices,
- cooperation and collaboration,
- business models,
- business agreements,
- policy framework,
- copy and access rights,
- authorization,
- others...

Altogether they will contribute in a way or another to a successful implementation of geospatial interoperability by reaching a consensus between the users’ need for compatibility with the autonomy and heterogeneity of the inter-operating systems (Open Geospatial Consortium 2005).

2.2.4 Standards

Standards are documented agreements, used in public contracts or international trade, containing technical specifications or other precise criteria to used consistently as rules, guidelines, or definitions of characteristics, to
ensure that materials, products, processes and services are fit for their purpose (Ostensen 2001). In other words, standardization means agreeing on a common system (Open Geospatial Consortium 2005).

The existence of non-harmonized standards for similar technologies contribute to the so-called “technical barriers to trade”, avoiding a user to share data, information or services.

Although developing standards is a long and complex process, involving many organizations, based on a consultative approach and aiming to find a consensus between all the parties involved (Ezigbalike 2004), organizations and agencies are increasingly recognizing that standards are essential for improving productivity, market competitiveness, export capabilities (Nebert 2005), and lowering maintenance and operation costs over time (Kaufmann and Dorfschmid 2001; Booz, Allen et al. 2005; European Commission 2006).

We can summarize the functions of standards as follow:

• help to ensure interoperability,
• promote innovation, competition, commerce and free trade,
• increase efficiency,
• make things work,
• affect every aspect of our life (widespread use of standards).

2.2.5 Benefits

After discussing what are standards and interoperability, we can give an overview of the (expected) benefits of a truly interoperable architecture.

• Allow sharing and reusing of data: gives access to distributed and heterogeneous sources of data.
• Avoid data duplication: data are collected and maintained at the most appropriate place.
• Reduce the costs: of maintenance, of operations and of course of production.
• Integration: As Mohammadi et al. (2007) shows multi-source data integration can only be achieved with an effective interoperability.
• Reduce the complexity: through common knowledge, standards offer a set of rules that every data provider can follow, understand and become familiar with. Moreover when a user shares a data in a standardized way, another will be immediately able to use it.
• Increase efficiency.
• Vendor neutral: avoid being locked in to one vendor.
• Improve decision-making: offering standardized access to a vast amount of data and information and to used them as effectively as they should.
• New opportunities and knowledge: open the doors to new activities and relations that were not foreseen before.

Finally, as OGC (2005) stated:

“Changing internal systems and practices to make them interoperable is a far from simple task. But the benefits for the organization and for those who make use of information it publishes are incalculable”.

- 51 -
2.3 Initiatives

Different initiatives at the regional and global level are influencing and promoting the creation of Spatial Data Infrastructures and the use of open standards. They all are concerned about data access, harmonization, standardization, interoperability, seamless integration and services. They coordinate actions that promote awareness and implementation of complementary policies, common standards and effective mechanisms for the development and availability of interoperable digital geographic data and technologies to support decision making at all scales for multiple purposes. These actions encompass the policies, organizational structures, data, technologies, standards, delivery mechanisms, and financial and human resources necessary to ensure that those working at the global and regional scale are not impeded in meeting their objectives.

2.3.1 Infrastructure for Spatial Information in the European Community (INSPIRE)


The Infrastructure for Spatial Information in the European Community, namely INSPIRE, is a European Directive (entered into force in May 2007 and fully operational by 2019) that aims to create a European Union Spatial Data Infrastructure. This will enable the sharing of environmental spatial information among public sector organizations and better facilitate public access to spatial information across Europe (European Commission 2007). When fully implemented, it will, theoretically enable data from one Member State to be seamlessly combined with data from all other States. This is particularly important for activities relating to the environment.

The main purpose of INSPIRE is to support the formulation, implementation, monitoring, and evaluation of Community environmental policies (European Commission 2010). Therefore spatial information considered under the directive is extensive and includes a great variety of topical and technical themes and will be based on Spatial Data Infrastructures established and operated by Member States.

This initiative wishes to overcome the barriers affecting data access and exchange in Europe, including (European Commission 2010):

- Inconsistencies in collection of georeferenced data: geospatial data are often missing and/or incomplete, or are collected twice by different organizations.
- Lacking of documentation, description (metadata) of the data.
- Geospatial data are often incompatible and thus cannot be combined.
- Infrastructures used to find, access and use geospatial data often function in isolation and are incompatible.
- Barriers to sharing: cultural, linguistic, institutional, financial and legal.

In order to overcome these barriers, it has been recognized that it would be necessary to develop a legislative framework asking Member States to coordinate their activities and to agree on a set of requirements, common standards and processes. In consequence, INSPIRE is unique in the sense that it
Chapter 2: Theoretical framework

is a major collaborative and participative process to formulate the directive, create implementing rules and develop relative specifications and services.

INSPIRE seeks to create a European SDI and the INSPIRE Directive defines it: “infrastructure for spatial information means metadata, spatial data sets and spatial data services; network services and technologies; agreements on sharing, access and use; and coordination and monitoring mechanisms, processes and procedures, established, operated or made available in accordance with this Directive” (European Commission 2007).

The end users of INSPIRE include policymakers, planners and managers at the local, national and regional levels, and citizens and their organizations.

INSPIRE is based on common principles (European Commission 2007):

1. Data should be collected only once and kept where it can be maintained most effectively.
2. It should be possible to combine seamless spatial information from different sources across Europe and share it with many users and applications.
3. It should be possible for information collected at one level/scale to be shared with all levels/scales; detailed for thorough investigations, general for strategic purposes.
4. Geographic information needed for good governance at all levels should be readily and transparently available.
5. Easy to find what geographic information is available, how it can be used to meet a particular need, and under which conditions it can be acquired and used.

A step-by-step approach is used to implement and develop the infrastructure because such an initiative cannot be built from one day to another and is asking Member States to drastically change their existing infrastructure. Thus the implementation of services has been started just after the adoption of the Directive, whereas the harmonization of INSPIRE data themes will be made in three phases up to 2013.

The European Commission Joint Research Center (JRC) plays a major role in this initiative as it has supported the development of the proposal and now endorses the responsibility of the overall technical coordination of the Directive, providing support to preparation of technical rules on implementation, data harmonization, documentation and required services to discover, view and download data.

The Directive provides five sets of Implementing Rules (IR) that set out how various elements of the system (metadata, data sharing, data specification, network services, monitoring and reporting) will operate and to ensure that spatial data infrastructures of the Member States are compatible and usable in a Community and transboundary context (European Commission 2010). The Drafting Teams that are currently working on these IRs are composed of international experts and the process includes open consultation – particularly with Spatial Data Interest Communities (SDIC) and Legally Mandated Organizations (LMO).

The Directive specifically states that no new data will need to be collected. However it does require that two years after adoption of the Implementing Rules for data sets and their related services each Member State will have to ensure that all newly collected spatial data sets are available in conformity with the IR.
Chapter 2: Theoretical framework

Other data sets must conform to the Rules within 7 years of their adoption. Implementing Rules will be adopted in a phased manner between 2008 and 2012 with compliance required between 2010 and 2019 (European Commission 2010).

![Data resources](image)

**Figure 6: Data and information flow within the INSPIRE framework (Source: INSPIRE).**

The envisioned interoperability in INSPIRE is a possibility offered to the user to combine geospatial data and services from different sources across the European Community in a consistent way without involving specific efforts of humans or computers (figure 6). Thus users will spend less time and efforts to integrate data delivered within the INSPIRE framework.

The Directive (European Commission 2007) defines 34 “spatial data themes” that have been defined in three Annexes sorted in order of priority. Annex 1 datasets cover the ‘basic’ spatial building blocks such as spatial referencing systems, geographic names, addresses, transport networks, hydrography and land parcels. Because of the range of data types involved, the impact of INSPIRE is comprehensive. Annex 1 datasets have to be prepared and made available from 2011, with the other Annexes at later dates. In order to enable full system interoperability across the EU, each spatial data theme is described in a data specification. As mentioned on the INSPIRE website “The process for developing harmonized data specifications is designed to maximize the re-use of existing requirements and specifications, in order to minimize the burden for Member States’ organizations at the time of implementation. The consequence of this is that the process of developing Implementing Rules for interoperability of spatial datasets and services may be perceived as being complex: it involves a large number of stakeholders, with many interactions and consultations”.

Finally, all the data, information and services shared within INSPIRE would be accessible through the INSPIRE Community Geoportal. For Bernard et al. (2005) because the geoportal does not store or maintain data and metadata, it could be seen as a gateway aggregating a number of instances of specific
geospatial information services distributed across the Europe and maintained by the organization responsible for the data. According to the INSPIRE network architecture (European Commission 2008), Member States shall establish, operate and provide access to the following network services (figure 7):

- **discovery services**: support discovery of data, evaluation and use of spatial data and services through their metadata properties

- **view services**: as a minimum, display, navigate, zoom in/out, pan, or overlay spatial data sets and display legend information and any relevant content of metadata.

- **download services**: enabling copies of complete spatial data sets, or parts of such sets, to be downloaded.

- **transformation services**: enabling spatial data sets to be transformed (projection and harmonization).

- **invoke spatial data services**: enabling data services to be invoked.

![INSPIRE network architecture](image)

**Figure 7: INSPIRE network architecture (Source: INSPIRE).**

### 2.3.2 Global Earth Observation System of Systems (GEOSS)

Website: [http://www.earthobservations.org](http://www.earthobservations.org)

GEOSS is being established by the intergovernmental Group on Earth Observations (GEO) and is a worldwide effort to build a system of systems on the basis of a 10-Year Implementation Plan for the period 2005 to 2015 (GEO secretariat 2005; GEO secretariat 2005). GEO is voluntary partnership of governments and international organizations where membership and participation is contingent upon formal endorsement of the Implementation Plan mentioned above.

GEOSS is an effort to connect already existing SDIs and Earth Observations infrastructures and thus will not create and/or store data but rather works with and build upon existing systems. GEOSS, through its developing GEOportal, is foreseen to act as a gateway between producers of
environmental data and end-users, with the aim of enhancing the relevance of Earth observations for global issues and to offer a public access to comprehensive, near-real time data, information and analyses on the environment (GEO secretariat 2007).

GEOSS aims to provide a broad range of data and information for so-called Societal Benefits Areas (GEO secretariat 2005; GEO secretariat 2005):

1. Reducing loss of life and property from natural and human-induced disasters,
2. Understanding environmental factors affecting human health and well-being,
3. Improving the management of energy resources,
4. Understanding, assessing, predicting, mitigating, and adapting to climate variability and change,
5. Improving water resource management through better understanding of the water cycle,
6. Improving weather information, forecasting and warning,
7. Improving the management and protection of terrestrial, coastal and marine ecosystems,
8. Supporting sustainable agriculture and combating desertification, and

The mechanisms for data and information sharing and dissemination are presented and described in the 10-Year Implementation Plan Reference Document (GEO secretariat 2005) where information providers must accept and implement “a set of interoperability arrangements, including technical specifications for collecting, processing, storing, and dissemination shared data, metadata and products. GEOSS interoperability will be based on non-proprietary standards, with preference to formal international standards. Interoperability will be focused on interfaces, defining only how system components interface with each other and thereby minimizing any impact on affected systems”. GEOSS is based on existing technologies using satellite and internet-based services.

Moreover members must fully endorse the following data sharing principles:

1. There will be full and open exchange of data, metadata, and products shared within GEOSS, recognizing relevant international instruments and national policies and legislation.
2. All shared data, metadata, and products will be made available with minimum time delay and at minimum cost.
3. All shared data, metadata, and products being free of charge or no more than cost of reproduction will be encouraged for research and education.

These principles push data owners to go “open” and to share their data using standards and thus becoming interoperable.

### 2.3.3 United Nations Spatial Data Infrastructure (UNSDI)

Website: [http://www.ungiwg.org/unsdi.htm](http://www.ungiwg.org/unsdi.htm)

The United Nations Spatial Data Infrastructure in an initiative conducted by the United Nations Geographic Working Group (UNGIWG) that aims at building an institutional and technical mechanism to establish a coherent system to exchange data and services concerning geospatial data and information within
the United Nations, and also supporting SDI development activities in the Member Countries.

As stated in the UNSDI Compendium (Henricksen 2007), “Historically, the production and use of geospatial data have been accomplished within the United Nations by its component organizations, in accordance with their individual needs and expertise. But concordant with the recent, rapid increase in the use of geospatial data for UN activities is the need for greater coherence in its management system-wide”.

This initiative aims to contribute to the general mission of the United Nations to maintain peace and security, to address humanitarian emergencies, to assist sustainable development and support achievement of the UN Millennium Development Goals. The hope is to facilitate efficient access, exchange and utilization of georeferenced information in order to make the UN system more effective, increase the system coherence and support its “Delivering as One” policies.

The UNSDI provides an institutional and technical foundation of policies, interoperable standards procedures and guidelines that enable organizations and technologies to interact in a way that facilitates spatial discovery, evaluation and applications (Henricksen 2007).

2.3.4 Global Monitoring for the Environment and Security (GMES)


GMES is a European programme, coordinated by the European Commission and European Space Agency, for the implementation of a European capacity for Earth observation with the objective to monitor and better understand the environment and thus contribute to the security of every citizen. This initiative aims at providing decision-makers and other users who rely on strategic information with regard to environmental and security issues an autonomous, independent and permanent access to timely, reliable and accurate data and services.

The objective is to integrate data on atmosphere, oceans and continental/land processes giving an overview of the state of health of our planet and to deliver information through five thematic areas (served by different services) covering:

- land,
- marine,
- emergency,
- atmosphere,
- security.

allowing policy and decision-makers to prepare legislation (at different level) on environmental topics and to monitor the implementation of such laws.

To gather data and information on Earth Observation, GMES proposes to build an infrastructure around four components:

- space: environmental satellites
- in-situ measurements: ground-based and airborne sensors
- data harmonization and standardization
- services to users.
Chapter 2: Theoretical framework

Like various other data sources, Earth-observation-based services already exist in Europe but are dispersed and fragmented at national and regional level avoiding a sustainable observation capacity (meaning that long-term availability of information is not guaranteed). In consequence, GMES is the answer of the European Commission to develop a reliable and sustainable Earth Observation system and also contributing the GEOSS initiative.

GMES stated that “By securing the sustainability of an information infrastructure necessary to produce output information in the form of maps, datasets, reports, targeted alerts, etc…, GMES helps people and organisations to take action, make appropriate policy decisions and decide on necessary investments. GMES also represents a great potential for businesses in the services market, which will be able to make use of the data and information it provides according to a full and open access principle.” (GMES website).

2.3.5 Global Spatial Data Infrastructure (GSDI)

Website: http://www.gsdi.org

The mission of the GSDI Association, a world-wide inclusive body of organizations, agencies, firms and people, is to “promote international cooperation and collaboration in support of local, national, and international spatial data infrastructure developments that will allow nations to better address social, economic and environmental issues of pressing importance” (Nebert 2005). Its purpose is to focus on communication, education, scientific, research and partnership efforts to support all societal needs for access to and use of spatial data.

This is an association, guided by a board and funded by membership fees, to:

- serve as a point of contact and effective voice for those in the global community involved in developing, implementing and advancing spatial data infrastructure concepts,
- foster spatial data infrastructures that support sustainable social, economic, and environmental systems integrated from local to global scales, and
- promote the informed and responsible use of geographic information and spatial technologies for the benefit of society.

The GSDI community aims to truly develop and achieve the goal of a Global Spatial Data Infrastructure relying on international and open standards, guidelines and policies to enhance data management and access, and support global economic growth, and associated social and environmental objectives, through interoperable standards-based services, systems, software and products that operate in a web-based environment.

This vision is guided by five goals:

- Continue to promote and develop awareness and exchanges on infrastructure issues for all relevant levels from local to global.
- Promote and facilitate standards-based data access/discovery through the Internet.
- Promote, encourage, support, and conduct capacity building.
- Promote and conduct SDI development research.
- Collaborate with others to accomplish its Vision and Goals.
Chapter 2: Theoretical framework

To support this vision, the GSDI association acts as a platform and offers a vast choice of publications, conferences, workshops, projects and programs allowing people interested in SDI to learn, exchange, share their knowledge and expertise, because capacity building is one of the key points of SDIs.

2.4. Standards organizations relevant for GIS/SDI

In the field of geomatics there are several organizations involved in publishing standards to effectively achieve the goal of interoperability. Such standards are increasingly important in the geospatial community allowing the increase of interoperability between systems and data and thus to “geo-enable” the Web.

2.4.1 Open Geospatial Consortium (OGC)

Website: http://www.opengeospatial.org

The Open Geospatial Consortium (OGC) is a non-profit, international, voluntary consortium of more than 380 companies, government agencies and universities that is leading the development of standards for the geospatial community. Its approach is based on a member-driven consensus process to develop open and publicly available standards and software application programming interface for the geospatial community (Henricksen 2007). These standards offer to the developers the possibility to create complex georeferenced information and services accessible to a wide variety of applications and share data in a standardized and interoperable way.

OGC standards are based on a generalized architecture presented in the Abstract Specification and Reference Model (Open Geospatial Consortium 2008). On top of the Abstract Specification, there is a set of standards that have been developed and/or proposed to serve specific needs of the Geographical Information community in order to be interoperable.

These standards are mostly based upon the use of the HTTP protocol to interact through messages over the Internet. In last years, OGC members have been looking with interest to a more common approach used in the Service Oriented Architecture using Simple Access Object Protocol (SOAP) protocol and Web Service Description Language (WSDL) bindings. There is also work in progress around the Representational State Transfer (REST) protocol for web services.

The OGC is also closely working with the International Organization for Standardization (ISO) through a partnership with the ISO Technical Committee 211 (TC211) to promote and endorse the OGC standards to a higher level of standardization becoming part of the ISO 19100 series. For example, the WMS or the GML are now ISO standards.

The OGC vision is the realization of full societal, economic and scientific benefits of integrating electronic location resources into commercial and institutional processes worldwide. Its mission is to serve as a global forum for the collaboration of developers and users of spatial data products and services, and to advance the development of international standards for geospatial interoperability.
More specifically the OGC aims to (http://www.opengeospatial.org/ogc/vision):

- Provide free and openly available standards to the market, tangible value to Members, and measurable benefits to users.
- Lead the creation and establishment of standards that allow geospatial content and services to be seamlessly integrated into business and civic processes, the spatial web and enterprise computing.
- Facilitate the adoption of open, spatially enabled reference architectures in enterprise environments worldwide.
- Advance standards in support of the formation of new and innovative markets and applications for geospatial technologies.
- Accelerate market assimilation of interoperability research through collaborative consortium processes.

It must be noticed that the OGC and its members want to help users and developers to make usage of OGC’s standards offering them different resources (e.g., technical documents, training, best practices) through the OGC Network (http://www.ogcnetwork.net/).

2.4.2 International Organization for Standardization (ISO)

Website: http://www.iso.org

The International Organization for Standardization (ISO), the world’s largest developer, publisher and promoter of international standards, is a non-governmental organization made of a network of more than 160 countries representatives (one per country) with a central secretariat based in Geneva (Switzerland) that coordinates the system.

Even if the main focus of ISO is the development of technical standards, they have an important impact on the economy and the society because many members are coming from a governmental structure or from the private sector. Therefore, ISO is an ideal place to build consensus and solutions that meet the needs and requirements from both the economy and the society.

Within the ISO system there is a Technical committee (http://www.isotc211.org/) whose main area of interest are the Geographical information and the Geomatics aiming to establish a structured set of standards for information concerning objects or phenomena that are directly or indirectly associated with a location relative to the Earth.

Currently, they have more than 55 standards in the field of Geographical information specifying methods, tools and services for data management (including definition and description), acquisition, processing, analyzing, accessing, presenting and transferring such data in digital/electronic form between different users, systems and locations.

2.4.3 The World Wide Web Consortium (W3C)

Website: http://www.w3c.org

The World Wide Web Consortium (W3C) is an international consortium where members, organizations, staff and the public work together to create and develop Web standards, protocols and guidelines. Since 1994, the W3C has
published more that 110 Recommendations (the W3C standards) aiming to “
lead the World Wide Web to its full potential by developing protocols and
guidelines that ensure long-term growth for the Web”, accommodating
the growing diversity of people, hardware and software and ensuring the core
principles and components of these standards would be supported by everyone.
For the W3C it is crucial to reach the goal of the “web interoperability” allowing
the Web to reach its full potential by using technologies that must be compatible
with one another and allowing any hardware and/or software to access the Web.
By publishing open and non-proprietary standards, the W3C seeks to avoid
market fragmentation and thus Web fragmentation.

Therefore the W3C engages in education and outreach, develops software
and interoperable technologies that support this mission and acts as an open and
vendor-neutral forum for discussion about the Web.

2.4.4 Organization for the Advancement of Structured Information Standards
(OASIS)

Website: http://www.oasis-open.org/

The Organization for the Advancement of Structures Information Standards
is a non-profit, global consortium (with 5000 members coming from more than
600 organizations in 100 countries) that drives the development, convergence
and adoption of open standards for the global information society, the so-called
e-business. OASIS produces different web standards concerning the following
Chain, Computing Management, Application Focus, Document-Centric, XML
Processing, Conformance/Interoperability, and Industry Domains.

2.5 Standards description

ISO and OGC are providing a lot of different specifications regarding
geographical information. The general aim of these standards is to abstract data
delivery mechanisms from physical storage formats and offer services that could
be consumed by applications through different interfaces.

The OGC defines a general architecture for the geoportal (Open Geospatial
provides the basis for an open, vendor-neutral portal that is intended to be a first
point of discovery for geospatial content in the context of designing and
implementing the Spatial Data Infrastructures. The Geospatial Portal Reference
Architecture is founded on the tenants of a Service Oriented Architecture (SOA).
An SOA is an architecture that represents software functionality as discoverable
services on a network yielding the following benefits:
• Easier extension of legacy logic to work with new business functionality
• Greater flexibility to change without the need to constantly re-architect for
growth
• Cost savings by providing straightforward integration.
The Geospatial Portal Reference Architecture specifies also four services that are
needed for creating a interoperable and standardized geoportal (figure 8):
Chapter 2: Theoretical framework

Figure 8: Geospatial Portal Reference Architecture (Source: GeoNetwork).

- **Portal Services**: provide the single point access to the geospatial information on the portal. In addition, these services provide the management and administration of the portal.

- **Catalog Services**: used to locate geospatial services and information wherever it is located and provide information on the services and information to the user.

- **Portrayal Services**: used to process the geospatial information and prepare it for presentation to the user.

- **Data Services**: used to provide geospatial content and data processing.

To implement and deploy these different service classes, the OGC propose to use web services that are applications running on a computer connecting to a remote web service via a URL allowing access to distributed data and services. As stated by the CGDI “Web service architectures provide a distributed environment in which you can deploy and invoke services using standard Internet protocols. In this context, a service is a collection of operations, accessible through one or more interfaces, that allows you to evoke a behavior of value to you.” ([http://www.geoconnections.org/publications/Technical_Manual/html_e/s4_ch10.html#10.1](http://www.geoconnections.org/publications/Technical_Manual/html_e/s4_ch10.html#10.1))

Using such a Service Oriented Architecture (SOA) provides a distributed computing platform over a network, typically the Internet, allowing to publish standardized services no matter how it is implemented or on which platform it is executed. This is leveraging the full potential of the interoperability and thus web services to be seamlessly coupled, reusable and available for a variety of applications.

A traditional open web service must have the ability to describe its capabilities and provide a standard way to communicate with it, enabling applications and other web services to communicate and interoperate. Through OGC standards, different GIS softwares and/or components can interoperate,
work together and exchange information over a network by means of agreed standards. For example, when two softwares implement the same OGC standard, they can immediately work together without the necessity to develop new components to translate from one file format (used in one software) to another file format (used in a second software). This means that in a SOA environment that implements OGC standards, a user can access in a transparent way the data stored in different databases, with different formats, and running on different Operating Systems.

Without interoperability and standardization, accessing and integrating different data sources is really difficult or in the worst case impossible. This leads to a fragmentation of geospatial data sources and limit organizations to work only within a single software package.

2.5.1 Catalogue Service for the Web (CSW)


The Catalogue Service defines an interface to publish, discover, search and query metadata about geospatial data, services and related resources. CSW uses queryable properties, which enable clients to search for geospatial resources by subject, title, abstract, data format, data type, geographic extent, coordinate reference system, originator, publisher, purpose, etc... (figure 9)

![Figure 9: GeoNetwork, a catalogue system using CSW.](image)

The CSW interface provides the following operations:

- **GetCapabilities**: allows CSW clients to retrieve service metadata from a server. The response to a GetCapabilities request is an XML document containing service metadata about the server.
Chapter 2: Theoretical framework

- **DescribeRecord**: allows a client to discover elements of the information model supported by the target catalogue service.
- **GetRecordById**: retrieves the default representation of catalogue metadata records using their identifier.
- **GetRecords**: allows querying the catalogue metadata records specifying a query in OCG Filter or CQL languages.
- **Transaction**: defines an interface for creating, modifying and deleting catalogue records.

Example of CSW UTL with a GetCapabilities request:

### 2.5.2 Web Map Service (WMS)

OGC Web Map Service Specification:
http://www.opengeospatial.org/standards/wms

The Web Map Service defines an interface that allows a client to retrieve maps of georeferenced data. In WMS context, a map means a graphical representation (e.g., jpeg, gif or png files) of a geospatial data meaning that a WMS service does not give access to the data itself. It is used for mapping purposes and can be combined with other WMS services.

A traditional WMS interface, invoked by a URL, consists of the following operations:
- **GetCapabilities**: answer to a client telling him what kinds of layers are available and which one are queryable.
- **GetMap**: produce a map as a picture showing selected layers,
- **GetFeatureInfo**: answer simple queries about the content of the map

As seen on the following examples, invoking a WMS service need to specify different parameters (mandatory or optional) in the URL. For the purpose of this guideline we will focus our attention on the basic operations of the service that provides map layers in predefined styles (made available by the data provider) thus we will not discuss the Styled Layer Descriptor (SLD) capabilities.

Example of a WMS URL with a GetCapabilities request:
http://metafunctions.grid.unep.ch:8080/geoserver/ows?service=WMS&request=GetCapabilities&version=1.3.0

The GetCapabilities operation (figure 10) returns to users an XML document describing service and data sets available from which either desktop and/or web clients may request maps. This operation is common for all OWS and is presented in details in the OpenGIS Web Service Common Implementation Specification (Open Geospatial Consortium 2008). To invoke this operation, users have only to define service and request parameters.
Chapter 2: Theoretical framework

- <wfs:WFS_Capabilities version="1.1.0" xsi:schemaLocation="http://www.opengis.net/wfs http://preview.grid.unep.ch:8080/geoserver/schemas/wfs/1.1.0/wfs.xsd" updateSequence="202">
  - <ows:ServiceIdentification>
    - <ows:Title value="enviroSDE Web Feature Service"/>
    - <ows:Abstract>
      enviroSDE is the Spatial Data Infrastructure of the UNEP/DEWA/GRID-Europe (http://www.grid.unep.ch). This is the reference implementation of WFS 10.0 and WFS 1.1.0, supports all WFS operations including Transaction.
    </ows:Abstract>
  - <ows:Abstract/>
  - <ows:Title/>
  - <ows:Keywords>
    - <ows:Keyword value="enviroSDE"/>
    - <ows:Keyword value="UNEP"/>
    - <ows:Keyword value="GRID"/>
    - <ows:Keyword value="EUROPE"/>
    - <ows:Keyword value="WFS"/>
    - <ows:Keyword value="GEOSERVER"/>
  </ows:Keywords>
  - <ows:ServiceType value="WFS"/>
  - <ows:ServiceTypeVersion value="1.1.0"/>
  - <ows: Fees value="NONE"/>
  - <ows:AccessConstraints value="NONE"/>
</ows:ServiceIdentification>
- <ows:ServiceProvider>
  - <ows:ProviderName value="UNEP/DEWA/GRID-Europe"/>
  - <ows:Contact>
    - <ows:IndividualName value="Gregory Giuliani"/>
    - <ows:PositionName value="enviroSDE coordinate"/>
    - <ows:Voice value="+41 22 917 84 17"/>
</ows:ServiceProvider>

Figure 10: Example of the XML file returned after a GetCapabilities request.

Example of a WMS URL with a GeMap request:

The GetMap operation returns to a client request a map of selected geospatial layers (figure 11).
In comparison of a GetCapabilities request that needs only two parameters, we can see on above example that GetMap operation needs several parameters (also mandatory or optional) that we describe hereafter:

Mandatory parameters for the GetMap operation:

- **BBOX:** coordinates of the bounding box following minx, miny, maxx, maxy,
- **STYLES:** list of style names separated by a comma. It's necessary to have an exact correspondence between the number of layers and the number of styles. If this parameter has a empty value, the default style provided by the data custodian will be applied.
- **FORMAT:** graphical format of the returned map (eg: image/png, image/gif, image/jpeg).
- **REQUEST:** value “GetMap”, this is the request itself to invoke the specific operation.
- **VERSION:** the version of the specification.
- **LAYERS:** list of selected layers separated by a comma.
- **WIDTH:** specify the width of the returned map (in pixels).
Chapter 2: Theoretical framework

- HEIGHT: specify the height of the returned map (in pixels).
- SRS: identifier of the Spatial Reference System.

![Map with GetMap and GetFeatureInfo requests](image)

Figure 11: Returned image after a WMS request.

Example of a WMS URL with a GetFeatureInfo request:

http://webmapping.mgis.psu.edu/geoserver/wms?version=1.1.1&request=getfeatureinfo&layers=topp:states&styles=population&SRS=EPSG:4326&bbox=-125,24,-67,50&width=400&height=200&format=text/html&X=100&y=100&query_layers=topp:states

The GetFeatureInfo operation is used to query the attribute table of a selected layer and get information on a specific feature (figure 12). For example, a user can click on point of a map (retrieved by a GetMap request) and he obtains more information.

Mandatory parameters for the GetFeatureInfo operation:

- VERSION: the version of the specification.
- REQUEST: value “GetFeatureInfo”, this the request itself to invoke the specific operation.
- LAYERS: list of selected layers separated by a comma.
- STYLES: list of style names separated by a comma. It's necessary to have an exact correspondence between the number of layers and the number of styles. If this parameter has a empty value, the default style provided by the data custodian will be applied.
- SRS: identifier of the Spatial Reference System.
- BBOX: coordinates of the bounding box following minx,miny,maxx,maxy.
- FORMAT: the format of the returned information (text/xml, text/html, text/plain)
- XY: coordinates of the clicked point on the map (in pixels). The origin is at the upper left corner.
Chapter 2: Theoretical framework

- QUERY_LAYERS: list of selected layers to query separated by a comma.

![Image](http://webmapping.mgis.psu.edu/geosearch/wms?version=1.1.1&request=getCapabilities&service=WMS&bbox=-100,-100,100,100&layers=layer1,layer2)

Results for FeatureType 'states':

the_geom = [GEOOMETRY [MultiPolygon] with 153 points]
STATE_NAME = Arizona
STATE_FIPS = 04
SUB_REGION = Mtn
STATE_ABBR = AZ
LAND_KM = 294333.462
WATER_KM = 942.772
PERSONS = 3665228.0
FAMILIES = 940106.0
HOUSEHOLD = 1368653.0
MALE = 1810691.0
FEMALE = 1854637.0
WORKERS = 1358263.0
ORVALONE = 1178320.0
CARPOOL = 239085.0
PUBTRANS = 32856.0
EMPLOYED = 1603896.0
UNEMPLOY = 1239320.0
SERVICE = 455896.0
MANUAL = 185109.0
P_NALZ = 0.494
P_FEMALE = 0.506
SAMPOP = 446178.0

Figure 12: Result of a GetFeatureInfo query.

2.5.3 Web Feature Service (WFS)

OGC Web Feature Service specification:
http://www.opengeospatial.org/standards/wfs

The Web Feature Service defines an interface that allows a client to retrieve and update features of georeferenced data encoded in Geography Markup Language (GML). The main difference between WMS and WFS is that WFS gives direct access to the geometry and attributes of a selected geospatial data, meaning that a user can work with a dataset provided by WFS. In brief, WFS is the specification to access vector datasets. Similar to the WMS, a WFS interface is invoked by a URL and can perform a certain number of operations allowing a client to manipulate data. Following the type of operations needed to manipulate data we can define two classes of WFS services:

- Basic WFS: a client can retrieve and/or query features,
- Transactional WFS: a client can create, delete or update a feature.

A transaction is defined as one or more data manipulation operations that form a logical unit.

The concept of a geographic feature is described in the OGC Abstract Specification (Open Geospatial Consortium 2008) and the retrieved or created data are encoded in the Geographic Markup Language (Open Geospatial Consortium 2007).
Example of a basic WFS URL:
&styles=&request=GetFeature&version=1.0.0
&typename=preview:cy_buffers &srs=EPSG:4326

Like the WMS, WFS service is supported by a set of defined operations:

- **GetCapabilities**: answer to a client describing its capabilities. It tells the client which kind of features are available and what operations are supported on each feature.
- **DescribeFeatureType**: describe the structure of a selected feature (point, line, polygon).
- **GetFeature**: retrieve a selected feature encoded in GML. The client can constrain the query both spatially and non-spatially and also specify the feature properties to fetch.
- **Transaction**: this type of request is made of operations that allow a client to modify features: create, delete and/or update operations. In addition, a client can invoke the LockFeature, in order to be sure that only one user is updating a specific feature, avoiding the risk of multi-edition at the same time.

### 2.5.4 Web Coverage Service (WCS)

OGC Web Coverage Service specification:
http://www.opengeospatial.org/standards/wcs

Like the WFS allows a client to access vector datasets, Web Coverage Service allows a client to access raster datasets. By raster we mean data that are represented as a matrix of cells in continuous space organized in rows and columns where each cells contains a value. Thus WCS service provides access to different types of gridded data such as Digital Elevation Model (DEM), remote sensing imagery, etc... It must be noted that WCS gives only access to the raw data and does not have transactional capabilities.

Like all the OGC Web Services, a WCS interface consists of different operations:

- **GetCapabilities**: answer to a client describing its capabilities. It tells the client which kind of raster data (or coverage) are available.
- **DescribeCoverage**: describe the structure of a selected coverage.
- **GetCoverage**: retrieve the selected coverage.

Example of a WCS URL with a GetCapabilities request:
http://preview.grid.unep.ch:8080/geoserver/ows?service=WCS&request=GetCapabilities

The GetCapabilities operation returns to a client an XML document describing service and data sets available from which either desktop and/or web clients may request coverages.

To invoke the operation, users have only to define service and request parameters.
Example of a WCS URL with a DescribeCoverage request:
http://preview.grid.unep.ch:8080/geoserver/wcs?service=WCS&request=DescribeCoverage&version=1.0.0&coverage=preview:cy_frequency

The DescribeCoverage operation returns to a client an XML document describing selected coverages (figure 13). The information provided must be sufficient for a client to assess the fitness for use of the data. It gives different useful pieces of information such as the supported raster formats, supported SRS, supported interpolation methods, etc...

Mandatory parameters for the DescribeCoverage request:
- SERVICE: value “WCS”, this is the name of the invoked service.
- REQUEST: value “DescribeCoverage”, this is the request to invoke the specific operation.
- VERSION: the version of the specification.
- COVERAGE: list of selected coverages separated by a comma.

Example of a WCS URL with a GetCoverage request:
Chapter 2: Theoretical framework

The GetCoverage request returns to a client the requested raster data. The syntax and the parameters of the URL are similar to those used in a WMS GetMap request.

Mandatory parameters for the GetCoverage request:

- **BBOX**: coordinates of the bounding box following minx,miny,maxx,maxy
- **SERVICE**: value “WCS”, this is the name of the invoked service.
- **STYLES**: list of style names separated by a comma. It is necessary to have an exact correspondence between the number of layers and the number of styles. If this parameter has an empty value, the default style provided by the data custodian will be applied.
- **REQUEST**: value “GetCoverage”, the request to invoke the specific operation.
- **VERSION**: the version of the specification.
- **COVERAGE**: name of a single selected coverage.
- **WIDTH**: specify the width of the returned coverage (in pixels).
- **HEIGHT**: specify the height of the returned coverage (in pixels).
- **CRS**: identifier of the Coordinate Reference System.
- **FORMAT**: the desired format to be used for returning the coverage (eg: GeoTiff, ARCGRID, GTOP30,...)

If a request is validated then coverage is extracted (using the BBOX, FORMAT and the different parameters set in the URL) from the selected coverage and sent to the client. If the client is a web browser then users can download the coverage file. If the request is sent through a Desktop GIS client like ArcGIS then users gets the coverage directly into it.

### 2.5.5 Web Processing Service (WPS)


The two previous discussed standards are focusing on data accessibility: WFS allows a client to access vector data while WCS allows a client to retrieve raster data.

Now we need to extend our capabilities in order to process data available using the recently introduced Web Processing Service (Open Geospatial Consortium 2007) that provides access to processing and calculations on geospatial data. A WPS service can offer, through a network access, a vast variety of GIS functionalities ranging from a simple calculation to complex models. It acts as a sort of middleware between the client and the process that runs the calculations. It allows users to know which processes are available, to select the required input data and their formats, to create a model and run it, to manage processes (status, storage for the output, ...) and to return the output once computation is completed.

Like the others OWS, WPS specification includes a set of operations:

- **GetCapabilities**: answer to a client describing its capabilities. It tells the client which kinds of process are available.
- **DescribeProcess**: describe the parameters a selected process.
- **Execute**: execute a selected process.
Chapter 2: Theoretical framework

The WPS differs a bit from the others OWS because these operations can be invoked either by SOAP or the traditional http-get and http-post.

Example of a WPS URL with a GetCapabilities request:
http://localhost/wps/wps.py?service=WPS&request=GetCapabilities

The GetCapabilities operation returns to a client an XML document describing service and processes available for execution. To invoke the operation, users have only to define service and request parameters.

Example of a WPS URL with a DescribeProcess request:
http://localhost/wps/wps.py?service=WPS&request=DescribeProcess&version=1.0.0&identifier=soil_process

The DescribeProcess operation returns an XML document describing what are mandatory, optional and default parameters needed for a selected process, as well as data formats for inputs and outputs. Mandatory parameters for this operation:
- SERVICE: value "WPS", this is the name of the invoked service.
- REQUEST: value “DescribeProcess”, this is the request to invoke the specific operation.
- VERSION: the version of the specification.
- IDENTIFIER: the name of the selected process to run.

Example of a WPS URL with an Execute request:

The Execute operation allows a client to run a selected process using values entered by the client for required parameters (if needed) and references the datasets location. Once the process is completed, result is returned to the client as a new dataset.

2.5.6 Sensor Observation Service (SOS)

OGC Sensor Observation Service specification:
http://www.opengeospatial.org/standards/sos

The OGC Sensor Web Service provides an interface for managing sensors and retrieving data from them. This standard is part of the suite of standards called Sensor Web Enablement (SWE) enabling:
- discovery of sensors, processes, and observations
- tasking of sensors or models
- access to observations and related streams
- publish/subscribe capabilities for alerts
- robust sensory system and process descriptions.
2.5.7 ISO 19115/19139

ISO 19115, Geographic information – Metadata:

ISO 19139, Geographic information – Metadata- XML schema implementation:

The ISO 19115 standard defines (with more than 400 metadata elements, 20 core mandatory elements) how to describe a geospatial data and providing information about content, identification, quality, spatial and temporal extent, access and rights and spatial reference. This standard is applicable to:

- cataloguing of datasets, clearinghouse activities, and the full description of datasets,
- geographic datasets, dataset series, and individual geographic features and features properties.

The main applicability of ISO19115 is for digital data but its principles can be extended and applied to other forms of geospatial data such as maps, charts and textual documents as well as non-geographic data (Nogueras-Iso, Zarazaga-Soria et al. 2005).

The ISO 19139 standard complements ISO19115 by defining an XML encoding schema implementation specifying the metadata record format and may be used to describe, validate, and exchange geospatial metadata prepared in XML.

2.5.8 ISO 19119

ISO 19119, Geographic information – Services:

The ISO 19119 identifies and defines the architecture patterns for service interfaces used for geographic information, defines its relationship to the Open Systems Environment model, presents a geographic services taxonomy and a list of example geographic services placed in the services taxonomy. It also prescribes how to create a platform-neutral service specification, how to derive conform platform-specific service specifications, and provides guidelines for the selection and specification of geographic services from both platform-neutral and platform-specific perspectives. In other words, ISO 19119 specifies the form and content of the XML document that describes the capabilities of a geospatial web service (e.g., the answer to the GetCapabilities request).
Chapter 2: Theoretical framework

2.5.9 Keyhole Markup Language (KML)

OGC Keyhole Markup Language specification:
http://www.opengeospatial.org/standards/kml

The KML format is an XML based language schema for describing geographical objects in web-based viewer (the so-called geobrowser). It has been developed and popularized by the Google Earth application and due to its success was turned into an OGC standard.

2.5.10 Geographic Markup Language (GML)

OGC Geographic Markup Language specification:
http://www.opengeospatial.org/standards/gml

The GML is a very complete XML based language used to describe all geographical features and objects and provides a standard mean for representing geographical features (properties, relationships, geometries, ...). GML differs from KML as it is not only used for data visualization (which is the main focus of the KML specification) but serves also as a modeling language as well as an open and interoperable exchange format over the Internet. It is mostly used in the Web Feature Service to send geographical features between servers and clients.

The GML core schema does not contain definitions of features (because it is impossible to describe all features) but could be rather seen as a grammar providing means to define concrete features through GML Application Schemas that are created by users. Although GML is easily readable, it will be mostly used and generated by software or web services answering a specific request and then receiving the result as a GML dataset (figure 14).

The real advantage of using GML is that XML technologies are available meaning that information stored in a GML file could be easily shared with other information and then specialized application domains could reuse, extend and/or refine GML components in an application schema in order to develop a specific data model.
Chapter 2: Theoretical framework

2.6 Tools

After reviewing the set of OGC and ISO standards relevant for our purpose it is important to present a selection of tools that implement those standards, allowing users to publish standardized and interoperable web services. Note that none of these software integrate all standards at once. Instead each of them is a building block of a Spatial Data Infrastructure following the OGC Reference Architecture (Open Geospatial Consortium 2004; Open Geospatial Consortium 2008).

2.6.1 OGC web services

2.6.1.1 Mapserver

Website: [http://www.mapserver.org](http://www.mapserver.org)
Supported OS: Windows/Linux-Unix/Mac

MapServer is an open source geographic data rendering engine and development environment for building WebGIS applications and sharing data.
through OGC standards. It can run as a CGI program or via Mapscript which supports several programming languages. MapServer is now a project of OSGeo and is maintained by a growing number of developers from around the world.

Mapserver main features are:

- Advanced cartographic output: scale, labels, reference map, classification, ...
- Support for different scripting and development environments: PHP, Python, Perl, Java and .NET
- Cross-platform support: Linux, Windows, Mac, Solaris, ...
- OGC web services: WMS (client/server), WFS (client/server), WCS, GML, SLD, SOS, ...
- Multitude of raster and vector formats: GeoTiff, shp, PostGIS, ArcSDE, ...
- Map projection support: up to 1000 projections through the Proj.4 library.

2.6.1.2 Geoserver

Website: [http://www.geoserver.org](http://www.geoserver.org)
Supported OS: Windows/Linux-Unix/Mac

Geoserver is an open source server designed to publish data from different major data sources using OGC standards and allowing the users to share their data. Unlike Mapserver, Geoserver has no mapping capabilities and is only used for publishing data in an interoperable and standardized way. Geoserver is a community-driven project sponsored by OSGeo.

Geoserver main features are:

- Java-based.
- Support of WMS, WFS, WCS and KML specifications.
- Various raster and vector formats: PostGIS, Oracle spatial, ArcSDE, DB2, MySQL, shp, GeoTiff, ECW, MrSID and Jpeg2000.
- Production of: KML, GML, shp, GeoRSS, PDF, GeoJSON, JPEG, GIF, SVG and PNG.
- Editing capabilities using WFS-Transactional.
- Includes an OpenLayers client for previewing data layers.

2.6.1.3 Deegree

Website: [http://www.deegree.org/](http://www.deegree.org/)
Supported OS: Windows/Linux-Unix/Mac

Deegree is an open source framework, sponsored by OSGeo, offering the main building blocks of an SDI. Its entire architecture is developed around OGC and ISO standards. Deegree main features are:

- Java-based.
- Support of WMS, WFS, WCS, WPS and CSW.
Chapter 2: Theoretical framework

- Storage formats: PostGIS, Oracle, shp, GML, jpeg, gif, png, bmp, geotiff.
- Simplified installation and configuration.
- SLD support.
- Envisioned support of Sensor Observation Service (SOS) and Web Terrain Service / Web Perspective and View Service (WTS/WPVS).
- Security implementation using Web Authentication (WAS) and Web Security Service (WSS).
- Built-in web-geoportal.

2.6.1.4 PyWPS

Website: http://pywps.wald.intevation.org/
Supported OS: Linux-Unix

PyWPS in an implementation of the Web Processing Service specification. The great advantage of PyWPS is that it has been written with a native support of GRASS GIS software, meaning that accessing the GRASS modules via web interface should be really easy. Process can be written using either GRASS or other programs like R, GDAL or PROJ.

PyWPS main features are:
- Support of WPS specification.
- Simple configuration files.
- Method for custom process definition.
- Support for multiple WPS servers.
- Python based

2.6.1.5 52 north WPS

Website: http://52north.org/maven/project-sites/wps/52n-wps-site/
Supported OS: Windows/Linux-Unix/Mac

52north WPS is another implementation of the WPS specification that aims to create and design an extensible framework (with plug-in mechanism) for providing, orchestrating and executing processes as well as Grid computing on the internet.

52north WPS main feature are:
- Java-based
- Support of WPS specification.
- Pluggable framework for algorithms and XML data handling and processing frameworks
- Build up on robust libraries (JTS, geotools, xmlBeans, servlet API, derby)
- Supports full logging of service activity (exception handling, storing execution results, ..)
- Clients: basic implementation for accessing the WPS & plug-in for uDig and JUMP.
- WPS invocation: synchronous/asynchronous, http-get, SOAP, WSDL
- Supported data types: GeoTiff, ArcGrid, GML.
Chapter 2: Theoretical framework

2.6.1.6 ArcGIS Server


Supported OS: Windows

ArcGIS Server is part of the ArcGIS family and is the component to provide web-oriented and OGC standardized spatial data services. Since version 9.2, ArcGIS Server also includes the Spatial Data Engine (ArcSDE) that geo-enables databases and allows the user to store their data into popular database system like PostgreSQL or Oracle.

ArcGIS Server main features:

- .NET and Java frameworks.
- ArcGIS Server services can be consumed by web browsers, mobile devices and desktop clients.
- Full implementation of WMS, WFS (basic and transactional), WCS, KML specifications.
- Services: mapping, geocoding, geodata management, geoprocessing, virtual globes, network analysis.
- SOAP and REST API.
- Additional SDKs to build web applications: JavaScript, Flex, Silverlight, ...

2.6.2 Metadata editor and catalog system

2.6.2.1 GeoNetwork Open Source

Website: [http://geonetwork-opensource.org/](http://geonetwork-opensource.org/)

Supported OS: Windows/Linux-Unix/Mac

GeoNetwork is an open source project, sponsored by the UNSDI (Henricksen 2007) initiative and supported by several UN agencies (FAO, UNEP, OCHA and WFP) as well as the OSGeo. GeoNetwork implements both the Portal component and the Catalog database of a Spatial Data Infrastructure (SDI) defined in the OGC Reference Architecture (Open Geospatial Consortium 2004) allowing a user to search, discover, evaluate, publish, manage and edit metadata on spatial data and related services through the internet.

The main goal of GeoNetwork is to improve the accessibility and thus enhance the data exchange and sharing in a standardized and consistent way between the organizations to avoid duplication, increase the cooperation and coordination of efforts in collecting data and make them available to benefit everybody, saving resources and at the same time preserving data and information ownership.

Main features of GeoNetwork are:

- Instant search on local and distributed geospatial catalogues
- Support of CSW, Z39.50 and OAI protocols.
- Uploading and downloading of data, documents, PDF’s and any other content
- An interactive Web map viewer that combines Web Map Services from distributed servers around the world
- Online map layout generation and export in PDF format
- Online editing of metadata with a powerful template system
Chapter 2: Theoretical framework

- Scheduled harvesting and synchronization of metadata between distributed catalogues
- Groups and users management
- Fine grained access control.

2.6.2.2 ESRI Geoportal

Website: http://www.esri.com/software/arcgis/geoportal
Supported OS: Windows/Linux

ESRI Geoportal Server is a free open source product that enables discovery and use of geospatial resources including datasets, rasters, and Web services. It helps organizations manage and publish metadata for their geospatial resources to let users discover and connect to those resources. The Geoportal Server supports standards-based clearinghouse and metadata discovery applications. Key features of ESRI Geoportal are:

- Data Cataloging
- Geoportal Administration
- Data Publishing
- Data Discovery

2.6.3 Data storage

2.6.3.1 PostgreSQL/PostGIS

Website: http://www.postgresql.org/ & http://postgis.refractions.net/
Supported OS: Windows/Linux-Unix/Mac

PostgreSQL is a popular and powerful open-source relational database management system (RDBMS) allowing the user to store data and their relations in the form of tables. A RDBMS itself cannot store geographical information and thus to geo-enable a database system like PostgreSQL it is necessary to install a middleware to add support for geographic objects into the database. Two softwares are able to work in conjunction with PostgreSQL and add specific tables and functions that extend the capacities of a traditional RDBMS:

- PostGIS: follows the OGC Simple Features Specification for SQL (OGC, 200x).
- ESRI ArcSDE: that implements the powerful geodatabase system.

PostGIS main feature are:

- Geometry types for points, linestrings, polygons, multipoints, multilinestrings, multipolygons and geometrycollections.
- Spatial predicates for determining the interactions of geometries
- Spatial operators for determining geospatial measurements like area, distance, length and perimeter.
- Spatial operators for determining geospatial set operations, like union, difference, symmetric difference and buffers.
Chapter 2: Theoretical framework

- Powerful spatial indexes for high speed spatial querying.
- Index selectivity support, to provide high performance of query plans for mixed spatial/non-spatial queries.
- No support (for the moment) of raster data.

2.6.3.2 PostgreSQL/ArcSDE

Supported OS: Windows/Linux-Unix

ArcSDE is the second software that could “geo-enable” a RDBMS and implements the powerful concept of the geodatabase allowing to store vector and raster data in a central data repository for easy access and data management. It can be leveraged in desktop, server and mobile applications and is the common data storage and management system of the ArGIS family of softwares products.
In the latest version 9.3 of ArcGIS Server, ArcSDE in now a component of that and offers an integrated environment. Geospatial data is managed as a database accessible by the users using a desktop client and can be easily published on the internet. It allows query, mapping, analysis and editing in a multi-user environment.
ArcSDE main features are:
- Store a rich collection of spatial data in a centralized location.
- Apply sophisticated rules and relationships to the data.
- Define advanced geospatial relational models (e.g., topologies, networks).
- Maintain integrity of spatial data with a consistent, accurate database.
- Work within a multiuser access and editing environment.
- Integrate spatial data with other IT databases.
- Easily scale storage solution.
- Support custom features and behavior.
- Support DB2, Informix, SQL Server, Oracle and PostgreSQL.

