Integration of Hyperbooks into the Semantic Web

FALQUET, Gilles, NERIMA, Luka, ZISWILER, Jean-Claude

Abstract

A crucial aspect of the Semantic Web is the capacity to add formalized meanings to information to enable non-human actors to process it. This is usually accomplished by linking