2.6.3.3 File system

The simplest way to store data is probably under file system arborescence. The inconvenience is that the arborescence must be well structured and self-explainable in order to rapidly find the desired data. We do not recommend using file system because it is a complex and not efficient system to manage and maintain geographical data. The only advantage is that a user can see a small increase in performance when accessing data but this advantage disappears as soon as he works in a environment where there is a concurrent accesses.

2.6.4 Web Mapping

2.6.4.1 Open Layers

Website: [http://openlayers.org/](http://openlayers.org/)
Supported OS: Windows/Linux-Unix/Mac
Chapter 2: Theoretical framework

OpenLayers is an open-source JavaScript API for creating web-mapping application. Main features are:
• load map data from many sources: WMS, WFS, GeoRSS, ...
• Support for displaying geographic features, with markers and popups
• Easy mouse/keyboard navigation.
• Layers selection.
• Easy build configuration, designed to help build OpenLayers into other applications
• Javascript API to allow full control over OpenLayers-powered map from within Javascript on a web page.

2.6.4.2 Mapfish

Website: http://www.mapfish.org
Supported OS: Windows/Linux-Unix/Mac

MapFish is JavaScript API and web-mapping framework using the latest web 2.0 technologies and integrates different components like OpenLayers, ExtJS and GeoExt.

2.6.4.3 Google Maps

Website: http://code.google.com/apis/maps/
Supported OS: Windows/Linux-Unix/Mac

Google Maps and its JavaScript API are free services provided by Google and allow developers to embed Google Maps into their web pages using their own data. The API provides a number of functionalities for manipulating maps, adding content and allowing creating simple and robust maps applications.

2.6.4.4 Mapserver

As already discussed in the section 2.6.1.1 MapServer can produce OGC webservices but has also cartographic capabilities using different scripting languages like PHP, Perl or Python.

Chapter 3:
Is there a need to access and process environmental data in a better and efficient way?

Based on:
Chapter 3: Is there a need to access and process environmental data in a better and efficient way?

3.1 Preamble

Kofi Annan, the seventh secretary general of the United Nations and 2001 Nobel Peace Price said, “Information and knowledge are central to democracy and conditions for development.... Knowledge is power. Information is liberating. Education is the premise of progress, in every society, in every family”. With the emergence of globalized world and new information and communication technologies (ICT) like the Internet, means of communication have dramatically expanded allowing us to access an enormous and continuous flow of information strongly influencing our society, our economy, and how people live and work. ICTs have enhanced globally the productivity of people and allowing more individuals, industries, and countries to participate in the development of a knowledge-based society (OECD 2004).

These changes are obviously influencing and transforming the world of scientific research. Indeed, science is driven by data and new technologies have dramatically increased the easiness of data capture and therefore the amount of data collected (Hanson, Sugden et al. 2011). Two of the challenges that scientists are facing today are:

1. dealing with the tremendous amount, complexity, and variety of data that is currently being generated (Hanson, Sugden et al. 2011);
2. being able to take full advantage of the knowledge and information produced by scientists and researchers (Arzberger, Schroeder et al. 2004).

Our society is increasingly relying on science and technology advancements allowing improving health and well being, as well as informing on environmental threats and sustainability or managing resources. Moreover, researchers and scientific community accountability and transparency have been sometimes criticized (e.g., climategate) (Science staff 2011). Hence, better data collection, management, storage and access are essential elements to improve this situation. A recent study made by the journal Science polling 1700 scientists showed that more than 50% of the responders store their data only in their laboratories. This obviously impedes wide availability/easy accessibility to these data and this is not an ideal long-term storage solution (Science staff 2011). Additionally, the lack of common metadata and repositories are prohibiting factors for using and storing data and up-to 80% of respondents have no sufficient funding to support data archiving and management. For most of the polled scientists, benefiting from improved data access and organization allowing using and reusing data as much as possible, opens considerable opportunities. The possibility to integrate data sets together can potentially improve understanding of complex, multidimensional and interdependent systems, addressing key societal and environmental problems, and finally paving the way to new research areas (GEO secretariat 2007).

One of the most important advancement brought by ICTs it the emergence of the so-called e-science (Arzberger, Schroeder et al. 2004; Science staff 2011) that can be defined as “increased access, through desktop or other interface via the Internet, to distributed resources, global collaboration, and the intellectual, historical, analytical, and investigative output of a range of communities”. One of the “hot topics” in e-science concerns the access to data that mainly serve as input to research (Arzberger, Schroeder et al. 2004). Indeed data availability and
accessibility are essential in scientific research. Consequently, opening data access can be seen as a benefit in order to enable wide usage of them (both for scientists and the public in general) and allowing specifically scientists to compare their results and methods more easily and enhancing scientific accountability, credibility, and potentially improve quality of data. For Arzberger et al. (2004), due to the fact that research is becoming global, there is a clear need to address routinely data access and sharing. For these authors “ensuring research data are easily accessible, so that they can be used as often and as widely as possible, is a matter of sound stewardship of public resources”. These authors stated that publicly funded data are a public good, produced in the public interest and thus should be freely available to the maximum extent possible. Currently in some specific scientific communities and well-established networks (e.g., geosciences, meteorology) data are already shared using state of the art technologies but their impacts remain low being restricted within these communities.

In environmental sciences, “Observe, Share, Inform” are three steps needed to address complex (local to global) challenges arising from the growing pressure caused by climate change, biodiversity losses, exposure to natural hazards and other environmental threats impacting our societies and influencing various aspects of our everyday life (GEO secretariat 2010). People, when trying answer and address problems like resource scarcity, food insecurity, pollutions, biodiversity conservation, and energy efficiency need to take many (small to important) decisions every day. However, they are all regularly facing the same problem: they need to take sound decisions with partial information and consequently cannot manage efficiently and effectively what they cannot measure (UNEP, Multidisciplinary Networks to Integrate Environment into Development Processes). Therefore it is essential to gather and integrate the vast amount of environmental data generated on a daily basis but often operated in isolation. Earth observations are used and compiled to answer specific questions, to understand or explain a trend, to confirm or invalidate a thesis, to make predictions or support management processes (GEO secretariat 2007). Such data represent quantitative measurements corresponding to a value at a certain time, its changing rate, its spatial distribution, and many other attributes allowing describing it as accurately as possible.

The establishment and implementation of programs such as the Global Earth Observation System of Systems (GEOSS) reflect a growing commitment to share data more openly (McKee 2011). It also indicates increasing recognition of the potential benefits for informed decision-making from evaluation, access, integration and processing of various environmental, economical, statistical and other data sources within a common framework (GEO secretariat 2007). As stated by the Group on Earth Observation (2007): “If information is power, shared information is even more powerful – and the more effectively data can be shared, the greater and more widely distributed are the benefits”.

Interoperability is the essential condition to develop an open science framework allowing scientists and researchers to publish, discover, evaluate and access data. Current technologies are suitable to match these requirements only if open software interfaces and standards are developed allowing these technologies to interoperate at a global scale (McKee 2010). The Open Geospatial Consortium (OGC) aims to develop and provide such standards enabling
Chapter 3: Is there a need to access and process environmental data in a better and efficient way?

Communication and exchange of information between different systems of different types operated with different software. Indeed, a non-interoperable system cannot share data and computing resources, inducing scientists to spend much more time than necessary on data discovery and transformations. McKee (2010) summarizes eighteen reasons for which interoperability and related standards can significantly help scientists in their daily researches and contributing to elementary values of science: (1) data transparency, (2) verifiability, (3) useful unification of observations, (4) cross-disciplinary studies, (5) longitudinal studies, (6) re-use, (7) planning, (8) return on investment, (9) due diligence, (10) maximizing value, (11) data discoverability, (12) data exploration, (13) data fusion, (14) service chaining, (15) pace of science, (16) citizen science and outreach, (17) forward compatibility, (18) timely intervention.

Therefore, by committing to interoperability any system allows to widely and effectively exchange data, to maximize the value and reuse of data and information under its control. Enabling the exchange of data and information with other interoperable systems, allows new knowledge to emerge from relationships that were not envisioned previously (Open Geospatial Consortium 2004; Reed 2004). The ultimate objective is to spend less time in searching data and more time in doing science by easily integrating, analyzing data, and communicating newly generated information and results. The environmental challenges we are collectively facing today do not only require useful science but used science.

These different considerations were the starting point to write the first paper of this thesis called “Sharing environmental data through GEOSS” and published in the International Journal of Applied Geospatial Research. The aim of this paper was to present experiences gathered through different United Nations and European research projects, and to discuss promises and challenges envisioned in participating to GEOSS, both in term of data sharing and data processing (Rogers 2008).
3.2 Sharing environmental data through GEOSS

Gregory Giuliani\textsuperscript{1,2}, Nicolas Ray\textsuperscript{1,2}, Stefan Schwarzer\textsuperscript{2}, Andrea De Bono\textsuperscript{1,2}, Pascal Peduzzi\textsuperscript{2}, Hy Dao\textsuperscript{2,3}, Jaap Van Woerden\textsuperscript{2}, Ron Witt\textsuperscript{2}, Martin Beniston\textsuperscript{1}, Anthony Lehmann\textsuperscript{1,2}

\textsuperscript{1}University of Geneva, Institute of Environmental Sciences, Climate Change and Climate Impacts, enviroSPACE laboratory, Battelle - Building D, 7 route de Drize, 1227 Carouge, Switzerland.
\textsuperscript{3}University of Geneva, Institute of Environmental Sciences, Human Ecology Group, Battelle - Building D, 7 route de Drize, 1227 Carouge, Switzerland.

3.2.1 Abstract

Understanding the complexity of earth-system processes is crucial to convey improved information on the environment to decision-makers and the general public. Addressing this need by sharing environmental data is challenging because it requires a common agreed framework that allows easy and seamless integration of data from different sources. In this regard, the Global Earth Observation System of Systems (GEOSS) portends major benefits through various sharing mechanisms and by giving access to services that could be linked together to process and generate new understandable knowledge and information. Various United Nations projects could greatly benefit from the GEOSS approach.

Keywords:
GEOSS, SDI, data sharing, interoperability, data integration, service chaining, capacity building, grid computing.

3.2.2 Introduction

Today we are living in a globalized world with rapidly evolving processes including climate change, population growth or environmental degradation. In parallel, means of communication have expanded to take on a remarkable place in our society, allowing us to access an enormous and continuous flow of information.

In the last 30 years, the availability of geospatial data has grown dramatically following the evolution of communication technologies supported by the rapid development of spatial data capture means such as remote sensing imagery, sensors and GPS (Philips, Williamson et al. 1999). One of the challenges we are facing today is to make sense of this vast amount of data in order to turn them into understandable knowledge (Gore 1998). Concrete actions can be taken only on the basis of knowledge and understanding, but often we know too little about the state of our planet’s environment to take informed and sound decisions about how it should be managed.

Our planet is a multi-dimensional system made of complex interactions highly interconnected and continuously evolving at many spatial and temporal scales (GEO secretariat 2007). This means that to understand these interactions,
Chapter 3: Is there a need to access and process environmental data in a better and efficient way?

we need to gather and integrate different sets of data about physical, chemical and biological systems. Altogether, these sets of data constitute environmental data sets or data related to the environment. These data are often georeferenced, describing a geographical location through a set of attributes and thus could be understood as being part of geospatial data. An environmental data set is seldom interesting in itself, but rather displays its full information potential when used in conjunction with other data sets, allowing one to monitor and assess the actual status of the global, regional or local environments, to discover complex relationships between them and to model future changes.

In 1998, the former vice-president of the United States, Al Gore, presented his visionary concept of a Digital Earth (Gore 1998), a representation of the Earth embedding a vast amount of geospatial data and allowing to make better sense of it. To achieve this vision, Gore highlighted the need for a collaborative effort (from government, industry, academia and citizens) and pointed out the different technologies required: computational power, mass storage, satellite imagery, broadband network, interoperability and metadata.

Despite the fact that administrations and governments are recognizing that geospatial data are an important component of an information infrastructure (such as e-government) that needs to be efficiently coordinated and managed for the interest of all citizens (Ryttersgaard 2001), this huge amount of geospatial data is stored in different places, by different organizations and the vast majority of these data are not being used as effectively as they should. In consequence, a framework allowing one to discover, access, publish, share, maintain and integrate geospatial data appears to be essential. Such a framework is commonly known as a Spatial Data Infrastructure (SDI).

Different initiatives at the regional and global levels are influencing and promoting the creation of SDIs allowing data providers to share and publish their data in an interoperable manner. These initiatives coordinate actions that promote awareness and implementation of complementary policies, common standards and effective mechanisms for the development and availability of interoperable geospatial data and technologies to support decision making at all scales for multiple purposes. These initiatives are related to data access, harmonization, standardization, interoperability, seamless integration and services. Such an initiative is the Global Earth Observation System of Systems (GEOSS), which is a worldwide voluntary effort, coordinated by the Group on Earth Observation secretariat, to connect already existing SDIs and Earth Observation infrastructures. GEOSS is foreseen to act as a gateway between producers of geospatial data and end users, with the aim of enhancing the relevance of Earth observations for the global issues and offering public access to comprehensive information and analyses on the environment (GEO secretariat 2005; GEO secretariat 2007; GEO secretariat 2007). The GEOSS Common Infrastructure (GCI) provides core capabilities that allow users to search, access and use data, information, tools and services, and is made of five components: GEO portal (web portal to access GEOSS and search registries), GEOSS clearinghouse (connects the different components), GEOSS components and services registry (catalogue of services and components), GEOSS standards and interoperability registry (catalogue of standards to use allowing users to set up and configure an interoperable system), and a best practices wiki (offers a single space to share, discuss, propose and exchange ideas and best practices within the
Chapter 3: Is there a need to access and process environmental data in a better and efficient way?

...These components are dependent on the voluntary contributions of members and participating organizations. To support the nine defined Societal Benefit Areas (SBAs) (disasters, health, energy, climate, water, weather, ecosystems, agriculture, biodiversity), the mechanisms for data sharing and dissemination are presented in a 10-year Implementation Plan Reference Document (GEO secretariat 2005) providing data sharing principles that any volunteer member must endorse. The key element to share data through GEOSS is to agree on “interoperability arrangements” (GEO secretariat 2007) allowing different components of the system to communicate with each other.

Turning data into understandable knowledge requires that data coming from different sources be easily and seamlessly integrated. With the capabilities offered by standards like the one proposed by the Open Geospatial Consortium (OGC), geospatial community can not only discover, access and publish interoperable geospatial data but also services that can be linked together, in chains of services, to process data and generate new information. Moreover, by registering services into GEOSS, these different resources are now accessible in a standardized way and are reusable for many different purposes.

The aim of this paper is to present experiences gathered through different United Nations (UN) and European research projects and to discuss promises and challenges envisioned in participating to an initiative like GEOSS, both in term of building chains of services and sharing data.

3.2.3 The need for data sharing and integration

Until very recently, the different systems used to acquire environmental data were mostly operating in isolation, which made it difficult to easily discover, access and use the data content of these systems due to incompatibilities and inconsistencies of formats and data models (Bernard and Craglia 2005). In addition, there is typically insufficient data exchange among different stakeholders, which is partially due to differing data policies. Other important impediments to the flow of data are the delays in accessing data that prevent timely use of information, duplication and redundancy of data acquisition, potential high costs associated with data creation and access, and unclear access rights and licensing policies (GEO secretariat 2005). Altogether, these difficulties lead to a fragmentation of data sources, impeding their effective and efficient use, requiring much more time than necessary for data collection (Open Geospatial Consortium 2004). All the previous considerations highlight the growing need to share data in an interoperable way and to ensure that data are easily accessible and discoverable, so that they can be used as often and widely as possible (Arzberger, Schroeder et al. 2004). Moreover, the adoption of the Agenda 21 resolution, a United Nations initiative proposing a set of actions to be taken at different scales to promote a sustainable development, fostered the importance of geospatial data to support decision-making and management related to degradation and threats affecting the environment (Nebert 2005). Availability and access to appropriate information, and the related development of interoperable databases, are the necessary conditions for creating the basis for supporting the information management needs of implementing and monitoring sustainable development policies and goals, such as the United Nations Millennium Development Goals (MDGs) (Henricksen 2007). The MDGs
are eight development objectives (eradicate extreme poverty and hunger, achieve universal primary education, promote gender equality, reduce child mortality, improve maternal health, combat different diseases, ensure environmental sustainability, and develop and global partnership for development) that all UN members have agreed to achieve by 2015 (United Nations 2010).

Over the last twenty years, the emergence and evolution of Geographic Information Systems (GIS) technology and the advent of applications such as Google Earth or OpenStreetMap (Craglia, Goodchild et al. 2008), allowed for a clear change on how geospatial data are handled and incorporated into regular workflows of organizations and agencies in the governmental, private and public sectors (Booz, Allen et al. 2005). Highlighting these changes, Masser (2007) stated that to realize the full potential and benefits of geospatial data, access must be maximized with the help of Spatial Data Infrastructures (SDIs) that allow users to share, discover, visualize, evaluate and retrieve geospatial data. Moreover, the vast amount of data needed to run a complex model (e.g., in climatology or ecology), and the recognition that organizations and/or agencies need more data than they can afford financially (Rajabifard and Williamson 2001), reinforce the concept that once a particular set of geospatial data has been created, it should be accessible to potential users in both the public and private sectors (Rytersgaard 2001). This reinforces the need to store such data in databases that are made widely accessible for various purposes (Philips, Williamson et al. 1999). As a consequence, geospatial data can be seen as a shared resource, which is maintained continuously.

To remove the barriers that block and impede a wide use of geospatial data and related information, Masser (2005, 2007) identified different needs such as eliminating or reducing restrictions on data access and availability (but protecting intellectual property rights), promoting interoperability between different data sets and different systems, and disseminating the information about data (metadata). Altogether these objectives are designed to create an environment that fosters activities for using, managing, producing and sharing geospatial data in which all stakeholders can cooperate with each other and interact with technology, to better achieve their objectives at different political/institutional levels (Rajabifard, Mansourian et al. 2004). In this sense, interoperability appears to be a key element enhancing data sharing, communication and efficiency.

The great advantage of interoperability is that it describes the ability of locally managed and distributed heterogeneous systems (different operating systems, different databases, different data formats) to exchange data in real time to provide a service (Open Geospatial Consortium 2004). The shift towards a processed-based infrastructure offering reusable and standardized components responsive to user needs and requests is supported by the Service Oriented Architecture (SOA) concept. In a SOA, services are the elementary components representing a set of operations that could be invoked by users allowing them to access, in the case of the geospatial community, distributed geospatial data as well as geoprocessing services. To implement and deploy geoenabled services, the OGC proposes a suite of standards that use services over the Internet, so-called web services, giving access to distributed data and services through Uniform Resource Locators (URLs). This allows data providers
Chapter 2: Is there a need to access and process environmental data in a better and efficient way?

to publish standardized services independently on how it is implemented and on which platform it is executed. This emphasizes the full potential of interoperability allowing an organization to maximize the value and reusability of data under its control and giving the ability to exchange these data with other interoperable systems. Using such OGC web services offers the possibility to seamlessly couple and reuse them in a variety of applications. By chaining together a series of web services, users can perform a set of operations to process data whereby new knowledge emerges from relationships that were not envisioned before (Open Geospatial Consortium 2004). Granell et al. (2009) define service chaining as a mechanism for combining individual geospatial web services to create customized web applications. Although current SDIs mostly offer the abilities to search, view and access data, with the support of interoperable services and SOA related concepts it is now possible to build new applications based on distributed services (Friis-Christensen, Ostländer et al. 2007; Diaz, Granell et al. 2008). When services are organized through a coherent chain, combined services can achieve a larger task (Di 2004). The International Organization for Standardization (ISO) through its ISO 19119 standard defines three types of chaining:

- Transparent (user-defined): the workflow is defined and managed by users.
- Translucent (workflow-managed): users invoke a service that manages the chain. Users are aware of atomic services that constitute the chain.
- Opaque (aggregated): users invoke an aggregated service that carries out the chain. Users have no awareness of the atomic services that constitute the chain.

In this paper, we will focus on the transparent chaining either by hard coding or by using OGC Web Processing Service (WPS) specification (Open Geospatial Consortium 2007).

Through its online catalogue of registered services, GEOSS is an interesting and promising entry-point to discover and access services that could be integrated into service chaining process. It offers a framework to share data, expose them through interoperable services and allow the production and dissemination of timely and accurate data needed by decision makers and the public (GEO secretariat 2005).

3.2.4 Serving data into GEOSS

In 1985, the United Nations Environment Programme (UNEP)/Division of Early Warning and Assessments/GRID-Europe was founded as one of the first two centres of the Global Resource Information Database (GRID) network to support environmental decision-making within UNEP and the UN system as a whole, by generating and disseminating information about the state of the world’s environment in a timely and understandable manner. To provide reliable environmental assessments and early warnings, GRID-Europe specialized in handling and analyzing spatial and statistical data on environmental and natural resource issues through computerized GIS and remotely-sensed imagery. Over the years, GRID-Europe has compiled an archive of global, European and other geospatial databases as part of its information management function. The experience and in-house capabilities of GRID-Europe offer a great potential to
Chapter 3: Is there a need to access and process environmental data in a better and efficient way?

make geospatial and tabular databases compiled over the years available to a large array of users. Since its foundation, the Geneva office has received considerable support from Swiss and local authorities as well. This supporting was significantly reinforced, and GRID-Europe's institutional base broadened, with the signing of a "Partnership Agreement" between UNEP, the Federal Office for the Environment (FOEN) and the University of Geneva in June 1998.

GRID-Europe closely monitors developments in information technologies and examines their utility for environmental monitoring and policy formulation and thus is extending and developing its field of activities using SDIs. Moreover, the “Partnership Agreement” provides a major opportunity to work at different geographic scales ranging from global, to regional (Europe) and national (Swiss) and finally local (Geneva). Such specificity allows GRID-Europe to participate to different applied research projects funded either by the United Nations or the European Commission. A common ground for these projects is to serve and share data through the European Directive on Infrastructure for Spatial Information in the European Community (INSPIRE) (European Commission, 2007), the United Nations Spatial Data Infrastructure (UNSDI) (Henricksen, 2007), as well as GEOSS.

3.2.4.1 PREVIEW Global Risk Data Platform

The PREVIEW (Project of Risk Evaluation, Vulnerability, Information, and Early Warning) Global Risk Data Platform (http://preview.grid.unep.ch) is a collaborative effort of UNEP, United Nations Development Programme (UNDP/BCPR), United Nations International Strategy for Disaster Reduction (UNISDR) and the World Bank to share geospatial data on global risk from natural hazards. Users can freely visualize, download or extract data on past hazardous events, human and economical hazard exposure and risk from natural hazards. The platform covers nine types of natural hazard: tropical cyclones and related storm surges, drought, earthquakes, biomass fires, floods, landslides, tsunamis and volcanic eruptions. The collection of data is made via a wide range of partners. This geoportal was developed as a support to the 2009 Global Assessment Report on Disaster Risk Reduction (UNISDR 2009), replacing the previous PREVIEW platform initially designed by UNEP/GRID-Europe and already available since 2000. The new PREVIEW platform is fully compliant with the OGC Web Services (OWS) to access data using Web Map Service (WMS), Web Feature Service (WFS), Web Coverage Service (WCS), geo-enabled Really Simple Syndication (GeoRSS) or Keyhole Markup Language (KML) as well as metadata using Catalogue Service for the Web (CS-W).

3.2.4.2 GEO Data Portal

The GEO Data Portal (http://geodata.grid.unep.ch) is the authoritative source for data sets used by UNEP and its partners in the Global Environment Outlook (GEO) report and other integrated environment assessments. Its online database holds more than 550 different variables, as national, sub-regional, regional and global statistics or as geospatial data sets (maps), covering themes such as Freshwater, Population, Forests, Emissions, Climate, Disasters, Health and Gross Domestic Product (GDP). The data can be displayed and explored on-
Chapter 2: Is there a need to access and process environmental data in a better and efficient way?

the-fly through maps, graphs, data tables, downloaded in various popular formats, or copied and pasted into word processors. All information products in the GEO Data Portal can be accessed and used as web services as well. The retrieval of statistical and country-wide information has been enabled via a Simple Object Access Protocol (SOAP) connection; data from the database can be retrieved as maps via WMS or WFS; graphs can be displayed via a direct Uniform Resource Locator (URL) usage.

3.2.4.3 EnviroGRIDS

EnviroGRIDS (http://www.envirogrids.net) is a European research project that will last from 2009 until 2013 and is funded under the seventh framework programme (FP7). The Black Sea Catchment is largely following an ecologically unsustainable pathway based on inadequate resource management that could lead to severe environmental, social and economical problems, especially in a changing climate (WWF 2008). The aim of the project is to build capacities in the Black Sea region to use new international standards to gather, store, distribute, analyze, visualize and disseminate crucial information on past, present and future states of this region, in order to assess its sustainability and vulnerability. EnviroGRIDS objective is to federate and strengthen existing Observation Systems to address several GEOSS Societal Benefit Areas within a changing climate framework. The expected result will be a shared information system that operates on the boundary of scientific/technical partners, stakeholders and the public. It will contain early warning systems able to inform in advance decision-makers and the public about risks to human health, biodiversity and ecosystems integrity, agriculture production or energy supply caused by climatic, demographic and land cover changes on a 50-year time horizon. To achieve and support the enviroGRIDS vision and objectives, a grid-enabled Spatial Data Infrastructure (gSDI) is under construction. The aim of the gSDI is to host and analyze the data for the assessment of GEOSS Societal Benefit Areas, as well as the data produced within the project. These data must be gathered and stored in an organized form and accessible in an interoperable way on the grid infrastructure in order to provide a high performance and reliable access through standardized interfaces.

3.2.4.5 ACQWA

ACQWA (http://www.acqwa.ch) stands for Assessing Climate impacts on the Quantity and quality of Water. It is also a FP7 European research project lasting from 2008 until 2013. As the evidence for human induced climate change becomes clearer, so does the realization that its effects will have impacts on natural environment and socio-economic systems. Some regions are more vulnerable than others, both to physical changes and to the consequences for ways of life. According to the description of work, the project will assess the impacts of a changing climate on the quantity and quality of water in mountain regions which are particularly affected by rapidly rising temperatures, prolonged droughts and extreme precipitation. Modeling techniques will be used to project the influence of climatic change on the major determinants of river discharge at various time and space scales. Regional climate models will provide the essential
Chapter 3: Is there a need to access and process environmental data in a better and efficient way?

Information on shifting precipitation and temperature patterns. Snow, ice, and biosphere models will feed into hydrological models in order to assess the changes in seasonality, amount, and incidence of extreme events in various catchment areas. Environmental and socio-economic responses to changes in hydrological regimes will be analyzed in terms of hazards, aquatic ecosystems, hydropower, tourism, agriculture, and the health implications of changing water quality. Attention will also be devoted to the interactions between land use/cover changes, and changing or conflicting water resource demands. Adaptation and policy options will be elaborated on the basis of the results. The chain of processes involved in climatic, cryospheric and hydrologic models is complex because each process impacts on different compartments of human and natural systems. Different types of data covering various geographical regions are therefore necessary to build different sets of scenarios, which translates into substantial amount of data.

3.2.5 Technical comparison and common grounds

All these projects have in common that they already share (or will share in a near future) their data and metadata into the GCI. As a pre-requisite all the registered services have to be interoperable using mainly standards proposed by the OGC, but also other protocols like the Simple Access Object Protocol (SOAP). A short comparison of these different projects is given in Table 1, indicating which services are available, which software are used to publish these services, what are the types of models used to chain these layers, and finally what are the challenges and difficulties raised while integrating these services.

<table>
<thead>
<tr>
<th>Project name</th>
<th>enviroGRIDS</th>
<th>ACQWA</th>
<th>GEO Data Portal</th>
<th>PREVIEW</th>
</tr>
</thead>
<tbody>
<tr>
<td>Services</td>
<td>WMS, WFS, WCS, CS-W, KML, GeoRSS, WPS, grid services</td>
<td>WMS, WFS, WCS, CS-W, WCS, KML, GeoRSS</td>
<td>WMS, WFS, WCS, CS-W, WCS, CS-W, SOAP</td>
<td>WMS, WFS, WCS, CS-W, KML, GeoRSS</td>
</tr>
<tr>
<td>Software</td>
<td>GeoServer, ArcGIS Server, PyWPS, GeoNetwork, gLite</td>
<td>GeoServer, GeoNetwork, PyWPS</td>
<td>GeoServer, GeoNetwork, MapServer</td>
<td>GeoServer, GeoNetwork, MapServer</td>
</tr>
<tr>
<td>Type of models</td>
<td>- Hydrological models</td>
<td>- Snow cover mapping</td>
<td>- providing base layers (socio-economic, ...)</td>
<td>- providing base layers (events, risk, ...)</td>
</tr>
<tr>
<td>Challenges &amp; difficulties</td>
<td>- linking SDI and grid infrastructure</td>
<td>- capacity building</td>
<td>- data integration</td>
<td>- data integration</td>
</tr>
<tr>
<td></td>
<td>- capacity building</td>
<td>- data integration</td>
<td>- different standard implementation</td>
<td>- data/metadata harmonization</td>
</tr>
<tr>
<td></td>
<td>- authorization/ authentication</td>
<td>- portal integration</td>
<td>- capacity building</td>
<td>- capacity building</td>
</tr>
</tbody>
</table>

Table 3: Technical comparison of enviroGRIDS, ACQWA, GEO Data Portal and PREVIEW projects.
Chapter 3: Is there a need to access and process environmental data in a better and efficient way?

Most of these projects make use of free and open source software (PostgreSQL/PostGIS, MapServer, GeoServer, GeoNetwork and PyWPS) because it can ease the portability and replicability of tools developed. Indeed, many countries with low to moderate incomes are often also affected by natural hazards, environmental threats or degradation, and these countries are especially interested to manage and share their geospatial data using free and open sources software. Having tools readily available to be deployed in these countries is a strong incentive for capacity building, knowledge transfer, and sharing of expertise.

These projects are also strongly related to capacity building in order to enhance an ‘‘open and sharing spirit’’. It is necessary to show and prove the benefits of data sharing through appropriate examples, to communicate best practices as much as possible and to develop guidelines and policies. Altogether this will help to reach agreement and endorsement on the use of new standards. Such a participative approach will certainly stimulate data providers to be more ‘‘open’’ and in consequence to share their data. The different projects presented before will organize different workshops and develop various teaching material allowing participants, ranging from students to members of government, to learn how to use the specific applications to share large amount of data. Rajabifard and Williamson (2004) believe that building capacities is an important challenge for SDIs concepts to be accepted and adopted at a large extent. For these authors, the best way to reach this objective is to establish a long-term commitment to education and research: otherwise the SDI vision will remain unclear and unachievable. Through these projects, the objective is to build the capacity of scientists to share and document their data in order to strengthen existing observation systems, the capacity of decision-makers to use it, and the capacity of the general public to understand the important environmental, social and economic issues at stake.

Through simple data integrating scenarios (integration into other web portals or applications) GEO Data Portal and PREVIEW have pointed out different issues. Integrating some socio-economic data sets coming from the GEO Data Portal with natural hazards maps of the PREVIEW project to compute, for example, economical exposition of a country to a specific hazard, was impossible. This problem comes from the different implementations of OGC specifications between Mapserver (used by the GEO Data Portal) and Geoserver (used by PREVIEW). Indeed, it appears that Mapserver use an argument "MAP" that is not standardized and not recognized by all clients. This problem will be solved by migrating to Geoserver all the data services of the GEO Data Portal. Thus implementation of a same specification can differ from one software to another and can impend a consistent integration of services.

Another issue raised by data integration process was raised by the United Nations High Commissioner for Refugees (UNHCR) while trying to integrate WMS data coming from the PREVIEW project in order to identify area that are not suitable to install refugees camps. It appears that the only projection available was EPSG:4326 (Geographic) whilst UNHCR geoportal makes use of Google maps in EPSG:900913 (Spherical Mercator). This experience showed us that it is important, while publishing data services, to support at least the most frequent projection types. Geoserver supports natively all types of projections and it is easy to reproject on-the-fly data stored in another projection so that it
can be integrated with data with other projections. In addition, following the size of the data set, an important processing overload has been observed caused by the on-the-fly reprojection process. This can slow the service chain and impend and efficient data integration.

In the ACQWA project, a specific constraint is the important number of partners involved and their different scientific backgrounds (climatologists, hydrologists, glaciologists, ecologists). It is quite challenging to raise awareness on new tools and way of gathering and exchanging data without strongly influencing the way these different communities are working with geospatial data. For that reason, the aim is to concentrate on the promotion of GEOSS as an interesting and useful framework to handle and discover scientific data. Obviously, a dedicated geoportal is under development to register the main outputs of the ACQWA project into GEOSS using OGC web services. Nevertheless to show the benefits of working with interoperable services, we are currently developing a scenario to make estimation of snow cover from remote sensing imagery using data coming from the Moderate Resolution Imaging Spectrometer (MODIS) and Shuttle Radar Topography Mission (SRTM). Project partners that are currently working to produce such estimations are working with PCI Geomatica, doing all the process chain manually. Our objective is to help our partners by publishing a WPS geoprocessing service that allows them to automatize this analysis (figure 15). Once retrieved by FTP, MODIS images are saved on a server that store also SRTM tiles. All data are in EPSG:4326 and will be available using WCS standard published by Geoserver. Finally, the WPS service, currently under development using PyWPS, will implement the different steps to process the data. A major difficulty encountered until now is to “translate” the PCI functionalities by finding the equivalent in Geographic Resources Analysis Support System (GRASS) software. Indeed, PyWPS does not process data by itself and instead uses GRASS as a backend to access all the geoprocessing functionalities.

**Figure 15: Data sources and processing steps for a geoprocessing service estimating snow cover.**

Once the snow cover process is successfully achieved, our hope is to convince other communities within the project to benefit from such an approach and to develop other scenarios especially making use of climate data.

In the process of turning data into understandable information and
knowledge by chaining data services a new challenge has emerged. The ever-increasing spatial and temporal resolution of geospatial data are causing a tremendous increase in term of data volumes and the limits of the processing capacities of traditional GIS and SDI are being reached. With the advent of grid computing and the progressive deployment of large grid infrastructure projects (e.g., Enabling Grids for E-sciencE) many scientific disciplines now have access to sizable computing resources and new opportunities are emerging. For Foster et al. (2008) grid aims to federate resource sharing in a dynamic and distributed environment across a network allowing to access unused CPUs and storage space to all participating computers. Currently, SDIs are lacking processing power and should therefore be made interoperable with grid infrastructures, which are offering large storage and computing capacities. Recent studies (Di, Chen et al. 2003; Muresan, Pop et al. 2008) applied a successful approach to extend grid computing to the remote sensing community and to make OGC web services grid-enabled. Both studies considered that the grid has a great potential for the geospatial disciplines. Padeberg and Greve (2009) have identified several differences between OGC-compliant SDIs and grid infrastructures concerning service description, service interface, service state and security. In particular, grid infrastructures are based on SOAP messaging protocol to invoke operations and Web Service Description Language (WSDL) to describe services. OGC-compliant do not support neither SOAP nor WSDL, except WPS, and thus chaining geospatial services with grid services could be problematic. In addition, OGC standards do no provide any security mechanisms (authentication, encrypted communication between resources), which is a major concern in grid infrastructures. Finally, Di et al. (2003) showed that the current grid metadata catalog system is not good enough to answer the needs of the geospatial community, especially the requirements of the ISO19115 standard. All these differences must be overcome in order to allow traditional SDIs to benefit from the power of grid computing, and consequently to offer new services to GEOSS.

The main scientific and technological challenge of the enviroGRIDS project will be to link an SDI with a grid infrastructure to benefit from the processing capacities offered by grids. Indeed, WPS appears to be an adequate candidate to be grid-enabled because, first, it supports SOAP protocol and, second, geospatial community has a growing processing need that current SDIs cannot deliver. A grid-enabled SDI will allow users to model high-resolution hydrological models (e.g., Soil and Water Assessment Tool) of the Black Sea catchment under various climate, land cover and demographic scenarios. In order to develop such a gSDI to support the development of Black Sea portal functionalities, the different components of the enviroGRIDS architecture are currently being defined to highlight the main issues emerging from different conceptual and technological solutions (figure 16).
Chapter 2: Is there a need to access and process environmental data in a better and efficient way?

Figure 16: EnviroGRIDS grid-enabled SDI components supporting Black Sea portal.

These issues concern the choices of software components, data repositories, data management, grid-oriented processing, grid portal, and interoperability between SDIs and grid infrastructures. Although the use of grid-enabled web services to access data sets stored in the SDI will also be explored (Maué and Kiehle 2009), bridging architectural gaps between grids and SDIs remains very challenging (Padberg and Greve 2009) without extensions and customizations. For example, an important question concerns the location of geospatial data repositories: inside or outside the grid? The answer is not trivial and will greatly influence services and in particular chains of services to process data. In the one hand, being outside the grid, all OGC-compliant services functionalities remain the same and grid services are only used to process the data. On the other hand, being inside the grid, all OGC web services have to be modified to support grid environment, becoming grid-enabled. The latter would allow benefiting from all the advantages of the grid (security, replicability, scalability, storage and processing capacities) but would obviously require a lot of developments for adapting already existing SDIs. In consequence, an incremental development and implementation strategy will be developed taking into account different integration scenarios aiming to hide the complexity of the grid while preserving OGC interfaces.

3.2.6 Challenges and promises

From the experience acquired, or being acquired, through these different projects, it is obvious that many challenges remain either tangible (e.g., technology) or less tangible (e.g., culture, behavior). Nevertheless, it is critical to
Chapter 3: Is there a need to access and process environmental data in a better and efficient way?

overcome them in order to improve our knowledge, share our experience and attempt to strive towards a society that is better informed. Achieving the goal of sustainable development requires the integration of a large number of different data from various sources. Through agreed common standards and a clear political will, these data can be integrated in an interoperable way, leading to a new collaborative approach to decision-making.

Having environmental data in digital form allows easy storage and dissemination, facilitate data exchange and sharing, faster and easier update and corrections, ability to integrate data from multiple source (see Fig.1), and customization of products and services (Henricksen 2007). In this sense SDIs appear to be a good choice to encompass the sources, systems, network linkage standards and institutional issues involved in delivering geospatial data from many data sources to the widest possible group of potential users (Coleman, McLaughlin et al. 1997). The fact that, during the last years, multiple SDIs initiatives have been developed all around the world, ranging from local to regional levels, is a good sign. It appears that there is a growing recognition that geospatial data is a critical element underpinning decision making in many disciplines (Rajabifard and Williamson 2001) and as such needs to be effectively managed.

The SDI hierarchy model proposed by Rajabifard (2002) is composed of inter-connected SDIs developed at different levels (from local to global). Each SDI of a higher level is formed by the integration of data developed and made available by the lower level. Such a hierarchy can be approached though two views: on one hand, it is an umbrella in which the SDI at a higher level encompasses all SDI components from lower levels. On the other hand, it can be seen as the building blocks supporting the access of data needed by SDIs at higher levels. This hierarchy allows creating an environment in which users working at any level can rely on data from other levels and integrate data from different sources (Mohammadi and Rajabifard 2009). Such a hierarchy is clearly envisioned in the concept of the system of systems on which GEOSS relies, integrating systems together into an information highway which both links together environmental, socio-economic and institutional databases and provides a movement of information from local to global levels. For Masser (2006), the SDI hierarchy poses the challenge of multi-stakeholder participation in SDI implementation, because the bottom-up approach differs a lot from the top-down approach. The top-down vision, common in the SDI literature, emphasizes the need for standardization and uniformity while the bottom-up view stresses the importance of diversity and heterogeneity caused by the different needs of the various stakeholders. As a consequence, it is necessary to find a consensus ensuring sufficient standardization and uniformity while recognizing the diversity and heterogeneity of the different stakeholders acting at different levels. In particular, building a system of systems like GEOSS is highly dependent on a clear governance structure that is understandable and acceptable by the volunteer participants in order to develop a shared vision of the system and to allow users to feel a common sense of ownership (Masser 2007). As it is reminded in the Strategic Guidance document (GEO secretariat 2007), the success of GEOSS will depend on interoperability arrangements that data providers agree to endorse.

As a provider of environmental data, GRID-Europe is continuously facing
Chapter 3: Is there a need to access and process environmental data in a better and efficient way?

the challenge of encouraging data providers to go “open” and to share their data in an interoperable and OGC-compliant way. At present, technology is no longer a problem because solutions based on a variety of software can be proposed and/or developed depending on the requirements and the technical capabilities available. The most difficult task is to create an environment allowing wide agreement on data sharing principles. In this particular regard, the GEOSS “best practices wiki” could be of great benefit to help people promote sharing principles. A lesson learned from our experience is that once users can discover data they need, their most important preoccupation is to know what is the quality of the data they are going to access and whether they can trust this data. We are convinced that sharing data is an efficient way to eventually recognize whether this data is of sufficient quality. By submitting/exposing the data to the judgment of the broader community, one can know if it is useful or not. Through data sharing, one can also benefit from the interaction with end users by receiving feedbacks and then improve the data sets accordingly. Sharing data and participating to GEOSS can therefore contribute to the improvement of data, which in turn allows better information and eventually better decisions.

In the current climate of economic constraints, interoperability and standardization have never been so important because a non-interoperable system impedes the sharing of data, information and resources, which increases the risk for a system to fail in delivering its expected benefits and to remain unused (Open Geospatial Consortium 2004). Geospatial data can be an expensive and time-consuming resource to produce, and for this reason, it is of high importance to improve accessibility and availability and promote its reuse. Many decisions that organizations need to make depend on good quality and consistent data, readily available and accessible (Rajabifard and Williamson 2001). The process of reuse does not only concern the data itself, but also encompasses the capabilities, skills developed, invested effort and capital. This process allows an organization to share the costs of data, people, and technology, which help realize more rapid returns on investment. By reusing data, one can avoid duplication of efforts and expenses and enable users to save resources, time and effort when trying to acquire or maintain data sets (Rajabifard and Williamson 2001).

Percivall (2006) claimed that in a distributed environment, the help of open standards such as OGC can help scientists to rapidly find and evaluate a lot of different data sets and processing approaches, providing a flexible and cooperative environment that foster collaboration in the different scientific communities that work with geospatial data. Thus, organizing the workflows using standard-based web services could provide a great benefit in term of productivity to address the nine SBAs of GEOSS. OGC standards provide a solid ground for interoperability between services within distributed geoprocessing environment offered by SDIs (Friis-Christensen, Ostländer et al. 2007). In particular, the fact that these services can be reused and chained within other applications is a very useful aspect offering the opportunity to solve specific problems in a more flexible way than with stand-alone applications. Nevertheless, some performance issues can appear with services that need to access and move large amount of data. This can negatively impact the execution time of this service (e.g., huge overload in gathering necessary data) especially if this service is chained with other services.
Chapter 3: Is there a need to access and process environmental data in a better and efficient way?

Consequently, GEOSS represents a very promising and potentially powerful framework to share and expose data. In particular, the fact that a good governance structure is already in place allows a clear vision that can be easily shared and endorsed by the participants. The fact that participating to GEOSS is on a voluntary basis could be seen either as a great opportunity or as a risk. Indeed, the voluntary aspect poses the threat that only a few data providers join such an initiative and, as a consequence, the system could miss its objectives. Nevertheless, the growing number of components and services registered through GEOSS is a good sign for optimism. In particular, we think that international organizations such as UNEP could play a major role by paving the way toward a broader acceptance by similar organizations. The fact that GEOSS is based on distributed systems that can operate, evolve and be managed in a relative independence appears to be a good choice to find a consensus ensuring sufficient standardization and uniformity, while recognizing the diversity and heterogeneity of the different stakeholders. Finally, GEOSS offers a unique characteristic that justifies by itself its existence, which is the possibility to see emergent properties. For Béjar et al. (2009), this emergence is the main objective of a system of systems, where users perform functions that cannot be made with any single component. This means that such a system is more than the sum of its parts and offers the possibility to better understand the complex relationships between the different components of the Earth system.

3.2.7 Conclusions

Geospatial data is a critical element underpinning decision-making for many disciplines and is indispensable to make sound decisions at all levels, from global to local. Experiences from developed countries show that more than two-thirds of human decision-making are affected by spatially-referenced data (Ryttersgaard 2001). Even if the technology exists, organizations and agencies around the world are still spending billions of dollars every year to produce, manage and use geospatial data, but they still do not have the information they need to answer the challenges our world is facing (Rajabifard and Williamson 2001).

The web service model proposed by the OGC appears to be suitable to allow users to combine different services to solve a specific problem in a scalable and flexible way. Nevertheless, through simple examples of services chaining, we have highlighted different issues that could potentially impede an easy integration: problems with different implementation of a same specification, problems regarding different projections used in different web applications, overload caused by on-the-fly reprojection using large data sets. Moreover, working with different communities that are not necessarily aware of the possibilities offered by OGC web services could limit the diffusion of such approach outside the geospatial community. These communities need to be convinced, through simple examples, that working with chained services can bring benefits in their own working flows. Finally, grid computing appears to be a promising complement of traditional SDIs capabilities to build WPS services for processing large data sets. To achieve this objective, implementation of SOAP protocol into OGC specifications is a pre-requisite in order to allow the two types of infrastructures to communicate (interoperability) and to ease the combination
Chapter 3: Is there a need to access and process environmental data in a better and efficient way?

of OGC and grid services in efficient chains.

Ten years after, GEOSS could be seen as an initial step to achieve Gore’s vision, because the relevant technologies are available and there is growing recognition that countries can benefit both economically and environmentally from better access to data. GEOSS has the potential to support the achievement of sustainable development initiatives such as the UN Millennium Development Goals and to offer a unique framework to share data and collaborate for a better society. In this sense, organizations such as UNEP can act as a "catalyst", contributing to GEOSS, building capacities and ensuring that environmental data are easily accessible. This is a necessary step to ensure better-informed decision-making for the more sustainable development of our planet.
Chapter 3: Is there a need to access and process environmental data in a better and efficient way?

3.3 Summary and lessons learned

- Achieving the objective of a sustainable development requires the integration of a large number of different data from various sources.
- Geospatial data is a critical element underpinning decision-making.
- Having environmental data in digital form allows easy storage and dissemination, facilitate data exchange and sharing, faster and easier update and corrections, ability to integrate data from multiple sources.
- Geospatial data can be a shared resource, maintained continuously and made available to the widest possible group of (potential) users.
- Interoperability and standardization are crucial elements when sharing data.
- SDIs appear a good choice for delivering geospatial data from many sources to the widest possible audience.
- On a technical level, publishing services to access data, metadata, and processing facilities is not a problem anymore due to the various solutions to easily generating them.
- Identified technical difficulties mainly concern: data integration, data/metadata harmonization, processing large data sets, and capacity building.
- Capacity building is a key element for large adoption, acceptation and commitment to SDI concepts inside and outside the GI community (e.g. climatologists, hydrologists, ecologists). In particular, showing and proving the benefits of sharing interoperable data/metadata through appropriate examples will help to strengthen (1) existing observation systems, (2) capacities to decision-makers to use it, (3) capacities of the general public to understand important environmental, social and economical issues, (4) scientists to work with it.
- GEOSS and INSPIRE offer a cooperative framework to share data and have the potential to foster collaboration in different scientific communities to support the achievement of sustainable development initiatives and better understanding of Earth processes.
- A unique characteristic that justifies sharing data is that users can perform functions that cannot be made with any single component. This offers the possibility to better understand the complex relationships between the different components of the Earth system.
Chapter 4:
How can SDI improve our capacity to discover, share, retrieve and integrate environmental data?

Based on:

4.1 Preamble

In chapter 3, we have seen that there is a clear need to manage, access and process environmental data in a much more effective and efficient way. Therefore, the next question we may ask is how well SDI allows discovering, sharing, retrieving and integrating environmental data.

Geospatial data are essential to make sound and timely decisions on the environment at all scales, from local to global (Nebert 2005). Governments and institutions must ensure that these data are available to those who need it (GEO secretariat 2005). These data are handle in general using Geographical Information Systems (GIS). Masser (2007) defines GIS as computer systems for capturing, managing, integrating, manipulating, analyzing, and displaying geospatial data. GIS software were available since 1960s, but it is during the 1980s with the advent of personal computers sufficiently powerful to handle the amount of data used into these systems that GIS gained considerable success and became widely used within different communities. Since that time, storage and computational capacities have dramatically increased while the success of the Internet allowed moving large volume of data between locations.

However, geospatial data are often expensive and time consuming to produce, and for this reason suitable information and resources to make use of that information are not always available especially in developing countries (Rajabifard and Williamson 2001; Nebert 2005). Even in countries characterized by sufficient (financial and human) resources, geospatial data are not effectively and efficiently used. This is mainly due to political and institutional barriers (e.g., copyright, licensing), as well as social habits of the past (e.g., not used to share data with other sectors, difficulties in exporting data from one system to another) (Nebert 2005; Masser 2007). As a result, suitable data are hard or costly to access, difficult to interpret, may be of inadequate quality, fragmented (i.e., partial data set), redundant (i.e., recreating/capturing an existing data set), and difficult to exchange due to incompatibilities. Moreover, this can cause delays in data access and prevent the timely use of information that can potentially save lives or minimize losses. Finally, opportunities to reuse already existing data are missed (GEO secretariat 2005). Additionally, it is frequent that organizations need more data than they can afford, especially when outside their area of authority. The Group on Earth Observations (GEO) stated in its 10 Year implementation plan (2005) that “despite laudable efforts in some domains, the current situation with respect to the availability of data is far from satisfactory, particularly in terms of coordination and data sharing between countries, organizations and disciplines, to meet the need of sustainable development”. Therefore, the ability to make sound decisions is dependent on a suitable framework that allows discovering, accessing, integrating and using geospatial data in a compatible way. This framework must (1) ensure that data coverage is comprehensive, (2) promote interoperability between data sets and systems, (3) reduce restrictions on data access and availability and protecting intellectual property rights, and (4) disseminate metadata allowing to know which data are available and where (Masser 2007). Nebert (2005) recognizes that only through commonly agreed conventions and technical agreements data will be easily discoverable, accessible and usable enhancing and supporting decisions-making processes.
Chapter 4: How can SDI improve our capacity to discover, share, retrieve and integrate environmental data?

The collection of relevant technologies, policies and institutional arrangements creating this suitable framework is commonly known as Spatial Data Infrastructure (SDI). An SDI can be thought as the infrastructure facilitating access and use to geospatial data, and allowing users (e.g., government, commercial sector, non-profit sector, academia, citizens) to benefit from geographical information. With an SDI all stakeholders can co-operate to achieve their own objectives (Rajabifard and Williamson 2001). The term “infrastructure” is of particular importance as it denotes the idea of providing an enabling and reliable environment that comprises the basic functionalities, services and equipment to fulfill the needs of a community of users. It is important to state that SDIs are much more than data repositories. They allow storing data but also metadata (e.g., data documentation) and offer means to search, discover, evaluate, visualize, access and integrate data. Beyond these technological aspects, in order to make an SDI functional, it is required to reach sufficient organizational agreements to coordinate and manage this infrastructure at various institutional scales (Ryttersgaard 2001; Nebert 2005).

For Rajabifard and Williamson (2001), SDIs must be part of basic infrastructures, such as roads or electricity networks, in order to support sustainable development fostering environmental management, social stability and economical development. To achieve the objective of supporting decision-making and sustainable development processes, it is essential to recognize the importance played by individuals. Nevertheless, Masser (2006) poses the challenge of a multistakeholder participation in SDI implementation because the bottom-up vision differs a lot from the top-down approach. The top-down approach emphasizes the need for standardization and uniformity while the bottom-up stresses the importance of diversity and heterogeneity due to the different aspirations of the various stakeholders. In consequence, it is necessary to find a consensus to ensure some measure of standardization and uniformity while recognizing the diversity and the heterogeneity of the different stakeholders performing different tasks at different levels. Consequently six key factors can be defined to realize the benefits brought by SDIs (Rajabifard and Williamson 2001): (1) awareness of the potential and advantages of GI and SDIs among all stakeholders, (2) cooperation and partnerships between all involved stakeholders, (3) involvement of concerned politicians, (4) knowledge about who’s got what and where (i.e. metadata), (5) providing access to data, and (6) widespread use of data sets. All these factors will contribute to justify the benefits of establishing a SDI to “enable users to save resources, time and effort when trying to acquire new data sets by avoiding duplication of expenses associated with the generation and maintenance of data and their integration with other datasets” (Rajabifard and Williamson 2004). However, it is important to denote that people, organizational structures, procedures, and cultural behaviors have a more important influence on the operational effectiveness of SDI than technological factors like software and standards (Henricksen 2007). Rajabifard et al. (2001) found that the first generation of SDIs, based on a product model, gave away to a second generation at the beginning of the years 2000, the latest being characterized by a process model. Indeed, the first generation of SDIs were product-based, aiming to link existing and future databases, while the second generation aimed to define a framework to facilitate the management of information assets allowing reuse of collected data by a wide range of people.
Chapter 4: How can SDI improve our capacity to discover, share, retrieve and integrate environmental data?

and/or organizations for a great diversity of purposes at various times. For Masser (2005) this evolution emphasis the shift from the concerns of data producers to those of data users, and the shift from centralized structures to decentralized and distributed networks like the web. Increasing usage and availability of web services is therefore an indicator of this shift.

In summary, geospatial data is the “fuel upon which the analytical power of GIS depends”, and with the support of SDIs to facilitate the exchange and integration of scientific and socio-economical data, new opportunities are emerging to monitor and assess our environment, allowing better informed decisions on anthropogenic and natural systems (Henricksen 2007).

Recognizing that SDI can be a suitable framework to share, retrieve and integrate environmental data, the second paper of this thesis entitled “The PREVIEW Global Risk Data Platform: a geoportal to serve and share global data on risk to natural hazards”, published in Natural Hazards and Earth System Sciences, aims to demonstrate the applicability of SDI concepts and methods targeting specifically the Disaster Risk Reduction community. Indeed, timely access and easy integration of geospatial data are essential to support efforts in reducing disaster risk and promoting a culture of disaster resilience. The development of such a platform has highlighted benefits and raised different issues. In particular, computational needs to process large data sets and efficient access to geospatial data through OGC services are two factors that will strongly influence the future success of SDIs implementation. Ensuring user satisfaction through sufficiently responsive services giving access to vector and raster data sets require to measure and monitor them to track latencies, bottlenecks and errors. Consequently the third paper, “OGC Web Feature and Web Coverage Services performance testing: towards an efficient access to geospatial data”, submitted to the International Journal of Digital Earth, gives some insights on this topic, presenting an approach to measure performances of different WFS and WCS services. It provides also some guidance to data providers for improving the quality of their services.
4.2 The PREVIEW Global Risk Data Platform: A geoportal to serve and share global data on risk to natural hazards

G. Giuliani1,2,3 and P. Peduzzi1,3,4

[1]{United Nations Environment Programme, Division of Early Warning and Assessment, Global Resource Information Database – Europe, Geneva, Switzerland}
[2]{University of Geneva, Institute of Environmental Sciences, Climatic Change and Climate Impacts, enviroSPACE, Geneva, Switzerland}
[3]{University of Geneva, Forel Institute, Geneva, Switzerland}
[4]{Institute of Geomatics and Risk Analysis (IGAR), University of Lausanne, Switzerland}

4.2.1 Abstract

With growing world population and concentration in urban and coastal areas, the exposure to natural hazards is increasing and results in higher risk of human and economic losses. Improving the identification of areas, population and assets potentially exposed to natural hazards is essential to reduce the consequences of such events. Disaster risk is a function of hazard, exposure and vulnerability. Modelling risk at the global level requires accessing and processing a large number of data, from numerous collaborating centres.

These data need to be easily updated, and there is a need for centralizing access to this information as well as simplifying its use for non GIS specialists. The Hyogo Framework for Action provides the mandate for data sharing, so that governments and international development agencies can take appropriate decision for disaster risk reduction.

Timely access and easy integration of geospatial data are essential to support efforts in Disaster Risk Reduction. However various issues in data availability, accessibility and integration limit the use of such data. In consequence, a framework that facilitate sharing and exchange of geospatial data on natural hazards should improve decision-making process. The PREVIEW Global Risk Data Platform is a highly interactive web-based GIS portal supported by a Spatial Data Infrastructure that offers free and interoperable access to more than 60 global data sets on nine types of natural hazards (tropical cyclones and related storm surges, drought, earthquakes, biomass fires, floods, landslides, tsunamis and volcanic eruptions) and related exposure and risk. This application portrays an easy-to-use online interactive mapping interface so that users can easily work with it and seamlessly integrate data in their own data flow using fully compliant OGC Web Services (OWS).

4.2.2 Introduction

The Hyogo Framework for action, priority two, states that “The starting point for reducing disaster risk and for promoting a culture of disaster resilience lies in the knowledge of the hazards and the physical, social, economic and environmental vulnerabilities to disasters [...] followed by action taken on the basis of that knowledge.” Parts of this priority consist in “producing risk assessment and maps, producing indicators on Disaster Risk Reduction (DRR) and vulnerability, data sharing” (UNDP/BCPR 2004). Moreover, the Agenda 21 resolution fosters the importance of having readily accessible, available, and
exchangeable appropriate information as a condition to create the basis of a sustainable development that supports information management needs to implement and monitor related policies and goals such as the Millennium Development Goals (Nebert 2005).

This means that there is a clear mandate to make data easily available and accessible in order to give users and stakeholders the opportunity to turn data into useable and understandable information. Achieving the objectives of both sustainable development and DRR requires the integration of a large number of different data types coming from various sources. Through agreed common standards and a clear political will, these data can be interchanged and integrated in an interoperable way, leading to a new collaborative approach in decision-making.

Although administrations and governments recognize that spatial information is important and must be effectively managed and coordinated for the interest of all citizens (Ryttersgaard 2001), geospatial data are often stored in different formats, based on different units and projections, located in different places, and managed by different organizations with different policies. In consequence, these issues impede an efficient and effective use of these data (Rajabifard, Mansourian et al. 2004; Alinia and Delavara 2009).

Spatial Data Infrastructures (SDIs) appear to be promising frameworks to facilitate interaction between users and geospatial data (Rajabifard, Mansourian et al. 2004; Alinia and Delavara 2009) allowing one to share and exchange required information on disasters. Mansourian et al. (2006) consider Geographical Information Systems (GIS) as a tool that facilitates decision-making and disaster risk reduction, allowing to capture, manage, integrate, manipulate, analyze and visualize geospatial data made available by SDIs (Masser 2007). In particular, web-based GIS portals, commonly known as geoportals, could be appropriate tools due to their high interactivity and accessibility, acting as gateways to relevant geospatial information (Tang and Selwood 2005).

In 1999, at the end of the International Decade for Disaster Risk Reduction (IDNDR), United Nations Environment Programme/Global Resource Information Database-Europe (UNEP/GRID-Europe) initiated the Project for Risk Evaluation, Vulnerability, Information and Early Warning (PREVIEW). One component of this project was a standalone web-GIS, called PREVIEW-Internet Map Server (IMS), an application aiming at visualizing data through an “easy to use” web interface and to disseminate data in common GIS formats (shapefile, grids2) with static zip files. It was based on the first generation of interactive mapping, developed with ESRI MapObject, (HTML and Visual Basic code). This application met with a good success, allowing users to access these data in a simple way; it ran without problem from August 2000 to May 2009.

Between 2001 and 2004, UNEP/GRID-Europe developed the Disaster Risk Index (DRI) for United Nations Development Programme/Bureau of Crisis Prevention and Recovery (UNDP/BCPR) (UNDP/BCPR 2004). This extensive study was the first to compute exposure and risk for four natural hazards (cyclones, droughts, floods and earthquakes) at the global level. These data sets were made available for visualization and download through PREVIEW-IMS and

2. http://resources.arcgis.com/content/kbase/fa-articleShow&d=30616
Chapter 4: How can SDI improve our capacity to discover, share, retrieve and integrate environmental data?

were used by different institutions such as the UN Office Coordination for Humanitarian Affairs (OCHA). Some of the data from PREVIEW (volcanic eruption and tropical cyclones) were also used by the World Bank for their study “Hotspots”.

With the availability of new data and more powerful computational capacity, a major improvement of these datasets was carry on by United Nations Environment Programme (UNEP/GRID-Europe), United Nations International Strategy for Disaster Reduction (UNISDR), Norwegian Geotechnical Institute (NGI) amongst others. This include the generation of new data sets for hazards distribution (cyclones, droughts, earthquakes, floods, biomass fires, landslides and tsunamis) with different intensities, computation of the human and economical exposures and computation of risk for cyclones, floods, earthquakes and landslides. All these data were included into a web-based GIS portal supported by a SDI that gives free and interoperable access to global data sets on natural hazards and related exposure and risk. These developments were made for the 2009 Global Assessment Report on Disaster Risk Reduction (UNISDR 2009).

This application is currently supported by UNEP and UNISDR. The data included in this platform is the result of a two-year research effort. These UN organizations have the mandate and the willingness to share this information in a free and open way, so that anybody may access data and hopefully contribute to its improvement.

This paper aims to describe the development of the web-based geoportal called PREVIEW Global Risk Data Platform (replacing PREVIEW-IMS) together with its associated SDI to support and facilitate geospatial data discovery, accessibility, visualization, and dissemination regarding past hazardous events, human and economical exposure, as well as risk maps from natural hazards. This application can be accessed at Http://preview.grid.unep.ch.

4.2.3 SDI and its role for Disaster Risk Reduction community

The Disaster Risk Reduction community works through several organizations/participants during its whole cycle of activities ranging from emergency units (e.g., fire, medical and police) up to government bodies. In an emergency situation, people involved need timely, reliable and up-to-date information in order to provide an efficient and effective response (Rajabifard, Mansourian et al. 2004; Mansourian, Rajabifard et al. 2006) while at the same time, returning new information on a specific situation. In other words, they are users as well as producers/updaters of important information. Moreover, because of the diversity of participants and the variety of information required for a disaster response, no single organization could collect and maintain all the required data sets (Mansourian, Rajabifard et al. 2006). Thus, recognizing that once data has been produced it could be used by different stakeholders (Ryttersgaard 2001) reinforced the need to store data into databases that are made largely accessible and available for different purposes (Philips, Williamson et al. 1999). This leads to the concept that geospatial data could be a shared resource that will be maintained continuously. In summary, having geospatial data in digital form allows easy storage into databases and file systems, facilitates data exchange/sharing, faster updates, gives the ability to integrate
data from multiple sources, and finally favours developing customized products and services (Nebert 2005).

From the previous considerations it could be argued that a collaborative environment based on the concept of partnership in data production, management, and integration would bring major benefits (Mansourian, Rajabifard et al. 2006; Alinia and Delavara 2009). As a result, the concept of SDI seems to be an interesting framework to facilitate and coordinate the exchange and sharing of geospatial data (Rajabifard and Williamson 2001; Masser 2007) encompassing data sources, systems, network linkages, standards and institutional issues in delivering geospatial data and information, from many different sources to the widest possible group of potential users (Coleman, McLaughlin et al. 1997). For Alinia and Delavar (2009) such a framework where people and geospatial data could interact offers an appropriate support for decision-making and disaster response objectives through improving availability, accessibility and applicability of geospatial data. SDIs intend to avoid duplication of efforts and expenses by enabling users to save resources, and time when trying to acquire or maintain data sets (Mansourian, Rajabifard et al. 2004). Finally SDIs can be seen as an integrated information highway which links together environmental, socio-economic and institutional geospatial data resources providing a movement of data from local to national and global levels (Masser 2005). Successful implementations of SDI frameworks for disaster management purposes have already been tested in different case studies such as wild fire risk assessment used by the Italian civil protection (Mazzetti, Nativi et al. 2009), evacuation scenario after a bomb threat (Weiser and Zipf 2007) or disaster management in Iran (Mansourian, Rajabifard et al. 2005; Mansourian, Rajabifard et al. 2006). All these authors consider and demonstrate that SDIs associated with tools like GIS and/or web-based services have a great potential in helping disaster community to effectively and efficiently manage disaster response, and allowing decision-makers to have permanent access to reliable and up-to-date geospatial data.

4.2.4 The PREVIEW Global Risk Data Platform

As part of Global Assessment Report on Disaster Risk Reduction3 (GAR) 2009 activities, the PREVIEW Global Risk Data Platform (http://preview.grid.unep.ch) was developed to allow different users to visualize risk distribution and provide easy access to the geospatial data sets on global risk from natural hazards.

With the initiation of the GAR process in 2007, it was clear that PREVIEW-IMS had to be updated and extended using recent advancements in geospatial technology. In the meantime, SDI concept has matured and was endorsed in different countries as well as in different communities dealing with geospatial data (Masser 2007). In consequence, SDI appeared to be a rationale choice to extend and reinforce capabilities of PREVIEW. Sharing all scientific data produced during the project was a primary objective since its beginning by making them visible to the largest possible audience and to promote the concept of data reusability. Moreover, to emphasize this “open and sharing spirit”

---

3 http://www.preventionweb.net/english/hyogo/gar/
**Chapter 4: How can SDI improve our capacity to discover, share, retrieve and integrate environmental data?**

PREVIEW was entirely developed using free and open source software, with them aim of facilitating its replication at national and/or local levels in developing countries. This choice was also justified by the fact that most countries experiencing high risk from natural hazards have medium to low human development and in general are looking with interest to free and open source technologies. In addition the Policies of the United Nations System Organizations Towards the Use of Open Source Software (OSS) for Development” JIU/REP/2005/7, stated that "OSS has been recognized in many instances to be a valid alternative to corresponding proprietary software, such recognition should be reflected in Member States’ ICT policies for development.” [...] "The open standards-based notional information architecture envisaged by UNEP is a vendor-neutral, interoperable platform that needs to work effectively with the services run by its partners inside and outside the UN." (Henricksen 2007).

In consequence, having such an open source platform readily available to be deployed in these countries is a strong incentive for capacity building, knowledge transfer, and sharing of expertise.
As already mentioned earlier, and highlighted by Asante et al. (2006), integrated geospatial data are extremely valuable and data integration may contribute to adoption of long-range actions to manage risk. The GAR 2009 report focused its attention on disaster risk patterns, trends and drivers at global scale allowing identifying geographical distribution and concentration of risk across countries (UN, 2009). In particular having means to easily communicate risk information to exposed communities will improve the different phases of disaster management. In these regards, the PREVIEW concentrates on visualization and dissemination of such data targeting more specifically prevention and awareness rising.

Considering SDIs as a collaborative environment that fosters activities for using, managing, producing and sharing geospatial data in which different stakeholders can co-operate and interact with technology to achieve their own objectives (Rajabifard and Williamson 2001), the first step is to define a conceptual model to clearly identify the different components involved. Indeed, Nebert (2005) reported that the existence of geospatial data does not alone ensure that it is used for decision-making. Other factors must be considered so that data will be effectively used. Users need to know that a specific data set exists and where to obtain it, they need to be authorized to access and use it, and finally they need to know its history allowing them to interpret it, trust it and integrate it meaningfully with data coming from other sources. Rajabifard (2002) identifies five core components (people, data, standards, policies and accessing network) divided into two categories depending on the nature of their interactions within SDI environment. People and data constitute the first category while the second is composed of technological components. This vision highlights the dynamic nature of these categories: evolving technology, changing data requirements, improved data sets, changes within user communities, and changes in policies. Altogether these different components interact, influencing each other’s, and evolving continuously. Describing these different components (figure 17) will ensure a good and consistent implementation of SDI addressing different issues (e.g., standards to use, policies/agreements and communication means) facilitating the relation between people and data (Mansourian, Rajabifard et al. 2005), enhancing availability, accessibility and usage of
Chapter 4: How can SDI improve our capacity to discover, share, retrieve and integrate environmental data?

geospatial data to support and improve decision-making process (Snoeren, Zlatanova et al. 2007).

4.2.5 PREVIEW SDI conceptual model

The developed PREVIEW SDI conceptual model is based in part on the experience and knowledge acquired through the first generation of the PREVIEW application. In particular, People and Data components were well known.

![PREVIEW SDI conceptual model](image)

**Figure 17:** PREVIEW SDI conceptual model (adapted from Mansourian et al., 2006; Rajabifard, 2002)

### 4.2.5.1 People component

Two categories (data providers and end-users) of the People component were clearly identified: data providers are the network of collaborating centres of the GAR and PREVIEW projects that give access to their data. End-users are already-known and/or are targeted users that could potentially use data produced by the project. This corresponds to United Nations development agencies (UNISDR, UNDP, World Bank), humanitarian action (e.g., OCHA, Red Cross) or specialized agencies (e.g., UNEP, World Meteorological Organization (WMO), United Nations Educational, Scientific and Cultural Organization (UNESCO), World Health Organization (WHO), Food and Agriculture Organization (FAO), United Nations High Commissioner for Refugees (UNHCR)), as well as decision-makers and governments that wish to have good and consistent data to take sound decisions. Finally such data are obviously useful for education in general and for scientists specialized in natural hazards and related risk in particular.

Currently most of the platforms that share and disseminate data on natural hazards use static download for a webpage or in the best case provide a File Transfer Protocol (FTP) access. Such capabilities are exemplified by the Hotspots project from the Center for Hazards and Risk Research (http://www.ldeo.columbia.edu/chrr/research/hotspots/). Moreover they
provide often access to hazard maps through PDF or image files. This is a common way within UN agencies to disseminate maps in this way like ReliefWeb from OCHA (http://www.reliefweb.int/) and UNOSAT (http://unosat.web.cern.ch/unosat/). This is certainly useful for people that are working on the field but clearly impede an interoperable access to data and create difficulties in updating information. Nevertheless there are a lot of web applications that allow visualization of data on hazards: some are concerned by a specific hazard type at a specific scale like the European Flood Alert System (EFAS, http://floods.jrc.ec.europa.eu/efas-flood-forecasts), some concentrate on the access to real-time information like the Global Disaster Alert and Coordination System (GDACS, http://www.gdacs.org/), and some are at the global scale but only for a few hazards types and without possibilities to access/download the data like National Geophysical Data Center (NGDC, http://www.ngdc.noaa.gov/hazard/) or the United States Geological Survey (USGS) Geologic Hazards Science Center (https://geohazards.cr.usgs.gov/). Finally, some platforms focus their attention on acquiring and disseminating data at the sub-national/local level. A good example of such a platform is DesInventar (http://www.desinventar.org) that provides users different web tools to visualize (maps, graphs, tables), query and analyze data. In summary, most of the current platforms concentrate on data visualization on specific hazard type and/or at a specific scale with limited capabilities to download/share data and metadata in an interoperable way.

In regard to the previous considerations and to our knowledge the unique value offered by PREVIEW is that it provides users an interactive and almost full access to harmonized, interoperable and fully documented data sets on multiple hazards at the global scale.

### 4.2.5.2 Data component

Data shared by PREVIEW are of two types: recorded past events in vectorial format (point, line or polygons) and results of the models used to compute hazards frequency, severity, human/economical exposures and risk in raster format (array of grid cells containing the value of the attribute to map). In absence of a commonly agreed data models within disaster community, one of the first step of the project was to develop a common set of specifications for each data category (e.g., cyclones, earthquakes, floods). After the collection of events data sets, a general assessment was conducted to evaluate their differences in term of metadata, data formats, data attributes, projections, terminology, and naming conventions. At this stage, it has been observed that these data sets were very heterogeneous. Much of them were not documented with metadata, projections were different, attributes names and content were very different from one data set to another, and finally sometimes reference systems were missing. Nevertheless, all data providers give access to their data in a common format (shapefile). Hence, a general conceptual data model was defined to harmonize these data sets that will further be used in the different models to compute frequency, exposures and risk. After several discussions with both data providers and end-users an agreement on the following points has been reached: (1) all data sets must have metadata using ISO19115/19139, (2) working at the global scale, the World Geodetic System (WGS) 84 (EPSG 4326)
Chapter 4: How can SDI improve our capacity to discover, share, retrieve and integrate environmental data?

has been used as reference system, (3) a common spatial resolution of 1km appears to be a reasonable choice (because of the resolution of input data sets, such as population distribution, and the fact that output data are at global scale), (4) a common temporal resolution for events data sets (ranging from 1970 to 2009), (5) a glossary has been developed defining unambiguously different terms (e.g., cyclones, exposure, risk), (6) a set of general attributes has been defined for both input data sets (events) and results of models (frequency, exposures and risk maps), (7) a naming convention for data, attributes and versions has been created allowing an easy identification of the hazard type, data format, category (e.g., events, frequency, exposure or risk), and version, (8) finally, in order to overcome the problem of data formats, it has been decided to share and disseminate the data using interoperable data formats through OGC web services. All these different items were documented and available among data producers. Reaching such level of agreement has allowed to ease the quality check both for completeness and accuracy, to update and harmonize input data sets (and create their metadata), to create a database accordingly specific constraints, to ease the production of new data from the different analytical models, and to ensure that all data producers use the same norms.

Data storage is in general achieved using database systems that allow a better management and control, and therefore will enhance data production and availability. It is important to note that following the choice of the database system, required functionalities and desired performance, it would be better to store raster data in a traditional file system.

Finally, data access are given through a web-GIS portal, along with different kind of web services that enable users to retrieve data in interoperable way, using both web-based clients or desktop clients (like ArcGIS4).

4.2.5.3 Standards component

Once People and Data components have been defined, technological components were clarified in order to facilitate the interaction between users and data. The Standards component consists of four identified categories (interoperability, metadata, data quality, guides and specifications) aiming to ensure interoperable access to good quality data and metadata. In particular, interoperability is an essential task allowing different systems or components to exchange data through a common agreed system and to use data that has been exchanged. When systems are interoperable, it gives users the ability to find what they need, access it, understand it, employ it and finally have tools and services responsive to their needs. As an answer to the need for interoperability, the Open Geospatial Consortium5 (OGC) has specified a suite of standards that allows interoperable access either to data or metadata. This allows users to retrieve, use and integrate geospatial data coming from different sources and stored in different formats using HTTP protocol to communicate. Web Feature Service (WFS) specification (Open Geospatial Consortium 2005) gives access to vector data sets, while Web Coverage Service (WCS) gives access to raster data (Open Geospatial Consortium 2006). In addition, Web Map Service (WMS) (Open

4 http://www.esri.com/software/arcgis
5 http://www.opengeospatial.org
Chapter 4: How can SDI improve our capacity to discover, share, retrieve and integrate environmental data?

Geospatial Consortium 2006) offers the ability to users to access georeferenced images of vectorial or raster data useful to display maps on the Web. To document data and related services through metadata, the International Organization for Standardization (ISO) 19115 (resource metadata) and 19139 (metadata encoding) standards are widely used. In addition, the OGC provides a Catalog Service for the Web (CSW) specification (Open Geospatial Consortium 2007) that complement these ISO standards, defining an interface to publish, discover, search and query metadata in an interoperable way. Metadata allows users to understand data quality, their history, interpret them, trust them and integrate them meaningfully with data coming from other sources (Nogueras-Iso, Zarazaga-Soria et al. 2005). In consequence, quality standards and guides also appears important in order to ensure that data are produced with consistent and documented procedures (allowing replicability of methodologies) and output data are of sufficient quality to be reliable (Mansourian, Rajabifard et al. 2006). Indeed, data quality is an essential element in any geospatial implementation or application ensuring users to obtain meaningful results. For PREVIEW, four categories have been identified: (1) data completeness (amount of missing features), (2) data precision (degree of details), (3) data accuracy (degree to which data reflects correctly the real object) and (4) data consistency (usability of the data). These categories are assessed at two levels. The data producer that checks the quality of data based on given data specifications makes the first level. The second level is based on the users side that provide feedbacks that are taken into account to update/correct data.

4.2.5.4 Network component

From a technical point of view, the network is a critical element allowing people to effectively use data. This component facilitates the access to different geospatial data resources through access networks and interoperable services for cataloguing, searching, visualizing, and downloading geospatial data. In the case of PREVIEW, data should be accessible through the Internet and thus Local Area Network (LAN) is an important building block for that need. The different set of services offered by the platform is another important element. They have been defined in the early stage of the development together with the suite of software used to publish data and metadata. In that sense, these services are basic components of a Service Oriented Architecture (SOA) that allow efficient access to spatially distributed data. The aim of an SOA is to promote loosely coupled, standard-based, protocol-independent distributed computing so that services can be reused as often as possible. These services are well-defined set of actions, self-contained, stateless, and does not depend on the state of other services. To support reusable deployment of services, OGC web services are based on the publish/find/bind pattern (Open Geospatial Consortium 2004). In a

\[\text{http://www.iso.org/iso/catalogue_detail.htm?csnumber=26020}\]

\[\text{http://www.iso.org/iso/catalogue_detail.htm?csnumber=32557}\]

\[\text{http://www.iso.org/iso/catalogue_detail.htm?csnumber=39890} \]
Chapter 4: How can SDI improve our capacity to discover, share, retrieve and integrate environmental data?

simple scenario, a service provider hosts a web service and “publishes” a service description to a service broker. The service requester uses a “find” operation to retrieve the service description and uses it to “bind” with the service provider and invoke the web service itself. Such approach, based on reusable and standardized services, allows application development to be more focused by providing users just the functionality they need. Finally, performances/quality of services and in particular response time to a query is essential as users need to access data in a timely manner. For that purpose, each data set has been tested under varying load and requests conditions in order to tune services so that they can deliver a good quality of service under high workload conditions.

4.2.5.5 Policies component

Finally, the Policies component is probably the most sensitive one as it must facilitate and encourage organizations to participate in the development of a SDI. Building a SDI is not only a matter of technology but first it relies on individuals and/or organizations. For Masser (2005), building a SDI is a long term process that depends on support and commitment. Moreover, because of its multi-stakeholder nature, it is necessary to find a consensus to ensure measures of standardization and uniformity while recognizing the diversity and heterogeneity of the different participants that have different requirements to perform different tasks (Masser 2006). As figure 17 shows, different categories are involved to fully describe this component. First of all, it is important to mention that building an efficient SDI is almost impossible without partnership because a single agency is unlikely to have all resources, skills or knowledge to undertake the development of all aspects of a SDI (Henricksen 2007). In the case of PREVIEW, financial support was shared between UNISDR, UNEP and UNDP/BCPR. The authors (UNEP/GRID-Europe) did the scientific development of the SDI and the content was produced by UNEP/GRID-Europe with inputs from other scientific institutions such as Norwegian Geotechnical Institute (NGI) for landslides and tsunamis hazard modeling, and Columbia University for drought hazard modelling. The process also benefited from inputs of the United States Geological Survey (USGS) on earthquakes, Dartmouth Flood Observatory for identification of areas affected by past flood events, and European Space Agency (ESA) for fires detection (high temperature events). Finally UNESCO and WMO provided a reviewing process by 24 independent experts, who reviewed the methodologies for modeling the hazards. The methodologies for the development of the datasets and the results are not within the scope of this paper and will be described in a forthcoming paper.

It was particularly important to reach agreements with data providers and partner’s agencies to develop a clear, coordinated, and committed vision to use, share and disseminate geospatial data (and related metadata) that were produced. This obviously covers issues like copyright, pricing and rights that influence directly access and use of data. In order to enhance the “open spirit” that guided the development of PREVIEW, almost all data sets on this platform are freely available and accessible. However, some institutions in the network still need to be convinced about providing free access to their data. The usual issue is financial: they might be willing to provide free access for research, while selling data to profit making institutions (e.g., re-insurance companies). Some
data may also be only temporary protected (e.g., prior to publication). Finally, in other cases, the data providers agree to freely distribute their data, but prefer to do so through their own website instead of through the PREVIEW application. This is usually for visibility purpose and justification to financial support. In total, only four datasets (floods, tsunamis and volcanoes events; earthquakes shakemaps) out of the 60 layers provided by PREVIEW are not freely accessible/downloadable. Nevertheless they can be visualized and their metadata are available. Regarding metadata, data providers and partners agencies agree that they must be all freely available as they are essential to discover data and associated services. Hence, to be sorted in the PREVIEW SDI, a data set must be documented with metadata (mandatory requirement).

We believe that previously mentioned situations are part of the transition phase that any new technology has to face. The sustained use of the platform should show the benefits to share data and reinforces the need to build new capacities by showing appropriate examples, sharing experiences and developing guidelines and policies. As the Group on Earth Observations (GEO) secretariat stated (GEO secretariat 2006), capacity building should be made at three levels: human (education and training of individuals), infrastructure (installing/configuring/managing of the needed technology) and institutional (enhancing the understanding within organization and governments of the value of geospatial data to support decision-making). All these actions will help to reach endorsement on the use of such technologies, raising and increasing awareness on the benefits of using and sharing geospatial data, and finally creating new commitments. In addition, cultural and political aspects may influence either positively or negatively the acceptance and adoption of SDIs as a framework. Finally, different international initiatives at the regional (INSPIRE) (European Commission 2007) and global scale (GEOSS, UNSDI, GMES) (GEO secretariat 2005; Henricksen 2007) had a great influence on the development of PREVIEW. In particular, all partner agencies are part of the United Nations. In consequence, making these data available in the United Nations Spatial Data Infrastructure (UNSDI) framework (Henricksen 2007) and registering to the Global Earth Observation System of Systems (GEOSS) (GEO secretariat 2008) is a de facto requirement for these agencies.

4.2.6 PREVIEW Global Risk Data Platform, the gateway to global natural disaster data

Key elements of any SDI are geoportals. Maguire and Longley (2005) define portals as gateways to an organized collection of resources (data, services, tools, links and documents) allowing an organization or a community to share specific content on the web. By extension, a geoportal can be seen as an entry-point to discover geospatial content. The OGC defines a common Geospatial Portal Reference Architecture (Open Geospatial Consortium 2004) to support data sharing, discovery, visualization and retrieval. This architecture specifies four classes of services that are required to implement a geoportal using OGC interoperability standards: (1) Portal service is the gateway to discover and access data as well as management and administration facilities; (2) Catalog service offers information about data and related services; (3) Portrayal service focuses on mapping and styling; and (4) Data service provides data access and
processing capabilities. To implement and deploy these different service classes, the OGC proposes to use web services technology giving access to distributed data and services through Uniform Resource Locators (URLs). This URL-based mechanism allows publishing standardized services over a network, typically the Internet, no matter how it is implemented (e.g., data format, storage) or on which platform it is executed. This leverages the real potential of interoperability by allowing web services to be seamlessly coupled, reusable and available for a wide variety of applications. In summary, the Geospatial Portal Reference Architecture mainly concerns the technological aspects (e.g., SOA, web services, standards) needed to implement a complete SDI model (data-standards-policies-network-people components).

4.6.2.1 System architecture

The PREVIEW Global Risk Data Platform is then the portal service supported by PREVIEW SDI that aims to offer a simple user interface to freely visualize, access, download and extract geospatial data on natural disaster at the global scale. This OGC-compliant platform serves data through interoperable web services into GEOSS, UNSDI, Google Earth or other clients (web or desktop-based). Data and metadata are exposed through a set of interoperable services to different kind of applications following a three-tier model (figure 18).

![Diagram of the PREVIEW Global Risk Data Platform](image)

*Figure 18: PREVIEW Global Risk Data Platform data and metadata flow.*
Chapter 4: How can SDI improve our capacity to discover, share, retrieve and integrate environmental data?

Data and metadata layer is the level where all data sets and their related documentation are stored into a PostgreSQL\(^{10}\) database system. This allows to store metadata and vectorial data (using PostGIS\(^{11}\) extension) in this relational system benefitting from a set of tools to manage them efficiently. Raster data are stored in a well-structured file system due to the current lack of support in PostGIS. Nevertheless, storing raster data in this way is not a limitation as they will be accessible through WCS. Service layer is an important element of the architecture as it implements most of the services defined by the Geospatial Portal Reference Architecture. Data and portrayal services are published using Geoserver\(^{12}\). This allows publishing data using OGC Web Services (OWS) such as WMS, WFS and WCS. In addition, Keyhole Markup Language\(^{13}\) (KML, to publish data into Google Earth) and GeoRSS\(^{14}\) (a geo-enabled RSS feed) are also provided. On the same basis, catalog service is exposed using Geonetwork\(^{15}\) that implements CSW specification allowing users to search, query, discover, publish, and manage metadata on the different data layers. This further enhances visibility and sharing capacities of PREVIEW. Finally, a mapping service is published using UMN Mapserver\(^{16}\) allowing the web-based application to access different GIS capabilities. The application layer is the last tier where all the data and metadata should be accessible through the set of services provided by the service layer. The portal service is exposed through a specific web-based GIS application that presents data, catalog and portrayal services in a good and consistent way, offering users a single entry-point where they can easily work with, and seamlessly integrate, data in their own dataflow. These interoperable services are also accessible through other web applications (such as partners or other UN/non-UN agencies) or desktop-GIS clients like ArcGIS. This allows users to work directly with data without the necessity to download and eventually change the formats to work with.

4.6.2.2 Geoportal functionalities

This web-GIS application, written in PHP\(^{17}\) with the UMN Mapserver mapscript extension, offers a highly interactive access to up to 60 data layers based on five core modules:

- **Mapping module**: provides traditional web-GIS tools (zoom-in/out, full extent, pan) and the possibility to create its own map, to export it as a PDF file or share it with a bookmark and/or URL to copy/paste into an email.

\(^{10}\) [http://www.postgresql.org](http://www.postgresql.org)
\(^{11}\) [http://postgis.refractions.net](http://postgis.refractions.net)
\(^{12}\) [http://www.geoserver.org](http://www.geoserver.org)
\(^{13}\) [http://www.opengeospatial.org/standards/kml/](http://www.opengeospatial.org/standards/kml/)
\(^{14}\) [http://georss.org/](http://georss.org/)
\(^{15}\) [http://geonetwork-opensource.org/](http://geonetwork-opensource.org/)
\(^{16}\) [http://www.mapserver.org](http://www.mapserver.org)
\(^{17}\) [http://www.php.net](http://www.php.net)
Chapter 4: How can SDI improve our capacity to discover, share, retrieve and integrate environmental data?

- **Graph module**: offers the ability to plot different graphs presented in the GAR report and interact with them (e.g., zoom, identify a specific country). Moreover, by clicking on a specific country supplemental information can be accessed through disaster risk country profiles provided by UNISDR system website.

- **Data download module**: offers the possibility to explore, discover and access data. Each layer is presented on a page providing a complete overview at once. This gives access to the metadata, a preview of the data and its attributes (in the case of vector data), the different URLs of associated OGC web services and finally a traditional download capability. Users can chose a specific GIS data format for download. The data will be dynamically generated in the selected format with its associated metadata in an ISO19115/19139-compliant XML file and made available for download in a zip file.

- **Data extraction module**: gives the possibility to extract either vector or raster data for a specific country or region. This ensures the same geographical extent to all selected data sets and allows users to work only within a specific area.

- **OGC web services module**: this module offers for each web service (WMS, WFS, WCS, KML and GeoRSS), the URL of the selected service to use in desktop clients and the URL generator for web-mapping application. Indeed, these URLs are in general complex and required different arguments (such as bounding boxes, type of request, format). Offering a simple access to these different request parameters should help users to implement these services into their own web applications.

![Global Risk Data Platform](image)

**Figure 19: Cyclones events from 2001 to 2008 in Asia as seen in the PREVIEW Global Risk Data Platform.**
4.6.2.3 User-testing phase

To ensure that this geoportal is easily usable, a two-day user-testing phase has been set up, before the official launch of the application with the help of UNISDR information unit. Different users from various organizations (academia, partner agencies, humanitarian agencies, decision-makers) involved in disaster management community were invited to test the PREVIEW through distinct proposed scenarios (e.g., create a map that shows the risk map for Cuba regarding tropical cyclones). This testing has shown the clear interest and benefits for participants in using such a platform. They have found it easy to use in producing maps and accessing data. Moreover, this phase also pointed out some inconsistencies, problems (e.g., legend, tools names) and bugs as well as the need for new functionalities (e.g., need of clear definitions, a tool to zoom on a specific country) that have been implemented and/or corrected immediately after the testing phase.

4.2.7 Uses of the PREVIEW and lessons learnt

The PREVIEW Global Risk Data Platform is accessible through the websites of the partner’s institutions (UNEP/GRI-Europe\textsuperscript{18}, UNISDR Preventionweb\textsuperscript{19}, UNDP/BCPR GRIP website\textsuperscript{20}). It is also accessible through GEOSS as well as through other geoportals (e.g., UNHCR). In light of the experience acquired since the launch of the platform in June 2009, different use cases and lessons have been identified.

4.2.7.1 PREVIEW in the disaster management cycle

During the major earthquake events that occurred in Haiti (January 2010) and in Samoa/ Indonesia (September 2009), an important increase in term of data download, map production and web access were observed on the platform. In particular, during the catastrophic event in Haiti, web statistics showed nearly four times more accesses on the portal than the daily mean (450 unique visitors vs. 100) and ten times more data downloaded (around 1Gb of data vs. 100Mo). This demonstrates that such a platform is probably used to provide general hazards, exposure and risk context for the disaster community and the media. Moreover, the quality of proposed services appears to be suitable as no interruptions have been reported meaning that even when numerous concurrent access occurred, data are still accessible and available. Such observations confirm and reinforce the idea that SDIs together with their geoportals could act as gateways to geospatial information (Masser 2005; Mohammadi and Rajabifard 2009). This allows maximizing the reuse of data and ensuring readily accessible to up-to-date data. Nevertheless, it should be noted that PREVIEW does not intend to be a real-time platform. Apart from the high data resolution,

\textsuperscript{18} http://preview.grid.unep.ch/index3.php?preview=map
\textsuperscript{19} http://www.preventionweb.net/english/maps/index.php
\textsuperscript{20} UNDP/BCPR: http://www.gripweb.org/grip.php?ido=1003

- 120 -
the analysis was conducted using global datasets, whose resolution are not relevant for in-situ planning and should not be used for life and death decisions. However, those global data sets can be used for comparison between countries, or for international agencies to help them prioritizing projects and investments in DRR.

4.2.7.2 Interoperable access to data and metadata

OGC web services allow users to work in collaboration and offer new opportunities to solve a specific problem. Thus, making data interoperable appears to be a key requirement (Sahin and Gumusay 2008). In this sense, the open data policy of PREVIEW could act as a catalyst, giving access to all its resources (data as well as metadata) in an interoperable and standardized way and proving the benefits to make data widely available. The act of sharing has the advantage of exposing data to the judgement of the broader community that could in consequence recognize whether or not these data are of sufficient quality. This also gives a possibility to users to provide some feedbacks and, if needed, data providers can improve their data accordingly.

Bernard and Craglia (2005) reported that, in Europe, the most frequent difficulties that could impede an efficient and effective use of data are: getting access to existing data, finding which data are available, data incompatibilities and quality issues. A positive sign is that the different OGC web services proposed by the PREVIEW Global Risk Data Platform are progressively used by different organizations: United Nations High Commissioner for Refugees (UNHCR) is accessing data within their own geportal focused on humanitarian activities, UNEP benefits from data and metadata services for its currently under development Data and Indicators Platform, the UN community through UNSDI activities can also benefit from the availability of these services and the Swiss government accessed PREVIEW metadata through its own metadata catalogue (GeoCat 21). These examples demonstrate that the difficulties mentioned previously can be mostly overcome by publishing interoperable services and potentially can benefit to other communities that was not envisioned before. Nevertheless a limitation must be highlighted concerning the interoperability. Indeed PREVIEW, by using OGC standards, is syntactically interoperable by allowing the platform to communicate and exchange data with other systems. However, to be fully interoperable, a system must be also semantically and schematically interoperable. Semantic and schematic interoperability gives the ability to interpret the information exchanged meaningfully and accurately, in order to produce useful results. To achieve these levels of interoperability, both sides of the system must define a common and agreed information exchange reference model. This way the content of the information exchanged is unambiguously defined: what is sent is the same as what is understood. The current absence of commonly agreed semantics and schemes to define and represent risk from natural hazards data sets can then impede a full integration. This can create data heterogeneities and result in a waste of time for data analyzers to homogenize data and in consequence negatively influence the quality of the disaster response phase. To achieve such level of interoperability,

21 http://www.geocat.ch
Chapter 4: How can SDI improve our capacity to discover, share, retrieve and integrate environmental data?

common agreed frameworks would help data providers to share their data in the most efficient way. The Global Earth Observation System of Systems (GEOSS) and the Infrastructure for Spatial Information in the European Community (INSPIRE) represent two promising frameworks to share and publish data. They provide guidance such as data sharing principles or data specifications that will help data providers to overcome the problems mentioned previously and push data owners to use more open policies and share their data in a fully interoperable way. As stated in the GEOSS 10 Year Implementation Plan (2005), disaster reduction requires a great effort on data sharing because data are time-sensitive, of known quality, long-term and global by nature. An initiative like GEOSS can facilitate such sharing, enhancing collaboration and coordination between countries and organizations involved in minimizing losses from natural disaster. Contributing to GEOSS not only increases data visibility and wide accessibility to PREVIEW data sets, but also supports and sustains a collaborative effort to move toward a better-informed society on reducing disaster risk.

4.2.7.3 Building capacities

In order to improve the visibility of PREVIEW, to promote the use of OGC standards, to prove the benefits of exchanging and sharing data in an interoperable manner, it is of high importance to build new capacities. For Rajabifard and Williamson (2004), building new capacities through education and research, is certainly the best way to achieve the objective of a wide acceptance and adoption of SDI concepts. This will help to increase endorsement and allows SDIs concepts and related technologies to be widely accepted and adopted. In this regard, PREVIEW is regularly presented and used for demonstrations to different audiences (students, scientists, UN collaborators) allowing people to use it and to gain familiarities with these emerging technologies. The user testing phase and the different demonstrations have clearly highlighted the fact that most users: (1) were looking for “traditional” download options (link to zip files), (2) were not aware of the possibilities offered by OGC web services. Consequently, it was important to offer users the possibility to download data either using a download module and OGC webservices. Nevertheless, once we have shown them what are OGC web services, users appear to be convinced by this new way of accessing data.

Another lesson learnt concerns the process developed to harmonized data. This has shown the necessity to agree on common data models and in our vision it is of high importance to bring such a process at higher level in order to develop standardized data models accepted and used by the overall community. In the case of PREVIEW, the data models were quite easy to develop as it was mostly done with partners that we already know and with which we have been working for a long time.

Finally, the selected open-source approach appears to be a good choice. First because this allows developing a specific mapping API for our partner agencies that could implement the mapping service into their own web portal. Second, different countries seem to be interested to replicate this platform to publish and share their own data on natural hazards at the national level. This will help UN agencies to support developing countries in learning and taking
advantage of SDI concepts and related technologies. It will also improve data accessibility for a better-informed and sustainable development.

4.2.7.4 Limitations and new developments

One of the issues in providing access to 60 different layers of information with a cartographic web dynamic interface lies in the multiple possibilities of combinations of displayed information. Thus the choice of colours legend schemes was a difficult task. It was sometimes needed to prevent the simultaneous display of different layers to avoid confusion.

The UMN Mapserver and all tested Map Servers do not provide the capacity to turn around the globe, unlike Google earth (sphere) and Google maps (cylinder). The data are provided in a rectangle shape, preventing the display centred, for example, over the Pacific Ocean. This could have been possible by doubling the size of data sets, but this option was finally abandoned (bad performance). This is something that should be improved in the future. Keeping the balance between a user-friendly interface and end-users tailored options is a difficult exercise. New developments will include query tools and a search interface for identifying specific events.

4.2.8 Conclusions

This paper presented the PREVIEW Global Risk Data Platform and described the potential usefulness of a geportal to facilitate and coordinate data sharing, access and use among different partners. This work also highlighted that sharing spirit, accessibility, availability, interoperability, and data harmonization are important issues that can be clarified with the help of a clearly defined SDI conceptual model. Such a model enhances collaboration and partnership among different participants and gives the possibility to agree on making these data freely available for the benefit of the whole DRR community and beyond.

Geospatial data are essential in disaster cycle of activities and timely access to such data could improve decision-making process and thus could save lives and/or minimize losses of property (GEO secretariat 2005). Even if damages could not be completely avoided, better coordination and collaboration between organizations as well as better data exchange will help to better identify risk and in the mid to long term, help to reduce losses. With systems that could deliver improved information on natural hazards, this will also help to prevent that such events become disasters. Feeney et al. (2001) argued, SDIs could be seen as an answer to the growing need to organize data across different disciplines and organizations, and helping them address the issue of supporting decision-making process. To achieve this objective, data sharing and open access policies appear to be an important issue allowing an easy and wider usage of data (Craglia, Goodchild et al. 2008). Having the ability to collect data once and reuse it many times is a clear incentive for such initiatives avoiding duplication of time, effort and expenses, and improving access to good quality data and in turn improving decision-making process. In this sense, the PREVIEW Global Risk Data Platform offers the possibility to freely and easily access data that could be useful for preparedness, response and mitigation phases. This work also highlights that facilitating access to data by sharing it in an interoperable and standardized way.
Chapter 4: How can SDI improve our capacity to discover, share, retrieve and integrate environmental data?

increases the ease of integration and use. This can support disaster activities and the achievement of a more sustainable development. Moreover, with the help of OGC standards, data are now available and accessible through a wide array of different stakeholders (partner agencies, humanitarian agencies, GEOSS) and by saving time in accessing and integrating data, this could positively influence the disaster response in an emergency situation. Nevertheless, it is urgent to develop data models commonly agreed by the community in order to be interoperable at the syntactic, semantic and schematic levels. It is an indispensable condition to leverage to full potential of interoperability avoiding heterogeneities and positively improving the quality of the disaster response phase.

Finally, it is important to keep in mind that data sharing and related SDI developments rely mostly on individuals that should have in common (1) a sense that better data will lead to better decisions, (2) a sharing spirit that they got something in return and are viewed as collaborative partners and (3) the fact that they are involved in a professional culture that honours serving society and cooperating with others (Craig 2005). For Arzberger et al. (2004), ensuring that data are easily accessible so that they can be used as often and widely as possible is a matter of sound stewardship of public resources. These authors stated that publicly funded data are a public good, produced in the public interest and thus should be freely available to the maximum extent possible. Based on these considerations, creating harmonized platforms for hazard and risk information to support decision-making processes and allowing data management; distribution and exchange appear to be a necessity.
Chapter 4: How can SDI improve our capacity to discover, share, retrieve and integrate environmental data?

4.3 OGC Web Feature and Web Coverage Services performance testing: towards an efficient access to geospatial data

**Gregory Giuliani**1,2,3, **Alain Dubois**4,5, **Pierre Lacroix**1,3

Addresses:
1 University of Geneva, Institute of Environmental Sciences, Climatic Change and Climate Impacts, enviroSPACE, Battelle – Building D, 7 route de Drize, CH-1227 Carouge.
3 University of Geneva, Forel Institute, 10 route de Suisse, CP416, CH-1290 Versoix, Switzerland.
4 University of Geneva, Institute of Environmental Sciences, Human Ecology Group, Battelle - Building D, 7 route de Drize, CH-1227 Carouge, Switzerland.
5 University of Applied Sciences Western Switzerland, HEPIA, 4 rue de la Prairie, CH-1202 Genève.

4.3.1 Abstract

Open Geospatial Consortium (OGC) Web Feature Service (WFS) and Web Coverage Service (WCS) specifications allow an interoperable access to distributed geospatial data made available through Spatial Data Infrastructures (SDIs). To ensure that a service is sufficiently responsive to fulfill users expectations and requirements, performances must be measured and monitored in order to track latencies, bottlenecks and errors that may negatively influence the service overall quality. Despite the importance of data retrieval and access, only little research has been published on this topic and mostly concentrate on the usability of services when integrating distributed data sources.

Considering these issues, this paper presents (1) an approach to measure the performances of different WFS and WCS services provided by various software implementations and provide (2) some guidance to data provider looking to improve the quality of their data services. Our results show that performances of tested implementations are generally satisfactory and memory tuning/data and storage optimization are essential factors to handle in order to increase efficiency and reliability of services.

4.3.2 Introduction

Facilitating access and making geospatial data interoperable are recognized as important factors in determining the future success of Spatial Data Infrastructures (SDI) (Rajabifard and Williamson 2001; Bernard and Craglia 2005; Masser 2005; Boes and Pavlova 2008). Indeed, geospatial data are extensively used in various domains like environmental monitoring, disaster management, and decision-making processes requiring seamless integration in Geographic Information Systems (GIS) and sharing of data across organizations and providers (Sahin and Gumusay 2008). Hence, geospatial data can be a shared resource maintained continuously and made accessible for different purposes both for public and private sectors (Ryttersgaard 2001). SDIs aim to support sharing and exchanging of geospatial data across institutional, regional, and/or national border mostly through Service Oriented Architecture (SOA) principles (Simonis and Sliwinski 2005; Granell, Diaz et al. 2009). SOA is defined as an
Chapter 4: How can SDI improve our capacity to discover, share, retrieve and integrate environmental data?

architectural approach where standardized interfaces give access to functionalities as a set of independent, interoperable, loosely coupled distributed services that can be reused. The rapid evolution of web-based communication technologies allow for an easier access to distributed geospatial data sources and related services results in the fact that geospatial data are increasingly used in many areas (Paul and Ghosh 2006).

To make an effective use and integration of geospatial data with other sources, systems must be interoperable (Sahin and Gumusay 2008). Interoperability means that two or more systems or components are able to transmit and exchange information through a common system and to use that exchanged information (Open Geospatial Consortium 2004). When systems are interoperable, it gives users the ability to (1) find what they need, (2) access it, (3) understand and employ it, (4) have goods and services responsive to their requirements. As of today, in a climate of economic constraint, interoperability and standardization have never been so important because a non-interoperable system impedes the sharing of data, information and computing resources (Open Geospatial Consortium 2004), which leads organizations to spend much more time than necessary on data, software and hardware. Moreover, being non-interoperable increases the risk for a system or an infrastructure to not deliver its expected benefit and in consequence leads to users disappointment and system failure.

To solve interoperability problems within the GIS community, the Open Geospatial Consortium (OGC) introduced various standard specifications covering data sharing, retrieval, processing, content, visualization, and cataloguing (Brauner, Foerster et al. 2009; Bulatovic, Ninkov et al. 2010). These standards are built around web services technology. A web service is a stateless application containing a collection of operations, exposed as a function, that is accessible through the web and that users can discover and invoke (Sahin and Gumusay 2008). In other words, interoperable web services aim to provide users just the functionality they need independently on the computing platform (e.g., Operating System, Data Management System) and programming language. Moreover, the composability and reusability of standard components offered by web services allows building applications specific to the needs of domains and/or users community overcoming disadvantages and inflexibilities of monolithic GIS (Simonis and Sliwinski 2005). OGC web services (OWS) are based on eXtended Markup Language (XML) to encode calls and HyperText Transfer Protocol (HTTP) for communicating. However, having an interoperable access to data is only a first step to gain credibility in showing the benefits of easy data integration. It is obviously also required to have services of sufficient quality in order to satisfy users expectations. Several studies (Booz, Allen et al. 2005; European Commission 2006; Craglia and Campagna 2009) have shown that projects that have adopted and implemented geospatial interoperability standards saved around 25% compared to those who rely on proprietary standards. Moreover, these reports showed that using standards lower the transaction costs for sharing data and information when exchange are facilitated by standardized interfaces enhancing flexibility and adaptability of projects over time. The recently published INSPIRE State of Play 2010 (Vandenbroucke 2010) highlighted that more and more European countries are focusing their attention on interoperability issues. View services appear to be very well developed and
download services are really starting to emerge now. Consequently we can assume that WMS is now well accepted among different communities to make their data viewable in an interoperable manner but this is not yet the case for making data downloadable using WFS and WCS standards. Only in few cases data have made available through these standards (Da Silva, Fidalgo et al. 2004; Christensen, Bernard et al. 2006; Khoumeri and Benslimane 2007; Luis 2009; Bruniecki, Kulawiak et al. 2010). 

Currently, SDIs are mainly concerned with catalog and data services (Baranski 2008; Schaeffer 2008) allowing data discoverability using Catalog Service for the Web (CSW) (Open Geospatial Consortium 2007), retrieval with Web Feature Service (WFS) (Open Geospatial Consortium 2005) and Web Coverage Service (WCS) (Open Geospatial Consortium 2006), and visualization through Web Map Service (WMS) (Open Geospatial Consortium 2006). Initiatives at the regional and global scale like the Infrastructure for Spatial Information in the European Community (INSPIRE) (European Commission 2007) or the Global Earth Observation System of Systems (GEOSS) (GEO secretariat 2005) are seeking to relate environmental data providers to the widest possible audience with the objective to enhance and improve decision-making. In this context, OWS are key enablers providing interoperable access to data in an efficient and timely manner. To ensure that a service is sufficiently responsive to fulfill users needs and requirements, performance of a given service must be measured and monitored in order to track latencies, bottlenecks and errors that may negatively influence its overall quality.

Despite the importance of data retrieval and access, only little research has been published on benchmarking and evaluating the quality of WFS and WCS services (Tu, Flanagan et al. 2004; Simonis and Sliwinski 2005; Zhang 2005; Lance, Georgiadou et al. 2006; Vanmeulebrouk, Bulens et al. 2009). These studies mostly concentrate on the usability of OWS when integrating distributed data sources but do provide neither framework nor guidance to measure performances of data services. Moreover, and to our knowledge, the only published paper specifically on this topic examines the WMS specification (Horak, Ardielli et al. 2009). Based on these considerations the aims of this paper are:

1) to present an approach to measure the performance of different WFS and WCS services provided by various software implementations.

2) to provide some guidance to data providers aiming at improving the quality of their services.

All developed material (benchmark scripts, data, and procedures) is freely accessible and downloadable at: http://www.unige.ch/sig/owsbenchmark.

4.3.3 Geospatial data interoperability

Bernard and Craglia (2005) reported that data accessibility, availability and compatibility are among the most frequent difficulties evidenced while preparing Strategic Environmental Assessments (SEA) and Environment Impact Assessment (EIA) in Europe. Moreover, it is estimated that up to 50% of users' time is spent in data discovery and transformation in order to make them compatible and integrable (Ma, Pan et al. 2005; Yaxing, Santhana-Vannan et al. 2009). These authors emphasize various reasons leading to such problems:
Chapter 4: How can SDI improve our capacity to discover, share, retrieve and integrate environmental data?

(1) Geospatial data are voluminous,
(2) Geospatial data are geographically distributed (e.g., various data collector/providers around the World),
(3) Geospatial data are heterogeneous (e.g., formats, schemas, coordinate systems)
(4) Geospatial data are complex (e.g., geometries, relationships, attributes),
(5) Institutional arrangements and policies (e.g., copyright) that are lacking or restricting access to (meta)data.

All these factors may influence the way that data providers store, publish and deliver geospatial data. Hence, making data interoperable can potentially improve the above-mentioned situation allowing data users to spend more time in data analysis than in data discovery, which would enable more people to benefit from using geospatial data. To be fully interoperable a system must be syntactically, semantically and schematically interoperable. However, OGC specifications concentrate only on syntactical interoperability allowing exchanging data with other components or systems. To reach semantic and schematic interoperability, both components of a system must agree on a common reference model providing the possibility to interpret accurately, unambiguously, and meaningfully the exchanged information. These levels of interoperability are important issues that are extensively studied (Fü, Jones et al. 2005; Lacasta, Nogueras-Isó et al. 2007; Lutz, Sprado et al. 2009; Zhao, Di et al. 2009). Hence, to enable effective and syntactically interoperable access to geospatial data, OGC WFS (for vector data) and WCS (for raster data) specifications are essential components.

4.3.3.1 Web Feature Service (WFS)

The OGC Web Feature Service specification (Open Geospatial Consortium 2005) defines an interface for accessing feature-based geospatial data, commonly known as vector (e.g., rivers, country borders, cities), encoded in Geography Markup Language (GML) (Sahin and Gumusay 2008; Yaxing, Santhana-Vannan et al. 2009). A WFS interface is invoked through a URL and can perform a certain number of operations (e.g., retrieving, creating/deleting/updating) allowing a client to handle data (Sayar, Pierce et al. 2005):

(1) GetCapabilities: provides metadata about capabilities offered by a WFS instance. It tells the client which kind of features are available and what operations are supported on each feature.
(2) DescribeFeatureType: describe the structure of a selected feature (point, line, polygon).
(3) GetFeature: retrieve a selected feature encoded in GML. The client can constrain the query both spatially and non-spatially, and also specify the feature properties to fetch.
(4) Transaction: this type of request is made of operations that allow a client to modify features: create, delete and/or update operations.
(5) In addition, a client can invoke the LockFeature, in order to be sure that only one user is updating a specific feature, avoiding the risk of multiple editions at the same time.
Chapter 4: How can SDI improve our capacity to discover, share, retrieve and integrate environmental data?

4.3.3.2 Web Coverage Service (WCS)

The OGC Web Coverage Service specification (Open Geospatial Consortium 2006) defines a web interface allowing a client to access raster datasets. A raster data represents a matrix of cell in continuous space organized in rows and columns where each cells contains a value. Thus, WCS services give access to different types of space and/or temporal-varying data like Digital Elevation Model (DEM) or remote sensing imagery. WCS delivers raw data and does not have transactional capabilities (Sahin and Gumusay 2008; Yaxing, Santhana-Vannan et al. 2009). Similarly to WFS, WCS offer different operations: GetCapabilities (e.g., service metadata), DescribeCoverage (e.g. raster data description), and GetCoverage (e.g., raster data retrieval).

4.3.4 INSPIRE Network Services

The Infrastructure for Spatial Information in the European Community (INSPIRE) is a legal framework (entered into force in May 2007 and fully operational by 2019) for the establishment of a European Union Spatial Data Infrastructure. The Directive provides five sets of Implementing Rules (IR) that set out how the various elements of the system (metadata, data sharing, data specification, network services, monitoring and reporting) will operate, and ensures that spatial data infrastructures of the Member States are compatible and usable in a Community and transboundary context (Bernard, Kanellopoulos et al. 2005). INSPIRE will enable sharing of environmental geospatial data among public sector organizations and better facilitate public access to spatial information across Europe (European Commission 2007). When fully implemented, it will theoretically enable data from one Member State to be seamlessly combined with data from all other States to support the formulation, implementation, monitoring and evaluation of environmental policies and to overcome barriers that are affecting availability and accessibility of data (Craglia and Campagna 2009; Horak, Ardielli et al. 2009).

According to the INSPIRE network architecture (European Commission 2008), Member States shall establish, operate and provide access to the following network services:

- **discovery services**: support discovery of data, evaluation and use of geospatial data and services through their metadata properties.
- **view services**: as a minimum, display, navigate, zoom in/out, pan, or overlay geospatial data sets and display legend information and any relevant metadata content.
- **download services**: enabling copies of complete geospatial data sets, or parts of such sets, to be downloaded.
- **transformation services**: enabling geospatial data sets to be transformed (projection and harmonization).
- **invoke spatial data services**: enabling data services to be invoked.

Network services are a key component in INSPIRE architecture and, following the Directive, vector and raster data will be made available through download services. Craglia and Campagna (2009) reported that network services, in particular discovery and view services, are well developed in European countries and are progressing rapidly in those who recently joined the
Chapter 4: How can SDI improve our capacity to discover, share, retrieve and integrate environmental data?

Union (e.g., in 2004). In comparison download, transformation, and invoke services are less developed. The commission regulation about download services (European Commission 2010) establishes a list of Quality of Service criteria (e.g. performance, capacity and availability) that must be measured ensuring an efficient access to data (European Commission 2007).

4.3.4.1 Quality of Service (QoS)

Satisfaction of users is a major objective of any service provider. However, evaluating and predicting user’s satisfaction is a complex task because quality can be based on quantitative and qualitative measurements (Lance, Georgiadou et al. 2006). According to Simonis and Sliwinski (2005) quality can be interpreted both as a measure that represents accessibility/performance and also as a perceived quality of a service. Despite the importance of qualitative perceptions (e.g., good description, easy access to a service) it is undeniable that insufficient technical quality of service will strongly affect users’ satisfaction (Menasce 2002). With the expected increasing diffusion of OGC Web Services (OWS), Quality of Service (QoS) will be an important factor to distinguish between reliable services and less reliable ones. Therefore QoS can be referred as a set of measurable attributes related to the individual behavior of a web service that guarantees sufficient availability and performance of a service (Simonis and Sliwinski 2005). This excludes all qualitative perceptions and concentrates on technical aspects that can be easily measurable and replicable in different conditions/environments. Horak et al. (2009) give a detailed overview of testing methods both on server and client sides. These authors recognize the importance of client-side testing to evaluate characteristics that identify measurable/quantifiable parameters.

To ensure sufficient availability and accessibility to data, the European Commission (European Commission 2009; European Commission 2010) has adopted regulations about Network services and specifically about download services. Annex I of the regulation on download Services sets out Quality of Service criteria and corresponding values that any services must achieve:

1) **Performance**: refers to how fast a request can be completed. The response time for sending the initial response shall be maximum 30 seconds in normal situation (e.g., periods out of peak load) and shall maintain a sustained response greater than 0.5 Megabytes (or 500 spatial objects) per second.

2) **Capacity**: is the limit of simultaneous requests that a service can handle with guaranteed performance (Horak, Ardielli et al. 2009). For a download service, the minimum number of served simultaneous service shall be 10 requests per second. The number of requests processed in parallel may be limited to 50.

3) **Availability**: measures the probability that a network service is available and it shall be 99% at any given time.

These quantitative measurements are important as they allow easy quantification, repeatability, comparability, and understandability (Mendes and Mosley 2006; Horak, Ardielli et al. 2009). QoS criteria can be measured either on server-side (e.g., number of visits, session duration, average time per page) or client-side (e.g., load testing, capacity testing). In general, the tested service is
Chapter 4: How can SDI improve our capacity to discover, share, retrieve and integrate environmental data?

considered as a black box receiving requests and providing responses without considering software implementation at all.

4.3.5 Methodology of testing

Currently, in absence of a commonly agreed framework to test download services, it is important to measure performances with approaches that (European Commission 2007):

(1) are sufficiently generic to be independent on infrastructure and application design;
(2) are independent on the communication (e.g transport network) between the service and client;
(3) are based on request-response pairs and avoid complex transactions. Therefore, the main objective of such a testing framework is to assess the usability and performances from client-side/end-user perspective of various network services and specifically download services.

Our proposed approach is based on the one developed by the Free and Open Source for Geospatial (FOSS4G) community\(^{22}\) to test WMS services. The aim is to extend this open approach to WFS and WCS services in order to test various OWS implementations on a common set of geospatial data located on a common platform. Obviously, in order to fulfill the requirements mentioned previously, it is important that these tests be based on realistic usage scenarios. For that purpose, load testing (also known as stress testing) is generally performed. Thus simulates multiple concurrent queries to a service. This approach allows analyzing the behavior (e.g., response to unusual loads) of a service under various conditions that are beyond normal usage patterns (Horak, Ardielli et al. 2009).

In our case, WMS testing was achieved through the use of the GetMap query, with the sole purpose of calibrating our resulting curves with those of FOSS4G 2009 shootout\(^{23}\). Regarding WFS and WCS testing, the use of the GetFeature and GetCoverage retrieval commands placed the emphasis on the real access to the data. Thus, overhead due to symbology and projection on the fly were not tested.

Our general methodology was to take one case as a base case, and to vary one by one a given set of parameters: geometry type, data resolution, complexity and number of attributes, field indexation, input and output formats. Random screen sized requests were performed on commonly used scales and extents corresponding to the bounding box of various countries. In order to represent realistic usage scenarios, two groups of users were simulated: (1) a small team (e.g., people working within the same laboratory) with a workload varying from 1 to 16 concurrent requests, and (2) a larger set of 150 users simultaneously retrieving data like in emergency situation (e.g., after Haiti or Japan earthquakes). Each test was performed three times but only the last measure, considered as the most stable one, was retained. We did not take the mean of the three measurements because the first run is sometimes not stable and thus does not reflect the normal behavior of the server (Menasce 2002).

\(^{22}\)http://wiki.osgeo.org/wiki/FOSS4G_Benchmark

Chapter 4: How can SDI improve our capacity to discover, share, retrieve and integrate environmental data?

4.3.5.1 Testing scenarios

Testing scenarios were developed in order to simulate user’s behavior when downloading data through WFS and WCS services based on:

1. common vector and raster input and output formats: shapefile, ESRI geodatabase, ArcSDE, GML, TIFF, GeoTIFF, and JPEG.
2. popular datasets, such as the Blue Marble New Generation and the Global Lakes and Wetlands Database, were chosen for their potential use in many fields and their availability at the global scale in WGS 84 spatial reference.

Various conditions were defined according to (a) our daily experience using OWS and (b) past experiments made during the development of the PREVIEW geoportal (Giuliani and Peduzzi 2011). The different testing scenarios developed to test WFS services are summarized in Table 4 and those concerning WCS in Table 5. Each scenario is tested both under ArcGIS Server and GeoServer platform.

<table>
<thead>
<tr>
<th>Case study for WFS testing</th>
<th>Tested parameter</th>
<th>Input format</th>
<th>Input geometry</th>
<th>Input attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>None</td>
<td>ArcSDE</td>
<td>Polygons</td>
<td>Base attributes</td>
</tr>
<tr>
<td>2</td>
<td>Input format</td>
<td>Shapefile</td>
<td>Same as case #1</td>
<td>Same as case #1</td>
</tr>
<tr>
<td>3</td>
<td>Input format</td>
<td>ESRI File</td>
<td>Same as case #1</td>
<td>Same as case #1</td>
</tr>
<tr>
<td>4</td>
<td>Input geometry</td>
<td>Same as case #1</td>
<td>Polylines</td>
<td>Same as case #1</td>
</tr>
<tr>
<td>5</td>
<td>Input geometry</td>
<td>Same as case #1</td>
<td>Points</td>
<td>Same as case #1</td>
</tr>
<tr>
<td>6</td>
<td>Complexity and number of attributes</td>
<td>Same as case #1</td>
<td>Same as case #1</td>
<td>No attributes</td>
</tr>
<tr>
<td>7</td>
<td>Complexity and number of attributes</td>
<td>Same as case #1</td>
<td>Same as case #1</td>
<td>Additional attributes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>- text 255</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>- five eight-byte double-precision columns</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Indexation of the id field</td>
</tr>
</tbody>
</table>

Table 4: Summary of the WFS testing scenarios. Case #1 is the “base case”.

As an example, the following WFS request executed under ArcGIS Server allows one to retrieve all rivers of Switzerland (i.e. “cntry_name” field equal to “Switzerland”) present in the data set “WFS_GDB:river”. This query uses the OGC filter encoding specification capabilities (Open Geospatial Consortium 2005) to extract data based on their attributes:

Chapter 4: How can SDI improve our capacity to discover, share, retrieve and integrate environmental data?

In GeoServer, the same request shows differences (e.g., “ogc” term mandatory in ArcGIS) in syntax for the expression of the filter, which already highlights a potential problem in terms of interoperability and universality of semantic:

```
http://<server>/geoserver/ows?service=WFS&styles=&request=GetFeature&version=1.0.0&srs=EPSG:4326&typename=owsbenchmark:AQ.SDE.river&filter=<Filter><PropertyIsEqualTo><PropertyName>cntry_name</PropertyName><Literal>Switzerland</Literal></PropertyIsEqualTo></Filter>
```

Table 5: Summary of the WCS testing scenarios. Case #1 is the “base case”.

Concerning WCS request, in ArcGIS Server the base case corresponds to the following syntax:

```
http://<server>/arcgis/services/<service>/MapServer/WCSServer?SERVICE=WCS&VERSION=1.0.0&REQUEST=GetCoverage&COVERAGEx=1&CRS=EPSG:4326&BBOX=0,0,40,20&WIDTH=9600&HEIGHT=4800&FORMAT=tiff
```

And the same request expressed in GeoServer:

```
http://<server>/geoserver/ows?SERVICE=WCS&VERSION=1.0.0&REQUEST=GetCoverage&COVERAGEx=1&CRS=EPSG:4326&BBOX=0,0,40,20&WIDTH=9600&HEIGHT=4800&FORMAT=tiff
```

4.3.5.2 Data sets

In order to fit the GIS community’s common users practices, we have selected widely used data sets with a global extent and a large number of features (e.g., attributes, resolution, formats). Therefore vector data sets (used for WFS tests) were extracted from:
Chapter 4: How can SDI improve our capacity to discover, share, retrieve and integrate environmental data?

(1) the Global Lakes and Wetlands Database – Level 2 (GLWD-2) (Lehner and Döll 2004) for Polygon data set. This very popular dataset, used by numerous governmental and non-governmental organizations, contains about 250’000 natural lakes and man-made reservoirs, and 8’500 rivers - respectively 4’000’000 and 350’000 vertices. By default, the polygonal layer includes a dozen of textual and numeric attributes.

(2) the European Space Agency (ESA-ESRIN) World Fires Atlas Program (ATSR) for Point data set. This represents an estimation of fires events around the world between 1997 and 2008 and available in a compiled form on the PREVIEW Global Risk Data Platform (Giuliani and Peduzzi 2011). This data set is composed of around 1’000’000 points.

While for the WCS testing the raster data set we have selected as base case:

(3) Blue Marble Next Generation (BMNG) (Stöckli, Vermote et al. 2005). This data set is provided into 240 GeoTiff tiles (e.g., separate images) of 12 MB and 15 Decimal Degree (DD) each, spatially referenced on the WGS 84 ellipsoid, with a 0.008 DD pixel size, and for a total amount of data of 2.87GB. The BMNG is traditionally used not only in education, and research, but also often as a base map, with a relatively high spatial resolution.

4.3.6 Technical architecture & software

One of the aims of the proposed approach is to give some guidance to service providers that are looking to share their data in an interoperable manner. In particular, vast amount of geospatial data are available within institutions that do not have means and/or do not want to develop complex infrastructures. Consequently, the proposed architecture to serve OWS must be simple, reflecting these conditions and showing that with little effort it is feasible to provide download services of sufficient quality. The technical architecture is based on a three-tier model: (1) a data layer (2) a service layer, and (3) a client layer.

Figure 20: Architecture used for testing different OWS implementations
In the data layer, vector and raster data can be stored directly either into a database or in directories. PostGIS 1.5.1\textsuperscript{24} and ArcSDE 9.3.1\textsuperscript{25} have been used enabling PostgreSQL 8.4.4\textsuperscript{26} Database Management System (DBMS) to support geospatial data. Interoperable access to data is provided through the service layer using GeoServer 2.0.2\textsuperscript{27} and ArcGIS Server 9.3.1\textsuperscript{28} that are able to manage different data sources like PostGIS, ArcSDE or files in folders. These different pieces of software have been selected because they are widely used in the GI community allowing testing “native” solutions (e.g., ArcGIS Server/ArcSDE, GeoServer/PostGIS), as well as “crossed” solutions (e.g. ArcGIS Server/PostGIS, GeoServer/ArcSDE).

The client requesting services will be simulated using JMeter\textsuperscript{29}, a Java open source software designed to test servers, networks, and services performances under various load conditions on static and dynamic resources (e.g., files, servlets, scripts, databases and queries, FTP servers). JMeter was originally designed for testing web applications but has since expanded to other testing functions using plugins. Despite the fact that this is a desktop application, it is important to note that JMeter is not a browser (e.g., it does not execute Javascript nor does HTML rendering). Its usage is very simple, requiring only the URL querying a specific service. Users can design their testing plan using variables, counters, logs or whatever parameter to be tested. A JMeter test then corresponds to a script that can be easily executed and maintained through the JMeter Graphical User Interface (GUI). Once tests have been executed, performance logs can be accessed in text files providing various information such as number of requests, number of threads, or execution time. This information can be used to draw results as graphs or tables. JMeter scripts provided by FOSS4G WMS 2009 benchmark have been used and extended/adapted to test our WFS and WCS instances.

4.3.7 Results

The tests realized with JMeter consist in sending requests for maps with random geographical extent to the different services and measuring the response time. The performance indicator is expressed in maps served per second. Different load conditions varying between 1 to 40 simultaneous threads allow evaluating the server response. Vector data sets have been used for testing WMS and WFS, and raster data sets for WCS.

---

\textsuperscript{24} http://postgis.refractions.net/
\textsuperscript{25} http://www.esri.com/software/arcgis/arcsde/
\textsuperscript{26} http://www.postgresql.org/
\textsuperscript{27} http://www.geoserver.org
\textsuperscript{28} http://www.esri.com/software/arcgis/arcgisserver/
\textsuperscript{29} http://jakarta.apache.org/jmeter/
Chapter 4: How can SDI improve our capacity to discover, share, retrieve and integrate environmental data?

4.3.7.1 WMS

Different file formats have been used with the same data set consisting in edge polylines (roads, rivers) for Texas: shapefile with spatial index, ArcSDE PostgreSQL, and ESRI file geodatabase.

<table>
<thead>
<tr>
<th>Threads [Maps /s]</th>
<th>Requests</th>
<th>Mapserver CGI SHP</th>
<th>MapServer FCGI SHP</th>
<th>MS4W CGI SHP</th>
<th>MS4W FCGI SHP</th>
<th>AGS SHP</th>
<th>AGS SDE</th>
<th>AGS FGDB</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100</td>
<td>5.0</td>
<td>8.2</td>
<td>3.6</td>
<td>5.7</td>
<td>2.4</td>
<td>6.6</td>
<td>7.2</td>
</tr>
<tr>
<td>10</td>
<td>200</td>
<td>11.0</td>
<td>17.1</td>
<td>12.2</td>
<td>17.5</td>
<td>3.3</td>
<td>18.2</td>
<td>20.7</td>
</tr>
<tr>
<td>20</td>
<td>400</td>
<td>11.0</td>
<td>16.8</td>
<td>11.0</td>
<td>16.9</td>
<td>3.5</td>
<td>17.2</td>
<td>21.5</td>
</tr>
<tr>
<td>40</td>
<td>800</td>
<td>11.5</td>
<td>17.6</td>
<td>11.5</td>
<td>17.3</td>
<td>3.3</td>
<td>17.1</td>
<td>22.1</td>
</tr>
</tbody>
</table>


Figure 21: WMS results under various conditions (OWS implementation and file formats)

The performance tests show that the results for shapefiles from FOSS4G WMS benchmark 2009 are similar for Mapserver CGI and FCGI on Linux platform.
and for Mapserver CGI and FCGI on Windows\footnote{Package Mapserver for Windows - MS4W 3.0 \url{http://www.maptools.org/ms4w/} containing Apache 2.2.15, Mapserver 5.6.3}. This means that the infrastructure used for the benchmark is comparable in terms of performance.

WMS service produced by ArcGIS Server for the same data set in shapefile is less efficient than Mapserver, which is equivalent to Mapserver FastCGI when storing data in ArcSDE. The best results are obtained using ArcGIS Server and file geodatabase format.

4.3.7.2 WFS

WFS services have been tested through different scenarios on ArcGIS Server and Geoserver including different data sets (e.g., points and polygons) stored in different formats including ArcSDE, Shapefile, ESRI File geodatabase and PostgreSQL/PostGIS.

The fires point data set (in shapefile format) containing 1’089’107 features for a size of 99 MB is about four times slower than the lake data set (shapefile of 117 MB) containing 244’892 polygons and 3’990’096 vertices. This means that the relationship between the storage size of a data set and the performance is not straightforward. “The larger the file size, the slower” is not always true. It depends also on the number of features.

On ArcGIS Server, using the data set “Lake polygon”, performances are only slightly affected by the storage format, with a little advantage for the ESRI file geodatabase format.

<table>
<thead>
<tr>
<th>Threads [Maps/s]</th>
<th>Requests</th>
<th>SDE Fires point</th>
<th>SHP Lake polygon</th>
<th>SDE Lake polygon</th>
<th>FGDB Lake polygon</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100</td>
<td>0.5</td>
<td>1.7</td>
<td>1.8</td>
<td>1.8</td>
</tr>
<tr>
<td>2</td>
<td>200</td>
<td>0.8</td>
<td>3.8</td>
<td>4.4</td>
<td>4.5</td>
</tr>
<tr>
<td>4</td>
<td>400</td>
<td>1.0</td>
<td>4.9</td>
<td>4.1</td>
<td>5.1</td>
</tr>
<tr>
<td>8</td>
<td>800</td>
<td>1.1</td>
<td>4.7</td>
<td>4.5</td>
<td>5.3</td>
</tr>
<tr>
<td>16</td>
<td>1600</td>
<td>1.1</td>
<td>5.0</td>
<td>4.6</td>
<td>5.3</td>
</tr>
</tbody>
</table>

Table 7: WFS ArcGIS Server tests results. Refer to table 3 for acronyms
Chapter 4: How can SDI improve our capacity to discover, share, retrieve and integrate environmental data?

**Figure 22: WFS ArcGIS Server tests results**

Compared to ArcGIS Server, Geoserver WFS instances are about three times faster. The polygon dataset representing lakes is more efficient when stored in PostGIS. A little decrease in performances is observed when the number of simultaneous threads is exceeding 4, but remains still very good. This phenomenon could maybe come from a disk access bottleneck (e.g., access times, queues). An increasing number of service errors have been observed when the simultaneous threads rise up.

<table>
<thead>
<tr>
<th>Threads [Maps /s]</th>
<th>Requests</th>
<th>SDE Fires point</th>
<th>SHP Lake polygon</th>
<th>SDE Lake polygon</th>
<th>PostGIS Lake polygon</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100</td>
<td>1.4</td>
<td>6.5</td>
<td>5.1</td>
<td>8.2</td>
</tr>
<tr>
<td>2</td>
<td>200</td>
<td>2.5</td>
<td>12.1</td>
<td>9.8</td>
<td>14.7</td>
</tr>
<tr>
<td>4</td>
<td>400</td>
<td>3.2</td>
<td>13.5</td>
<td>14.0</td>
<td>17.0</td>
</tr>
<tr>
<td>8</td>
<td>800</td>
<td>3.5</td>
<td>13.2</td>
<td>14.8</td>
<td>16.9</td>
</tr>
<tr>
<td>16</td>
<td>1600</td>
<td>3.3</td>
<td>12.6</td>
<td>13.9</td>
<td>16.0</td>
</tr>
</tbody>
</table>

*Table 8: WFS Geoserver tests results. Refer to table 3 for acronyms.*
Chapter 4: How can SDI improve our capacity to discover, share, retrieve and integrate environmental data?

4.3.7.3 WCS

The performance of WCS is tightly bounded to the size and resolution of the data set. The Blue Marble 8 bits color image was stored in low-resolution 3600x1800 pixels and high-resolution. Medium resolution has been produced by directly resampling the high-resolution image through the WCS instance by requesting a smaller number of pixels in height and width. Different storage options have been tested for the high-resolution image consisting in (1) one big geotiff image 45’688x22’509 pixels (2.87 Gb), and (2) 240 tiles of geotiff images 1876x1876 pixels (10 Mb each tile).

Setting up a responsive WCS instance involves accessing efficiently the raster data set. On ArcGIS Server platform, the WCS is faster when the data set is stored in ESRI file geodatabase raster than in flat Tif file. This is true for low-resolution, while in medium-resolution the file geodatabase is exactly the same than the flat Tif file. In high-resolution the performance of the two storage formats are about the same. One hypothesis to explain this pattern is that on heavy load the server is limited by other factors like CPU or memory rather than data access. The overall best result is obtained with ArcGIS Image Server with tif tiled in high resolution. This result is explained by the overview images created automatically when the ArcGIS Image Service is built. These overview images speed up the service for the WCS requests that cover a large area at small scale.
Chapter 4: How can SDI improve our capacity to discover, share, retrieve and integrate environmental data?

**Table 9:** WCS Geoserver (GS) and ArcGIS Server (AGS) tests results. An ArcGIS Server Image Service (IS) has also been tested. Tests have been executed on Linux and Windows (win) operating systems. Lr: Low-resolution, Mr: Medium-resolution, Hr: High-resolution.

<table>
<thead>
<tr>
<th>Threads [Maps /s]</th>
<th>Requests</th>
<th>GS lr tif</th>
<th>GS lr fgdb</th>
<th>S lr tif</th>
<th>GS mr tif</th>
<th>GS mr fgdb</th>
<th>S mr tif</th>
<th>Swin mr geotiff</th>
<th>GS lr tif</th>
<th>GS hr tif</th>
<th>GS hr fgdb</th>
<th>GS IS hr geotiff</th>
<th>Swin hr geotiff</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100</td>
<td>0.0</td>
<td>0.5</td>
<td>0.2</td>
<td>0.3</td>
<td>0.8</td>
<td>0.01</td>
<td>0.08</td>
<td>0.10</td>
<td>0.9</td>
<td>0.11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>200</td>
<td>0.9</td>
<td>0.3</td>
<td>0.3</td>
<td>0.2</td>
<td>0.2</td>
<td>0.01</td>
<td>0.07</td>
<td>0.10</td>
<td>0.5</td>
<td>0.02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>400</td>
<td>0.9</td>
<td>0.3</td>
<td>0.8</td>
<td>0.3</td>
<td>0.1</td>
<td>-</td>
<td>0.11</td>
<td>0.10</td>
<td>0.7</td>
<td>0.04</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>800</td>
<td>0.9</td>
<td>0.4</td>
<td>0.5</td>
<td>0.3</td>
<td>0.3</td>
<td>-</td>
<td>-</td>
<td>0.14</td>
<td>0.10</td>
<td>0.8</td>
<td>0.02</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>1600</td>
<td>0.9</td>
<td>0.4</td>
<td>0.7</td>
<td>0.4</td>
<td>0.4</td>
<td>-</td>
<td>-</td>
<td>0.23</td>
<td>0.20</td>
<td>0.8</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

**Figure 24:** WCS ArcGIS Server tests results

Geoserver has been tested on Linux with Tomcat and on Windows with Jetty. It seems that it performs faster on Linux. When the load on the server rises up in terms of simultaneous threads and with higher resolution, the number of service errors increases and the throughput decreases to very small values. It takes more than one minute to produce each map when the WCS server is overloaded.
Chapter 4: How can SDI improve our capacity to discover, share, retrieve and integrate environmental data?

4.3.8 Discussion

Geospatial data are valuable in various domains and particularly in multi-disciplinary research activities that require integrating different data sets from different sources with different formats. Therefore, having these data readily available and easily accessible is a key requirement. OGC standards and especially WFS and WCS, are of high interest because these can increased interoperability of heterogeneous data sources.

Our aim was not to crystalize around the debate of proprietary versus open source solutions. Instead we wanted to give a first insight around WFS and WCS implementations in two of the most widely used solutions within the GI community and contribute (1) in improving the quality of these services independently on the platform used, (2) in discussing the possibilities to test performances of these services, and finally (3) in sharing our testing scripts and used data to the community (FOSS4G and other testers) in order to improve the proposed testing framework.

Our tests have highlighted that globally these different implementations provide already good general performance right out of the box. However, they clearly show that memory is the most critical factor to control and a lot of elements can influence the response of a given service. Therefore, simplifying the instance (e.g., turning off extraneous services or options), configuring suitable memory parameters (e.g., map rendering, type of data served) and managing the number of request (e.g., limiting the number of concurrent requests that could prevent timely responses, workload manager) are factors that can improve the overall quality of web services (ESRI 2010; OpenGeo team 2010). Various other factors may affect performances in particular if implementations use containers.
like Java Virtual Machine (JVM)\textsuperscript{31} or Jetty\textsuperscript{32}. Tuning the different options provided by these containers may significantly improve performances of the proposed services.

Serving large amount of data efficiently requires optimizing data and storage (ESRI 2010; OpenGeo team 2010). From our results it appears that ESRI file geodatabase (FGDB) appears a suitable format, both for vector and raster data, as it provides good performances compared to flat files or ArcSDE. It is an interesting result and FGDB can potentially become a common and cross-platform format because ESRI has recently released an open Application Programming Interface (API)\textsuperscript{33} meaning that this format can be now virtually used in any GIS software. Native solutions like ArcGIS Server/ArcSDE and GeoServer/PostGIS also give globally good results and may be reliable solutions to serve efficiently data using WFS and WCS standards. When storing vector data into databases, indexing geometry and attributes significantly improve performances by accelerating performances and data retrieval. This is particularly important also because there is not a completely linear relationship between storage size of the data set (e.g., number of feature, attributes) and overall performance of the download service. Regarding raster data, building pyramids, caches and overviews can improve the performance of the proposed service. These operations typically decrease the amount of data sent when panning or zooming by tiling and down sampling an image into a standardized size, and creating overviews as separate image files in a hierarchy (ESRI 2009).

In the same direction, on-the-fly reprojection can be CPU intensive and storing data in the most frequently requested projection may also improve performances. Finally, symbology can also influence the performance of services. In this study we have decided not to test this parameter because (1) this is not the most sensitive parameter (compared to memory), and (2) lots of factors need to be tested in order to achieve a reliable symbology (e.g., classes, forms, texture, size).

In a more general context, WFS and WCS specifications are suitable standards to share and access data in an interoperable manner. However, even if data and storage are tuned and optimized, some bottlenecks may appear when transferring large volume of data (e.g., geographical extent, attributes, features, or resolution). Indeed, WFS uses Geographic Markup Language (GML) encoding (Peng and Zhang 2004; Open Geospatial Consortium 2007; Amirian and Alesheikh 2008) to serve data. Due to its verbose nature, transferring large amount of data may be problematic and leading to high latencies and low performances. Similarly, WCS is very sensitive to data resolution. Consequently, these standards are more suited to share local medium-resolution data than global high-resolution data because of limitations caused by file formats, file size and also network bandwidth. Additionally, differences in OGC specifications have been noted in the various implementations tested (e.g., filter encoding, parameters). This may lead to interoperability problems, especially if clients do not implement the different flavors of these specifications, and may limit

\footnotesize
\begin{itemize}
\item \textsuperscript{31}\url{http://www.java.com/en/}
\item \textsuperscript{32}\url{http://jetty.codehaus.org/jetty/}
\item \textsuperscript{33}\url{http://resources.arcgis.com/content/geodatabases/10.0/file-gdb-api}
\end{itemize}
Chapter 4: How can SDI improve our capacity to discover, share, retrieve and integrate environmental data?

seamless data integration capabilities. Although it was not the purpose of this study, we noticed differences between clients (e.g. ArcGIS, QGIS, uDig) when accessing a layer provided by a given service. Indeed, when downloading a layer in different desktop clients, we realized that some read data faster than others. This is also something to investigate in the future because this can also potentially have an influence on the way users perceive the quality of a service. In this study, we have focused our attention on the performance criteria, which is the most critical factor to test in order to have a sufficient Quality of Service. To get a complete picture of the quality of a given service, it is still required to measure two other parameters: capacity and availability. Capacity is about the limit of simultaneous requests that a service can handle with guaranteed performance. Managing the number of requests can do this in a very effective way. We noticed that the tested implementations manage threads with queuing functionality. After a certain number of threads, the throughput is stable and the service seems to wait until a worker is free and then handles the request. Availability (i.e. the probability that a network service is available) is more difficult to measure as it is strongly related to the architecture of a system. Having multiple instances, load balancers and high availability routers will help ensuring high availability of a service. Monitoring and measuring the status of these components require further investigations.

4.3.9 Conclusions

Spatial Data Infrastructures are seeking to facilitate the access and integration of geospatial data coming from various sources. To achieve this objective systems must be interoperable. OGC specifications are key enablers providing interoperable access to data in an efficient and timely manner. To ensure that a given service is sufficiently responsive to fulfill users expectations and requirements, performance of a given service must be measured and monitored in order to track latencies, bottlenecks and errors that may negatively influence its overall quality. The objectives of this study were to (1) propose an approach to measure the performance of different WFS and WCS implementations, (2) provide some guidance to data provider looking to improve the quality of their services, and (3) share our testing scripts and used data to the community (FOSS4G and others) in order to improve the proposed framework of testing.

Overall performances of the tested implementations are globally satisfactory already without tuning different parameters. This is a positive sign for data providers that are reluctant to install these kinds of tools because of too much technicalities. However, our tests have shown that to achieve reliable services tuning memory is an essential and critical factor – even if memory gets more and more optimized with default software installation. Additionally, optimizing data and storage are factors that can easily increase efficiency of services. Some differences have been highlighted regarding the various implementations of WFS and WCS specifications. This can potentially limit data integration if clients do not implement the different flavors of these specifications. Finally, these specifications are by nature not well suited to transfer large volume of data, and the current specifications are more appropriate to share local medium-resolution data than global high-resolution
Chapter 4: How can SDI improve our capacity to discover, share, retrieve and integrate environmental data?

data. This can be a potential issue, especially given the ever-increasing volume of high-resolution data availability.
4.4 Summary and lessons learned

- Geospatial data sources are often fragmented and metadata are often missing and/or incomplete.

- Cultural behavior, political aspects, and policies have a strong influence in “opening”/sharing data.

- Policies are the most sensitive component. They must encourage and facilitate participation and sharing of data. SDI is a long term process that depends on support and commitment.

- Multi-stakeholder nature of participants requires consensus to accommodate their diversity and heterogeneity while ensuring standardization and interoperability.

- Involving all stakeholders in early development stage of SDI is important in order to reach agreements and develop a conceptual model that fosters sharing spirit, accessibility, availability, interoperability and data harmonization.

- Data quality is very important for users.

- Sharing implies exposing data to the judgment of a community that can recognize whether or not data are of sufficient quality or not.

- Data models are important but currently often missing. They will ease quality check procedures accuracy, completeness, data/metadata harmonization, production, and integration.

- In absence of agreed data models, semantic and schematic interoperability cannot be achieved. This can negatively influence data integration.

- International frameworks (like INSPIRE or GEOSS) may help in providing guidance and reaching commitments to develop widely agreed data models.

- Performances/quality of services are essential because users want to access data in a timely and responsive manner.

- Currently, most users are not aware of possibilities offered by OGC web service. Instead they are looking for traditional downloads.
Partnerships are essential because a single agency is unlikely to have all resources, skills or knowledge to undertake the development of all aspects of a SDI.

Geoportal can be thought as a gateway to geospatial data, maximizing reusability of data, improving availability, accessibility and applicability of geospatial data.

“Open Data” approach is promising.

Data may be useful for other communities that where not envisioned initially.

Major barriers identified in opening data: financial and visibility.

Capacity building must be accomplished at three levels: human, infrastructure, institutional.

Facilitating access and making geospatial data interoperable are recognized as important factors in determining the future success of SDIs.

WFS and WCS are important standards giving access to data.

View services are more developed than download services.

To ensure that services are sufficiently responsive and reliable to fulfill users needs and expectations, performance must be measured and monitored in order to track latencies, bottlenecks and errors.

There are currently no frameworks to measure performances of data services.

Generally software implementations give already good results “out of the box”. However they have differences in syntax that may influence interoperability.

The most critical factor to control and manage is computer memory (RAM).

Data and storage optimization are also important factors.

File Geodatabase appears a promising format.

WFS and WCS are suitable to share and access data in an interoperable way. However they are not well suited to share high-resolution data. This is an issue given the ever-increasing volume of high-resolution data.
Chapter 5:
Can SDI take advantage of distributed computing power to process the increasing amount of high-resolution data?

Based on:

Chapter 5: Can SDI take advantage of distributed computing power to process the increasing amount of high-resolution data?

5.1 Preamble

A recent study (Hilbert and Lopez 2011) showed that human beings were able to store approximately 295 exabytes (or 296 billion gigabytes) of data using digital and analog devices over the past 20 years. These authors consider that 2002 corresponds to the beginning of digital age where digital storage capacity overcame the analog one. Currently, 94 percent of storage devices are in digital form: 52 percent of it on hard disk, 28 percent on optical devices and 11 percent on tapes. In comparison, paper-based storage corresponds to 0.007 percent. Moreover, they estimate that people all around the world have exchanged in 2007 the equivalent of 65 exabytes of data through telecommunications (e.g., Internet, mobile phones). This corresponds to sharing the content of six newspapers every day by every person in the world. In term of processing, all the computers around the world have computed 6.4x10^{10} instructions every second. In 20 years, the total computing capacity increased by 58 percent annually (corresponding to 10 times the US GDP) while telecommunications capacity grew of 28 percent per year and storage capacity by 23 percent. This study showed also that we reached the point where we collect more data than we are physically able to store (Hilbert and Lopez 2011; Science staff 2011). This gap will inevitably increase, especially in data intensive domains such as environmental sciences. Therefore, decisions need to be taken on which data to archive and which to delete.

A related problem is how to access and handle data sets that are becoming too large to be downloaded through the Internet and processed on single desktop computers or servers. A survey made by the journal Science showed that approximately 31 percent of responders are regularly working with data sets ranging from 1 to 100 gigabytes, 20 percent use and analyze data from 100 gigabytes to 1 terabyte, and up to 7 percent exceeding 1 terabyte. Therefore, data management, distribution, heterogeneity, volume and processing are becoming increasingly challenging (Fiore and Aloisio 2011). For these authors, revolutionary and evolutionary approaches should be investigated to improve understanding on (1) dealing with these emerging issues and challenges and (2) offering new perspectives to scientific communities who will be influenced by this data deluge. In the field of environmental sciences, different research areas can be prospected: (a) spatial data infrastructures, (b) distributed computing environments, (c) large scale access and integration of data repositories, (d) workflow management, and (e) metadata management.

As seen previously, geospatial data are widely used in many activities like environmental management, monitoring and assessment, and have taken a remarkable place in our everyday life. As of today, it is estimated that geospatial data is the most important data type in terms of volume accounting for more than 80% of data collected by humankind (Di 2004). Due to their complexity and multi-disciplinary nature, environmental data are very diverse, collected and operated by various data centers around the world. Extracting information often requires complex analyses and processing procedures together with handling vast amount of (high-resolution) data. In such conditions, most users of geospatial data are often experiencing lack of resources to analyze these data. Therefore, the essential task of transforming raw data into understandable information cannot currently fully deliver. Typically, environmental sciences are
Chapter 5: Can SDI take advantage of distributed computing power to process the increasing amount of high-resolution data?

data and computing-intensive domains where data are in general discovered, visualized and access through SDIs, but processed on desktop computers. This clearly limits the types of analyses that can be conducted due to their reduced power and given an ever-growing size of data that need to be analyzed. Hence, SDIs need to go “one step further” by extending their analysis capabilities (Kiehle, Greve et al. 2006), and by adapting solutions to manage and process in a more efficient way the vast amount of geospatial data that are generated and collected every day.

The increasing computational power and network capabilities enable processing of distributed geospatial data over the web (Brauner, Foerster et al. 2009) using Service Oriented Architecture (SOA) principles and web services technologies. Web-based geoprocessing services can therefore be seen as the next logical step to extend SDI capabilities (Friis-Christensen, Ostländer et al. 2007; Kiehle, Greve et al. 2007) by providing access to a collection of geospatial calculations (like in a standalone desktop GIS software) delivering some concrete functionalities (Granell, Diaz et al. 2009; Granell, Gould et al. 2009; Granell, Abargues et al. 2010). The recently introduced OGC Web Processing Service (WPS) specification and the promises of high storage, data management and computing capacities offered by distributed computing infrastructures like Grids and Clouds, offer new opportunities within the environmental communities (Koon Leai and Turner 2006).

At this point, it is important to differentiate between Distributed and Parallel computing. Parallel computing can be defined as the concurrent use of shared-memory with multiple processors (CPUs) or distributed-memory clusters to execute calculations (Foster, Yong et al. 2008). This type of computation is suited for operations that need to exchange intermediate results like climate models (Shujia, Cruz et al. 2010). In comparison, Grid computing can be thought as cluster of shared (unused) computing and storage resources to perform large-scale computing tasks that cannot be completed with single systems (e.g., local server or cluster). Grid can be used for calculations that can be executed independently and requiring many iterations/simulations (e.g., sensitivity analysis) (Lecca, Petitdidier et al. 2011).

Therefore, the paper “Grid-enabled Spatial Data Infrastructure for environmental sciences: challenges and opportunities” published in Future Generation Computer Systems discusses (1) actual status of technologies used to describe, catalog, share and process an ever-growing set of high resolution geospatial data, (2) current analytical limitations of SDIs and presents, (3) promises and challenges offered by Grid to extend analytical capabilities of SDIs. To extend processing capacities of SDIs several attempts to implement the WPS specification in a distributed computing environment have been successfully made. Nevertheless, they are in general dependent on the middleware used by the distributed computing infrastructure meaning that with the current diversity of computing environments it is very difficult to execute and reuse a given WPS process on different computing backends. This constrains service providers to develop a different implementation for any specific backend. This situation can potentially limit the development, adoption, and diffusion of WPS. Consequently, the last paper of this thesis, entitled “WPS mediation: an approach to process geospatial data on different computing backends” submitted to Computers and
Chapter 5: Can SDI take advantage of distributed computing power to process the increasing amount of high-resolution data?

Geosciences presents a possible approach to use WPS specification on distributed computing infrastructures. This approach allows one to execute a given geoprocessing calculation independently of the computing backends (local servers, clusters or different Grids/Clouds) avoiding the need to rewrite processes by making WPS processes as scalable and flexible as possible.
Chapter 5: Can SDI take advantage of distributed computing power to process the increasing amount of high-resolution data?

5.2 Grid-enabled Spatial Data Infrastructure for environmental sciences: challenges and opportunities

Gregory Giuliani¹,², Nicolas Ray¹,², Anthony Lehmann¹,²

Addresses:  
¹University of Geneva, Climate Change and Climate Impacts, enviroSPACE laboratory, Battelle - Building D, 7 route de Drize, CH-1227 Carouge.  

5.2.1 Abstract

Spatial Data Infrastructures (SDIs) are being widely used in the environmental sciences to share, discover, visualize and retrieve geospatial data through Open Geospatial Consortium (OGC) web services. However, SDIs have limited analytical capabilities, an essential task to turn data into understandable information. Geospatial data are typically processed on desktop computers, but their limited power limits the types of analyses that can be conducted given ever-increasing amounts of high-resolution data. With the recently introduced Web Processing Service and the availability of large storage and computing facilities offered by Grid infrastructures, new opportunities are emerging within the environmental sciences communities. The enviroGRIDS project, funded by the European Commission “Seventh Framework Programme” (EU/FP7), will target these issues.

5.2.2 Introduction

Understanding the complex, highly interconnected and continuously evolving processes of Earth-system is a challenging task requiring to gather and integrate different data sets about physical, chemical, biological and anthropic systems (Mohammadi and Rajabifard 2009). These environmental data sets need to be processed to turn them into understandable information before they can be disseminated to appropriate decisions-makers, stakeholders, and the general public.

Environmental data are often spatially referenced and are thus part, in a broader context, of geospatial data. Geospatial data typically describe geographical location giving through various attributes knowledge about their spatial and/or temporal extents. Geospatial data, also known as geodata, are extremely valuable as users can build spatial relationships between feature and data (Nebert 2005). If previously geospatial data were mostly presented in the form of paper maps, they are now used and analyzed within a Geographical Information System (GIS). This computer-based system is capable of assembling, storing, manipulating and displaying geospatial data (Ezigbalike 2004). A GIS gives the ability to merge existing data from different sources, facilitating collaboration in creating and analyzing them. This collaborative approach highlights the need to have harmonized data sets in digital form to store them in databases that allow easy storage and dissemination, facilitating data exchange, sharing and updating, and finally improving the accessibility for multiple
purposes (Philips, Williamson et al. 1999; Henricksen 2007).

As a result of the previous considerations, the concept of Spatial Data Infrastructure (SDI) was developed to facilitate and coordinate the exchange of geospatial data (Rajabifard and Williamson 2001). A SDI encompasses data sources, systems, network linkages, standards and institutional issues involved in delivering geospatial data and related information from many sources to the widest possible group of potential users.

5.2.2.1 Geospatial Service Oriented Architecture

Today's effort on the technical development of SDI components clearly focuses on the exchange of geospatial data and processes in a way that ensures interoperability (Bernard and Craglia 2005) through services that allow efficient access to spatially distributed resources. The shift towards an infrastructure offering services, rather than a stand-alone system allowing one to find, view and analyze geospatial data, is highlighted by the growing importance of the distributed model based on independent, specialized, and interoperable services (Nebert 2005). This shift is driven by the increasing role that geospatial data are playing in our every day life, the maturity of web technologies, and the need for organizations to work efficiently by reusing data, capabilities and invested effort and capitals. Hence, geospatial technologies are evolving from monolithic GIS systems toward Service Oriented Architecture (SOA) (Lee and Percivall 2009), an IT architectural approach defined as “a paradigm for organizing and utilizing distributed capabilities that may be under the control of different ownership domains” (Papazoglou and van den Heuvel 2007). The aim of this architecture is to promote loosely coupled, standard-based, protocol-independent distributed computing so that its components can be reused (Papazoglou and van den Heuvel 2007; Sahin and Gumusay 2008).

In a SOA, the key components to build applications are services. These services are well-defined sets of self-contained and stateless actions that do not depend on the state of other services. In other words, a service is simply a collection of operations that a user can discover and invoke. In the case of the geospatial domain, an operation can be a simple request to create a map or a complicated geoprocessing routine applied to a remote sensing image. These services are defined in a standard manner, have a published interface, and can communicate with other services to achieve a specific process or task. Therefore, an SOA provides the framework and rules for services description, discovery, interaction, and execution (Lee and Percivall 2009). To support reusable deployment of services a common pattern defines three components: service provider, service requestor and service broker associated with three operations: publish, find and bind. A SOA relates these three components to the three operations to allow automated discovery and use of services. It must be highlighted that such an automated process depends on syntactic conventions used by the service(s) involved (i.e. matching inputs, preconditions and outputs). In a traditional scenario, a service provider hosts a service and “publishes” a service description to a service broker. The service requestor uses a “find” operation to retrieve the service description and uses it to “bind” with the service provider and to invoke the service itself. This approach can deliver more flexible and agile systems that are easier to maintain and adapt to evolving
Chapter 5: Can SDI take advantage of distributed computing power to process the increasing amount of high-resolution data?

technologies and requirements than stand-alone software. Indeed, most users of GIS systems are using only a small portion of their software functionalities. Consequently, the open and interoperable environment provided by SOA, based on reusable and standardized services, allows application development to be more focused by providing users just the functionality they need (Alameh 2003).

As of today, implementations of SOA application are mostly realized through the use of web services (Friis-Christensen, Ostländer et al. 2007; Krishnan and Bhatia 2009). Web services support interoperable application-to-application communication over a network (e.g., Internet) and are based, in general, on a maturing set of open standards that are widely accepted and used like eXtended Markup Language (XML), Simple Object Access Protocol (SOAP), and Web Service Description Language (WSDL) (Cömert 2004). This broad acceptance provides a common approach to define, publish, and use web services and permits the development of platform and programming language independent services accessible over standard Internet protocols (Papazoglou and van den Heuvel 2007; Bosin, Dessi et al. 2011). Web services emphasize the necessity that systems involved must communicate with each other meaning that web services rely on interoperability. The core functionalities of this type of service are: communication with Internet protocols (commonly HTTP) and data exchange with formatted XML documents. In addition, it is generally accepted to describe a service using WSDL and to use SOAP to transport XML document over HTTP.

At this point, it is important to distinguish between (1) the generic term of web services that refers to any service provided through the web and (2) the specific web services solutions that are services conforming to a well-defined set of specifications. Hereafter we use the second definition.

Within the geospatial community, SOA is the underlying concept for an interoperable environment based on reusability and standardized components, and thus it is fundamental for SDIs to allow applications and related components to exchange data, share tasks, and automate processes over the Internet (Open Geospatial Consortium 2004). Based on SOA principles, the Geospatial Portal Reference Architecture (Open Geospatial Consortium 2004) specifies four classes of services that are required to implement an efficient SDI using related interoperability specifications: (1) Portal services offer an entry point to discover and access data as well as management and administration capabilities; (2) Catalog services provide information about data and services; (3) Portrayal services focus on mapping and styling capabilities; and (4) Data services concentrate on data access and processing.

To deploy these different service classes, the OGC proposes to use web services relying on XML-Remote Procedure Call (RPC), a protocol that uses XML to encode its calls and HTTP as transport mechanism. Indeed, OGC's first specifications were proposed in late 90's and actual standards (like SOAP/WSDL) were not available back then. However, OGC specifications are continuously evolving and they are now defining new approaches to extend their standard capabilities to use SOAP/WSDL (Open Geospatial Consortium 2005; Open Geospatial Consortium 2006; Open Geospatial Consortium 2006; Open Geospatial Consortium 2007) as well. This is an important requirement so that OGC Web Services (OWS) can be combined with other types of web services and web-based applications (that are not necessarily geo-enabled). Moreover, the
Chapter 5: Can SDI take advantage of distributed computing power to process the increasing amount of high-resolution data?

SOAP/WSDL layer provides useful information for discovering and chaining services. Thus, it allows users to create and manage workflows much more easily using, for example, Business Process Execution Language (BPEL) standard (Fleuren and Muller 2008).

In summary, OWS provides an interoperable framework for web-based discovery, access, integration, analysis, processing and visualization of geospatial data and allows users to build new applications to achieve a specific task based on this set of reusable services.

5.2.2.2 The need for interoperability

Being interoperable means that two or more systems or components are able to transmit or exchange information through a common system and to use that information. The great advantage of interoperability, and the reason why it is an essential building block for a SDI, is that it describes the ability of locally managed and distributed heterogeneous systems to exchange data and information in real time to provide a service (Open Geospatial Consortium 2004).

Following the OGC (Open Geospatial Consortium 2004) there are two levels of interoperability:
- **syntactic** (or technical): when two or more systems are capable of communicating and exchanging data. Specified data formats and communication protocols are fundamental. Syntactical interoperability is required for any attempts of further interoperability.
- **semantic**: the ability to automatically and accurately interpret the information exchanged to produce useful results as defined by the end users of both systems. To achieve semantic interoperability, both sides must defer to a common information exchange reference model, so that what is sent is the same as what is understood. Being syntactically and semantically interoperable can leverage the full potential of interoperability by allowing services to be seamlessly coupled, reusable and available for a variety of applications.

OGC web services share the ability to return an XML document that describes their capabilities (e.g., data available, formats, and functions) using a standard communication method, enabling applications and other web services that implement OGC standards to interoperate. This means that in an OGC-compliant SDI, a user can access data stored in different databases, in different formats, and running on different operating systems.

5.2.2.3 The need for processing capacities

Current SDIs are essentially supporting data discovery, visualization and retrieval, but have typically limited analysis capabilities (Kiehle, Greve et al. 2006). This means that the processing of geospatial data is done in general on the client’s desktop computer, which is an inhibiting factor when very large, and high-resolution data sets must be processed. With the recently introduced Web Processing Service (WPS) and the promises of high storage and computing capacities offered by Grid (Lee and Percivall 2009) and Cloud (Baranski, Schäffer et al. 2009) infrastructures, new opportunities are emerging within the environmental communities (Padberg and Greve 2009). Following Foster et al.
Chapter 5: Can SDI take advantage of distributed computing power to process the increasing amount of high-resolution data?

(Foster, Yong et al. 2008) a Grid is a parallel processing architecture in which computational resources are shared across a network allowing accessing unused CPU and storage space to all participating machines. Resources can be allocated on demand to consumers who wish to obtain computing power. In other words, a Grid aims to harness resources in a dynamic, distributed environment. Recent studies have exemplified a successful approach to extend Grid technology to the remote sensing community (Di, Chen et al. 2003; Muresan, Pop et al. 2008), as well as to the field of disaster management (Kurzbach, Pasche et al. 2009; Mazzetti, Nativi et al. 2009) making OGC web services grid-enabled.

In the last years, Cloud computing has increased in popularity. The term Cloud represents the Internet or whatever large networking infrastructures in which data storage and processing are performed directly on distributed resources provided by third-party storage and processing facilities (Foster, Yong et al. 2008; Buyya, Yeo et al. 2009; Myerson 2009). According to Foster et al. (2008), an evolution from Grid computing gave rise to Cloud computing, which has been a result of a shift in focus from an infrastructure that delivers storage and compute resources (such as Grids) to one that is economy based aiming to deliver more abstract resources and services (such as Clouds). Cloud computing has a business model in which computing resources are packaged as metered services similar to a physical public utility, such as electricity. Popular examples of Cloud infrastructures ran by large companies are the Amazon Elastic Compute Cloud (EC2) and Simple Storage Service (S3), Microsoft Azure platform, and Google App Engine.

Grid and Cloud infrastructures are mostly dealing with the same issues (manage large facilities, methods to discover and use resources). Nevertheless they differ in different aspects such as security, programming model, business model, compute model, data - application models (Foster, Yong et al. 2008; Myerson 2009). Moreover, the targeted communities are different. Grid is mainly used within the scientific community that runs large-scale models and resources-time consuming applications (e.g., climate simulations, particle physics, molecular docking) whereas Cloud targets small to medium companies that wish to scale on-demand their web-based applications without the need to invest in large computational infrastructure (Foster, Yong et al. 2008; Baranski, Schäffer et al. 2009; Myerson 2009).

Although Clouds user interfaces are typically more user friendly than standard Grid user interfaces, Clouds cannot satisfy all the needs of Grid users today. Aspects of collaboration, result-sharing in virtual organizations, and complex data management are still not well covered by Clouds. A first attempt to bring an OGC web service into the Cloud has been successfully achieved highlighting some promises (response time, publish-find-bind pattern not modified, economical aspects) and also some bottlenecks (data allocation, high traffic on servers) (Baranski, Schäffer et al. 2009).

Consequently, Grids and Clouds appear to be promising facilities to extend SDIs capabilities at least for processing large geospatial data sets. In our view, it is therefore too early to consider Cloud computing as a sustainable alternative to access the distributed computing resources within the enviroGRIDS framework. The aims of this paper are (1) to give from a SDI perspective an overview of the actual status of technologies used to describe, catalog, share and process an ever-growing set of high resolution geospatial data, (2) to discuss promises and
challenges offered by Grid to extend analytical capabilities of SDIs, and (3) to present the approaches of the EU/FP7 enviroGRIDS project to make SDIs and Grids interoperable.

5.2.3 Background

The aim of the EU FP7 enviroGRIDS project (hereafter enviroGRIDS) is to build capacities in the Black Sea region to use new international standards, like those proposed by the OGC and the International Organization for Standardization (ISO), to gather, store, analyze, visualize and disseminate crucial information on past, present and future states of the environment of this region to assess its sustainability and vulnerability (see http://www.envirogrids.net). To achieve its objectives, enviroGRIDS will build a grid-enabled Spatial Data Infrastructure (gSDI) serving data, information and services in global and regional initiatives like the Global Earth Observation System of Systems (GEOSS) (GEO secretariat 2005) and being compatible with both the European directive on Infrastructure for Spatial Information in the European Union (INSPIRE) (European Commission 2007) and the United Nations Spatial Data Infrastructure (UNSDI) (Henricksen 2007). The overarching scientific aim of the enviroGRIDS project is to start building an observation system that will address the nine GEO Societal Benefit Areas (SBAs) (disasters, health, energy, climate, water, weather, ecosystems, agriculture, biodiversity) within a changing climate framework. This observation system will contain an early warning system that will inform in advance the decision makers and the public about risks to human health, biodiversity and ecosystems integrity, agriculture production or energy supply provoked by climate, demographic and land cover changes on a 50-year time horizon. This system will allow systematic monitoring and assessment of GEOSS SBAs in the Black Sea region and aims to serve as a decision-making support tool to assist stakeholders to attain sound decisions in a timely manner based on valid scientific information. To support the development of this observation system, the gSDI (currently under development) will provide interoperable and standardized data storing, discovery, accessibility and retrieval as well as processing capabilities based on the Grid infrastructure of the Enabling Grids for E-Science (EGEE) project. Hence, one of the key challenges of the enviroGRIDS project is to bridge the technological gap between SDIs and Grids infrastructures and to make these two infrastructures interoperable.

5.2.3.1 Why do we need Grids?

A Grid infrastructure will be important to address several objectives during the four-year timeframe of enviroGRIDS:
- Running a high-resolution (sub-catchment spatial and daily temporal resolution) water balance model will be applied to the entire Black Sea catchment (2.1 millions square kilometers) using the Soil Water Assessment Tool (SWAT) (Arnold, Srinivasan et al. 1998).
- Adequate sensitivity and uncertainty analysis will be performed on the Black Sea SWAT model. A gridified version of the SWAT-CUP (Abbaspour, Vedjani et al. 2007) tool will be used for that purpose.
- Access to real time data from sensors and satellites will provide early warning
Chapter 5: Can SDI take advantage of distributed computing power to process the increasing amount of high-resolution data?

and decision support tools to policy-makers and citizens. These data may be streamlined into the gSDI to ensure fast computation and dissemination of results.

Because spatial data is very heterogeneous in format and quality across the European community, urgent efforts are needed to organize and standardize spatial data to improve its interoperability. The gSDI will rely on the development of policies, technologies, data, common standards, standard practices, protocols and specifications such as those of the OGC, GEOSS and INSPIRE. Through cataloguing, the Grid infrastructure will help implementing and sharing standardized data sets.

The strong Grid component of the project will foster data interoperability and will certainly trigger new directions of research or alternative ways of analyzing high-resolution data sets. In terms of analysis capacities, this will offer the possibility to shift from a traditional single desktop computer to sizable computing resources, allowing environmental scientists to leverage the full potential of high-resolution spatio-temporal data sets. Moreover, the large worldwide user community of SWAT may greatly benefit from a gridified version of the software and associated tools. The EnviroGRIDS gSDI will be a distributed system built on SOA paradigm that allows a flexible use of services over heterogeneous components and technologies (OGC and Grid services). The functionality provided by web services could be used anywhere over the computing infrastructure by open standards and communication protocols. Making OGC and Grid services interoperable in a SOA is therefore a key requirement for the project. The enviroGRIDS gSDI should be very innovative due to its implementation in a trans-national framework.

5.2.4 Describing and cataloguing geospatial data

Administrations and governments recognize that spatial information is a critical element underpinning decision making for many disciplines (Zhao, Chen et al. 2004; Masser 2007) and must be part of the information infrastructure that needs to be efficiently coordinated and managed for the interest of all citizens (Ryttersgaard 2001; Masser 2007). Nevertheless, these geospatial data, stored in different places and managed by different organizations, are often poorly documented (Nogueras-Iso, Zarazaga-Soria et al. 2005). Therefore, the vast majority of these data are not being used as effectively as they should due to issues such as the lack of awareness of their availability, a poor documentation, and numerous data inconsistencies (Open Geospatial Consortium 2004; Nebert 2005).

Nebert (2005) highlights that data documentation, commonly known as metadata, is an essential requirement for locating/evaluating data and associated services. Moreover, Masser (2005) reinforces this need by asserting that without appropriate metadata services, a SDI fails its main objective of promoting greater and efficient use of geospatial data. Ideally, each newly created data set (e.g., a map, a single file or a collection of data) must be described by metadata allowing users to determine whether the shared data set is useful to meet their needs. In a networked environment, web-based metadata services can act as gateways to geographic information (Masser 2005; Nogueras-Iso, Zarazaga-Soria et al. 2005; Tang and Selwood 2005; Masser 2007) providing
Chapter 5: Can SDI take advantage of distributed computing power to process the increasing amount of high-resolution data?

an entry point to SDIs and allowing users to search for a specific data, to know where to obtain it, and to understand access constraints and the history of data capture. This allows one to interpret correctly the information about data, to trust it and eventually to meaningfully integrate it with data coming from other sources.

Describing a geospatial data set through metadata is thus an essential task, but it is not sufficient to ensure wider knowledge and usage (Ezigbalike 2004). The collected metadata must be accessible, searchable and query-able, which means that metadata must be stored in a catalog system made of a database with an interface that has the required functionalities (Henricksen 2007). In addition, users should not have to individually access different catalogs but rather have the possibility to query from a single entry point collections of metadata stored and maintained in different places. Such capacities could enhance geospatial data access and sharing within and between distributed organizations, avoiding duplication, increasing cooperation and coordination of data collection and, at the same time, preserving data and information ownership (Henricksen 2007).

In the geospatial community, international standards such as OGC and ISO form the basis for most catalog implementations (Senker, Voges et al. 2004). The ISO standard 19115 (Geographic Information – Metadata) defines the schema required for describing geospatial data and services. It provides different information such as identification, spatial and temporal extent, quality, distribution rights or spatial reference system. This standard is complemented by ISO 19139 (Geographic Information - Metadata - Implementation Specification) that defines the XML encoding schema for describing, validating, storing and exchanging georeferenced metadata and by ISO 19119 (Geographic Information – Services) that describes associated geospatial web services. In addition, the OGC Catalog Service for the Web (CSW) specification (Open Geospatial Consortium 2007) was developed to define a standard interface to publish, discover, search and query metadata about geospatial data and related services. This specification allows an independent and interoperable access to geospatial metadata (Nogueras-Iso, Zarazaga-Soria et al. 2005) defining a set of operations like GetCapabilities (retrieves capabilities and characteristics of a service), DescribeRecord (discover the information model and definitions), GetRecords (search the registry and retrieve results) and GetRecordById (retrieve a result by an identifier). In summary, metadata and interoperable catalogs are the basic components of any SDI to facilitate access to data and related resources.

Searching data or related services directly in a search engine like Google is difficult because it will return hundreds of potential documents in response to a simple query like “land cover data”. Fortunately, geospatial data (and services) described through metadata can give information through their coordinates, place names, reference date, or capabilities. Such descriptions can therefore give a solution to refine user's queries by offering a common vocabulary that describes data and services that can be used for searching and retrieval (Nebert 2005). The pre-requisite to search geospatial data and services is that metadata must be stored in catalogs that can support functionalities to search, query and access. These functionalities are commonly known as “catalog services” proposed in the OGC Geospatial Portal Reference Architecture (Open Geospatial Consortium 2004). A catalog service and its user interface allow users to query
Chapter 5: Can SDI take advantage of distributed computing power to process the increasing amount of high-resolution data?

distributed sets of geospatial data or services through their metadata descriptions. A user aiming to locate a specific service needs to access a search user’s interface to fill out a search form and to build a query for a service with certain properties. The search is then sent to a gateway that queries one or more registered catalogues. Each of these catalogues manages their own collection of metadata. By using a common descriptive vocabulary provided by standards like ISO19115/19139/19119, a common search and retrieval protocol like OGC CSW, and a registry of metadata collections, an interoperable search across different catalogues is possible. The Global Earth Observation System of Systems (GEOSS) is a good example of such an interoperable system that can query multiple catalogues registered in its system.

Nevertheless, this set of specifications (CSW and ISO19115/19119/19139), targets essentially data (and related services) discoverability. As of today, processing services based on the Web Processing Service specification does not require the use any metadata standards, such as ISO19115. GetCapabilities and DescribeProcess operations offer a possibility to access some metadata about the WPS service. In addition, WPS standard recommends including WSDL documentation (Sancho-Jimenez, Bejar et al. 2008) but such metadata does not specify the content of the input and output data involved in the process (Yang, Hong et al. 2009). In other words, the lack of adequate service metadata impedes users to discover, evaluate and use WPS processes. Users have to locate a WPS by themselves and then perform GetCapabilities and DescribeProcess requests to determine if a specific service can satisfy their requirements. To overcome this barrier and facilitate WPS discovery, different approaches (Sancho-Jimenez, Bejar et al. 2008; Yang, Hong et al. 2009) have been proposed to enrich the metadata model of WPS’s WSDL document with additional information automatically retrieved from GetCapabilities and DescribeProcess requests. In particular, process description (e.g., input and output data required) can be directly embedded into the generated WSDL document.

In enviroGRIDS, the approach proposed by Yang et al. (2009) for improving WPS discoverability appears to be promising. The authors suggest that clients can get “enriched” WPS’s WSDL documents from a catalog service, with process description and data types, allowing them to find appropriate processing service specified constraints (e.g., data type) as search criteria.

5.2.5 Accessing and sharing geospatial data

Discovering and evaluating data through their metadata is the first functionality that users can expect from a SDI. Once they know the existence of a specific data set, users want to have the possibility to access it either by direct download or through web services.

Many of the decisions that organizations need to make depend on good, consistent, and readily accessible geospatial data to support decision making processes (Rajabifard and Williamson 2001; Stollberg and Zipf 2007).

The OGC has specified a suite of standards supporting the data service class of the Geospatial Portal Reference Architecture (Open Geospatial Consortium 2004) and two of them are of particular interest for data providers and users: the Web Feature Service (WFS) (Open Geospatial Consortium 2005) that provides a web interface to access vectorial geospatial data (e.g., country
Chapter 5: Can SDI take advantage of distributed computing power to process the increasing amount of high-resolution data?

Borders, GPS points or roads) encoded in Geographic Markup Language (GML) and the Web Coverage Service (WCS) (Open Geospatial Consortium 2006) that defines a web interface to retrieve raster geospatial data of spatially distributed phenomena such as population maps or digital elevation models. In addition, the Web Map Service (WMS) (Open Geospatial Consortium 2006) defines an interface to serve georeferenced map images suitable for displaying purpose based on either vector or raster data. A map served through WMS is only a graphical representation of a geospatial data and does not give access to the data itself. These OWS are invoked using URL and each service supports different sets of standardized operations like GetCapabilities (to describe the service) or GetFeature/GetCoverage/GetMap (to retrieve a selected feature/raster data set).

EnviroGRIDS aims at building the capacity of scientists of the Black Sea catchment to publish and use data/metadata using OWS. In consequence, the first crucial step is to teach them how to install, configure and publish their data as well as their metadata in an interoperable manner. This will be done organizing workshops along the whole duration of this 4-year project, covering interoperability, hands-on experience with web portals, information access, open source software (GeoNetwork\(^{34}\) and GeoServer\(^{35}\)) and data/metadata sharing through web services and GEOSS registries.

5.2.6 Processing geospatial data

A key feature of OGC-compliant Service Oriented Architecture, as proposed in the Geospatial Portal Reference Architecture (Open Geospatial Consortium 2004), is that it provides a set of functionalities composed of independent services allowing dynamic integration and composition (Kiehle, Greve et al. 2007). The ability to turn data into understandable information is then dependent on the capacity to acquire data on a specific problem, to apply processing algorithms and then to visualize the result. Chaining web services is the solution envisioned by the OGC to transform raw data into new information by integrating different data sources and different processing steps. As Stollberg and Zipf (2007) mentioned, web services orchestration is a central concept of SOA, and it adds great value through the possibility to re-use “simple” services to solve “complex” tasks.

Currently, users can find and evaluate data using SDIs but once they have identified the required data they have to download it on their desktop computer and process it on specific GIS software (like ArcGIS or GRASS). These pieces of software have the ability to process and concatenate made available either with OGC standards (WFS, WCS, and WMS) or in proprietary formats (like shapefiles). SDIs need to go “one step further” to, first, extend their analysis capacities by providing standardize way to access GIS calculations and, second, to allow complex chaining and orchestration in order to process data and generate new information (Kiehle, Greve et al. 2006).

The recently introduced Web Processing Service specification (WPS) (Open Geospatial Consortium 2007) aims to close such a gap by offering geoprocessing functionalities in a web service environment (Kiehle, Greve et al. 2007). This will

\(^{34}\)http://geonetwork-opensource.org

\(^{35}\)http://www.geoserver.org
allow one to process distributed geospatial data via Internet on remote servers. Through WPS a user can “offer” the possibility to process data to users that do not have such capacities.

Like all OGC standards, WPS provides a set of traditional operations accessible via a URL that allow service description and processes availability (GetCapabilities request), process description and input/output parameters (DescribeProcess request) and finally execution of a selected process (Execute request). The service description request returns a metadata under the form of an XML file that is both readable by humans and machines allowing an automatization of integration procedures using the returned description. After selecting a process that meets the requirements for a specific task, its description interface gives a detailed view of input and output parameters required to execute this process. These parameters could be either complex (like geometries) or literal values. The complex value data type is interesting because it gives the ability to reference remote locations (Kiehle, Greve et al. 2007) in order to access, for example, a WFS service provided by another organization. Finally, the execution interface allows monitoring the progress of geoprocessing task using simple status message. Once the process is finished, the result can be either returned directly to the client or stored on a server.

WPS, like any other OGC standards, relies on XML-RPC specification using HTTP as the messaging protocol and XML as the encoding schema to answer. This contrasts with more traditional service-oriented architectures based on SOAP protocol to exchange structured information. This difference leads to an important problem of communication and thus greatly limits the orchestration and chaining capacities of OGC web services (Stollberg and Zipf 2007). Currently, selection and coordination by a central orchestration engine (semi-automated orchestration) is a widely used approach aiming to provide flexibility and efficiency when building chains of services (Kiehle, Greve et al. 2007; Deelman, Gannon et al. 2009). Workflows are valuable because real scenarios (e.g., creating an earthquake risk map, analyzing the geographical distribution of species) rarely involve few simple tasks. Transforming data into information requires, in general, the sequencing and organization of processes. Hence, workflows can be seen as a series of coordinated analytical and processing steps. They can be described as web-based scripts that automate tasks. W3C standards (e.g., SOAP and WSDL) are important for creating and managing workflows using widely accepted standards such as BPEL (Coleman, McLaughlin et al. 1997; OASIS 2006). BPEL is a workflow description language that models the behavior of web services in a process interaction (Peltz 2003; OASIS 2007) and is used by many service-chaining tools such as Apache Orchestration Director Engine (ODE), Orchestra, EasyBPEL, Kepler, Scientific Dataflow (SciFlo) or Taverna (Elmroth, Hernandez et al. 2010). BPEL defines an XML-based grammar to describe the control logic needed to coordinate the sequence of web services involved in a workflow. It uses WSDL as component model and XML as data model (Foster, Yong et al. 2008). Despite the fact that Kepler, SciFlo and Taverna are engines natively designed to support Grid services, recent studies (Ezenwoye, Sadjadi et al. 2007; Ma, Wu et al. 2008; Folino, Forestiero et al. 2010) applied a successful approach, using the Web Service Resource Framework (WSRF), to extend BPEL capabilities to orchestrate Grid and non-Grid services. This appears to be a promising approach to orchestrate OGC services and Grid
services (Koon Leai and Turner 2006; Fleuren and Muller 2008).

Another problem is that XML does not support raw binary data, which implies that WCS cannot be directly integrated into SOAP messages (Kiehle, Greve et al. 2007). Different approaches have been identified to overcome these problems and to try to successfully orchestrate OWS (Foerster and Schaffer 2007; Schaffer and Foerster 2008) by encapsulating them into SOAP message.

During the second half of enviroGRIDS project time frame, once scientists are sufficiently comfortable using OWS, we will propose new workshops to show them the possibilities offered by workflow engines building chains of services to create customized solutions for solving specific analysis tasks. Following the incremental approach proposed to progressively implement OWS as Grid services we will first use “non-grid” orchestration engines like Apache ODE. “Grid specific” engines (e.g., SciFlo) will then follow when more and more OWS as Grid services are implemented. To reach the project objectives we will rely on a continuous cycle of enhancement by which enviroGRIDS is evaluated (e.g., quality of service, usability, performance) and validated according to user needs. These improvements are based on results of research activities, taking into account developments and constraints in the areas of standardization, technologies, and policies.

Processing high resolution distributed data on a networked environment raises the issue of computational performance, especially when working on a single instance of a single server causes an important decrease in calculation speed and provokes high latencies. Thus, to leverage the full potential of OWS and related processing capabilities, a high performance-computing environment is required. Grid and Cloud computing (Foerster and Schaffer 2007; Baranski 2008; Lee and Percivall 2009) appear to be interesting candidate to empower SDIs.

Environmental sciences are a data-intensive domain in which applications typically produce and analyze a large amount of geospatial data. For example, the recently accessible Digital Elevation Model (DEM) from ASTER (Hayakawa, Oguchi et al. 2008) covers the world at 30m resolution and is composed of 22,600 tiles for a total size of 1.2 TB. Analyzing such a data set at global scale (e.g., running a global flood model) is currently impossible with a single desktop computer due to, first, the huge computation time required to run such a model and, second, the size of the data set itself that cannot be assimilated by current GIS software. Therefore, distributed computing infrastructures like a Grid can be helpful. In the case of a global flood model, the possibility to split this DEM in tiles and run the model on each individual tile through independent jobs can speed up the process of analysis and allow consuming this data set in an efficient manner.

Another good example where Grid computing can be useful is the computation of risk map on natural hazards. Modeling risk at the global scale requires to access and process a large number of data (DEM, population, economical) distributed all around the world in different data centers. Experiences acquired from the development of such data sets in the context of the Global Assessment Report on Disaster Risk Reduction\(^\text{36}\), showed that covering nine types of hazards (cyclones and related storm surges, droughts,

\(^{36}\) http://www.preventionweb.net/english/hyogo/gar/
Chapter 5: Can SDI take advantage of distributed computing power to process the increasing amount of high-resolution data?

earthquakes, biomass fires, floods, landslides, tsunamis and volcanic eruptions) requires about 6,000 CPU-hours of computation (i.e. about 250 CPU-days) on a single desktop computer. These risk maps (and related data sets such as vulnerability and exposure) need to be easily updated when new events occur (e.g., updating the risk map on earthquake after major events). Without access to distributed computing facilities, these data sets can currently only be updated once a year. This negatively influences the timely access to reliable and up-to-date information, which is mandatory for efficient and effective responses required in emergency situations.

5.2.7 Benefits and challenges to use Grids within SDIs

In the previous section, we have seen that OGC web service specifications provide standards to implement interoperable and distributed geospatial data systems following the OGC Reference Model (Open Geospatial Consortium 2008), but at the same time they do not provide secure mechanism to share computing resources to process data. To access a Grid infrastructure, users belonging to different administrative organizations are typically grouped into a specific user community, called a Virtual Organization (VO). A VO is therefore defined as a group of people who share a data-intensive goal. This group of users wants to share geographically distributed resources in a secure way. Users as well as resources must be authenticated by a certification authority before acceptance in the VO (for users) or in the Grid infrastructure (for resources). The acceptance in a VO authorizes users to access resources based on the policies of the VO. Security and confidentiality is of great importance in that context. In environmental sciences, complex data policies typically govern the access to data. As an example, local or regional data concerning the water management may be very sensitive. In their current form, most Grids use encryption and advanced authentication mechanisms, such as public key certificates and different user roles in the VO, to protect data confidentiality. Thus, the Grid paradigm appears to be the ideal candidate to fill this technological gap allowing SDIs to access high performance computing resources.

The main challenge is to be able to use secure sharing and processing mechanisms provided by Grids while using widely adopted OGC interfaces and services within the geospatial community. As Di et al. stated (Di, Chen et al. 2008), it is difficult to connect Grids and SDIs without extensions and customizations. These authors highlight different reasons for that: first, geospatial data differ significantly from other disciplines (e.g., complexity, diversity and volume). Second, there are already widely adopted sets of standard within the geospatial community. Finally, Grid technology focuses on sharing computational resources thus it is not well calibrated for SDIs requirements. Two possibilities are envisioned to combine Grids and SDIs technologies. The simplest approach is to encapsulate the required OGC Web Service and to use Grids only as a backend for processing or accessing resources. The main advantage of this solution is that existing geospatial services are unchanged, gridification process is easy and implementation is independent of the underlying Grids middleware (Baranski 2008). The second approach is the full integration of OGC services in a Grid environment, creating a gridded SDI. It requires extending Grids middleware capabilities to support geospatial data
characteristics and requirements. Several studies have already successfully implemented such approach to benefit from distributed processing and data accessing tasks (Yanfeng, Jack Fan et al. 2006; Di, Chen et al. 2008; Muresan, Pop et al. 2008; Mazzetti, Nativi et al. 2009). All these studies highlight the fact that such a gridification process is not easy to implement, is always middleware-dependent, and is dependent on a broker to allow compliance with the OGC web services. Typical implemented architectures use Grid infrastructures as a foundation for SDIs (Di, Chen et al. 2003; Mazzetti, Nativi et al. 2009). They integrate the web service approach in the upper functional layers allowing easy communication with other systems based on web-oriented technologies (eg. SDIs, e-Government) and Grids as the lower functional layer to support data-intensive computation and large data sets storage. The main objective is to hide the complexity of Grid while preserving OGC interfaces and thus allowing OGC-compliant clients to access and process geospatial data in a Grid environment.

Following Yanfeng et al. (2006), the Open Grid Services Architecture (OGSA) is promising and of high interest to combine SDIs and Grids. Indeed, OGSA aims to make different Grid systems interoperable by introducing the Service Oriented Architecture and related web services concepts (based on SOAP protocol) into Grids (Ghimire, Simonis et al. 2005). Through OGSA-DAI (Data Access and Integration) and OGSA-DQP (Distributed Query Processor), a standardized and uniform service interface is provided allowing data access and integration over a Grid deployment. Padberg and Kiehle (2009) provide an overview of actual incompatibilities between Grids and SDIs. For these authors, service description, service interfaces, service states and security differ on many points. Especially, Grid infrastructures are based on SOAP to invoke operations and WSDL to describe services while OWS are based on HTTP-GET/POST and XML-RPC. This means that without support for SOAP protocol and automated creation of WSDL description document, integration in a Grids workflow could be problematic. Moreover, OGC services are stateless and thus cannot give any information on their state. WPS is the only standard that supports SOAP and could partially send information about its state. Finally, OGC specifications do not include security mechanisms, a key requirement in Grid architectures. Hence, these limitations must be overcome to leverage the full benefit of Grids to the geospatial community and WPS standard appears as an interesting candidate to be grid-enabled. Recognizing these limitations, the OGC and the Open Grid Forum (OGF) have signed in late 2007 a Memorandum of Understanding to collaborate on the integration of WPS specification into Grid environments and workflow management tools, as well as on the integration of federated catalogs/data repositories with Grid data movement tools like GridFTP (Lee and Percivall 2009).

### 5.2.7.1 Grid-enabling catalogs of geospatial data

In Grid infrastructures, middleware like gLite or Globus Toolkit have their own metadata catalog to describe and localize the distributed data generated by Grid applications. In general, these catalogs associate simple descriptive attributes to the files and suppose a hierarchical, file-based data model. As such, they do not cover the requirements of a geospatial data catalog service in terms of spatial, temporal, and other parameters for data discovery (Di, Chen et al.
Chapter 5: Can SDI take advantage of distributed computing power to process the increasing amount of high-resolution data?

2008). This means that Grid metadata catalogs appear to be inadequate to deal with geospatial metadata (Zhao, Chen et al. 2004) and their functionalities need to be extended to support more complex data types and relationships. Different solutions have been explored to grid-enable geospatial metadata catalog, ranging from a simple wrapper to extend Grid metadata catalog capabilities (Zhao, Chen et al. 2004) to a full integration of CSW standard and ISO schema as a Grid Service (Di, Chen et al. 2008). The last solution seems promising as it converts a web service into a Grid service, while preserving interfaces as well as request and response messages of the OGC CSW specification. Nevertheless, this approach still requires a lot of development to overcome the barriers mentioned previously and to make geospatial metadata catalogs grid-enabled. An interesting work done by Sandoval (2006) has shown a great potential in linking GeoNetwork geospatial metadata catalog with AMGA37 (ARDA Metadata Grid Application) used in gLite middleware. This author has successfully extended traditional XML schema used in GeoNetwork (Sandoval 2006) to take into account Grid’s specific information such as Logical File Name (LFN) or Grid Unique IDentifier (GUID) used to localize data in Grid environment. Sandoval (2006) has also shown the benefits of Grid environment to store and process satellite images.

5.2.7.2 Grid-enabling geospatial data services

Currently, no Grid services are equivalent, in terms of functionalities, to WFS and WCS specifications (Di, Chen et al. 2008). Moreover, data Grids appear to be an interesting approach (Coetzee and Bishop 2009) to deal with large amount of distributed data, benefiting from secure controlled sharing and management capabilities offered by a Grid. Padberg and Greve (2009) have grid-enabled OGC data services using OGSA-DAI data store implementation allowing users to invoke OGC-compliant WFS and WCS services and accessing data stored inside a Grid infrastructure. Di et al. (2008) have applied the same approach they used for grid-enabling CSW to successfully access OGC-compliant data sources over a Grid, making them traditional Grid services. In term of performance, these authors noticed that the Grid services offer a performance overhead compared to the Web services, due principally to the authentication cycles (Padberg and Greve 2009) and the size of the request and associated response payload. Mazzetti et al. (2009) have also grid-enabled WCS specification extending gLite middleware functionalities and they highlighted benefits both in term of scalability (capacity to deal with multiple requests, sending multiple jobs in parallel) and interoperability (between OGC WCS and gLite middleware).

By its distributed nature and characteristics, Grid environment is potentially an interesting choice for a data management system (Muresan, Pop et al. 2008). It offers robustness (distribution storage and data replication capabilities), efficiency (data stored as close as possible to components that access them) and transparency (hiding Grid complexity to users). In addition, data moving protocol like GridFTP are interesting as well. Indeed, one possible bottleneck when dealing with large data sets is that data access strategy in SDIs is not location-based. This means that current SDIs have limited replication and

37http://amga.web.cern.ch/amga/
data transfer capabilities to minimize the access time to a selected data set. Often, geospatial data providers do not offer possibilities to have different replicas distributed in several data centers. Therefore, users must access data directly at its source, which implies that network distance and a potential large number of concurrent accesses can impede users to retrieve data in an efficient way. Moreover, if a geospatial data source disappears, data access to this source is definitively lost if data replication was not enforced. In summary, data replication mechanisms proposed by Grids are promising to avoid single points of failure, to enhance data availability (Lee, Xu et al. 2009; Scholl, Bauer et al. 2009; Sanchez-Artigas and Garcia-Lopez 2010), and to ensure that data will be as close as possible to the worker nodes (avoiding high latencies produced by the movement of large size data before the beginning of a processing task).

5.2.7.3 Grid-enabling geospatial processing services

The Memorandum of Understanding between OGC and OGF primary focuses on the integration of WPS into Grid environments aiming to make high-performance computing available to a wider community (Werder and Krüger 2009). Proofs of concept of such implementations have already been made in different Grid middleware such as Unicore (Baranski 2008), Globus Toolkit (Di, Chen et al. 2008; Padberg and Greve 2009) and gLite (Mazzetti, Nativi et al. 2009). These different studies have shown clear improvements in processing performance and speed, by dividing a given task into smaller subtasks that can be processed in parallel and merged together at the end. Werder and Krüger (2009) stated that the real processing benefit of Grids comes from the development of efficient strategies to parallelize tasks. Indeed, not all environmental models can be parallelized (e.g., climate models) due to their high interdependence and process logic, and are better deployed through supercomputers. For these authors, other factors like system architecture, tiling strategies and orchestration (Fleuren and Muller 2008) could influence the overall result of a geoprocessing task and must be taken into account. Particular attention must also be payed to time investment for parallelizing a specific process. Padberg and Kiehle (2009) highlight the importance to define a “break-even-point” where the gain in computational speed outweighs the overhead induced by Grid technologies and implementation. A promising approach in grid-enabling WPS is proposed by Padberg and Greve (2009). First, they try to address the differences described previously between OWS and Grid services (service description, service interface, statefulness and security) and, second, to preserve traditional WPS interface when connecting to a Grid. The gridification is made at the level of the process ensuring that WPS requests are not modified. For that purpose, the authors suggest to split a WPS process into two parts, one inside the Grid infrastructure (containing the process logic, methods and functions) and one through a traditional WPS interface (that only invokes the Grid service). Another possibility to grid-enable WPS is represented by the encapsulation of Grid processing services within a standard WPS request (Woolf and Shaon 2009) by encoding directly Job Submission Description Language (JSDL) to describe job and resource requirements (disk space, CPU and other parameters) directly into the Execute request of WPS interface. All these approaches have shown that grid-enabling WPS is feasible and could increase
Chapter 5: Can SDI take advantage of distributed computing power to process the increasing amount of high-resolution data?

processing capabilities of SDIs.

5.2.8 EnviroGRIDS approaches to interoperability between SDIs and Grids

Interoperability is a great challenge for the successful implementation of the enviroGRIDS gSDI. Such a technology can significantly reduce problems associated with archiving, manipulating, analyzing, and utilizing large volumes of geospatial data at distributed locations. EnviroGRIDS gSDI is a distributed system built on a SOA that allows a flexible use of services over heterogeneous architectural components and technologies. The OGC Web Services and the gLite middleware must be able to communicate and interact with each other in order to combine the complex specialized geospatial functionalities with the computation capacities of the Grid.

In the enviroGRIDS project, several applications are intended to be ported on the grid, through a so-called "gridification" process. This process aims to generate a Grid application that can be defined as "software that interacts with Grid services to achieve requirements that are specific to a particular VO or user". The gSDI will be the core of the Grid activities within enviroGRIDS. The different components of the gSDI (figure 26) will be implemented throughout the project, and it is likely that many challenges will emerge that are not foreseen at this early stage of the project. The need for sustainable access to the Grid infrastructure stems from the need of a continuous offering of web and Grid services, and for the future EG web portal(s).

![Figure 26: EnviroGRIDS gSDI components supporting web portals.](image)

EnviroGRIDS architecture consists of three main functional layers (figure 27):
Chapter 5: Can SDI take advantage of distributed computing power to process the increasing amount of high-resolution data?

data layer, Grid layer and service layer.

![Diagram of EnviroGRIDS functional layers]

(1) **Data layer**: consists of stored data (data repositories) and functionalities required to manage the repositories. They store raw data (e.g., geospatial data) as well as processed data (e.g., output data such as maps or tables). They also store application data, which are specific for each application type and instance (e.g., hydrology, climate, soil, etc). The register stores metadata catalogues that support the searching, discovering and using of distributed data by the user applications, and processing services.

(2) **Grid layer**: provided by the EGEE infrastructure and the gLite middleware and giving access to the basic resource management and data processing services available over the Grid infrastructure as secure and persistent services and over the Web as stateless services. The services encapsulate the basic functionality provided to user applications:

- **Data Management**: provides the basic operation on data repositories (e.g., data access, transfer, replication, metadata storage).
- **Security and User Management**: provides the functionality needed to work with VOMS database, to support user authentication, authorization, and credential management as well as the implementation of particular policies for data access and use.
- **Scheduling**: provides optimal resource allocation and sharing through static or dynamic load balancing.
- **Monitoring**: supports evaluation of the execution performance, and statistical analysis.
- **SWAT Management and Execution**: provides the functionality to control the execution over the Grid of the SWAT modules and related data.
Chapter 5: Can SDI take advantage of distributed computing power to process the increasing amount of high-resolution data?

- Workflow Management: supports the graph description of the processing, service composition, Grid mapping, workflow interpretation and execution, and fault recovering.
- Spatial Data Acquisition: supports the working with sensors by supervising sensor status, data acquisition and transformation, store, and processing.
- Visualization and GIS Mapping: supports data visualization in graphical user interfaces.

(3) Services layer: The enviroGRIDS Portal will expose to the user a set of tools and applications of which functionality is composed of the services (data management, geospatial functionality, security and user management, scheduling, monitoring, SWAT related management and execution, workflow management, spatial data acquisition, and visualization) provided by the below level. There are four types of interactive applications and tools available through the enviroGRIDS Portal:

- Applications/ SWAT Scenarios Development Tools: the user may develop various scenarios/workflows for natural phenomena and use cases, perform their execution over the Grid, and finally visualize the results and analyze statistical data.
- Data Management Tools: data administrators and providers may access, upload, update, and organize spatial data.
- Decision Maker Tools: provide the possibility to develop and execute various scenarios on different data series, in order to analyze and make predictions on the phenomenon evolution.
- Citizen Tools: provides the citizen, as an Internet visitor, the ability to execute a given set of scenarios by limited set of data, and graphical visualization of the results.

EnviroGRIDS applications will be in the Web domain while data repositories and resources management and processing services will be progressively implemented in the Grid following an incremental approach enabling communication between SDI and Grid infrastructures (figure 28).

![Figure 28: Scenarios to implement SDI/Grid interoperability in the enviroGRIDS project.](image_url)

- The first scenario is a file-based communication between SDI and Grid infrastructures. A user sends a request to the portal application that forwards the request to a proxy server. The proxy server identifies the different calls (for SDI calls it executes OGC services directly to extract...
Chapter 5: Can SDI take advantage of distributed computing power to process the increasing amount of high-resolution data?

required data from the data repository, for Grid calls it uses Ganga functionalities to submit the job to the Grid). Finally, the proxy waits for the results from the SDI and Grid environments, merges the final result and sends it to the client.

- The second scenario is an extension of the first one where the data repository is still in the SDI but Grid accesses data directly using modified OGC web services (grid-enabled) that replace the proxy of the first scenario. This allows Grid services to extract data using WFS or WCS that are outside the Grid.

- The third scenario is where data repositories are not in the SDI part but directly integrated in the Grid infrastructure. Geospatial applications (e.g., geoportals, GIS software) search and use data through Grid services that are compliant with OGC specifications. In such an incremental approach, complexity will increase following the level of adaptation of OGC and Grid services (e.g., modifications of the Grid middleware and OGC’s implementation). In the first scenario, no modification of OGC standards is required. What needs to be done is to write a proper proxy component that can divide SDI and Grid calls, to manage Grid security, and to merge the results obtained under Grid and SDI environments. In the second scenario, different solutions have already been presented in the previous section (grid-enablement of OGC services) and as of today the 52North implementation appears to be a promising one. The last scenario needs a lot of adaptation, as this requires extending gLite middleware capabilities to make it spatially aware. In other words, specific libraries need to be written and added directly into the middleware. Such an implementation is under development within the gLite-OWS (G-OWS) working group. Indeed, in the first two scenarios Grids are only used as a backend to process large data sets (both in term of data size and computation time required) provided by a SDI through OGC web services. In the last scenario, data repositories are completely part of the Grid infrastructure and they can benefit from Grids high level of performance and reliability (e.g., data replication, data security). By making geospatial data and related services available in the Grid environment, this can potentially pave the way to the development of innovative workflows involving large data sets obtained from different disciplines. As an example, we can mention the very promising idea of linking environmental data with genetic data (Breckenridge, Pierson et al. 2003).

Following Mazzetti et al. (2009) the major benefits of making Grid technology spatially aware consist of:

- Scalability: the grid infrastructure provides high processing and storage capabilities to: (1) improve output data resolution, (2) improve model complexity, (3) widen the covered area (from local to national/global scales), (4) improve the time-of-response (from hours to minutes), and (5) time-of-response is almost independent of resolution/coverage.

- Flexibility and interoperability: the geospatial service layer allows one to (1) integrate new and heterogeneous input data, (2) integrate outputs in a higher level application chain, (3) facilitate models interoperability and composition, and (4) be interoperable with other standard based infrastructures (e.g., GEOSS).

38 http://ganga.web.cern.ch/ganga/
39 http://52north.org/maven/project-sites/wps/52n-wps-site/
40 https://www.g-ows.org/
Chapter 5: Can SDI take advantage of distributed computing power to process the increasing amount of high-resolution data?

5.2.8.1 EnviroGRIDS use-case: satellite image processing

Remote sensing gives the opportunity to have access to continuous data collection. In the context of this project, satellite images will be useful to monitor changes and trends in the Black Sea region/watershed (e.g., land use, deforestation, water quality). Monitoring of land cover / land use is an important element for quantifying land surface characteristics for environmental management. Processing high-resolution/high volume of remotely sensed data requires high computation resources and massive data storage capacity. The main processing consists of imagery classification that can be defined as a search of information through various combinations of multispectral bands. The data exploration, analysis, and interpretation are a multivariable process considering satellite image types (e.g., MODIS, Landsat, and QuickBird), geographical areas, soil composition, vegetation cover, and context (e.g., clouds, snow, and season). All these specific and variable conditions require flexible tools and friendly user interfaces to support an optimal research for the appropriate solutions.

The following user requirements have been highlighted: (1) satellite image visualization tool (search images in a database, zoom in/out, scale, metadata), (2) flexible description and execution of complex processes (workflows), (3) satellite image processing with simple algorithms (indices calculation, map algebra), (5) output visualization (pseudo color rendering, classes), (6) save images in different formats, (7) crop images to an area of interest, (8) display information about the image.

To answer enviroGRIDS user needs, the proposed solution for satellite image processing will be based on the Environment oriented Satellite Data Processing Platform (ESIP) (Gorgan, Bacu et al. 2009) that has been developed through the South East Europe-GRID-eInfrastructure for regional eScience41 (SEE-GRID-SCI) project. ESIP is a suite of interactive toolset supporting the flexible description, instantiation, scheduling and execution of the processing over the Grid infrastructure. ESIP layer is built on top of the gProcess platform, a collection of Grid services and tools providing the functionalities mentioned previously and allowing the development and execution over the Grid of workflow to process remotely sensed images. It supports the exploration of optimal solutions for Grid processing and information searching in multispectral bands of the satellite images. The architecture of gProcess is based on a client-server model developed for the gLite middleware. The services exposed by the server side supports the access to Grid infrastructure resources and distributed databases, while the client side (web and desktop applications) accesses the services of gProcess through SOAP web services. The set of implemented operators can be used in the definition of various vegetation or water indices or other satellite image processing algorithms. The current operators work on GeoTiff images and operate on 512 x 512 tile dimension. Images are processed by dividing them into smaller pieces, applying the desired operator and then reconstruct the final image by merging all of the smaller pieces.

EnviroGRIDS offers new possibilities to further refine ESIP platform as well as implementing the support to OGC web services to get an interoperable access

41 http://www.see-grid-sci.eu/
Chapter 5: Can SDI take advantage of distributed computing power to process the increasing amount of high-resolution data?

to data coming from different sources. Finally, this will give partners the possibility to test some already existing geospatial-oriented grid services and to become familiar to build processing workflows.

5.2.9 Conclusions and outlook

In our everyday life, geospatial data have taken a remarkable place allowing us to continuously access a large amount of data ranging from a car position using a GPS to results of complex simulations (such as climate models). In other words geospatial data are omnipresent. One of the challenges we are facing today is to make sense of this vast amount of data in order to turn them into understandable information to support decision-making processes. This requires analysis capabilities that current Spatial Data Infrastructures cannot fully provide. Moreover, the increasing spatial and temporal resolution of geospatial data causes a tremendous challenge for their computation, with which traditional SDIs cannot cope. To address these challenges, the environmental science community is looking with interest to Grid computing infrastructures because these can satisfy the increasing need for processing power and storage capacity, can improve accessibility to distributed storage and computing resources, and can provide a reliable and secure infrastructure. In other words, Grids have the potential to underpin SDIs services and resources.

To achieve the goal of linking Grids and SDIs, interoperability appears to be a key requirement and an important and challenging task. In particular, the implementation of SOAP messaging protocol into OGC standards is a necessity. This will greatly enhance the gridification process of OGC Web Services as well as allow easier workflows integration using orchestration engine to combine OWS and Grid services (Fleuren and Muller 2008).

Using Grid as a computational backend represents only a first step, and currently there is no agreed and common solution to gridify OGC Web Services while remaining OGC compliant (Baransi 2008; Baranski, Schäffer et al. 2009). The integrative approach proposed by the G-OWS Working Group is very promising, extending gLite middleware capabilities with OGC specifications and thus making Grid infrastructures based on gLite spatially aware. This will allow SDIs users to rely on existing standards while hiding the complexity of the Grid. Such a gridded SDI approach could provide benefit to both environmental science and Grid communities by enhancing discoverability, accessibility, processing and retrieval of geospatial data. As a result, new opportunities and collaborations could emerge.

Invoking grid-enabled OWS through mainstream desktop GIS application like ArcGIS\(^{42}\) or GRASS\(^{43}\) would also be a major achievement allowing users to access seamlessly different resources depending on their needs (e.g., data retrieval, processing or map making).

The first generation of SDIs, based on a product model, gave way to a second generation at the beginning of the years 2000 that is characterized by a process model (Rajabifard 2002; Rajabifard and Williamson 2004; Masser 2005). For Masser (2007) this evolution emphasizes the shift from the concerns of data

\(^{42}\)http://www.esri.com/arcgis
\(^{43}\)http://grass.osgeo.org/
Chapter 5: Can SDI take advantage of distributed computing power to process the increasing amount of high-resolution data?

producers to those of data users and the shift from centralized structures to decentralized and distributed networks. In our view, connecting Grids and SDIs could potentially mark the advent of a third generation of SDIs extending their capacities to, and benefiting from, Grid infrastructures. These grid-enabled SDIs have the potential to become a powerful tool within the multi-disciplinary field of environmental sciences, empowering researchers to explore new venues to better understand the vast complexity of the interactions between anthropic and natural systems.
Chapter 5: Can SDI take advantage of distributed computing power to process the increasing amount of high-resolution data?

5.3 WPS mediation: an approach to process geospatial data on different computing backends

Gregory Giuliani1,2,3, Stefano Nativi4, Anthony Lehmann1,2,3, Nicolas Ray1,2,3

Addresses:
1 University of Geneva, Institute for Environmental Sciences, Climatic Change and Climate Impacts, enviroSPACE Lab., Battelle – Building D, 7 route de Drize, CH-1227 Carouge, Switzerland.
3 University of Geneva, Forel Institute, 10 route de Suisse, CP 416, CH-1290 Versoix, Switzerland
4 Institute of Methodologies for Environmental Analysis, C.da S. Loja - Zona Industriale, I-85050 Tito Scalo, Italy.

5.3.1 Abstract

The OGC Web Processing Service (WPS) specification allows generating information by processing distributed geospatial data made available through Spatial Data Infrastructures (SDIs). However, current SDIs have limited analytical capacities and various problems emerge when trying to use them in data and computing-intensive domains such as environmental sciences. These problems are usually not or only partially solvable using single computing resources. Therefore, the Geographic Information (GI) community is trying to benefit from the superior storage and computing capabilities offered by distributed computing (e.g. Grids, Clouds) related methods and technologies. Currently, there are no commonly agreed approaches to grid-enable WPS. No implementation allows one to seamlessly execute a geoprocessing calculation following users requirements on different computing backends, ranging from a stand-alone GIS server up to computer clusters and large Grid infrastructures.

Considering this issue, this paper presents a proof of concept by mediating different geospatial and Grid software packages, and by proposing an extension of WPS specification through two optional parameters. The applicability of this approach will be demonstrated using a Normalized Difference Vegetation Index (NDVI) mediated WPS process highlighting benefits and issues that need to be further investigated to improve performances.

5.3.2 Introduction

Spatial Data Infrastructure (SDI) is a widely accepted concept to facilitate and coordinate the exchange and sharing of geospatial data among different organizations through network technologies (Kiehle, Greve et al. 2006). A SDI realizes a spatially enabled Service Oriented Architecture (SOA) in which standardized interfaces provide access to functionalities as a set of independent and interoperable services (Granell, Diaz et al. 2009). The objective of this architectural approach is to promote loosely coupled, standard-based distributed computing so that developed components can be reused. Different standards proposed by the Open Geospatial Consortium (OGC), the International Organization for Standardization (ISO), the World Wide Web Consortium (W3C)
and other standardization bodies are used enabling interoperability between geospatial data and services. Brauner et al. (2009) subdivide the services that handle geospatial data into three categories: catalog, data, and processing services.

Currently SDIs are mainly concerned with catalog and data services allowing data discoverability, retrieval and visualization (Baranski 2008; Schaeffer 2008). However, the real added value in geospatial data handling is to turn data into usable information to answer a complex query or support a decision. This requires: finding and retrieving data, applying specific calculations, and finally visualizing the result. Commonly, users still process data on their desktop computers using Geographic Information Systems (GIS) software, like ArcGIS44 or GRASS45 (Kiehle, Greve et al. 2006).

The increasing computational power and network capabilities enable processing of distributed geospatial data over the web (Brauner, Foerster et al. 2009) using SOA principles and web services technologies. Web-based geoprocessing services can therefore be seen as the next logical step to extend SDI capabilities (Friis-Christensen, Ostländer et al. 2007; Kiehle, Greve et al. 2007) by providing access to a collection of geospatial calculations (like in a standalone desktop GIS software) delivering some concrete functionality (Granell, Diaz et al. 2009). Li et al. (2010) have successfully developed a prototype to make available GRASS modules and algorithms using Simple Object Access Protocol (SOAP)-based web services. These authors highlighted that: (a) the interoperability of web services improves the sharing of geospatial data by applications on different platform, (b) the modularity of web services enables the sharing of specific geospatial processes by a wide range of users.

In 2007, the OGC has introduced the Web Processing Service (WPS) specification with the aim to propose a standardized interface for publishing and performing geoprocessing tasks in a web services environment (Open Geospatial Consortium 2007). In the last years, different implementations have been proposed that demonstrated the applicability of the WPS approach (Kiehle, Greve et al. 2006; Stollberg and Zipf 2007; Brauner and Schaeffer 2008; Diaz, Granell et al. 2008). In particular, the reusability and the possibility to chain processing services and solve specific and complex problems have been emphasized. In addition, these authors showed that servers are in general more powerful than desktop computers allowing users: (a) to process more rapidly a given data set, (b) to process larger data sets (in term of spatial resolution, spatial extent or file size).

However, users can experience a lack of computing power when they process large data sets -such as the global ASTER Digital Elevation Model (DEM) (Hayakawa, Oguchi et al. 2008) at 30m resolution- or run complex simulations (e.g., dynamic climate models) requiring several CPU hours or days of calculations. In such situation the use of distributed computing appears to be an interesting solution (Lee and Percivall 2009). Distributed computing is a form of computation in which many calculations are carried out simultaneously on several computing elements linked over a network. The term “distributed” should be distinguished from “parallel” computing that commonly refers to

44 http://www.esri.com/software/arcgis
45 http://grass.osgeo.org

Chapter 5: Can SDI take advantage of distributed computing power to process the increasing amount of high-resolution data?
processing tasks that are executed simultaneously on multiple processors on a single computer. Various distributed computing platforms are available such as Grids, Clouds, and Clusters.

Different approaches have been developed to extend SDIs capabilities to use either Grids (Di, Chen et al. 2003; Gorgan, Bacu et al. 2009; Mazzetti, Nativi et al. 2009; Folino, Forestiero et al. 2010) or Clouds (Baranski, Schäffer et al. 2009). All these authors showed benefits in term of high calculation performance and improved availability of services but also highlighted differences (e.g. service description, service interface, service state, security) between SDIs and distributed computing infrastructures (Padberg and Kiehle 2009).

Several attempts to implement the WPS specification in a distributed computing environment have been successfully made. Nevertheless, they are in general dependent of the middleware used by the distributed computing infrastructure: some implementations are working on gLite46 (Muresan, Pop et al. 2008; Mazzetti, Nativi et al. 2009), Globus47 (Di, Chen et al. 2008), or Unicore48 (Baranski 2008). In theory, a developed process might be reused accross different WPS frameworks. However, in practice, this is limited due to different programming languages and Application Programming Interfaces (APIs). In other words, a service provider who wants to share a geoprocessing task using the WPS specification must develop a specific version of that process for each specific backend supported by a dedicated WPS implementation.

This means that the scalability in term of execution and reusability of a given WPS process on different computing backends is currently restricted. This situation can potentially limit the development, adoption, and diffusion of WPS.

The aim of this paper is to present a proof of concept to enhance WPS usability allowing one to execute a given geoprocessing task, with a dedicated WPS implementation, independently of the computing backends (e.g., local server, cluster or different Grids/Clouds) avoiding the need to rewrite processes by making WPS processes as scalable and flexible as possible.

5.3.3 Web Processing Service and distributed computing

The OGC Web Processing Service specification (Open Geospatial Consortium 2007) aims to provide a standardized way to access geo-processing algorithms in a web service environment, which consequently extends SDIs analysis capabilities (Kiehle, Greve et al. 2006; Schaeffer 2008).

Brauner et al. (2009) reported that performance and processing power are crucial in the context of geoprocessing services, especially in the case of large-scale data sets. To leverage the full potential of WPS, a high performance-computing environment is consequently a key requirement. Distributed computing promises to support SDIs, especially Grids and Clouds (Foerster and Schaffer 2007; Baranski 2008; Baranski, Schäffer et al. 2009; Lee and Percivall 2009).

---

46 http://glite.web.cern.ch/glite/
47 http://www.globus.org/toolkit/
48 http://www.unicore.eu
5.3.3.1 Web Processing Service

In respect of traditional geo-processing implementations, WPS-enabled processes are flexible and remotely accessible algorithms available through web services (Kiehle, Greve et al. 2006) that can be reused in different scenarios.

The principal element of a WPS is the notion of process that is a geospatial calculation with defined inputs and outputs (Granell, Diaz et al. 2009). This implies the following steps: (a) to find suitable geospatial data needed to run the algorithm, (b) to initiate the process, (c) to control the output, (d) to make the results available to the client.

To work in a web service environment, a WPS instance must offer various operations accessible through web communication. In particular, descriptions of geoprocessing tasks with the help of metadata that are accessible, usable and understandable both by humans and other web services (Kiehle, Greve et al. 2006) are key elements to build chains of services (Schaffer and Foerster 2008). The interface is based on three operations (figure 29) which can be called using either HTTP-GET and key-value pair (KVP), or HTTP-POST and eXtended Markup Language (XML)-encoding (Schaffer and Foerster 2008).

![Figure 29: Communication pattern between a WPS-client and a WPS-instance.](image)

The GetCapabilities operation provides an XML document defining service metadata (e.g., server provider, contact information), abstracts and a list of available processes offered by the queried WPS instance. Once users have selected a required process, they can perform a DescribeProcess operation to retrieve process metadata within an XML document including parameters descriptions (e.g., input and output parameters). Finally, with the Execute operation it is possible to initiate the selected geoprocessing service with all necessary input data. The WPS instance will run the calculation and sends back the result to the client.
Chapter 5: Can SDI take advantage of distributed computing power to process the increasing amount of high-resolution data?

These three operations are invoked through Uniform Resource Locator (URL) as shown in following examples:

(1) *GetCapabilities* request:

```plaintext
http://localhost/cgi-bin/wps?service=WPS
&request=getcapabilities
```

(2) *DescribeProcess* request for the “buffer” process:

```plaintext
http://localhost/cgi-bin/wps?service=WPS
&version=1.0.0&identifier=buffer&request=describeprocess
```

(3) *Execute* request using “cities.gml” data set as input:

```plaintext
http://localhost/cgi-bin/wps?service=WPS
&version=1.0.0&identifier=buffer&request=execute&data=http://foo.bar/cities.gml; width=5
```

The required input data can be delivered across a network or made available on a server. The WPS specification discerns three input/output data types:

1. *LiteralData*: can be any character string like string, float, integer or boolean.
2. *ComplexData*: used for vector and raster data by sending directly the data to the server or by referencing a remote data source.
3. *BoundingBox*: to be used within a specific area.

The output of a geoprocessing service can be obtained either by a direct response to the request (e.g., result sent directly to the client) or can be stored on the server as resource accessible through the web using URLs (Schaeffer 2008). In the latter case, the *Execute* response will be an XML document providing the URLs to access each stored output.

To ensure that processes are reusable, the specification defines WPS Application Profiles enabling optimization of interoperable client user interface behavior, as well as the semantic discovery (i.e., publish/find/bind pattern) and orchestration (Open Geospatial Consortium 2009; Lanig and Zipf 2010). Such level of interoperability can be realized only if each process is described in a dedicated Application Profile. Consequently, a WPS Application Profile is a document describing how WPS shall be configured to serve a process that is recognized as an OGC WPS process (Open Geospatial Consortium 2009; Lanig and Zipf 2010). An application profile consists of: (1) an OGC Uniform Resource Name (URN) to uniquely and unequivocally identify a process (mandatory), (2) a reference response to a DescribeProcess request for that specific process (mandatory), and optionally (3) a human-readable document describing the process and its implementation, as well as (4) a Web Service Definition Language (WSDL) description. Such documentation allows one to formally describe inputs/outputs/semantic of the various geoprocessing operators maintained in dedicated repositories (e.g., web service registries) that are structured following a semantically defined hierarchy of processes, each defined by a URN. This allows defining each unique process within the repository and each WPS instance to refer to that URN (Open Geospatial Consortium 2009; Lanig and Zipf 2010).

5.3.3.2 Distributed computing

The lack of performance (i.e., calculation speed, latencies) when integrating and processing large-scale data sets on a single computer backend
limits the types of analyses that can be carried out. Therefore a high performance computing environment is required to extend SDIs capabilities. Distributed computing related methods and technologies have been reported as promising to support SDIs (Brauner, Foerster et al. 2009). A distributed computing infrastructure can be thought of various autonomous computers that communicate with each other through a network to achieve a common task. When applicable, an incoming task is divided into several smaller sub-tasks executed simultaneously that are merged into an overall result at the end of the process.

There are several categories of distributed computing systems and the most commonly used are: Grids, Clouds, and Clusters. Grid can be defined as a parallel processing architecture in which computational resources are shared across a network offering access to unused CPU and storage capacities to all participating servers (Foster, Yong et al. 2008). Resources can be provided dynamically to users that are looking for computing power. In a Grid environment, users are grouped into specific communities, called Virtual Organization (VO) where they have similar data-intensive goals. Secure authentication is performed using a certificate that allows identifying users unequivocally and grant access to resources only to those how are authorized to use them. Cloud represents the Internet or whatever large network infrastructures where data computation and storage are moved away from local computers being “outsourced” and operated by third-party distributed facilities (Foster, Yong et al. 2008; Baranski, Schäffer et al. 2009). This allows users dynamically (i.e., “on-demand”) accessing resources (e.g., computational power, storage) without the need to manage the underlying infrastructure. Clouds have an economy-based model focusing on delivering computing resources as services. Finally, a computer cluster can be seen as a group of locally managed and linked computers acting like a single powerful computer (Foster, Yong et al. 2008). Despite that Cloud computing is gaining popularity, Grids have been specifically targeting the scientific community and consequently are more employed to support SDIs so far. To our knowledge, no significant attempts have been made to use WPS on a cluster. Therefore we concentrate our effort on Grid; it is the most complex infrastructure (for instance in terms of security and non-uniformity of resources), is widely used in the scientific community, and is potentially available to any users that have a valid certificate and do not have access to local facilities like clusters.

To group distributed computing elements in a Grid infrastructure, a piece of software called “middleware” is required. This middleware enables sharing resources and acts as a layer between heterogeneous hardware and specific user applications. Various Grid middleware are available (e.g., gLite, Globus, Unicore); they must all deal with: job submission, VO management, data management, and security. Applications that use and benefit from a Grid infrastructure are known as “grid-enabled” (or gridified) applications. For Baranski (Baranski), the term “gridification” means “the adaptation of existing applications and services to the requirements and expectations of a grid environment”. Although this term refers clearly to the Grid environment, we extend this definition in this paper to any application or services that might be adapted to distributed computing environments (e.g. Grids, Clouds, Clusters).
Chapter 5: Can SDI take advantage of distributed computing power to process the increasing amount of high-resolution data?

Within the GI community, a clear challenge is to make use of the secure sharing and processing mechanisms offered by Grids while implementing the necessary OGC specifications in terms of interfaces and services for the interoperability sake. Due to their respective nature (e.g. targeted audience, technicalities, standards) connecting Grids and SDIs is not a trivial task, and it requires extensions and customizations (Di, Chen et al. 2008). In consequence, it is needed to find a solution to grid-enabled OWS and especially WPS, while hiding the complexity of the Grid (and other distributed computing infrastructures) to SDI users.

5.3.4 Gridification approaches

Currently, two types of gridification processes have been recognized: encapsulation and integration (Open Geospatial Consortium 2009; Shaon and Woolf 2009). Encapsulation is recognized as a “low-level gridification” meaning that applications or services remain unchanged and can interact with distributed computing resources in the backend. For the Integration process, applications and services are resources fully embedded into the Grid middleware. Table 10 gives a comparison of the two approaches.

<table>
<thead>
<tr>
<th>Implementation:</th>
<th>Encapsulation</th>
<th>Integration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gridification level:</td>
<td>Easy</td>
<td>Difficult</td>
</tr>
<tr>
<td>Middleware:</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>Proxy:</td>
<td>Independent</td>
<td>Dependent</td>
</tr>
<tr>
<td>Level of OWS adaptation:</td>
<td>Not needed</td>
<td>Needed</td>
</tr>
<tr>
<td>Level of OWS interface extension:</td>
<td>Low, Not a Grid service</td>
<td>High, Become a Grid service</td>
</tr>
</tbody>
</table>

Table 10: Encapsulation vs. Integration approaches.

The encapsulation approach allows distributing calculation tasks and accessing data in a Grid environment but applications or services are not intended to become a Grid service as a whole (e.g. no security mechanism). To improve the encapsulated services scalability, the WPS interface must be extended as exemplified by Woolf and Shaon (2009) who encapsulate Job Submission Description Language (JSDL) within a WPS Execute operation. Although this approach is interesting and promising, it requires users to additionally understand and deal with JSDL language. Nevertheless, this kind of gridification is relatively easy to perform and the implementation can be considered as independent of the underlying Grid middleware (Baranski 2008). In comparison, the integrative approach is a more complex task requiring implementing the WPS specification directly into Grid middleware. A good example of such an implementation is the work done by the gLite-OWS (G-OWS) Working Group (Nativi, Verlato et al. 2009) that is implementing OGC Web Services (OWS) as Grid services within the gLite middleware used by the
Enabling Grid for E-sciencE (EGEE) project\(^{49}\). The main advantage of this gridification approach is that OWS are implemented as native Grid services and may offer all qualities of any services in a Grid environment (e.g. security, information and monitoring, job management, data management). Moreover, this facilitates and allows easier compliance with standards like the Open Grid Forum (OGF), Open Grid Services Architecture (OGSA), the Organization for the Advancement of Structured Information Standards (OASIS), and the Web Service Resource Framework (WSRF) (Yanfeng, Jack Fan et al. 2006; Nativi, Verlato et al. 2009).

Despite the fact that integrative gridification is the most promising one to support SDIs, this approach remains rather complex to implement because it requires important developments to transform OWS into Grid services (Yanfeng, Jack Fan et al. 2006; Di, Chen et al. 2008; Muresan, Pop et al. 2008; Mazzetti, Nativi et al. 2009). Moreover, the dependency of this approach to various middleware results in specific OWS middleware implementations. This situation might be potentially solved in the future with the adoption of interoperability standards like OGSA and WRSF. The objective of OGSA is to enable interoperability among various Grid middleware by introducing a service-oriented approach into the Grid (Yanfeng, Jack Fan et al. 2006). OGSA defines a Grid service as “a web service that provides a set of well-defined interfaces and that follows specific conventions” (Ghimire, Simonis et al. 2005). Moreover, a grid service is typically described by the Grid Web Service Description Language (GWSDL), it allows service discovery through Universal Description Discovery and Integration (UDDI) registries, and it uses SOAP to exchange information. The Open Grid Services Infrastructure (OGSI), the recommendation describing the infrastructure layer to support OGSA, is evolving towards web services standards through WSRF (Yanfeng, Jack Fan et al. 2006). One particularity of Grid services is that they interact continually (e.g Job Management Service consulting the Resource Discovery Service to find a suitable computing element matching job requirements). Hence, information about their state is required. In other words Grid services are stateful while web services are usually stateless. Consequently, WSRF specifies how to make a web service stateful as required by OGSA. However, the encapsulation approach does not provide Grid services functionalities, WSRF support is lacking, and possibilities to use other computing backends (like Clouds) is difficult, as these might for instance require the implementation of other job submission languages.

5.3.4.1 Mediation: a new approach to gridification

The integration of heterogeneous resources in distributed systems is not a recent problem (Wiederhold 1992) and it is gaining importance with the emergence of large-scale distributed (web-based) applications (Herault, Thomas et al. 2007). As information systems evolves and extend their scope they are increasingly depending on different heterogeneous resources (e.g., databases, web technologies, computing infrastructures) (Wiederhold and Genesereth 1997). All these resources are in general developed and maintained separately. Dealing with the diversity and heterogeneity of different resources may in

\(^{49}\)http://eu-egee.org/
Chapter 5: Can SDI take advantage of distributed computing power to process the increasing amount of high-resolution data?

consequence be an obstacle to develop high-level applications such as decision-making tools. Hence, integrating disparate resources requires a layer of mediation between clients and different pieces of software or services.

For Hérault et al. (2007) the concept of mediation has emerged as an answer to the lack of interoperability between clients and services in Information Systems (IS). The mediation approach requires identifying situation of particular heterogeneity, and implementing adaptation logic (Nativi and Bigagli 2009). A specialized and lightweight component, called a mediator, does the execution. A mediation layer can be composed of one or several mediators to possibly form chains managing requests and answers between a client and a service (figure 30).

![Mediation layer](image)

**Figure 30: Mediation layer (adapted from Hérault et al. 2005).**

Mediators can accomplish tasks dealing with control (e.g., routing, filtering, aggregation), transformation (e.g., format translation, ontology matching, semantic enrichment), quality of service (e.g., security, transaction), service level agreement (e.g., contract negotiation/management) (Herault, Thomas et al. 2007). In other words, mediators are used to bind clients and resources implementing some of the binding tasks required to SOA clients (i.e. service consumers).

In the present study, using a mediation concept appears interesting not only to integrate different pieces of software, but mainly in term of scalability (e.g., adding new computing backends and/or functionalities) and (potential) independency to middleware.

### 5.3.5 Implementation and architecture

To enable the concepts of the mediation approach, a WPS interface implementation was developed. The proposed WPS implementation through the mediation approach was built on an intermediate gridification level and can offer possibilities to overcome some of the shortcomings highlighted previously by integrating and benefiting from the following functionalities offered by various pieces of software:

1. Simple implementation of WPS specification.
2. Use of WPS Application Profiles to accommodate the Grid environment.
Chapter 5: Can SDI take advantage of distributed computing power to process the increasing amount of high-resolution data?

(3) No dependency on Grid (or other) middleware.
(4) Service providers to write scalable WPS process and users to select backends following their computing power requirements.
(5) The management of data tiling and merging when using a geospatial data set.
(6) Potential access and use of other distributed computing infrastructures such as Clouds (e.g., Amazon Elastic Compute Cloud\(^{50}\), Google App Engine\(^{51}\), Clusters (e.g., Oracle Grid Engine\(^{52}\)), and Desktop Grids (e.g., XtremWeb\(^{53}\) that exploits an array of desktop computers spread over a network to execute distributed tasks).

Traditionally, a WPS instance runs on a local web GIS server using third-party geospatial libraries and/or software (e.g., GRASS, ArcGIS). However, in order to ensure the portability and scalability of WPS on various distributed computing infrastructures, it is needed to:

(1) use a suitable implementation of WPS, ideally open source, to adapt the code and run it on distributed and heterogeneous Linux platforms;
(2) use a suitable software package allowing the development of tiling and merging modules for raster and vector data sets;
(3) use a suitable software package that deals with security/access (e.g. authentication, authorization), job submission, and job management in a flexible way on different computing infrastructures.

After reviewing different WPS implementations, tools and software, the following packages were selected:

(1) PyWPS\(^{54}\) for the WPS implementation.
(2) GDAL/OGR\(^{55}\) for data handling (i.e. tiling and merging).
(3) Ganga\(^{56}\) for computational task management.

Therefore, the presented WPS-mediated implementation is dependent on the selected software packages and on Linux Operating System. However, this specific implementation aims to show the possibilities to overcome, through the general concept of mediation, some issues limited to encapsulation or integration approaches.

PyWPS (Cepicky 2007) is an OGC WPS 1.0.0 implementation written in Python. PyWPS does not process the data by itself but uses GRASS to access geospatial functionalities. This is interesting because it offers the possibility to access the full list of GIS algorithms provided by GRASS, avoiding the need to rewrite the code of each basic core GIS capabilities (e.g., clip, intersect, buffer) like in other WPS implementations (e.g., Deegree\(^{57}\)). PyWPS is a Free and Open Source Software (FOSS) distributed under GNU General Public License (GPL). In
addition, PyWPS was recommended by the GEOSS, INSPIRE and GMES an Action In Support (GIGAS\textsuperscript{58}) project because it is an up-to-date WPS implementation, easy to install on most Linux platforms, and Python is recognized as a good choice to easily write processes (GIGAS consortium 2010). Furthermore, PyWPS offers features that other WPS implementations do not provide. It does not require complex installation procedures (e.g. compilation), and can be adapted to work as a standalone package. The Geospatial Data Abstraction Library (GDAL) and OGR Simple Feature Library are open source libraries for handling raster and vector geospatial data formats, respectively. Theses libraries are widely used in the GI community, offering various capabilities to manipulate data, and offer Python bindings. They are used to implement data tiling and the corresponding data merging. Finally, Ganga is a Python frontend for jobs specification, submission, management and monitoring on various distributed resources (Harrison, Tan et al. 2006; Moscicki, Brochu et al. 2009). Ganga provides a simple and consistent environment for processing data on heterogeneous resources; a wide community of users already exists. Ganga allows transparent switching between running test jobs on a local server up to large-scale processing on the Grid (Elmsheuser, Brochu et al. 2008; Maier 2008).

The aim of Ganga is to hide Grid (and other distributed computing infrastructures) technicalities by offering the possibility to switch computing backend by changing a single parameter of a job (Moscicki, Brochu et al. 2009). Job is the central concept within Ganga as it contains the complete description of a processing task (e.g., code to execute, input data, specification of computing environment). Another notable capability offered by Ganga is that it can be easily extended and customized through pluggable modules to access different Grid middleware such as Globus, Condor, and Unicore, as well as different dedicated backends such as PanDA (Vanderster and et al. 2010) or ARC (Read and et al. 2008). A first attempt to provide a Cloud backend to Ganga (Diaz, Ramo et al. 2011) is currently under development in the CLOBI\textsuperscript{59} project. In consequence, this tool allows using different computing backends with a common interface acting as an interoperable layer (Moscicki, Brochu et al. 2009).

It is important to mention that the implemented solution is restricted to a Linux environment because large Grid infrastructures rely on middleware developed for this Operating System (OS).

The proposed WPS implementation based on the mediation approach relying on PyWPS/GRASS, GDAL/OGR and Ganga is presented in figure 31. This implementation supports optional HTTP GET and mandatory HTTP POST methods.

\textsuperscript{58} \url{http://www.thegigasforum.eu/}

\textsuperscript{59} \url{http://code.google.com/p/clobi/}
Chapter 5: Can SDI take advantage of distributed computing power to process the increasing amount of high-resolution data?

Figure 31: Technical architecture of the proposed WPS mediation. The WPS mediation layer receives and handles the Execute query, evaluates parameters, and according to them is running the geoprocessing task on the local server or on a selected distributed computing infrastructure. The first option “backend=LCG” will send a list of files to be processed in parallel, while the second option “backend=LCG; jobs=50” allows tiling a large data set according to the number of desired calculation jobs and then submit jobs to the selected backend.

Through the use of WPS Application Profiles, the current WPS specification provides a mechanism to choose where to execute a process and how to tile/merge data. In consequence, this allows users to choose a computing backend and select the number of jobs (corresponding to the number of tiles) to be sent on working nodes. The term tile refers either to vector or raster data that are divided into smaller chunks. When executing distributed geoprocesses, each job corresponds to a “regular” WPS request sent through Ganga by the mediation layer along with input data (tiled by GDAL/OGR) and a modified version of PyWPS/GRASS binaries. During processing phase, the WPS mediation layer waits that all calculations from individual nodes are finished and then merges results.

The proposed parameters are:

1. **backend**: for choosing the computing backend (using the codification used in Ganga). For instance, local server (Local), Sun Grid Engine (SGE) and LHC Computing Grid (LCG) are implemented.

2. **jobs**: for selecting the desired number of tiles and corresponding jobs to be submitted.

Figure 32 gives a more detailed insight of the chain of processes involved within the mediation layer. When submitting a WPS query, GetCapabilities and DescribeProcess are processed regularly: WPS queries the instance and PyWPS sends corresponding XML documents to the client. In the case of an Execute request that does not contain the optional backend and jobs parameters, the query is handled by the mediation layer as a “regular” WPS and is run locally on
Chapter 5: Can SDI take advantage of distributed computing power to process the increasing amount of high-resolution data?

the web GIS server. However, if the request contains the proposed parameters (in DataInput), then two alternatives are offered to users:

(1) add backend argument:

```
http://localhost/cgi-bin/wps?service=WPS
&version=1.0.0&identifier=buffer&request=execute&dat
ar/rivers.gml;width=5,3;backend=LCG]
```

(2) add backend and jobs arguments:

```
http://localhost/cgi-bin/wps?service=WPS
&version=1.0.0&identifier=buffer&request=execute&dat
ainputs=[data=http://foo.bar/cities.gml;width=5;
backend=LCG;jobs=5]
```

In the former case, users can select a computing backend and provide a list of data inputs (separated by a comma) that need to be processed in parallel (i.e., task parallelism). In the latter case, data are first tiled according to the number of required jobs and possible dependencies at tile borders (i.e., data parallelism), and then the mediator writes corresponding Python and Shell scripts to submit jobs on different nodes. Once all sub-tasks are finished, Ganga sends back tiled results to a GDAL/OGR Python script for merging into the final result sent back to the client. Different tiling/merging strategies can be applied and may strongly affect performances (Werder and Krüger 2009). The adopted solution will be further explained under section 5.3.5.1.
Chapter 5: Can SDI take advantage of distributed computing power to process the increasing amount of high-resolution data?

5.3.5.1 Development

The three main components that need to communicate for handling mediated and “traditional” WPS requests are written in Python or support Python bindings. Hence, developing a dedicated WPS mediator with this scripting language to bind the different software involved appeared a reasonable choice. Moreover, Python is generally installed by default on Linux distributions and ensures that processes can be executed on remote computing nodes. The following mediation layer components were developed/adapted enabling a distributed environment.
Chapter 5: Can SDI take advantage of distributed computing power to process the increasing amount of high-resolution data?

5.3.5.1.1 PyWPS/GRASS package adaptation

The most important step in the implementation process was to adapt the PyWPS/GRASS making it executable on different Linux platforms as a standalone package. Indeed, in distributed computing environments users typically do not know which distribution of Linux are available and do not have administrative rights to install software on remote resources. In consequence, a software package needs to be executable without the need to install any third-party components.

For that purpose, GRASS binaries were downloaded directly from GRASS website. Unwanted or useless components (e.g., documentation files, widgets) were removed and some static libraries were added. Finally, PyWPS source was adapted (e.g., configuration files, functions). The resulting modified PyWPS/GRASS package (available as a tarball) can be run anywhere and independently on the following tested Linux systems (with Python version 2.5 or above): Scientific Linux, Ubuntu, Debian, SUSE, and Fedora.

5.3.5.1.2 Process implementation

To implement the process logic, PyWPS uses stand-alone Python scripts that have to be written in the PYWPS_PROCESSES directory. A PyWPS process has one class (Process) and two methods: __init__() and execute(). Within the __init__() method, users have to provide some general information (e.g. title of the process, identifier, data input and outputs) while in the execute() method the process itself is implemented calling GRASS commands.

To execute a process on a remote computing node, it is required (1) to access locally stored data and binaries and (2) to be able to reference them with path and/or filename. When a job is submitted data and binaries will be sent together to remote computing nodes on which users do not know where they will be processed. Therefore in the process logic itself it is important to make path to data and binaries independent of their location by creating suitable variables.

5.3.5.1.3 Supported Data formats and sources

GeoTiff, Geography Markup Language (GML) and shape files are currently implemented. However, the framework can be easily extended to all the formats supported by GDAL\textsuperscript{60} /OGR\textsuperscript{61}.

5.3.5.1.4 Data tiling and merging

One of the main advantages of distributed computing is the possibility to divide a given task into several subtasks, which can be executed in parallel. In the case of processing a large vector or raster data set, this means dividing it into smaller sub-sets. For that purpose, the parameter \textit{jobs} has been proposed. Indeed, the input data set will be subdivided according to the number of

\textsuperscript{60} \url{http://www.gdal.org/formats_list.html}

\textsuperscript{61} \url{http://www.gdal.org/ogr/ogr_formats.html}
calculation jobs specified by a user in the mediated WPS request. Werder and Krüger (2009) give detailed descriptions about various issues that can arise when tiling/merging data. These authors highlighted that to efficiently execute parallel geospatial calculations on the Grid it is required to know which geospatial algorithms can potentially benefit from parallelization, which type of parallelism is suitable (i.e., data or task), how parallelization can affect the results and how data are distributed. Hence all these considerations will affect the tiling/merging strategies. In the mediation layer, users choosing between the two types of query (presented under section 5.3.5) make the choice between task and data parallelism. If users submit data that need to be tiled (i.e., case 2), it is important to take into account possible correlations at tile borders to obtain continuous results (Werder and Krüger 2009) (e.g., NDVI process is not influenced by border effects while slope calculation is). The envisioned solution within the mediation layer is to develop WPS Application Profiles describing required parameters to take into account specificities introduced by distributed computing infrastructures and to support efficient tiling/merging strategies. From our point of view, the following parameters are a minimum set that must be considered: (1) whether the algorithm can be parallelized or not, (2) the type of correlations at tile border, (3) the extent of the boundaries, and (4) the smoothing functions to be applied. Depending on the data type (raster or vector, defined in process metadata), the potential inter-dependencies between created tiles, and the values of the previously mentioned parameters, the mediation layer first calls the required module (Python script using GDAL or OGR) to subdivide the data accordingly. Second, it writes the needed Ganga scripts with the correct references to data. When all processing subtasks are finished, the mediation layer calls the corresponding module to merge data and send it back to the requestor. It should be noted that input and output must be geospatial data.

To achieve this tiling/merging tasks before entering a Grid and submitting jobs, it is required to access the data. Hence, a specific module was developed supporting different data sources: OWS (WFS and WCS) servers, remote data repositories (web accessible folder), and locally stored data (uploaded by File Transfer Protocol (FTP)).

5.3.5.1.5 Job submission

The mediation layer writes corresponding Ganga scripts to submit jobs on distributed computing infrastructures with the command:

```
> ganga <name_of_script>.py
```

Each job consists of (1) a PyWPS/GRASS binaries package, (2) data input, (3) a shell script to execute the WPS request on the remote computing node.

The following example shows a typical Ganga script (Example 1):

```python
j=Job()
j.inputsandbox=['egMediator.tar.gz','TM3.tif','TM4.tif']
j.application=Executable()
j.application.exe=File('testWPS.sh')
j.backend=File('testWPS.sh')
j.backend.requirements=LCGRequirements()
j.outputsandbox=['*']
j.submit()
```

**Example 1: Ganga script**
Chapter 5: Can SDI take advantage of distributed computing power to process the increasing amount of high-resolution data?

The corresponding shell script (Example 2) is used to execute the calculation through a WPS request directly on the remote computing node.

```
#!/bin/bash
abspath=$(cd ${0%/*} & echo $PWD/${0##*/})
path_only=`dirname "$abspath``
tar -xzf egMediator.tar.gz
mv TM3.tif $path_only"/egMediator/datainput/TM3.tif
mv TM4.tif $path_only"/egMediator/datainput/TM4.tif
chcon -t texrel_shlib_t $path_only"/egMediator/WPS/grass/lib/*.so
chcon -t texrel_shlib_t $path_only"/egMediator/WPS/grass/lib/*.so.1
python $path_only"/egMediator/egmediator.py
'service=WPS&version=1.0.0&request=execute&identifier=ndvi&datainputs=[red=TM3.tif;nir=TM4.tif]'```

**Example 2: Shell script**

Once jobs are sent to the Grid, users can monitor jobs status using Ganga (figure 33).

```
In [2]: jobs
Out[2]:
Registry Slice: jobs (6 objects)

<table>
<thead>
<tr>
<th>fqid</th>
<th>status</th>
<th>name</th>
<th>subjobs</th>
<th>application</th>
<th>backend</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>new</td>
<td></td>
<td></td>
<td>Executable</td>
<td>LCG</td>
</tr>
<tr>
<td>300</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>301</td>
<td>running</td>
<td></td>
<td></td>
<td>Executable</td>
<td>Local</td>
</tr>
<tr>
<td>302</td>
<td>running</td>
<td></td>
<td></td>
<td>Executable</td>
<td>Local</td>
</tr>
<tr>
<td>303</td>
<td>running</td>
<td></td>
<td></td>
<td>Executable</td>
<td>Local</td>
</tr>
<tr>
<td>304</td>
<td>completed</td>
<td></td>
<td></td>
<td>Executable</td>
<td>Local</td>
</tr>
<tr>
<td>305</td>
<td>completed</td>
<td></td>
<td></td>
<td>Executable</td>
<td>Local</td>
</tr>
</tbody>
</table>

Figure 33: Jobs monitoring with Ganga (in command line)
```

When a job is finished, Ganga retrieves the result by moving it outside the Grid and writing it on a dedicated output folder on the GIS server.
Chapter 5: Can SDI take advantage of distributed computing power to process the increasing amount of high-resolution data?

5.3.5.1.6 Security

To access resources and submit jobs in a Grid environment, users must have a valid long-lived certificate (for authentication and authorization) issued by a certificate authority (CA) trusted by the resource provider. When submitting a job, grid users must first generate a short-lived proxy certificate (i.e., valid in general 12 hours) derived from their personal credentials that will be used to authenticate them on the remote resource (Robinson, Gemmill et al. 2005). From a user point of view, acquiring a certificate is time-consuming, and management of certificates may be difficult especially when accessing various types of computing resources (e.g., Grids, Clouds) requiring different authentication processes (Crampton, Lim et al. 2007; Crampton, Lim et al. 2011). Therefore, offering (web-based) single sign-on capabilities can be very useful. This is currently a “hot topic” within the computing infrastructures community that is exploring new venues to manage credentials through federation of CAs (Di Stefano, Morana et al. 2009; Jie, Arshad et al. 2011), Shibboleth (Wang, Jones et al. 2009), or completely new authentication/authorization mechanisms (Cornwall, Jensen et al. 2004).

However, the current WPS specification does not provide security mechanisms (Shaon and Woof 2009). To solve this issue and allow a mediated WPS request to submit jobs on a Grid infrastructure, a common solution is to store the proxy certificate on a central repository protected by a username/password. This consent any Grid client that has access to these credentials to retrieve the proxy certificate (Robinson, Gemmill et al. 2005; Menglong, Yong et al. 2009). A widely adopted solution to handle proxy certificates is the online credential repository called MyProxy (Novotny, Tuecke et al. 2001). Users can create a proxy certificate on the MyProxy repository using their long-lived Grid certificate and delegate a username/password combination to the created proxy certificate (Padberg and Greve 2009). The WPS Application Profile enables users to embed MyProxy parameters (e.g., host, port, username, password) within the WPS request as a special DataInput parameter and then the mediation layer manages username/password combination to retrieve short-lived credentials generated by the MyProxy Server to allow submitting jobs on a Grid. Adding such a security component will introduce an overhead that should be negligible if the request takes a long time to process (Di, Chen et al. 2008). However, this approach brings major flexibility and offers the possibility to access Grid resources through simple authentication mechanism (e.g., username, password) and consequently enable users to access to their credentials (e.g., web browser, mobile devices) seamlessly to submit jobs on a Grid through a WPS request.

5.3.5.1.7 Service metadata

The extension of the WPS interface with the possibility to execute geoprocessing calculations on different backends requires modifying accordingly the metadata of the service available through the GetCapabilities operation. In fact, the objective is to describe an extended capability of the WPS instance. This

---

62 http://grid.ncsa.illinois.edu/myproxy/
Chapter 5: Can SDI take advantage of distributed computing power to process the increasing amount of high-resolution data?

implies that such extended capabilities must be valid for every process that users will run using the mediated WPS. Hence, the XML document sent to the client must provide a list of supported backends as suggested:

```xml
<wps:SupportedBackends>
  <wps:Default>
    <ows:Backend>local</ows:Backend>
  </wps:Default>
  <wps:Supported>
    <ows:Backend>local</ows:Backend>
    <ows:Backend>sge</ows:Backend>
    <ows:Backend>lcg</ows:Backend>
  </wps:Supported>
</wps:SupportedBackends>
```

Example 3: Proposed extension for the GetCapabilities XML document providing a list of supported computing backends.

Users can declare different computational backends directly in the PyWPS configuration file that is used by the GetCapabilities operation to create the corresponding XML response. Obviously, clients that may query capabilities offered by the mediated WPS instance must support such an extension.

5.3.5.1.8 Process metadata

The possibility to execute geoprocessing algorithms on distributed computing infrastructures not only requires modifying accordingly the GetCapabilities operation but also that these additional capabilities and dependencies be reflected in the description (including inputs and outputs) of each WPS process obtained through a DescribeProcess operation. WPS Application Profiles allow enhancing the description of each process by defining inputs/outputs parameters and their values. This mechanism offers the possibility to specify dedicated parameters required to accommodate WPS request to distributed computing environments. Such “grid-enabled WPS Profiles” allow, in the current state of the presented implementation, to select a computing backend and the number of jobs to be executed by defining these parameters in the profile (see section 4). Consequently, this enables users to embed backend and jobs parameters within a mediated WPS request as special values of the DataInput parameter. Finally, the mediation layer will handle the request to efficiently negotiate with the different pieces of software involved. Furthermore, using grid-enabled WPS Application Profiles will allow to defining required parameters to manage a secure access to Grid resources (see section 5.3.5.1.6), as well as developing efficient strategies for data tiling/merging (see section 5.3.5.1.4).

5.3.6 Use case: NDVI computation

The proposed WPS mediation layer has been developed and will be tested in the context of the enviroGRIDS project, funded under the European Commission (EC) Seventh Framework Programme. This project focuses on the
Chapter 5: Can SDI take advantage of distributed computing power to process the increasing amount of high-resolution data?

ecologically unsustainable development and the inadequate resource management that is often observed in the Black Sea hydrological catchment area. A large catalogue of environmental data sets (e.g. land use, hydrology, and climate) has been gathered and is used to perform distributed spatially explicit simulations to build scenarios of key environmental changes. The project is also developing a Spatial Data Infrastructure (SDI) to feed its regional data into the Global Earth Observation System of Systems (GEOSS), while being linked to the Enabling Grid for E-sciencE infrastructure (EGEE). During its 4-year timeframe, the enviroGRIDS project aims at the following objectives for which a Grid infrastructure will be important:

1. A high-resolution (sub-catchment spatial and daily temporal resolution) water balance model will be applied to the entire Black Sea (BS) catchment using the Soil Water Assessment Tool (SWAT) (Arnold, Srinivasan et al. 1998). This tool will be gridified and used on the EGEE infrastructure.
2. Adequate sensitivity and uncertainty analysis will be performed on the BS SWAT model. A gridified version of the SWAT-CUP (Abbaspour, Vedjani et al. 2007) tool will be use for that purpose.
3. Access to real time data from sensors and satellites will provide early warning and decision support tools to policy-makers and citizens. These data may be streamlined into the grid-enabled enviroGRIDS SDI to ensure fast computation and dissemination of results.

The strong Grid component of the project will foster data interoperability, and is triggering new directions of research or alternative ways of analyzing high-resolution data sets.

One of these directions is the proposed WPS mediation that will be tested to compute Normalized Difference Vegetation Index (NDVI) at the catchment scale. Indeed, vegetation indices time series might be useful data sets to refine SWAT models calibration (Arnold, Srinivasan et al. 1998). Vegetation monitoring by remotely sensed data is carried out by means of vegetation indices, which are mathematical transformations designed to assess the spectral contribution of green plants to multispectral observations (Maselli, Gilabert et al. 1998). The NDVI is calculated from Near InfraRed (NIR) and Red (R) bands as follows (Rouse, Schell et al. 1974):

\[
NDVI = (NIR - R)/(NIR + R)
\]

The corresponding WPS geoprocessing task as been developed according to the different requirements mentioned and discussed previously. To study the benefits of the NDVI WPS, we show here its application on two large data sets (Table 11). The first one is a medium resolution remote sensing images (250m resolution) of 1.2 GB provided by MODe rate resolution Imaging Spectroradiometer\(^{(63)}\) (MODIS) covering the full Black Sea catchment (2 millions square km). This first data set will be used to test the NDVI process on a local GIS server and if needed to be executed on the Grid. The second one is a set of nine high-resolution orthophotos (10cm resolution) of Geneva (Switzerland) covering an area of one square kilometer each, for a total size of 3.6 GB. These images are

\(^{(63)}\) http://modis.gsfc.nasa.gov/
Chapter 5: Can SDI take advantage of distributed computing power to process the increasing amount of high-resolution data?

provided by the “Système d’Information du Territoire Genevois (SITG)”. This second data set will be used to make a comparison between different executions of the NDVI process:

- on a merged version of the nine orthophotos on a local GIS server,
- on a merged version of the nine orthophotos on the Grid with 50 jobs (i.e., use of backend and jobs parameters to tile and merge data),
- on a local GIS server with already tiled orthophotos and submitted as a file list in the datainputs parameter using only the backend parameter,
- on the Grid with already tiled orthophotos and submitted as a file list in the datainputs parameter using only the backend parameter.

<table>
<thead>
<tr>
<th>Tiling</th>
<th>Submission</th>
<th>Execution</th>
<th>Merging</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>MODIS, local</td>
<td>--</td>
<td>11’51”</td>
<td>--</td>
<td>11:51</td>
</tr>
<tr>
<td>MODIS, grid (j=10)</td>
<td>2’20”</td>
<td>8’40”</td>
<td>3’00”</td>
<td>19’10”</td>
</tr>
<tr>
<td>SITG – merged, local</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>SITG- merged, grid (j=50)</td>
<td>72’20”</td>
<td>51’10”</td>
<td>66’40”</td>
<td>230’35”</td>
</tr>
<tr>
<td>SITG – 9 tiles, local</td>
<td>--</td>
<td>43’20”</td>
<td>11’20”</td>
<td>54’40”</td>
</tr>
<tr>
<td>SITG – 9 tiles, grid</td>
<td>--</td>
<td>30’50”</td>
<td>11’20”</td>
<td>47’30”</td>
</tr>
</tbody>
</table>

Table 11: Indicative comparison of mediated WPS processes executed locally or on the Grid (times are expressed in minutes/seconds).

From the obtained indicatives results several observations can be made:

- Despite the fact that the used MODIS data set might be considered as voluminous (1.2 Gb), the developed NDVI WPS process executed on a single GIS server (with 3Ghz CPU/4Gb RAM) can easily handle this data set avoiding the need to use the Grid.
- However, a high-resolution merged image (30’000x30’000 pixels at 10cm resolution) cannot be processed locally on a single WPS instance due to lack of memory.
- Processing the same data set on the Grid is feasible but important overheads are introduced. In particular, subdividing and merging such a data set in 50 tiles is time consuming. This probably comes from GDAL internal process implementation (e.g., multiple load of the file in memory). Submission and execution on the different nodes can also be quite expensive and dependent of the workload of each computing elements.
- Despite the small overhead introduced by jobs submission, executing calculations on the Grid on already tiled data sets is the most efficient way. In this way, and in contrary to a local execution that processes jobs sequentially, the nine submitted jobs are executed simultaneously. In theory, such parallel execution can improve the overall processing phase by a factor corresponding to the number of submitted jobs (nine in our example). Nevertheless it is not the case here due principally to varying workload conditions on different nodes. The improvement is around 15% and should

---

64 http://www.sitg.ch
increase with growing number of jobs (i.e., processing 10'000 tiles will probably show a much larger improvement).

5.3.7 Discussion & perspectives

This tool was developed as a proof-of-concept of the mediation approach to grid-enable OGC WPS specification. The implementation was successful and first results show both benefits and limitations. In particular, this approach enables the possibility to further develop WPS implementation (with the help of Application Profiles), offering some of the advantages of a Grid service (e.g. secure access to resources, SOAP-based messaging, statefullness, process scalability).

5.3.7.1 WPS mediation benefits

The major benefit of this approach is to be able to flexibly and seamlessly execute WPS on different computing backends, ranging from a local GIS server up to large scale computing infrastructures like a Grid. This scalability is important because it enables users to develop their geoprocesses locally and run them later on various distributed computing infrastructures if more processing power is required. It should be noted that the developed WPS mediator was tested only with the EGEE infrastructure and a local GIS server. Nevertheless, experiments in High Energy Physics successfully used Ganga to submit jobs on various backends (Elmsheuser, Brochu et al. 2008; Moscicki, Brochu et al. 2009), and even added new computing backends (Read et al. 2008; Vanderster and et al. 2010; Diaz, Ramo et al. 2011). The presented mediated WPS instance should therefore run without problem on any other infrastructure for which a backend plugin exists in Ganga. In addition, the modular and pluggable framework of Ganga allows one to easily develop new plugins for accessing infrastructures that are not yet implemented such as Clouds or Desktop Grids. Hence, Ganga appears as a suitable solution to hide the complexity of various computing infrastructures by acting as an interoperability layers between different backends and users. Finally, the mediation approach is sufficiently general to be applicable to other WPS implementations and job submission software.

From the results on NDVI processing, it shows that accessing Grid resources allows processing larger data sets than it is possible on a single computer. However, submitting very large data sets (i.e., several GB) is not very efficient as it introduces several important overheads (in particular for tiling and merging data). Nevertheless, this can be seen as additional/optimal functionality offered to users that do not have sufficient memory on local resources. A more realistic usage and common practice when serving large data sets (such as those from the SITG) is to make them available already tiled in order to reduce download time and to improve data access to specific areas (i.e., avoiding the need to download a complete data set if not needed). Therefore, submitting these tiles in file lists as data inputs that can be processed in parallel on the Grid appears to be an efficient and promising possibility.

Finally, WPS Application Profile is an interesting feature offered by the current WPS specification. Our specific implementation has shown that it is an
efficient way to accommodate WPS request to distributed computing environments. This allows benefiting from the capabilities of the Grid with the simplicity of WPS.

5.3.7.2 WPS shortcomings

Actual WPS specification is not sufficient to fully benefit from distributed computing infrastructures capabilities. We showed that service metadata available through a GetCapabilities operation must introduce this extension by providing a list of supported backends and process metadata must be extended through WPS profiles to take into account specificities of distributed computing environments.

Additionally, some aspects of the current WPS specification are currently not sufficient to handle different types of references to ensure sufficient scalability within the process logic. Additionally, this issue highlighted the fact that currently available data types in WPS specification are too generic and need to be enhanced (Nash 2008). Nevertheless, the ComplexData type, used to handle vector and raster data in the current WPS specification, does not provide a possibility to reference locally stored data as inputs (i.e., it only addresses remote data sources). Although not foreseen for this usage, the LiteralData type helps to overcome this issue. This data type accepts any character strings and then allows introducing path and filename to local resources. If data inputs are declared as such then it is possible to reference locally stored data sets.

5.3.7.3 Technical limitations

The current implementation of WPS mediator suffers from different technical limitations that may be overcome with further developments.

At this stage, the mediator only accepts three widely available formats as input: GeoTiff, GML and shape files. However, further development can easily extend the input formats to other vector and raster formats supported by the GDAL/OGR libraries.

When processing data, real scenarios (e.g., generating a flood risk map, analyzing habitat distribution of birds) involve in general various complex tasks/calculations. The process to turn data into information requires the organization of these tasks into sequences or chains of processes. In a web services environment, building flexible and efficient workflows coordinated by an orchestration engine can do this. Orchestration engines are generally based on SOAP and WSDL to exchange structured information while OGC standards rely on XML-RPC (Fleuren and Muller 2008). These two different ways of representing data may limit the chaining capabilities of OWS. However, WPS has a very limited support of SOAP/WSDL in its current specification. Hence, at this stage of development, it was impossible to effectively test the integration of mediated WPS processes within a chain of services. The release of PyWPS 3.2 will introduce an extended support of SOAP and should therefore facilitate service chaining in the mediation layer.

Finally, OGC standards are in general stateless (i.e., lacking capabilities to give information about their state). However, monitoring jobs status (e.g., submitted, running, complete) with Ganga and some Python bindings should make it
possible to develop suitable interfaces to retrieve information about the state of submitted jobs. Additionally, the WPS interface has an optional status parameter that partly enables reporting on the status of execution (Open Geospatial Consortium 2007).

5.3.7.4 Performances

Granell et al. (2009) stated that performance is one of the main problems when implementing geoprocessing tasks in a distributed environment. These authors showed that network bandwidth, data transportation and data validation are potential bottlenecks when dealing with large geospatial data sets. Except in a few cases (e.g., dedicated high performance network), the Internet network bandwidth will always be a limiting factor, as data will be transported over the Internet. Thus, minimizing data transportation (i.e., having data stored physically close to the processing elements) and data validation (i.e., parsing data used within a process) is of high importance.

In order to decrease the overall overhead, the mediation approach might be potentially useful as it allows moving some of the components directly into the Grid. Indeed, a job using the Mediation layer is constituted by (1) process logic, (2) data to be processed, and (3) a PyWPS/GRASS binary package (engine) to execute the process (currently weighting around 27MB). If the latter package is directly accessible on the nodes of a Grid infrastructure and data is available in Storage Elements (SE), the mediator will be used only to forward WPS requests and send process logic through Ganga. This may considerably decrease the overhead caused by the size of PyWPS/GRASS binaries and data. Such a “moving-code” approach (Müller, Bernard et al. 2010) can be advantageous as it allows moving the algorithm code, rather than data, to a processing instance. Data is then processed as close as possible to its source, which reduces bandwidth and data transportation time.

Currently Ganga is only used within the mediation layer to submit geoprocessing task on different backends. It does not consider the workload on different computing nodes. In other words, if a node is available then a job is submitted. When submitting multiple jobs at the same time, this can negatively impact the overall processing time. In such situation, Ganga will wait until sufficient nodes are available to submit these jobs. However, in the vision of integrating Grid capabilities into SDIs, responsiveness appears to be an important condition. In order to overcome this issue, one possible solution is to upgrade the mediation layer using a tool called DIANE (for Distributed ANalysis Environment) that can be used in conjunction with Ganga. DIANE aims to improve quality of services using automatic control, workload management and jobs scheduling on a set of distributed worker nodes (Moscicki 2003; Moscicki 2004). This may (1) reduce the application execution time by using the resources more efficiently, (2) provide fully automatic execution and failure management, (3) efficiently integrate local and Grid resources, and (4) provide mechanisms to efficiently send jobs on nodes that are less loaded (Moscicki 2004).

Furthermore, it can be interesting to avoid the necessity to let users determine the computing backend and/or the number of jobs that need to be

65 http://it-proj-diane.web.cern.ch/it-proj-diane/
Chapter 5: Can SDI take advantage of distributed computing power to process the increasing amount of high-resolution data?

submitted. In the currently implemented solution, users need to know whether the computational requirements of the geoprocessing task require to be distributed. However for an optimum use of computing resources, it would be good to delegate this choice to the mediator. A possible solution to implement this functionality necessitates developing an algorithm to optimize the job submission and can be used in conjunction with DIANE. Such algorithm must consider various parameters such as file size, data complexity (e.g., resolution, geometry), and number of available worker nodes per backend. Based on these parameters (and potentially others), the mediator can then evaluate the resources required to efficiently process the data, select the suitable backend (e.g., local, cluster, Grid), and calculate the number of jobs/tiles to be used. Several attempts (Teo, Low et al. 2003; Hu, Xue et al. 2005; Yang, Wu et al. 2011) have been made to optimize jobs calculations and processing of remote sensing images. All these studies highlight that achieving the objective of high performance computing of geospatial data is not an easy task, but investigated solutions of optimization offer good potential.

Other factors that may influence performances of the mediated WPS service are the ones linked to download and tiling modules. If data are not stored locally they need to be downloaded and this may require some time depending on file size and bandwidth. Once download is finished, this dataset is tiled according to the number of jobs requested by the user. In such a case, these tiles can still have a size of several megabytes and must be uploaded to the different computing nodes introducing a data transportation overhead. If tiled data are directly available in SE this allows bypassing the need to download and tile data, this reduces the induced overhead close to zero. Nevertheless, the extension of the mediation layer to take into account Grid specificities such as data accessible through Logical File Name (LFN) remains to be investigated further. Other possibilities to achieve the objective of fast access to large and widely distributed amounts of data are (1) Peer-to-Peer (P2P) technologies (Yanfeng, Jack Fan et al. 2006; Sanchez-Artigas and Garcia-Lopez 2010) or (2) development of new protocols like GridJet that accelerate data transport (Wang, Helian et al. 2009).

Finally, parallelization is an important factor to take into account because it potentially strongly affects the quality of proposed services. For Padberg and Greve (2009) there is no generic parallelization method that fits every geospatial process. Parallelization can indeed be done at the task (i.e., task parallelism) or at the data level (i.e., data parallelism) (Werder and Krüger 2009). In the former case, a task can be split into various independent subtasks working on the same or different data set, while in the latter case each subtask operates on a subset of the whole data set. Data parallelism introduces the necessity to efficiently tile and merge data (to lower induced overhead) and to consider effects (e.g., interdependencies) at tile borders to obtain adequate result (Werder 2010). Moreover, not all geoprocessing tasks can be subdivided into independent calculations (e.g., climate models requiring communication among jobs). Consequently, suitable strategies are required. Werder and Krüger (2009) propose possible alternatives to tile/merge data: by object, by geometry, by operation, by spatial criteria, and by level of detail. In the mediation approach, further investigations are required to find effective and efficient solutions using grid-enabled WPS Application Profiles to manage and offer such capabilities to users.
5.3.8 Conclusions

WPS is a promising specification to handle data and a key element to enable SDIs as web-based geoinformation environment. Nevertheless, various issues emerge when trying to use WPS in data and computing-intensive domains like environmental sciences. To overcome these problems distributed computing paradigm and especially Grid computing appears to be an interesting candidate to empower SDIs. However, SDIs and Grids are technologically different and matching these two types of infrastructures is challenging. For that purpose, OGC WPS specification needs to be improved and extended in order to benefit from superior storage and computing capacities offered by distributed computing.

The presented conceptual approach, of mediating different GIS and Grid software components, has been exemplified by a specific implementation (i.e., using PyWPS/GRASS, GDAL/OGR and Ganga), offering the possibility to:

1. potentially execute WPS request on any computing backend implemented within Ganga (e.g., local servers, Clusters, Grids, Clouds) while hiding the technicalities of these infrastructures.
2. add two optional parameters (through WPS Application Profile): backend and jobs, allowing users to specify where to execute a selected process and how many sub-tasks are required.

A NDVI WPS process highlighting benefits both in term of performance and scalability has successfully exemplified this implementation. Nevertheless, several issues and shortcomings have been raised and require further investigations in order to build responsive, efficient and reliable grid-enabled SDIs. Some of these shortcomings (e.g., statefullness, asynchronosity, data types, process management) will be tackled in the future WPS 2.0.0 specification that is expected later in 2011. The hope it that the new version improves the grid-enablement process and facilitates the use of distributed computing resources.

The proposed mediation concept appears a promising alternative to encapsulation and integration approaches, opening new perspectives for the grid-enablement process. The mediation concept also enriches the discussion concerning future improvements of the WPS specification along with supporting Grid community in finding new and promising directions of research and development (Schwiegelshohn, Badia et al. 2010).
5.4 Summary and lessons learned

- Turning raw data into understandable information requires analysis capabilities to process data.
- SDIs can be thought as geo-enabled Service Oriented Architecture.
- SDIs have limited analytical capabilities. Processing of geospatial data is done in general on client’s desktop computer, which is an inhibiting factor when processing very large and high-resolution data sets.
- High performance computing environment may be required. Distributed computing, in particular Grid architecture, appears an interesting candidate to extend processing capacities of SDIs and leverage the full potential of WPS.
- OGC Web Processing Service specification aims to propose a standardized interface for publishing and performing geoprocessing functionalities in a web service environment.
- OGC web services rely on XML-RPC while “traditional” web services are based on SOAP/WSDL. This leads to potential difficulties in chaining OWS and traditional web services.
- Connecting SDIs and Grids is challenging because they are technologically different (e.g., service description, service interfaces, service states, security). Hence, it is difficult to make them interoperable without extension and customization.
- The main objective is to hide the complexity of Grid while preserving OGC interfaces (i.e., making OGC and Grid services interoperable).
- Grids may also be interesting for data storage, data management (e.g., replication, security), and data movement.
- To adapt OGC and Grid services as well as identifying benefits and limitations, an incremental approach of grid-enablement may be appropriate.
- Several successful attempts have been made to implement WPS and other OGC specification in distributed computing environments.
Chapter 5: Can SDI take advantage of distributed computing power to process the increasing amount of high-resolution data?

- Currently there are no commonly agreed approaches to grid-enable WPS. Gridification is mostly realized through “encapsulation” (low-level) or “integration” (high-level) approaches.

- No implementations allow to seamlessly execute a geoprocessing task following users requirements on different computing backends, ranging from a stand-alone GIS server up to computer clusters and large computing infrastructures like Grids or Clouds.

- The newly proposed “mediation” approach offers: (1) simple implementation of WPS specification, (2) no dependence on Grid (or other) middleware, (3) scalability of WPS process execution following users computing power requirements, (4) possibility to tile/merge large data within the process, and (5) potential seamless access and use of other computing infrastructures such as Clouds, Clusters and Desktop Grids.

- Using WPS Application Profiles allows to propose two parameters: *backend* (i.e. selecting computing backend) and *jobs* (i.e., numbers of tiles and corresponding jobs to be submitted).

- Current WPS specification is not sufficient to handle different types of data to ensure sufficient scalability within the process logic.

- Ganga is a useful tool to submit jobs on different computing backends.

- It is important to define what is a large data set.

- Important overheads can be introduced if an unsuitable parallelization strategy is applied.

- Processing several sub-tasks in parallel can improve performances (e.g., ability to process larger data set) and calculation speed (processing the same data set faster).

- Mediation approach has demonstrated flexibility, scalability and seamless execution.

- Further investigations on the mediation approach need to be done in order to build responsive, efficient and reliable grid-enabled SDIs. In particular, orchestration with different OWS must be tested and performances (e.g., workload, parallelization, job submission, Grid integration) must be improved.

- WPS Application Profiles are useful to adapt WPS to distributed computing frameworks (e.g., backend, jobs, security, parallelization strategies).
Chapter 6: Conclusions & recommendations
6.1 Conclusions

In the last decades, the amount of geospatial data available has grown dramatically following the evolution of communication means and the rapid development of spatial data capture technologies such as GPS, remote sensing, sensors or mobile devices (e.g., mobile phones, tablets). However this huge amount of geospatial data is stored in different places, managed by different organizations and often operated in isolation. Consequently, the vast majority of this data is not being used as efficiently and effectively as it should.

Geospatial data are essential to support sound and reliable decisions-making processes at all scales (from local to global) and in many disciplines. In the field of environmental sciences, to understand the complex interactions among the different components of the Earth system, it is required to gather and integrate physical, chemical, biological and socio-economical data. Currently, data accessibility, availability, compatibility, and lack of sufficient resources to analyze these data are among the most frequent difficulties that are negatively influencing the way that scientists, researchers, decision-makers and general public are accessing and using environmental data.

Based on this situation, we have assumed that making data interoperable and providing access to computing resources may potentially improve the above mentioned situation allowing users of environmental data to spend more time in data analyses and less in data discovery and integration. Therefore, the aim of this study was to:

Examine how well Spatial Data Infrastructure concepts, methods and related technologies are useful to support and facilitate environmental data discovery, accessibility, visualization, dissemination and analysis.

To support the objective of this study, three research questions have been formulated and answered within specific chapters that constitute the core of this thesis.

1. Is there a need to access and process environmental data in a better and efficient way?

Chapter 3 reported on and discussed the current situation within environmental sciences on data management, exchange, dissemination, and processing. It showed that there is a growing needs to better organize and share environmental data in order to understand the complexity of earth-system processes and to convey improved information on the environment to decision-makers and the general public. Addressing this need by sharing environmental data is challenging because it requires a common agreed framework that allows easy and seamless integration of data from different sources giving access to services that could be linked together to process and generate new understandable knowledge and information.

People, when trying to address environmental problems and threats need to take various decisions every day. However, they are facing the same problem: they need to take sound decisions with only partial information and consequently cannot manage efficiently and effectively what they cannot
Chapter 6: Conclusions & recommendations

measure. Therefore, gathering and integrating the vast amount of environmental data generated on a daily basis (but often operated in isolation) is an essential and fundamental effort to be taken in order to make sound decisions at all levels, from global, to local ones. Earth observations are used and compiled to answer a specific question, to understand or explain a trend, to confirm or invalidate a thesis, to make predictions or support management processes.

Additionally, ICTs have enhanced globally the productivity of people and allowed more individuals, industries, and countries to participate in the development of a knowledge-based society. These changes are obviously influencing and transforming the world of scientific research. Indeed, science is driven by data, and new technologies have dramatically increased the easiness of data capture and therefore the amount of data. Two of the challenges that scientists are facing today are:

1) dealing with the tremendous amount, complexity, and variety of data that is currently being generated,
2) taking full advantage of the knowledge and information produced by scientists and researchers.

Therefore, making sense of the vast amount of data and information and turning them into understandable information is challenging. It is then essential to make these data easily available and accessible in order to give the opportunity to the users to turn them into usable information. From a scientific perspective, data serve as input to research, relying on data availability and accessibility. Consequently, opening data access can be seen as a benefit in order to enable wide usage of them (both for scientists and the public in general) and allowing specifically scientists to compare their results and methods more easily, and enhancing scientific accountability, credibility and potentially improve quality of data. Having environmental data in digital form allows an easy storage and dissemination, facilitates data exchange and sharing, faster and easier update and corrections, and gives the ability to users to integrate data from multiple sources. Consequently geospatial data can be thought as a shared resource.

Interoperability is the essential condition to develop an open science framework allowing scientists and researchers to publish, discover, evaluate and access data. The Open Geospatial Consortium (OGC) aims to develop and provide such standards enabling communication and exchange of information between different systems of different types operated with different software. Therefore, by committing to interoperability, any system allows to widely and effectively exchange data, to maximize the value and reuse of data and information under its control. It becomes possible to exchange these data and information with other interoperable systems, allowing new knowledge to emerge from relationships that were not envisioned previously. The ultimate objective is to spend less time in searching data and more in doing science by improving integration and analyses, and by better communicating newly generated information and results. Environmental challenges we are collectively facing today do not only require "useful science, but used science". From a technical point of view, publishing services to access data, metadata, and processing facilities is no more a problem, because various solutions are available. The model proposed by the OGC, based on web services, appears to be suitable to allow users to combine different services to solve a specific problem in a scalable and flexible way. Nevertheless, through simple examples of services chaining, we
have highlighted different issues that could potentially impede an easy and efficient integration:

- problems with different implementation of a same specification,
- problems with data/metadata harmonization
- problems regarding different projections used in different web applications,
- overload caused by on-the-fly reprojection using large data sets,
- Each component of a chain of services must be efficient otherwise it will negatively influence the overall execution chain.

Through these different scenarios of service chaining and data sharing, we can also emphasize different challenges:

- working with different scientific communities that are not necessarily aware of the possibilities offered by OGC web services could limit the diffusion of such approach outside the geospatial community. Therefore, capacity building appears a key element for large adoption, acceptance and commitment to data/metadata sharing concepts. These communities need to be convinced, through simple examples, that working with chained services can bring benefits to their own working flows. This will help to strengthen (1) existing observation systems, (2) capacities of scientists and decision-makers to use them, (3) capacities of the general public to understand important environmental, social and economical issues, and (4) capacity of data providers to share their data in an interoperable way.

- sharing and giving free and interoperable access to data is an efficient way to submit/expose data to the judgment of a specific community of users who can recognize whether this data is of sufficient quality. Moreover, this might increase interaction between data providers and users by receiving feedbacks and improving their data sets accordingly. Ultimately, this can potentially improve the overall quality of data, and then improve information that can lead to eventual better decisions.

- in the current climate of economical constraints (e.g., financial crisis), interoperability and standardization have never been so important. Due to the fact that geospatial data are often expensive and time-consuming to produce, it is then important to promote data use and reuse to the widest extent. This can potentially avoid duplication of efforts and expenses, and enable users to save resources, time and effort when trying to acquire and manage data sets.

- finding a consensus to ensure sufficient standardization and uniformity while recognizing the diversity and heterogeneity of stakeholders involved when sharing data/metadata.

- improving processing capacities of current SDIs. Distributed computing paradigm appears to be promising to complement SDIs capabilities using WPS services for processing large data sets. To achieve this objective, implementation of SOAP protocol into OGC specifications is a prerequisite in order to allow the two types of infrastructures to communicate (interoperability) and to ease the combination of OGC and grid services in efficient chains.

- having tools readily available to be deployed in developing countries is a strong incentive for capacity building, knowledge transfer, and sharing of
expertise. Therefore free and open source software can be potentially interesting to ease the development, portability, and replicability of tools especially for low to moderate incomes countries that are also often affected by natural hazards or environmental degradation.
• participating on a voluntary basis to data sharing initiatives could be seen either as a great opportunity or as a risk. Indeed, the voluntary aspect poses the threat that only a few data providers join such an initiative and, as a consequence, the system could miss its objectives.

In consequence, we can conclude that the establishment and implementation of initiatives like the Global Earth Observation System of Systems (GEOSS) reflects a growing commitment to (1) better and efficiently manage data, and (2) share data more openly by making them interoperable. This also indicates increasing recognition about the potential benefits for informed decision-making from evaluation, access, integration and processing of various environmental, economical, statistical and other data sources within a common framework. However, the full potential and benefits must be still proven, because currently, we are at an early stage of implementation. Finally, a unique added value of interoperable data and processing services is that it allows users to perform functions that cannot be made with any single component. By integrating/composing different services, new properties are emerging and offer possibilities to better understand the complex relationships between the different components of the Earth system.

2. How can SDI improve our capacity to discover, share, retrieve and integrate environmental data?

Based on the recognition that there is a clear need to share data among different environmental disciplines, Chapter 4 addressed some of the challenges and issues identified in the previous chapter. We demonstrate the applicability of SDI concepts, methods and technologies targeting the Disaster Risk Reduction community. Timely access and easy integration of geospatial data is needed to support efforts in reducing disaster risk and to promote a culture of disaster resilience. Geospatial data are essential in the disaster cycle of activities, and timely access to such data could improve decision-making process, and thus could save lives and/or minimize losses of property. Even if damages can not be completely avoided, better coordination and collaboration between organizations as well as better data exchange are helping to better identify risk and in the mid to long term, and to reduce losses. Systems that can deliver improved information on natural hazards will help preventing disasters.

SDIs could be seen as an answer to the growing need to organize data across different disciplines and organizations, and helping them address the issue of supporting decision-making process. To achieve this objective, data sharing and open access policies appear to be an important issue allowing an easy and wider usage of data. Having the ability to collect data once and reuse it many times is a clear incentive for such initiatives, avoiding duplication of time, effort and expenses, and improving access to good quality data and in turn improving decision-making process.
Chapter 6: Conclusions & recommendations

In this sense, the PREVIEW Global Risk Data Platform offers the possibility to freely and easily access data that could be useful for preparedness, response and mitigation phases. This has also highlighted that facilitating access to data by sharing it in an interoperable and standardized way increases the ease of integration and use. With the help of OGC standards, data can be available and accessible through a wide array of different stakeholders (partner agencies, humanitarian agencies, GEOSS) and by saving time in accessing and integrating data, this could positively influence the disaster response in an emergency situation. Nevertheless, this research has also shown that it is urgent to develop data models commonly agreed by the community in order to be interoperable at the syntactic, semantic and schematic levels. It is a necessary condition to leverage to full potential of interoperability avoiding heterogeneities and enhancing the quality of the disaster response phase.

Sharing spirit, accessibility, availability, interoperability, and data harmonization are important issues that can be clarified with the help of a clearly defined SDI conceptual model. Such a model enhances collaboration and partnership among different stakeholders and gives the possibility to agree on making these data freely available for the benefit of a whole community and beyond. To build an efficient and comprehensive model, it is required to identify and clarify five core components (people, data, standards, policies and accessing network) divided into two categories depending on the nature of their interactions within SDI framework. People and data constitute the first category, while the second is composed of technological components. This vision highlights the dynamic nature of these categories: evolving technology, changing data requirements, improved data sets, changes within user communities, and changes in policies. Altogether these different components interact, influencing each others, and evolving continuously.

While developing the PREVIEW SDI conceptual model, several benefits and different issues have been raised requiring finding (when possible) suitable agreements/solutions ensuring a good and consistent SDI implementation:

- Absence of a commonly agreed data models within disaster community. Therefore, it was required to develop a common set of specifications for harmonizing each category based on an assessment of the diversity and heterogeneity of various collected data sets (i.e. metadata, data formats, attributes, projections, terminology, naming conventions).
- Geospatial data sources are often fragmented and metadata are often missing or incomplete.
- Data quality is a key requirement for users in order to obtain meaningful results. Four categories can be identified: data completeness (amount of missing features), data precision (degree of details), data accuracy (degree to which data reflects correctly the real object) and data consistency (usability of the data). These categories can be assessed at two levels. The data producer that checks the quality of data based on given data specifications makes the first level. The second level is based on the users side that provide feedbacks that are taken into account to update/correct data. Metadata can be also useful to track data quality.
- The act of sharing can be thought as exposing data to the judgement of others.
Chapter 6: Conclusions & recommendations

• The policies component is the most sensitive one, as it must facilitate and encourage organizations to participate in the development of a SDI. It should be noted that building a SDI is not only a matter of technology, but rather first relies on individuals and/or organizations. Hence, it is a long-term process that depends on support and commitment.

• Cultural behaviour and political aspects may strongly influence the opening/sharing of (meta)data.

• The multi-stakeholder nature of participants requires consensus to accommodate their diversity and heterogeneity, while ensuring sufficient standardization and interoperability. Hence, involving all stakeholders in early development stage of SDI is crucial in order to reach agreements and develop a conceptual model that reflects a coordinated and common vision and sense of ownership, and fosters sharing, accessibility, availability, interoperability and data harmonization.

• Partnership is essential because a single agency is unlikely to have all resources, skills or knowledge to undertake the development of all aspects of a SDI.

• Some partners are still reluctant in sharing openly their data, mostly for financial and visibility reasons: they might be willing to provide free access for research purpose, while selling data to profit making institutions (e.g., re-insurance companies). Some data may also be only temporary protected (e.g., prior to publication). Finally, in other cases, the data providers agree to freely distribute their data, but prefer to do so through their own website instead of through a third-party application.

• In general, data providers easily agree to freely share their metadata because it gives visibility to their data and associated services.

• The previously mentioned situations are part of a transition phase that any new technology has to face. The sustained use of the platform should show the benefits to share data and reinforces the need to build new capacities by showing appropriate examples, sharing experiences and developing guidelines and policies. Hence, capacity building should be made at three levels: human (education and training of individuals), infrastructure (installing/configuring/managing of the needed technology) and institutional (enhancing the understanding within organization and governments of the value of geospatial data to support decision-making). All these actions will help to reach endorsement on the use of such technologies, raising and increasing awareness on the benefits of using and sharing geospatial data, and finally creating new commitments.

• The Geospatial Portal Reference Architecture based on OGC services appears to be a suitable and reliable architecture to support data sharing, discovery, visualization and retrieval. It provides a solid ground to develop a geoportal using OGC interoperability standards through four classes of services (e.g., portal, catalogue, portrayal, and data). It is leveraging the real potential of interoperability by “storing once, reusing many times” and allowing services to be (seamlessly) coupled, reusable and available for a wide variety of applications.

• Organizing a user-testing phase is very valuable in order to ensure that a geoportal is easily usable and meets user expectations. Engaging different
users with different backgrounds helps identifying benefits, weaknesses, inconsistencies, and bugs that need to be solved/corrected.

- During major events such as in Haiti (January 2010) or Japan (March 2011), an important increase in term of data download, map production and web access was observed on the platform. This demonstrates that such a platform is used to provide general hazards, exposure and risk data for the disaster community and the media. Moreover, the quality of proposed services appears to be suitable as no interruptions have been reported, which means that even when numerous concurrent accesses occurred, data are still accessible and available. Such observations confirm and reinforce the idea that SDIs together with their geportals could act as gateways to geospatial information maximizing the reuse of data and ensuring readily accessible to up-to-date data.

The development of such a platform has highlighted several benefits and raised various issues when sharing interoperable data. In particular, computational needs to process large data sets and efficient access to geospatial data through OGC services are two factors that will strongly influence the future success of SDIs. Ensuring user satisfaction through sufficiently responsive services, giving access to vector and raster data sets require to measure and monitor them to track latencies, bottlenecks and errors. Despite the importance of data retrieval and access, only little research has been published to benchmark and evaluate the quality of WFS and WCS services. Most of the studies concentrate on the usability of OWS when integrating distributed data sources but do provide neither framework nor guidance to measure performances of data services. However, evaluating and predicting user's satisfaction is a complex task because quality can be based on quantitative and qualitative measurements. Quality can be interpreted both as a measure that represents accessibility/performance and also as a perceived quality of a service. Despite the importance of qualitative perceptions (e.g., good description, easy access to a service) it is undeniable that insufficient technical quality of service will strongly affect users' satisfaction. With the expected increasing diffusion of OGC Web Services (OWS), Quality of Service (QoS) will be an important factor to distinguish between reliable services and others. Client-side testing may help to evaluate characteristics that identify measurable/quantifiable parameters such as performance, capacity and availability. Considering these issues, we have proposed: (1) an open approach to measure the performance of different WFS and WCS services provided by various software implementations, and (2) some guidance to data provider aiming at improving the quality of their services.

Currently, in absence of a commonly agreed framework to test such download services, it is important to measure performances with approaches that are (1) sufficiently generic to be independent of infrastructure and application design, (2) independent of the communication (e.g., transport network) between the service and client, and (3) based on request-response pairs and avoid complex transactions. Therefore, the main objective of such a testing framework was to assess the usability and performances from client-side/end-user perspective of various network services, specifically download services. The proposed approach is based on the one developed by the Free and Open Source for Geospatial (FOSS4G) community to test WMS services. The aim is to extend this open approach to WFS and WCS services in order to test various
Chapter 6: Conclusions & recommendations

OWS implementations on a common and widely used set of geospatial data (e.g. GLWD, Blue Marble), on a common and widely used platforms (e.g., ArcGIS Server, GeoServer) and architecture model, and on common and widely used input/output formats (e.g., shapefile, file geodatabase, ArcSDE, PostGIS, GeoTiff). Obviously in order to fulfill the requirements mentioned previously, it is relevant that these tests are based on realistic usage scenarios.

The results of our tests have highlighted:

• generally the different tested implementations provide already good general performance right out of the box.
• computer memory (RAM) is the most critical factor to control and many factors can influence the response of a given service.
• simplifying the instance (e.g., turning off extraneous services or options), configuring suitable memory parameters (e.g., map rendering, type of data served), and managing the number of request (e.g., limiting the number of concurrent requests that could prevent timely responses, workload manager) are factors that can improve the overall quality of web services.
• various other factors may affect performances in particular if implementations use containers like Java Virtual Machine (JVM) or Jetty. Tuning the different options provided by these containers may significantly improve performances of the proposed services.
• optimizing data and storage can significantly increase service efficiency and reliability.
• ESRI file geodatabase can be a suitable format, both for vector and raster data, as it provides good performances compared to flat files or even ArcSDE. It can potentially become a common and cross-platform format because ESRI has recently released an open API, meaning that this format can be now virtually used in any GIS software.
• native solutions like ArcGIS Server/ArcSDE and GeoServer/PostGIS give also globally good results and may be reliable solutions to serve efficiently data using WFS and WCS standards.
• indexing geometry and attributes significantly improve performances by accelerating performances and data retrieval of vector data stored in databases.
• building pyramids and overviews improve the performance of services serving raster data.
• there is not a completely linear relationship between storage size of the data set and overall performance of the download service.
• Similarly, on-the-fly reprojection can be CPU intensive and storing data in the most frequently requested projection may also improve performances.
• In a more general context, WFS and WCS specifications are suitable standards to share and access data in an interoperable manner but even if data and storage are tuned and optimized some bottlenecks may appear when transferring large volume of data, potentially leading to high latencies and low performances. By their nature WFS and WCS are very sensitive to resolution. Consequently, these standards are more suited to share local medium-resolution data than global high-resolution data.
Chapter 6: Conclusions & recommendations

- differences in OGC specifications have been noted in the various implementations tested (e.g., filter encoding, parameters). This may lead to interoperability problems especially if clients do not implement the different flavors of these specifications and may limit seamless data integration capabilities.

The development of the PREVIEW Global Risk Data Platform has exemplified that having geospatial data in digital form allows easy storage into databases and file systems, facilitates data exchange/sharing, faster updates, gives the ability to integrate data from multiple sources, and finally favors developing customized products and services. SDI concepts, methods and technologies, provide a solid ground to facilitate and coordinate the exchange and sharing of geospatial data. Supported by interoperability arrangements SDIs offer the possibility to link together environmental, socio-economic and institutional geospatial data resources, providing a movement of data from local to national and global levels. However, to fully realize the benefits promised by SDIs, human component is the essential element to be taken into account because cultural, political, and financial aspects may influence either positively or negatively the acceptance and adoption of SDIs. This is on this component that future endorsements depend, because besides details at the technological level, it is no more a problem to share interoperable data and metadata. At this stage, it is only a matter of human/political will to make it happen or not. In our view, capacity building (at human, institutional and technical levels) will certainly help to reach endorsement on the use of such technologies, raising and increasing awareness on the benefits of using and sharing geospatial data, and finally creating new commitments.

3. Can SDI take advantage of distributed computing power to process the increasing amount of high-resolution data?

As of today, it is estimated that geospatial data is the most important data type in terms of volume accounting for more than 80% of data collected by humankind. Due to their complexity and multi-disciplinary nature, environmental data are very diverse, collected and operated by various data centers around the world. Extracting information often require complex analysis and processing procedures together with handling vast amount of potentially (high-resolution) data. In such conditions, most users of geospatial data are often experiencing lack of computing resources to analyze these data. Therefore, the essential task of transforming raw data into understandable information cannot currently fully delivered. Typically, environmental sciences are data and computing-intensive domains where data can be discovered, visualized and access through SDIs, but processed on desktop computers. This clearly limits the types of analyses that can be conducted due to their reduced power and given an ever-growing size of data that need to be analyzed. Hence, SDIs need to go “one step further” extending their analysis capabilities and find solutions to manage and process in a more efficient way the vast amount of geospatial data that are generated and collected every day. It is estimated that more than 50 percent of scientists are working regularly with data sets larger than 1 gigabyte and up to 1 terabyte. Therefore data management, distribution, heterogeneity, volume and
processing are becoming increasingly challenging, requiring new approaches to improve understanding on (1) dealing with these emerging issues and challenges and (2) offering new perspectives to scientific communities who will be influenced by this data deluge. In the field of environmental sciences, different research areas can be prospected: (a) spatial data infrastructures, (b) distributed computing environments, (c) large scale access and integration of data repositories, (d) workflow management, and (e) metadata management.

The increasing computational power and network capabilities enable web services technology to extend SDI capabilities by providing access to a collection of geospatial calculations (like in a standalone desktop GIS software) delivering some concrete functionality. The OGC Web Processing Service (WPS) specification and the promises of high storage, data management and computing capacities offered by distributed computing infrastructures like Grids and Clouds, hence offer new opportunities within the environmental communities.

Chapter 5 presented and discussed (1) the actual status of technologies used to describe, catalog, share and process an ever-growing set of high resolution geospatial data analytical limitations of SDIs and presents (2) promises and challenges offered by Grid to extend analytical capabilities of SDIs. To extend processing capacities of SDIs several attempts to implement the WPS specification in a distributed computing environment have been successfully made. Nevertheless, they are in general dependent on the middleware used by the distributed computing infrastructure meaning that with the current diversity of computing environments it is very difficult to execute and reuse a given WPS-enabled process on different computing backends. This constrains service providers to develop a different implementation for any specific backend. This situation can potentially limit the development, adoption, and diffusion of WPS. Consequently, we presented (3) an approach to extend WPS specification on distributed computing infrastructures, allowing users to execute a given geoprocessing task independently of the computing backends (local servers, clusters or different Grids/Clouds), avoiding the need to rewrite processes by making WPS processes as scalable and flexible as possible.

Using a distributed computing paradigm like the Grid can bring major benefits to support SDIs but a lot of challenges and incompatibilities between these two types of infrastructures have been identified and need to be overcome in order to fully benefit from the potential offered by the Grid:

- Service description and interfaces: OWS are based on HTTP-GET/POST and XML-RPC while Grid services are based on SOAP/WSDL.
- Service states: OGC services are stateless and thus cannot give any information about their state while Grid services are statefull.
- Service security: OGC specifications do not include security mechanisms that are a key element of Grid infrastructures.
- Current Grid metadata catalogs are not well suited to handle geospatial metadata based on ISO19115/19139/19119 and OGC CSW standards.
- XML does not support raw binary data implying that WCS cannot be used directly into SOAP messages.
- No Grid services are equivalent in terms of functionalities to WFS and WCS.

Chapter 6: Conclusions & recommendations
Chapter 6: Conclusions & recommendations

- WPS is a good candidate to be grid-enabled but current implementations are often specific to Grid middleware.
- WPS is lacking of adequate service metadata impeding users to discover WPS services.
- WPS 1.0.0 specification has a limited SOAP/WSDL support that restricts orchestration and chaining capacities of OGC services with other Grid or non-Grid services within workflows.

To achieve the goal of linking Grids and SDIs, interoperability appears to be a key requirement, as well as an important and challenging task. In particular, the implementation of SOAP messaging protocol into OGC standards is a necessity. This will greatly enhance the gridification process of OGC Web Services as well as allowing easier workflows integration using orchestration engine to combine OWS and Grid services. One of the main challenge is to be able to use secure sharing and processing mechanisms provided by Grid infrastructures while using and preserving widely adopted OGC interfaces and services within geospatial community.

Consequently, Grid infrastructures may provide solid components to empower SDIs in term of scalability, flexibility and interoperability for:

1. **Data processing**: providing a high-performance computing environment is among the first objective of a Grid environment. Different studies have shown clear improvements in processing performance and speed, by dividing a given task into smaller subtasks that can be processed in parallel and merged together at the end of the calculation.

2. **Data management**: by its distributed nature and characteristics, Grid environment is potentially an interesting choice for a data management system. It offers robustness (distribution storage and data replication capabilities), efficiency (data stored as close as possible to components that access them), transparency (hiding Grid complexity to users), and security (secure access to resources through certificates). In addition, data moving protocol like GridFTP are interesting as well. Indeed, one possible bottleneck when dealing with large data sets is that data access strategy in SDIs is not location-based. This means that current SDIs have limited replication and data transfer capabilities to minimize the access time to a selected data set. Often, geospatial data providers do not offer possibilities to have different replicas distributed in several data centers. Therefore, users must access data directly at its source, which implies that network distance and a potential large number of concurrent accesses can impede users to retrieve data in an efficient way. Moreover, if a geospatial data source disappears, data access to this source is broken if data replication was not enforced. In summary, data replication mechanisms proposed by Grids are promising to avoid single points of failure, to enhance data availability, and to ensure that data will be as close as possible to the worker nodes (avoiding high latencies produced by the movement of large size data before the beginning of a processing task).

The limited analytical capabilities currently provided by SDIs and the need for an efficient and powerful processing environment (i.e. to process large data sets) explain the strong interest on processing data using WPS specification and Grid computing.
Currently, there are no commonly agreed approaches to grid-enable WPS. However, two types of gridification processes have been recognized: encapsulation and integration. Encapsulation is recognized as a “low-level gridification” meaning that applications or services remain unchanged and can interact with distributed computing resources in the backend. For the Integration process, applications and services are resources fully embedded into the Grid middleware. Despite the fact that integrative gridification is the most promising one to support SDIs, this approach remains rather complex to implement because it requires important developments to transform OWS into Grid services. Moreover, the dependency of this approach to various middleware results in specific OWS middleware implementations. This situation might be potentially solved in the future with the adoption of interoperability standards like OGSA and WRSF. Therefore, with the diversity of computing environments no current implementations allows users to seamlessly execute a geoprocessing calculation following their requirements on different computing backends, ranging from a stand-alone GIS server up to computer clusters and large Grid infrastructures.

The proposed WPS mediation approach is built on an intermediate gridification level and offers:

- Simple implementation of WPS specification.
- No dependence on Grid (or other) middleware.
- Service providers to write scalable WPS process and users to select backends following their computing power requirements.
- Potential access and use of other distributed computing infrastructures such as Clouds (e.g., Amazon Elastic Compute Cloud, Google App Engine), Clusters (e.g., Oracle Grid Engine), and Desktop Grids (e.g., XtremWeb that exploits unused resources such as CPU and storage of desktop computers spread over a network to execute tasks in parallel).
- The management of data tiling and merging when using a geospatial data set.

To fulfill these requirements, the mediation approach was selected for developing a specific WPS implementation that integrates and benefits from the functionalities offered by different pieces of software (e.g., PyWPS, GDAL/OGR, Ganga).

A NDVI WPS mediated process was developed highlighting benefits both in term of performance (processing larger data sets; processing data faster) and scalability (using the same process on different computing backends) successfully exemplified this approach. From the obtained indicative results several observations can be made:

- Despite the fact that a MODIS data set can be considered voluminous (19200X14400, 250m resolution, 1.2Gb), a WPS process executed on a single GIS server (with 3Ghz CPU/4Gb RAM) can easily handle this data set avoiding the need to use the Grid.
- However, a high-resolution merged image (30’000x30’000 pixels at 10cm resolution, 3.6Gb) cannot be processed locally on a single WPS instance due to lack of memory.
- Processing the same data set on the Grid is feasible but important overheads are introduced. In particular, tiling and merging such a data set in 50 tiles is time consuming. This probably comes from GDAL internal...
process implementation (e.g., multiple load of files in memory). Submission and execution on the different nodes can also be quite expensive and dependent of the workload of each computing elements.

- Despite the small overhead introduced by jobs submission, executing calculations on the Grid on already tiled data sets is the most efficient way. In this way, and in contrary to a local execution that processes jobs sequentially, the nine submitted jobs are executed simultaneously. In theory, such parallel execution can improve the overall processing phase by a factor corresponding to the number of submitted jobs (nine in our example). Nevertheless, it is not the case here due principally to varying workload conditions on different nodes. The improvement is around 15% and should increase with growing number of jobs (i.e., processing 10'000 tiles will probably show a much larger improvement).

Nevertheless, several issues and shortcomings have been raised and require further investigations in order to build responsive, efficient and reliable grid-enabled WPS instance:

- Actual WPS specification is not sufficient to fully benefit from distributed computing infrastructures capabilities. However, using WPS Application Profiles, we can introduce two parameters to extend WPS interface. They allow users to select the number of processing jobs and where to execute these jobs.

- Service metadata available through a GetCapabilities operation must reflect the above mentioned extension by providing a list of supported computing backends.

- Some aspects of the current WPS specification are currently not sufficient to handle different types of references to ensure sufficient scalability within the process logic.

- Currently available data types in WPS specification are too generic and need to be enhanced. Nevertheless the ComplexData type, used to handle vector and raster data in the current WPS specification, does not provide a possibility to locally reference stored data as inputs (i.e., it only addresses remote data sources). Although it is not foreseen for this usage the LiteralData type helps to overcome this issue. This data type accepts any character strings and then allows introducing path and filename to local resources. If data inputs are declared as such, it is then possible to locally reference stored data sets.

- Current implementation of the WPS mediator only supports GeoTiff, GML and shapefiles. However, it can be easily extended to all other supported formats in GDAL/OGR.

- Presently, it was impossible to test the integration of mediated WPS processes with a chain of services due to the current lack of SOAP/WSDL support in PyWPS. The next version of PyWPS will introduce an extended support of SOAP/WSDL and should therefore solve this issue.

- At the present stage of development, the mediation layer does not manage Grid certificate accessing a WPS instance through the web. It requires having a valid certificate installed on the server. Nevertheless, the WPS mediator could be adapted with a module that can check required credentials and authentication to access a Grid (e.g., WPS Profiles & MyProxy server).
OGC standards are stateless (i.e., lacking capabilities to give information about their state). However, it is straightforward to monitor jobs status (e.g., submitted, running, complete) with Ganga, and some Python bindings should make it possible to develop suitable interfaces to retrieve information about the state of submitted jobs.

Performance is one of the main problems when implementing geoprocessing tasks in a distributed environment. Network bandwidth, data transportation and data validation are potential bottlenecks when dealing with large geospatial data sets. Except in a few cases, the Internet network bandwidth will always be a limiting factor, as data will be transported over the Internet. Thus, minimizing data transportation (having data stored physically close to the processing elements) and data validation is of high importance.

In order to decrease the overall overhead, the mediation approach might be potentially useful as it allows moving some of the components directly into the Grid. Indeed, a job using the Mediation layer is constituted by (1) process logic, (2) data to be processed, and (3) a PyWPS/GRASS binary package (engine) to execute the process (currently weighting around 27MB). If the latter package is directly accessible on the nodes of a Grid infrastructure and data is available in Storage Elements (SE), the mediator will be used only to forward WPS requests and send process logic through Ganga.

Ganga is currently only used within the mediation layer to submit geoprocessing task on different backends. It does not consider the workload on different computing nodes. In other words, if a node is available then a job is submitted. When submitting multiple jobs at the same time, this can negatively impact the overall processing time. In such situation, Ganga will wait until sufficient nodes are available to submit these jobs. However, in the vision of integrating Grid capabilities into SDIs, responsiveness appears to be an important condition. One possible solution is to use, in conjunction with Ganga, a tool called DIANE (DIstributed ANlysis Environment) that aims to improve the quality of services using automatic control, workload management and jobs scheduling on a set of distributed worker nodes.

Another element that may influence performances of the mediated WPS service is the download and tiling modules. If data are not stored locally they need to be downloaded and this may require some time depending on file size and bandwidth. Once download is finished, this dataset will be tiled according to the number of jobs requested by the user. In such a case, these tiles can still have a size of several megabytes and must be uploaded to the different computing nodes, introducing a data transportation overhead. If tiled data are directly available in SE this allows bypassing the need to download and tile data, and this reduces the induced overhead close to zero. Nevertheless it needs further investigation to extend the mediation layer to take into account Grid specificities (e.g., Logical File Name). Other possibilities to achieve the objective of fast access to large and widely distributed amounts of data are (1) Peer-to-Peer (P2P) technologies or (2) development of new protocols like GridJet that accelerate data transport.
Finally, parallelization is an important factor to take into account because it potentially strongly affects the quality of proposed services. Parallelization can indeed be done at the task (i.e., task parallelism) or at the data level (i.e., data parallelism). In the former case, a task can be split into various independent subtasks working on the same or different data set, while in the latter case each subtask operates on a subset of the whole data set. Data parallelism introduces the necessity to efficiently tile and merge data (to lower induced overhead) and to consider effects (e.g., interdependencies) at tile borders to obtain adequate result. Moreover, not all geoprocessing tasks can be subdivided into independent calculations (e.g., climate models requiring communication among jobs and then more suited for parallel computing). Consequently, suitable strategies are required. In the mediation approach, further investigations are required to find a efficient solution to manage and offer such capabilities to users and extend the solution currently implemented.

- WPS Application Profiles allows to find suitable solutions: (1) to offer secure access to Grid resources, (2) efficient tiling/merging strategies, (3) optimize the use of computing resources (e.g., jobs submission, backend selection, number of tiles to be generated). These solutions can be afterward implemented in the mediator in order to reach the objective of high performance computing of geospatial data.

In our vision, grid-enabled SDIs have the potential to become a powerful tool within the multi-disciplinary field of environmental sciences, empowering researchers to explore new venues to better understand the vast complexity of the interactions between anthropic and natural systems. The first generation of SDIs, based on a product model, gave way to a second generation at the beginning of the years 2000 that is characterized by a process model. This evolution emphasizes the shift from the concerns of data producers to those of data users and the shift from centralized structures to decentralized and distributed networks. In our view, connecting Grids and SDIs could potentially mark the advent of a third generation of SDIs extending their capacities to, and benefiting from, Grid infrastructures both in term of data processing and data management. Using Grid as a computational backend represents only a first step towards this vision. The proposed mediation approach opens new perspectives for the grid-enablement process and enriches the discussion concerning future improvements of the WPS specification along with supporting Grid community in finding new and promising directions of research and developments.

### 6.2 Limitations/Constraints to SDI diffusion and utilization

Despite the fact that SDIs methods, concepts and technologies can bring major benefits they still need to show their real benefits/added-value at large scale and therefore it is important to consider the factors that currently limit the diffusion and utilization of SDIs.

Henricksen (2007) identifies two classes of factors that can influence successful implementation and operation of SDIs: (1) tangibles (e.g., technology, tools, frameworks, methods, systems), and (2) intangibles (e.g., behaviors, resistance, commitment, accountability, self-interest, culture). The relative
influence is about 20% for technological factors and 80% for intangibles ones over SDI effectiveness. This author concludes that to achieve the objective of a successful SDI implementation, the key requirement is that institutions and people must be willing to work together and share a common vision. Results from various assessments and consultations (Crompvoets and Bregt 2003; Henricksen 2007; Craglia and Campagna 2009; Vandenbroucke 2010) showed that besides technological aspects the main issues are related to (1) political/cultural context, (2) policies, (3) organization, (4) people, and (5) resources.

Among the most frequent obstacles, lack of institutional and political wills to publish and share data is frequent (Nebert 2005). Indeed, data providers tend to hide data mostly for confidentiality, national security or “misuse prevention” reasons. This inevitably leads to duplication of activities, duplication and fragmentation of data, overlaps between initiatives and projects, lack of coordination, insufficient flow of information, and inadequate resources management. Additionally, insufficient staff skills can cause a lack of standardization (e.g., data, metadata, procedures) and lack of documentation on who is doing what and what is available. This incoherent, inconsistent and unshared vision of a SDI induces: (1) difficulties in finding/accessing required data, and (2) lack of knowledge from data providers about the value of what they have. Altogether, these reasons preclude a reliable organization for data dissemination purposes and prevent enhancement of the value of integrated information.

Various studies (Rajabifard, Feeney et al. 2002; Budhathoki and Nedovic-Budic 2006; Masser 2006; Masser 2006) highlighted potential weaknesses and threats that may also contribute to restrict the development and implementation of SDIs:

- Diversity of stakeholders (e.g., public sector, private sector) requiring taking into accounts a variety of (potentially conflictual) requirements and management structures.
- Changes in existing organizational structures in order to facilitate access and sharing of geospatial data to broad array of users. To overcome reluctance to changes, this requires a shift on the perception of the value of data, and to reach commitment on promoting data sharing.
- Developing and maintaining a consensus over time among the stakeholder involved because implementing a SDI is a long-term process.
- Dependencies to political support and commitment.
- Insufficient human and financial resources.
- Dependency on opportonites offered by the sociopolitical context (e.g., stability)
- Legal context & policies.
- Economical considerations (e.g., costs/benefits).
- No clear definition and consensus on SDI and its constitutive elements and principles. This contributes to give an incoherent and inconsistent vision of SDIs.
- Different models of SDIs (e.g., hierarchical, spider net) that do not necessarily take in consideration all the stakeholders involved.
Chapter 6: Conclusions & recommendations

- Too much standards create confusion for the user that cannot clearly identify which one to use and for which purpose.
- Lack of monitoring and evaluation mechanisms (e.g., indicators) helping to evidence who the users are, what they are using and how well SDIs can serve their needs.
- Finding the right balance between technology and social. Indeed, the risk is to concentrate too much on the technological component and neglect the social component.
- Need of multi/interdisciplinary approaches because SDIs rely on various disciplines such as sociology, cognitive science, political science, economics, informatics, and information science.

Developing and implementing a SDI implicitly ask for cooperation that often requires new forms of collaboration both internally and externally of an institution (de Vries 2009). Cooperation is strongly influenced by political-organizational factors such as institutionalised historical practices or dominating legacy systems and standards. Moreover, resistance can appear if the perceived complexity of (top-down) SDI initiatives do not directly relate to the own/local objectives of an institution, organization or a country. Therefore it is important to consider appropriately requirements of each stakeholder that need first to fulfill their own mandate. De Vries (2009) showed that despite INSPIRE is a legal framework, local policy motives, local politics, local interests, local alternatives to national or regional initiatives can negatively influence its implementation. Limited financial resources can also lead organization to select other, potentially non-optimal, solutions or technologies, which still match their local and immediate needs. Consequently, organizations are intending to cooperate only if they can perceive benefits. To make cooperation successful in SDIs, various elements are essential: a shared vision between all participants, necessity to solve a dominant problem, incentives, and sufficient resources (Salzmann 2011). Actions can be initiated not only by legal regulations (like in USA or in Europe) but also by protecting/creating some interests, or minimizing some financial costs (de Vries 2009). Finally, for Salzmann (2011) collaboration is also strongly influenced by the maturity of the environment (e.g., staff skills, shared problem) in which the SDI implementation will take place. However, collaborations are a complex set of dynamic formal and informal relationships and are consequently difficult to assess (Warnest, Rajabifard et al. 2003). Organizational, institutional and economic obstacles can be overcome at the condition of better collaborations and partnerships. Nevertheless, Warnest et al. (2003) stated that our current understanding on how to create and develop effective and efficient collaborations and partnerships is still limited and then bringing an important restriction. Participatory approaches can be useful to understand the nature of interactions between the different stakeholders, their motivations, their requirements and finally sustain effective SDI implementations (Warnest, Rajabifard et al. 2003; Strobl, Belgiu et al. 2010).

Craglia et al. (2008) stated that to justify the initial investment and ensure long-term sustainability of an infrastructure, appropriate methodologies to assess social and economic impacts of geospatial information are needed. Currently such frameworks are still little developed (Booz, Allen et al. 2005; European Commission 2006; Garcia Almirall, Moix Bergada et al. 2008; Craglia and Campagna 2009) and required more work to be done in order to show
positive financial impacts and, for example, associated costs of non-actions. As Nebert (2005) said: “...the longer the harmonisation of stand-alone databases is post-poned, the more difficult it will be to make them interoperable. Costs for integrating standalone systems into a SDI concept are increasing exponentially with time and the number of data sets. This suggests that a co-ordinated initiative based on SDI principles should be considered as soon as possible. A feasibility study carried out in Malaysia prior to the implementation of a national SDI concluded that a SDI would present an opportunity with dynamic benefits that would grow over time, culminating in accelerated socio-economic development the nation combined with a reduction in delays in the implementation of projects”. A recent report of the Joint Research Center of the Europen Commission (Craglia and Campagna 2009) indicates that regional and local levels are interesting scales to study socio-economical impacts of geoinformation. In particular it is at the local level that SDIs can bring large benefits, directly influencing, affecting and supporting millions of citizens and local businesses in their daily activites. For the authors of this report “think global but act locally” is the only possibility to realize at large extent the (potential) economic and social benefits of geospatial information, as well as engaging various stakeholders and reaching commitments.

Rajabifard et al. (2002) suggested that to improve support and commitment to SDI initiatives it is important: (1) to increase the level of awareness about the nature and value of SDIs (e.g., capacity building), (2) assess and understand the dynamic nature of collaboration and partnership in order to support a culture of sharing, (3) improve SDI models to better match the needs of various communities, (4) improve SDI definition to giver a clearer vision, and (5) identifying the key factors (in a given context) that can facilitate interactions between social, economical and political issues. Due to the importance of their social dimension, SDIs can be thought as social networks of people and organisations supported by data and technology (Craglia and Campagna 2009). Indeed, developing the technological component is rather simple but building and maintaining the social one is much more difficult requiring important human and financial resources as well as collaboration, partnership, commitment and trust. For Craglia et al. (2009): “The technology is cheap, data is expensive, but social relations are invaluable”.

6.3 Recommendations/Perspectives

Based on the experience acquired through this research that has shown benefits and limitations of SDIs methods, concepts, technologies and raised various issues on different topics both at scientific, technological and institutional levels, we aim to provide some recommendations and suggest directions of research. The aim is to contribute to make operational the different building blocks of an interoperable web-based geospatial information framework ensuring that data providers and end-users can access, share and disseminate data and information about the environment in a more open and efficient way.
Chapter 6: Conclusions & recommendations

**Recommendation #1:**
Find simple ways to document geospatial data.

Metadata are recognized as an essential element in SDI initiative for managing and discovering geospatial data. However, metadata documenting geospatial data sets are often missing, or are incomplete or fragmented. This is mainly explained by the following facts: (1) current metadata scheme and standards are complex and difficult to handle, (2) documenting geospatial data is often not perceived as an important task in projects (rather focusing on creating data sets but not taking into account the need to document them), (3) lack of resources both in term of time and financial support, and finally (4) documenting data is not really a enjoying task (and being honest, it is rather a boring task). Therefore, finding flexible and easy ways to document data would be very valuable otherwise the primary objective of SDIs will not be realized and SDIs would miss their essential purpose of making data discoverable. We can imagine systems that allow capturing relevant information when producing data reducing to the minimum the need to users to fill forms with tenth of fields. Such metadata generation framework is essential in particular with the huge amount of geospatial data generated continuously. Such framework must offer various capabilities such as automatic creation, automatic update, and automatic enrichment (Kalantari, Olfat et al. 2009). Additionally, having metadata as close as possible to the produced data set is also very important. Indeed, it is common to see data and metadata completely separated, and this probably also explains why metadata are often missing. Separation of storage creates two independent components (i.e., a data set and its associated metadata) that must be managed and maintained. This will inevitably lead to inconsistencies, redundancies and lack of reliability. In this sense, the netCDF format (recently accepted by the OGC as a standard) appears interesting as it allows having data and metadata in the same file. It is designed to represent space and time-varying data and netCDF and provides mechanisms to support multi-dimensional data. However, it requires implementing and supporting common metadata standards such as ISO19115/19139/19119 and also extends its supports to other data types (e.g., vector). In consequence, having a system where once a geospatial data set is updated, its metadata is also immediately (and automatically) updated will bring major benefits (Rajabifard, Kalantari et al. 2009).

**Recommendation #2:**
Manage and communicate about data quality

An important challenge that is not yet achieved concerns the quality of data. As seen during the development of the PREVIEW Global Risk Data Platform, once users have found required data, one of the first question they may ask regards the quality of data. Quality can be assessed at quantitative and qualitative levels: (1) quality control describing data completeness (amount of missing features), data precision (degree of details), data accuracy (degree to which data reflects correctly the real object) and data consistency (usability of the data), (2) managing uncertainty associated to data and models (quantifying uncertainty and encode it into metadata), (3) users feedback (e.g., user view of data utility). In our vision, the last point is rather important because even if a

---

**66** http://www.opengeospatial.org/standards/netcdf
Chapter 6: Conclusions & recommendations

data is complete, precise, accurate and consistent, sometimes it might be not useful from a user point of view. Hence, we think that the act of sharing data is really important in the sense that exposing data to the judgement of a community may also contribute to assess the quality of data. Communicating about the uncertainty of data and models obviously appears a necessity. Interesting efforts to manage and communicate on data quality are exemplified by projects such as GeoViQua\textsuperscript{67} and UncertWeb\textsuperscript{68}. GEO/GEOSS is seeking to define a GEO Label that will inform users about quality, relevance and acceptance on data and services.

**Recommendation #3:**
**Exploring the potential of mobile devices**

The growing adoption of smartphones and tablets (with GPS and wireless communication for internet access) has changed the way users are accessing the Web and consequently offers new possibilities to stream information to users (e.g., data consumers) and to capture information (e.g., data producers). Although desktop computers will still exist we can already see a clear shift to mobile devices. Therefore, we can imagine that most of tomorrow's applications will be accessed on such devices allowing users to seamlessly access data and information at anytime and everywhere. The current convergence between telecommunication and data communication is a sign in that direction. This will raise constraints in term of processor speed, memory space, screen size, network bandwidth, data visualization, and diversity of devices (e.g, consistent functionalities and understandable interfaces). This new mobile environment will obviously influence GIS and offer new possibilities particularly for field data collection and editing allowing data capturers to introduce data directly into a database in real time. This may also offer the possibility to work in either connected or disconnected environment by accessing large amount of geospatial data and cache it locally. That way, users can perform their task in the field without the need to go back to the office, only synchronizing new or update data in the field. Finally, this will allow users to discover content, display relevant information and eventually capture new information on real time. This possibly extends geospatial data to a wider audience and allows data consumers to interact with data more easily.

**Recommendation #4:**
**Exploring the potential of Volunteered Geographic Information**

Related to recommendation #3 and the new possibilities offered to users to generate content, Volunteered Geographic Information (VGI) is defined by Goodchild (2007) as “the harnessing of tools to create, assemble, and disseminate geographic data provided voluntarily by individuals”. Good examples of such participative approaches are Wikimapia\textsuperscript{69}, OpenStreetMap\textsuperscript{70} and Google

\textsuperscript{67} http://www.creaf.uab.es/projectes/geoviqua/
\textsuperscript{68} http://www.uncertweb.org/
\textsuperscript{69} http://wikimapia.org/
\textsuperscript{70} http://www.openstreetmap.org/
Chapter 6: Conclusions & recommendations

MapMaker\textsuperscript{71}. VGI is enabled by the Web 2.0 paradigm allowing users to generate content, populate websites, and interact with other participants. Additionally simple and widely diffused means such as mobile phone (equipped with a GPS) allow users to capture information (e.g., pictures, messages) with a geographical reference. VGI can be thought as a participative approach to GIS, allowing users to generate data/content by providing information on specific locations where events may occur (e.g., damages to buildings after an earthquake event) or features that can exist but are not available on publicly available data sets (e.g., road network in developing countries). Consequently any (potential) users can be thought as a sensor. Goodchild (2007) distinguished three types of sensors: (1) static sensors used to acquire measurements on a bounded location, (2) mobile sensors carried by humans, or vehicles (e.g., air pollution monitoring, noise monitoring), and (3) humans themselves (with their senses and intelligence) that are able to gather and interpret what they perceive on their local environment. If static and mobile sensors can provide quantitative values it is not the case for human sensors that generate more qualitative data. Consequently, the main issue it has to deal with, regards data accountability (e.g., data quality, data comparability) and credibility. In our opinion, VGI can be seen as a complement of traditional SDI components, built upon the collective observations and expertise of citizens in their everyday life providing useful information on their surroundings. However, to gain credibility VGI needs to develop mechanisms to acquire, integrate, share that information while assuring its quality. This can be an interesting and promising tool to empower citizens informing about environmental phenomena occurring in their region.

Recommendation #5:
Enable new tools to communicate spatially-enabled data within social media (LinkedIn, Twitter)

Closely related to the last two recommendations, the emergence and fast growing of social media/networks like Facebook, LinkedIn, Twitter, Flickr or YouTube allows users to share information and rapidly proliferate content. Indeed, users can now very easily share any reports, images, maps or whatever information about their communities with their entire network. Therefore, it can be expected that this kind of medium will become increasingly important to communicate information on the environment and potentially also generate new data as exemplified by the geotagging capabilities of pictures published on Flickr (e.g., identifying pictures on a map) or mapping of tweets (e.g., Twitter messages) in order to spatially analyze this data and make sense of it (e.g., in crisis situation). This may potentially foster communication, sharing and collaboration around emerging problems or current situation/event on a specific location and influence the development of new widgets, mash-ups, etc... to consume this information.

Recommendation #6:
Assessing the value and impact of geographical information

Section 6.2 and all the previously mentioned recommendations highlight the need to assess the value and impact of geographical information. Indeed,

\textsuperscript{71} http://www.google.com/mapmaker

- 223 -
geospatial data are playing an increasing role for citizens, public and private organizations as well as governments. Therefore suitable frameworks for justifying and quantifying investments and benefits is of high importance in order to ease the commitment to SDIs methods, concepts and technologies.

**Recommendation #7:**
**Building capacities and promoting an open and sharing spirit**

We strongly believe that capacity building at human (education and training of individuals), infrastructure (installing/configuring/managing of the needed technology) and institutional (enhancing the understanding within organization and governments of the value of geospatial data to support decision-making) levels is a major element for large adoption, acceptance and commitment to SDI concepts inside and outside the GI community (e.g., climatologists, hydrologists, ecologists). In particular, showing and proving the benefits of sharing interoperable data/metadata through appropriate examples, best practices and guidelines will help to strengthen (1) existing observation systems, (2) capacities to decision-makers to use it, (3) capacities of the general public to understand important environmental, social and economical issues. Altogether this will help to reach agreement and endorsement on the use of new standards. Such a participative approach will certainly stimulate data providers to be more "open" and consequently to share their data. Probably the best way to reach this objective is to establish a long-term commitment to education and research, otherwise the SDI vision will remain unclear and unachievable. All these actions will help to reach endorsement on the use of such technologies, raising and increasing awareness on the benefits of using and sharing geospatial data, and finally creating new commitments.

**Recommendation #8:**
**Promote and engage stakeholders in initiatives such as GEO/GEOSS and INSPIRE.**

At the global and regional scales there are well known initiatives such as GEO/GEOSS or INSPIRE that need the support and engagement of different stakeholders of various communities involved. There is two major differences between these initiatives: (1) GEO/GEOSS is based on voluntary contributions while INSPIRE is a directive, and (2) GEO/GEOSS is primarily targeting the scientific community while INSPIRE targets environmental policies. However they both wish to enhance the relevance of earth observations and geospatial information to support decision-making processes. Consequently, these differences will noticeably influence the means of engaging stakeholders. INSPIRE is more restrictive in the sense that Member States do not have the choice and must be compliant with the legislation by 2020. Therefore compliance with the various Implementing Rules (e.g., data harmonization, metadata harmonization, view-download-transform services) will be of primary interest for stakeholders. In GEO/GEOSS, that is aiming to provide a framework where voluntary participants can develop new projects and coordinate their strategies and investements. Hence, providing access to good-reliable-accountable data, best practices documentation, tutorials (for data providers and data users), and sets of interoperability arrangements will help enabling stakeholders to participate and obtaining new commitments to GEO/GEOSS.
Finally, for both initiatives, capacity building, promoting/raising awareness and showing their related benefits are essential elements to strengthen engagement and support coordination (and not scattering) of efforts, investments and energies otherwise GEO/GEOSS and INSPIRE visions will remain unclear and unachievable.

**Recommendation #9: Developing thematic SDIs**

The development of the PREVIEW Global Risk Data Platform has demonstrated that addressing specific needs of a targeted community can help to facilitate the endorsement of SDI concepts and methods. Indeed, clarifying issues with a clearly defined SDI conceptual model can enhance collaboration and partnership among different participants and gives the possibility to agree on making data interoperable and potentially freely available. Focusing attention on the specific requirements of a community may help also to answer their specific needs and show them that this can have a positive influence on their own working flow. Users are often reluctant to changes and therefore the risk exists to strongly influence/modify the way they are working with data, and this may result in difficulties and impediments in reaching the objective of commitment. Instead, it is important to let them experience the benefits of these new possibilities of efficient access to shared and interoperable data and this may help addressing their needs. This may be a way to spread out the benefits of interoperability, enhance open and sharing spirit, and reach endorsement and commitment.

**Recommendation #10: Test the applicability of SDIs concepts and methods in different contexts**

SDIs concepts and methods have been developed in the context of high-income (i.e. developed) countries where access to computing infrastructures (providing access to broadband networks and computing resources like servers) is easy. Consequently, the risk is that SDIs is perceived as techno-centric and oriented to support information infrastructures of developed countries, that in turn may lead to increase the digital divide between developed and under-development countries. However, one of the main concerns of SDIs is to provide an environment that facilitates access to data and information and can act as one of the pillars of sound and informed decisions, supporting a more sustainable development all around the world. Consequently, by the distributed nature of data sources and with the support of (1) SOA principles and (2) geospatial standards, it is possible to support emerging countries and facilitate their access to decentralized resources. Nevertheless, the different socio-economic-politico-technological contexts may have a strong influence. For example in Africa, many countries do not have access to Internet through high bandwidth connections. Electricity supply may not be guaranteed, and often people are using old generation mobile phone to communicate. Then, if we want to support these countries in benefiting of an improved access to data and information to make sound decision we need to confront and test SDIs concepts, methods and technology to these realities. This will allow evidencing what are the issues and challenges to face in order to match their requirements and constraints.
Chapter 6: Conclusions & recommendations

Otherwise this will preclude widespread use and implementation of SDIs and they will remain a tool essentially used by developed countries.

Recommendation #11:
Continue to evaluate the distributed computing paradigm to support SDIs

We are convinced that distributed computing paradigm, especially Grids, can bring major benefits to support SDIs and we have extensively discussed this topic in Chapter 5 of this research. However, if we want to take full advantage of this technology several issues must be tackled: (1) test the mediation approach on different computing backends such as Clouds or clusters (allowing to control all the resources and building responsive/efficient processing interfaces), (2) develop plugins for various clients (such as ArcGIS and GRASS) giving access to grid-enabled WPS capabilities and allowing users to access seamlessly different resources depending on their needs (e.g., data retrieval, processing or map making), (3) evaluate the capabilities offered by Desktop Grids providing a simple access to unused resources of Desktop computer (CPU, storage and network), (4) build efficient workflow based on grid and non-grid services using orchestration engine, and (5) evaluate the potential of Grid in term of data management, because Grids can offer capabilities that are currently not (or only partially) provided by SDIs: distribution storage, data replication, data stored as close as possible to components that access them, security, and efficient data moving protocols (e.g., GridFTP). This last point requires making Grid middleware spatially-enabled and then implementing OGC standards and interfaces directly into them.

Recommendation #12:
Improve and extend interoperability experiments

We have seen that interoperability is an essential element to facilitate data discovery, access and integration. However, to be fully interoperable it is required to be syntactically and semantically interoperable. If syntactic interoperability is mostly achieved using OGC and ISO standards, semantic interoperability is currently not fully realized and has not shown its real added value. Therefore, we need to find suitable practical applications and benefits. Another issue concerns capabilities of models to exchange data with each other, and with other modeling tools on a time step basis as they run. This requires a standard interface like OpenMI 72 to facilitate the modeling of process interactions (coming from different suppliers, representing processes from different domains, based on different concepts, with different spatial and temporal resolutions and having different representations). We also discussed in details in Chapter 5 the issue of making different infrastructures (SDI and Grid) interoperable. This is a strong requirement in order to make these infrastructures mutually beneficiary from the capabilities offered.

Recommendation #13:
Move towards a real multi-disciplinary framework

Understanding the complexity (e.g., interactions, multi-dimensionality, continuously evolving, spatial and temporal scales) of Earth system requires

72 http://www.openmi.org/
gathering and integrating different data sets about physical, chemical and biological system. Consequently, it necessitates a collaborative and multidisciplinary effort to provide an integrated access to a wide range of services and resources on the environment. This obviously raises the challenge of a multistakeholder participation. The top-down approach emphasizes the need for standardization and uniformity, while the bottom-up stresses the importance of diversity and heterogeneity due to the different aspirations of the various stakeholders. In consequence, it is necessary to find a consensus to ensure some measure of standardization and uniformity while recognizing the diversity and the heterogeneity of the different stakeholders performing different tasks at different levels.

Therefore, a suitable approach to tackle this issue can be the so-called System of Systems (SoS) exemplified by GEOSS or INSPIRE. These initiatives underpin a multi-disciplinary framework built on existing systems. This allows recognizing the heterogeneity of systems reflecting the diversity of stakeholders involved while specifying arrangements in order to federate these systems. Such framework provides interesting features: (1) each component can operate independently (e.g., in order to match their own objective) and can be connected to others component by agreeing and specifying interoperability arrangements. This provides flexibility (the overall framework will not fail if one or more components disappear), (2) increase the capacity to turn data into information by sharing resources, (3) provides a holistic approach, (4) supplement but not supplant existing systems, (5) enhance composability of resources, (6) avoid single point of failure, (7) based on SOA principles, and (8) can incorporate incrementally new components/systems. This approach is also interesting as it lowers the entry-level barrier, which translates into implementing only a few standards to achieve at least syntactic interoperability. However, many challenges on interoperability can be highlighted like semantic, different interoperability levels, scalability, extensibility, evolution, flexibility and diversity of standards. The latter is very interesting due to the fact that in the GI community there are hundreds of standards making it impossible for clients to implement all of them. Hence we need solutions that may act as gateways, exposing to clients only a few common and widely accepted standards. This brokering approach (proposed for example by EuroGEOSS\textsuperscript{73}) appears promising and need further investigations in order to test this applicability in a wider context. In our vision, SoS framework and brokering approach can ease commitment and facilitate endorsement and acceptance on interoperability allowing interconnecting different systems and providing a real multidisciplinary framework.

6.4 Concluding remarks

To conclude this research, we are convinced that SDIs and related concepts, methods and technologies are suitable and can bring major benefits to support and facilitate environmental data discovery, accessibility, visualization, dissemination and analysis. On a technical level, all the building blocks are available, supported by OGC and ISO standards, allowing data providers to start

\textsuperscript{73}http://www.eurogeoss.eu/broker/default.aspx
sharing and disseminating their data and metadata in an interoperable way. Our work has shown that this is feasible, it is not difficult to develop, software implementation are reliable and facilitating access, integration and use of geospatial data can answer the requirements of a specific community, as well as making these data available to the widest possible audience. SDIs have the potential to support the exchange and management of environmental data and information like what it is done since a long time in airport (e.g., management of passengers, flight and luggage) or banking (e.g., financial market) activities. Indeed, such private activities rely on a few, commonly accepted standards and protocols implemented in a few, widely used software. This contrast with the variety of standards, protocols, software implementations, data types and data formats that are available within the geospatial community. Despite the fact that this situation is probably caused by the diversity and heterogeneity of stakeholders involved in geospatial data creation and management, it would be good to rationalize and coordinate efforts (like in the OGC approach) for the development of standards, protocols and implementations.

An important lesson learnt is that we do not need to convince people. Instead, we need to keep all these concepts simple, make them understandable, and let people experience by themselves the benefits of working in an interoperable context and the potential positive influence it can have on their daily business. This will ease commitment to use standards, to endorse SOA and SoS principles, and to support/engage in initiatives like GEOSS and INSPIRE.

In our vision, it is time to make all these components operational otherwise SDI will remain only an innovative concept. Let the dream comes true because we cannot wait! The challenges that humankind is facing require acting now, and therefore we need to provide decision-makers with tools that allow them accessing rapidly and efficiently good and reliable environmental information. SDIs have clearly the potential to be a part of the answer to bridge the gap between science and policy-making. It is obvious that to achieve this objective in the shortest term possible it will mostly depend on political, social, and economical constraints. We have shown in this research that the human/social component is probably the most influencing one, and discussing and reaching agreements are important. Nevertheless, it must not be an excuse not to start sharing data and metadata. Scattering efforts and energies in discussions that sometimes concern only details can block entire process, resulting in lost of (precious) time and motivation, and finally leading to the risk of disappointment and disengagement. Therefore, finding the right balance between actions/decisions and discussions is an essential factor to address.

The System of Systems approach appears to be a valuable and promising concept in order to lower the entry-level barrier to a real multi-disciplinary framework where each component can perform independently their own tasks, while ensuring sufficient flexibility in connecting other components by agreeing and specifying interoperability arrangements. One of the major benefits of the SoS approach is to allow users to perform functions that cannot be made with any single component. This means that such a system is more than the sum of its parts, and offers the possibility to better understand the complex relationships between the different components of the Earth system. Consequently, such a framework can offer possibilities for SDIs to extend, complement and benefit from capabilities offered by other type of infrastructures. In this regard,
distributed computing infrastructures and especially Grids can be really beneficial/advantageous for data processing and data management of ever-increasing amounts of high-resolution data.

Finally, as mentioned by Craig (2005) and Arzberger et al. (2004), it is important to keep in mind that data sharing and related SDI developments rely mostly on individuals that should have in common: (1) a sense that better data will lead to better decisions, (2) a sharing spirit that they got something in return and are viewed as collaborative partners, and (3) the fact that they are involved in a professional culture that honours serving society and cooperating with others. Hence, ensuring that data are easily accessible so that they can be used as often and widely as possible is a matter of sound stewardship of public resources. These authors stated that publicly funded data are a public good, produced in the public interest and thus should be freely available to the maximum extent possible. In a general scientific context, sharing and documenting data is part of the elementary scientific approach, allowing scientists to compare their results and methods more easily, and then enhancing scientific accountability, credibility and potentially improve quality of data for the benefit of everyone.

**WITHOUT SHARING DATA: DOING SCIENCE IS DIFFICULT, TAKING SOUND DECISIONS CAN BE PROBLEMATIC, AND ENVISIONING A SUSTAINABLE DEVELOPMENT CAN BE COMPLICATED.**
References


References


References


References


Luis, B. (2009). Web feature service (WFS) and sensor observation service (SOS) comparison to publish time series data.


References

References


References


Simonis, I. and A. Sliwinski (2005). Quality of Service in a Global SDI. Proceedings of the 8th International Conference for Global Spatial Data Infrastructure, Cairo (Egypt).


References

References


Annexes
Annexes

A.1 INSPIRE themes

Annex 1

1. Coordinate reference systems
Systems for uniquely referencing spatial information in space as a set of coordinates (x, y, z) and/or latitude and longitude and height, based on a geodetic horizontal and vertical datum.

2. Geographical grid systems
Harmonised multi-resolution grid with a common point of origin and standardised location and size of grid cells.

3. Geographical names
Names of areas, regions, localities, cities, suburbs, towns or settlements, or any geographical or topographical feature of public or historical interest.

4. Administrative units
Units of administration, dividing areas where Member States have and/or exercise jurisdictional rights, for local, regional and national governance, separated by administrative boundaries.

5. Addresses
Location of properties based on address identifiers, usually by road name, house number, postal code.

6. Cadastral parcels
Areas defined by cadastral registers or equivalent.

7. Transport networks
Road, rail, air and water transport networks and related infrastructure. Includes links between different networks.

8. Hydrography
Hydrographic elements, including marine areas and all other water bodies and items related to them, including river basins and sub-basins.

9. Protected sites
Area designated or managed within a framework of international, Community and Member States' legislation to achieve specific conservation objectives.

Annex 2

1. Elevation
Digital elevation models for land, ice and ocean surface. Includes terrestrial elevation, bathymetry and shoreline.

2. Land cover
Physical and biological cover of the earth's surface including artificial surfaces, agricultural areas, forests, (semi-)natural areas, wetlands, water bodies.
Annexes

3. Orthoimagery
Geo-referenced image data of the Earth’s surface, from either satellite or airborne sensors.

4. Geology
Geology characterized according to composition and structure. Includes bedrock, aquifers and geomorphology.

Annex 3
1. Statistical units
Units for dissemination or use of statistical information.

2. Buildings
Geographical location of buildings.

3. Soil
Soils and subsoil characterized according to depth, texture, structure and content of particles and organic material, stoniness, erosion, where appropriate mean slope and anticipated water storage capacity.

4. Land use
Territory characterized according to its current and future planned functional dimension or socio-economic purpose (e.g., residential, industrial, commercial, agricultural, forestry, recreational).

5. Human health and safety
Geographical distribution of dominance of pathologies (allergies, cancers, respiratory diseases, etc.), information indicating the effect on health (biomarkers, decline of fertility, epidemics) or well-being of humans (fatigue, stress, etc.) linked directly (air pollution, chemicals, depletion of the ozone layer, noise, etc.) or indirectly (food, genetically modified organisms, etc.) to the quality of the environment.

6. Utility and governmental services
Includes utility facilities such as sewage, waste management, energy supply and water supply, administrative and social governmental services such as public administrations, civil protection sites, schools and hospitals.

7. Environmental monitoring facilities
Location and operation of environmental monitoring facilities includes observation and measurement of emissions, of the state of environmental media and of other ecosystem parameters (biodiversity, ecological conditions of vegetation, etc.) by or on behalf of public authorities.

8. Production and industrial facilities
Annexes

9. Agricultural and aquaculture facilities
Farming equipment and production facilities (including irrigation systems, greenhouses and stables).

10. Population distribution — demography
Geographical distribution of people, including population characteristics and activity levels, aggregated by grid, region, administrative unit or other analytical unit.

11. Area management/restriction/regulation zones and reporting units
Areas managed, regulated or used for reporting at international, European, national, regional and local levels. Includes dumping sites, restricted areas around drinking water sources, nitrate-vulnerable zones, regulated fair ways at sea or large inland waters, areas for the dumping of waste, noise restriction zones, prospecting and mining permit areas, river basin districts, relevant reporting units and coastal zone management areas.

12. Natural risk zones
Vulnerable areas characterized according to natural hazards (all atmospheric, hydrologic, seismic, volcanic and wildfire phenomena that, because of their location, severity, and frequency, have the potential to seriously affect society), e.g., floods, landslides and subsidence, avalanches, forest fires, earthquakes, volcanic eruptions.

13. Atmospheric conditions
Physical conditions in the atmosphere. Includes spatial data based on measurements, on models or on a combination thereof and includes measurement locations.

14. Meteorological geographical features
Weather conditions and their measurements; precipitation, temperature, evapotranspiration, wind speed and direction.

15. Oceanographic geographical features
Physical conditions of oceans (currents, salinity, wave heights, etc.).

16. Sea regions
Physical conditions of seas and saline water bodies divided into regions and sub-regions with common characteristics.

17. Bio-geographical regions
Areas of relatively homogeneous ecological conditions with common characteristics.

18. Habitats and biotopes
Geographical areas characterized by specific ecological conditions, processes, structure, and (life support) functions that physically support the organisms that live there. Includes terrestrial and aquatic areas distinguished by geographical, abiotic and biotic features, whether entirely natural or semi-natural.
Annexes

19. *Species distribution*
Geographical distribution of occurrence of animal and plant species aggregated by grid, region, administrative unit or other analytical unit.

20. *Energy resources*
Energy resources including hydrocarbons, hydropower, bio-energy, solar, wind, etc., where relevant including depth/height information on the extent of the resource.

21. *Mineral resources*
Mineral resources including metal ores, industrial minerals, etc., where relevant including depth/height information on the extent of the resource.
"Bringing GEOSS services into practice" aims at teaching participants how to install, configure and deploy a set of open source software to publish and share data and metadata through the Global Earth Observation System of Systems (GEOSS) using OGC web services & ISO standards.

GEOSS has been created as an international voluntary effort that connects geospatial and Earth Observation and information infrastructures, acting as a gateway between producers of environmental data and end users. The aim of GEOSS is to enhance the relevance of Earth observations and to offer public access to comprehensive, near-real time data, information and analyses of the environment.

This workshop is held within the framework of the enviroGRIDS project that addresses the environmental problems surrounding the Black Sea Catchment. This area is known as one of ecologically unsustainable development, where inadequate resource management has led to severe environmental, social and economic problems. enviroGRIDS aims at improving this by building capacities for sharing and exchanging environmental data through GEOSS.

All the teaching material (tutorial, software, data) is available on the enviroGRIDS website:
http://www.envirogrids.net/index.php?option=com_content&view=article&id=67&Itemid=80

and also in the GEOSS registry:
https://geossregistries.info:443/geosspub/component_details_ns.jsp?compld=urn:uuid:b96c48ad-1e85-49df-bbaa-d1bf04c0bbaf

and finally on the OGC website:
http://www.ogcnetwork.net/tutorials

This workshop has been already delivered in Romania (May 2010), Georgia (November 2010), the Netherlands (April 2011) and will be replicated in each country around the Black Sea within the enviroGRIDS project timeframe.

Next workshop: Istanbul – Turkey, 23rd of September 2011.
A.3 With or without SDI

The table presented below is a simple, optimistic and general comparison between what the world can be if diffusion and implementation of SDI methods, concepts and technologies are successful or not.

<table>
<thead>
<tr>
<th>Data is:</th>
<th>Without SDIs</th>
<th>With SDIs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scattered</td>
<td>Distributed</td>
<td></td>
</tr>
<tr>
<td>Disconnected</td>
<td>Integrated</td>
<td></td>
</tr>
<tr>
<td>Unused</td>
<td>Usable</td>
<td></td>
</tr>
<tr>
<td>Not validated</td>
<td>Validated through use</td>
<td></td>
</tr>
<tr>
<td>Difficult to find</td>
<td>Quickly accessible</td>
<td></td>
</tr>
<tr>
<td>Not documented</td>
<td>Documented</td>
<td></td>
</tr>
<tr>
<td>Expensive to create</td>
<td>Easier to justify</td>
<td></td>
</tr>
<tr>
<td>Boring stats</td>
<td>Exciting visualization tools</td>
<td></td>
</tr>
<tr>
<td>Not georeferenced</td>
<td>Spatially explicit</td>
<td></td>
</tr>
<tr>
<td>Targeting experts</td>
<td>For everyone</td>
<td></td>
</tr>
<tr>
<td>Data gap</td>
<td>Informed decisions</td>
<td></td>
</tr>
<tr>
<td>Secretocracy</td>
<td>Democracy</td>
<td></td>
</tr>
<tr>
<td>Hidden</td>
<td>Searchable</td>
<td></td>
</tr>
<tr>
<td>Long and difficult processing</td>
<td>Fast and easy processing</td>
<td></td>
</tr>
<tr>
<td>From labs, fieldworks</td>
<td>From sensors, remote sensing</td>
<td></td>
</tr>
<tr>
<td>Low resolution, small extent</td>
<td>Large/high resolution data sets</td>
<td></td>
</tr>
</tbody>
</table>

Table 12: With or without SDIs diffusion and implementation.

This change can happen only supported by various drivers that can push people to share their data:

- Environmental crisis
- Need for faster & better decisions
- Inform the public with trustable information
- Save money and time
- Publicly funded data should be publicly available
- Accessible through modern and portable devices
- Need for spatially explicit information and scenarios
- Others....
Lists of figures, tables & websites
Lists of figures, tables & websites

Figures

Figure 1 (38): GIS and economy (Source: Geoconnections).
Figure 2 (40): EnviroGRIDS vision of improved data access and geoprocessing
Figure 3 (42): Nature and relations between SDI components (Source: GISCafe).
Figure 4 (43): SDI hierarchy. (Source: GISCafe)
Figure 5 (45): Basic operations in SOA (Source: IBM).
Figure 6 (54): Data and information flow within the INSPIRE framework (Source: INSPIRE).
Figure 7 (55): INSPIRE network architecture (Source: INSPIRE).
Figure 8 (62): Geospatial Portal Reference Architecture (Source: GeoNetwork).
Figure 9 (63): GeoNetwork, a catalogue system using CSW.
Figure 10 (65): Example of the XML file returned after a GetCapabilities request
Figure 11 (66): Returned image after a WMS request.
Figure 12 (67): Result of a GetFeatureInfo query
Figure 13 (69): Result of a DescribeCoverage request.
Figure 14 (74): Example of GML returned after a WFS request
Figure 15 (94): Example of the XML file returned after a GetCapabilities request
Figure 16 (96): EnviroGRIDS grid-enabled SDI components supporting Black Sea Portal
Figure 17 (111): PREVIEW SDI conceptual model (adapted from Mansourian et al., 2006; Rajabifard, 2002)
Figure 18 (117): PREVIEW Global Risk Data Platform data and metadata flow
Figure 19 (119): Cyclones events from 2001 to 2008 in Asia as seen in the PREVIEW Global Risk Data Platform
Figure 20 (134): Architecture used for testing different OWS implementations
Figure 21 (136): WMS results under various conditions (OWS implementation and file formats)
Figure 22 (138): WFS ArcGIS Server tests results
Figure 23 (139): WFS GeoServer tests results
Figure 24 (140): WCS ArcGIS Server tests results
Figure 25 (141): WCS GeoServer tests results
Figure 26 (167): EnviroGRIDS gSDI components supporting web portal
Figure 27 (168): EnviroGRIDS functional layers
Figure 28 (169): Scenarios to implement SDI/Grid interoperability in the enviroGRIDS project
Figure 29 (177): Communication pattern between a WPS-client and a WPS-instance
Figure 30 (182): Mediation layer (adapted from Hérault et al. 2005)
Figure 31 (185): Technical architecture of the proposed WPS mediation. The WPS mediation layer receives and handles the Execute query, evaluates parameters, and according to them is
Lists of figures, tables & websites

running the geoprocessing task on the local server or on a selected distributed computing infrastructure. The first option “backend=LCG” will send a list of files to be processed in parallel, while the second option “backend=LCG;jobs=50” allows tiling a large data set according to the number of desired calculation jobs and then submit jobs to the selected backend.

Figure 32 (187): WPS mediation layer
Figure 33 (190): Jobs monitoring with Ganga (in command line)

Tables

Table 1 (47): Integration issues (Williamson et al., 2006)
Table 2 (49): Different types of interoperability.
Table 3 (92): Technical comparison of enviroGRIDS, ACQWA, GEO Data Portal and PREVIEW projects
Table 4 (132): Summary of the WFS testing scenarios. Case #1 is the “base case”.
Table 5 (133): Summary of the WCS testing scenarios. Case #1 is the “base case”.
Table 7 (137): WFS ArcGIS Server tests results. Refer to table 3 for acronyms
Table 8 (138): WFS GeoServer tests results. Refer to table 3 for acronyms
Table 9 (140): WCS GeoServer (GS) and ArcGIS Server (AGS) tests results. An ArcGIS Server Image Service (IS) has also been tested. Tests have been executed on Linux and Windows (win) operating systems. Lr: Low-resolution, Mr: Medium-resolution, Hr: High-resolution.
Table 10 (180): Encapsulation vs. Integration approaches
Table 11 (194): Indicative comparison of mediated WPS processes executed locally or on the Grid (times are expressed in minutes/seconds)
Table 12 (250): With or without SDIs diffusion and implemention.

Websites

ACQWA: http://www.acqwa.ch
Amazon EC2: http://aws.amazon.com/ec2
AMGA: http://amga.web.cern.ch/amga/
ArcGIS: http://www.esri.com/software/arcgis
ArcGIS Server: http://www.esri.com/software/arcgis/arcgisserver
Blue Marble: http://earthobservatory.nasa.gov/Features/BlueMarble/
CSW: http://www.opengeospatial.org/standards/cat
Lists of figures, tables & websites

Deegree: http://www.deegree.org/
DesInventar: http://www.desinventar.org
EGEE: http://www.eu-egee.org
EnviroGRIDS: http://www.envirogrids.net
File GeoDB: http://resources.arcgis.com/content/geodatabases/10.0/file-gdb-api
FOSS4G: http://wiki.osgeo.org/wiki/FOSS4G_Benchmark
Ganga: http://ganga.web.cern.ch/ganga/
GAR: http://www.preventionweb.net/english/hyogo/gar/
GDACS: http://www.gdacs.org/
GDAL: http://www.gdal.org
GeoCat: http://www.geocat.ch
GEO Data Portal: http://geodata.grid.unep.ch
GEOS: http://www.earthobservations.org
GeoNetwork: http://geonetwork-opensource.org/
GeoRSS: http://georss.org/
GeoServer: http://www.geoserver.org
GIGAS: http://www.thegigasforum.eu
GMES: http://www.gmes.info/
GML: http://www.opengeospatial.org/standards/gml
Globe: http://glite.web.cern.ch/glite
Globus: http://globus.org/toolkit
GLWD: http://www.worldwildlife.org/science/data/item1877.html
Google App Engine: http://code.google.com/appengine
Google Maps: http://code.google.com/apis/maps/
GRASS: http://grass.osgeo.org/
Grids: http://resources.arcgis.com/content/kbase?fa=articleShow&d=30616
GRIP: http://www.gripweb.org/
G-OWS: https://www.g-ows.org/
GSDI: http://www.gsdi.org
HotSpots: http://www.ldeo.columbia.edu/chrr/research/hotspots/
INSPIRE: http://inspire.jrc.ec.europa.eu/
ISO: http://www.iso.org
Jetty: http://jetty.codehaus.org/jetty/
JMeter: http://jakarta.apache.org/jmeter
KML: http://www.opengeospatial.org/standards/kml
MapFish: http://www.mapfish.org
MapServer: http://www.mapserver.org
MODIS: http://modis.gsfc.nasa.gov
Lists of figures, tables & websites

MS4W: http://www.maptools.org/ms4w/
MyProxy: http://grid.ncsa.illinois.edu/myproxy/
NGDC: http://www.ngdc.noaa.gov/hazard/
OASIS: http://www.oasis-open.org/
OGC: http://www.opengeospatial.org
OpenLayers: http://openlayers.org/
PHP: http://www.php.net
PostgreSQL: http://www.postgresql.org/
PostGIS: http://postgis.refractions.net/
PreventionWeb: http://www.preventionweb.net/
PREVIEW: http://preview.grid.unep.ch
PyWPS: http://pywps.wald.intevation.org/
ReliefWeb: http://www.reliefweb.int/
SEE-GRID: http://www.see-grid-sci.eu/
SITG: http://www.sitg.ch
SOS: http://www.opengeospatial.org/standards/sos
Unicore: http://www.unicore.eu
UNOSAT: http://unosat.web.cern.ch/unosat/
UNSDI: http://www.ungiwg.org/unsdi.htm
USGS: https://geohazards.cr.usgs.gov/
W3C: http://www.w3c.org
WCS: http://www.opengeospatial.org/standards/wcs
WFS: http://www.opengeospatial.org/standards/wfs
WMS: http://www.opengeospatial.org/standards/wms
WPS: http://www.opengeospatial.org/standards/wps
XtremWeb: http://www.xtremwebch.net
52 north: http://52north.org/maven/project-sites/wps/52n-wps-site/
Abbreviations & acronyms
### Abbreviations & acronyms

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACQWA</td>
<td>Assessing Climate impacts on the Quantity and quality of WAter</td>
</tr>
<tr>
<td>AMGA</td>
<td>ARDA Metadata Grid Application</td>
</tr>
<tr>
<td>BCPR</td>
<td>Bureau for Crisis Prevention and Recovery</td>
</tr>
<tr>
<td>BMP</td>
<td>BitMaP</td>
</tr>
<tr>
<td>BPEL</td>
<td>Business Process Execution Language</td>
</tr>
<tr>
<td>CERN</td>
<td>Centre Européen pour la Recherche Nucléaire</td>
</tr>
<tr>
<td>CSW</td>
<td>Catalog Service for the Web</td>
</tr>
<tr>
<td>DBMS</td>
<td>DataBase Management System</td>
</tr>
<tr>
<td>DEM</td>
<td>Digital Elevation Model</td>
</tr>
<tr>
<td>DIANE</td>
<td>Distributed Analysis Environment</td>
</tr>
<tr>
<td>DRR</td>
<td>Disaster Risk Reduction</td>
</tr>
<tr>
<td>EC</td>
<td>European Commission</td>
</tr>
<tr>
<td>EC2</td>
<td>Amazon Elastic Compute Cloud</td>
</tr>
<tr>
<td>ECW</td>
<td>ERMapper Compress Wavelets</td>
</tr>
<tr>
<td>EGEE</td>
<td>Enabling Grids for E-scienceE</td>
</tr>
<tr>
<td>EFAS</td>
<td>European Flood Alert System</td>
</tr>
<tr>
<td>EO</td>
<td>Earth Observation</td>
</tr>
<tr>
<td>ESA</td>
<td>European Space Agency</td>
</tr>
<tr>
<td>ESIP</td>
<td>Environment oriented Satellite Data Processing Platform</td>
</tr>
<tr>
<td>FAO</td>
<td>Food and Agriculture Organization</td>
</tr>
<tr>
<td>FOEN</td>
<td>Federal Office for the ENvironment</td>
</tr>
<tr>
<td>FOSS</td>
<td>Free and Open Source Software</td>
</tr>
<tr>
<td>FOSS4G</td>
<td>Free and Open Source Software for Geospatial</td>
</tr>
<tr>
<td>FTP</td>
<td>File Transfer Protocol</td>
</tr>
<tr>
<td>GCI</td>
<td>GEOSS Common Infrastructure</td>
</tr>
<tr>
<td>GAR</td>
<td>Global Assessment Report on disaster risk reduction</td>
</tr>
<tr>
<td>GEO</td>
<td>Group on Earth Observations</td>
</tr>
<tr>
<td>GEO</td>
<td>Global Environment Outlook</td>
</tr>
<tr>
<td>GEOSS</td>
<td>Global Earth Observation System of Systems</td>
</tr>
<tr>
<td>GEORSS</td>
<td>GEO Really Simple Syndication</td>
</tr>
<tr>
<td>GDACS</td>
<td>Global Disaster Alert and Coordination System</td>
</tr>
<tr>
<td>GDAL</td>
<td>Geospatial Data Abstraction Library</td>
</tr>
<tr>
<td>GDP</td>
<td>Gross Domestic Product</td>
</tr>
<tr>
<td>GI</td>
<td>Geographic Information</td>
</tr>
<tr>
<td>GIF</td>
<td>Graphics Interchange Format</td>
</tr>
<tr>
<td>GIGAS</td>
<td>GEOSS, INSPIRE, and GMES an Action In Support</td>
</tr>
<tr>
<td>GIS</td>
<td>Geographic Information System</td>
</tr>
<tr>
<td>GMES</td>
<td>Global Monitoring for Environment and Security</td>
</tr>
<tr>
<td>GML</td>
<td>Geographic Markup Language</td>
</tr>
<tr>
<td>G-OWS</td>
<td>gLite-OWS</td>
</tr>
<tr>
<td>GPL</td>
<td>General Public License</td>
</tr>
<tr>
<td>GPS</td>
<td>Global Positioning System</td>
</tr>
<tr>
<td>GRASS</td>
<td>Geographic Resources Analysis Support System</td>
</tr>
<tr>
<td>GRID</td>
<td>Global Resource Information Database</td>
</tr>
<tr>
<td>GSDI</td>
<td>Global Spatial Data Infrastructure</td>
</tr>
<tr>
<td>GUID</td>
<td>Grid Unique IDentifier</td>
</tr>
<tr>
<td>HTTP</td>
<td>HyperText Transfer Protocol</td>
</tr>
<tr>
<td>IDNDR</td>
<td>International Decade for Disaster Risk Reduction</td>
</tr>
<tr>
<td>IMS</td>
<td>Internet Map Server</td>
</tr>
</tbody>
</table>
Abbreviations & acronyms

INSPIRE: Infrastructure for Spatial Information in the European Community
ISO: International Organization for Standardization
IR: Implementing Rules
IT: Information Technology
JPEG: Joint Photographic Experts Group
JSDL: Job Submission Description Language
JSON: JavaScript Object Notation
KML: Keyhole Markup Language
KVP: Key-Value Pair
LCG: LHC Computing Grid
LFN: Logical File Name
MDG: Millennium Development Goal
MODIS: MODerate resolution Imaging Spectroradiometer
NDVI: Normalized Difference Vegetation Index
NGDC: National Geophysical Data Center
NGI: Norwegian Geotechnical Institute
NIR: Near InfraRed
OAI: Open Archive Initiative
OASIS: Organization for the Advancement of Structured Information Standards
OCHA: UN Office Coordination for Humanitarian Affairs
ODE: Apache Orchestration Director Engine
OGC: Open Geospatial Consortium
OGF: Open Grid Forum
OGSA: Open Grid Services Architecture
OGSA-DAI: OGSA Data Access and Integration
OGSA-DQP: OGSA Distributed Query Processor
OSS: Open Source Software
OWS: OGC Web Services
P2P: Peer-to-Peer
PDF: Portable Document Format
PNG: Portable Network Graphics
PREVIEW: Project on Risk Evaluation, Vulnerability, Information and Early Warning
PTB: Persistent TestBed
QoS: Quality of Service
RDBMS: Relational DataBase Management System
REST: Representational State Transfer
RPC: Remote Procedure Call
RSS: Really Simple Syndication
S3: Amazon Simple Storage Service
SBA: Societal Benefit Area
SciFlo: Scientific Dataflow
SDE: Spatial Data Engine
SDI: Spatial Data Infrastructure
SDK: Software Development Kit
SEE-GRID: South East Europe GRID
SHP: Shape File
SGE: Sun Grid Engine
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SITG</td>
<td>Système d’Information du Territoire Genevois</td>
</tr>
<tr>
<td>SLD</td>
<td>Styled Layer Descriptor</td>
</tr>
<tr>
<td>SOA</td>
<td>Service Oriented Architecture</td>
</tr>
<tr>
<td>SOAP</td>
<td>Simple Object Access Protocol</td>
</tr>
<tr>
<td>SOS</td>
<td>Sensor Observation Service</td>
</tr>
<tr>
<td>SQL</td>
<td>Structured Query Language</td>
</tr>
<tr>
<td>SRTM</td>
<td>Shuttle Radar Topography Mission</td>
</tr>
<tr>
<td>SVG</td>
<td>Scalable Vector Graphics</td>
</tr>
<tr>
<td>SWAT</td>
<td>Soil and Water Assessment Tool</td>
</tr>
<tr>
<td>SWE</td>
<td>Sensor Web Enablement</td>
</tr>
<tr>
<td>TIFF</td>
<td>Tagged Image File Format</td>
</tr>
<tr>
<td>UN</td>
<td>United Nations</td>
</tr>
<tr>
<td>UNEP</td>
<td>United Nations Environment Programme</td>
</tr>
<tr>
<td>UNECA</td>
<td>United Nations Economic Commission for Africa</td>
</tr>
<tr>
<td>UNESCO</td>
<td>United Nations Educational, Scientific and Cultural Organization</td>
</tr>
<tr>
<td>UNGIWG</td>
<td>United Nations Geographical Information Working Group</td>
</tr>
<tr>
<td>UNHCR</td>
<td>United Nations High Commissioner for Refugees</td>
</tr>
<tr>
<td>UNISDR</td>
<td>United Nations International Strategy for Disaster Reduction</td>
</tr>
<tr>
<td>UNSDI</td>
<td>United Nations Spatial Data Infrastructure</td>
</tr>
<tr>
<td>URM</td>
<td>Uniform Resource Management</td>
</tr>
<tr>
<td>URL</td>
<td>Uniform Resource Locator</td>
</tr>
<tr>
<td>USGS</td>
<td>United Stated Geological Survey</td>
</tr>
<tr>
<td>VGI</td>
<td>Volunteered Geographic Information</td>
</tr>
<tr>
<td>VO</td>
<td>Virtual Organization</td>
</tr>
<tr>
<td>VOMS</td>
<td>Virtual Organization Membership Service</td>
</tr>
<tr>
<td>WAS</td>
<td>Web Service Authentication</td>
</tr>
<tr>
<td>W3C</td>
<td>World Wide Web consortium</td>
</tr>
<tr>
<td>WMS</td>
<td>Web Map Service</td>
</tr>
<tr>
<td>WFS</td>
<td>Web Feature Service</td>
</tr>
<tr>
<td>WCS</td>
<td>Web Coverage Service</td>
</tr>
<tr>
<td>WHO</td>
<td>World Health Organization</td>
</tr>
<tr>
<td>WMO</td>
<td>World Meteorological Organization</td>
</tr>
<tr>
<td>WSDL</td>
<td>Web Service Description Language</td>
</tr>
<tr>
<td>WSS</td>
<td>Web Security Service</td>
</tr>
<tr>
<td>WSRF</td>
<td>Web Service Resource Framework</td>
</tr>
<tr>
<td>WPS</td>
<td>Web Processing Service</td>
</tr>
<tr>
<td>WPVS</td>
<td>Web Perspective and View Service</td>
</tr>
<tr>
<td>WTS</td>
<td>Web Terrain Service</td>
</tr>
<tr>
<td>XML</td>
<td>eXtended Markup Language</td>
</tr>
</tbody>
</table>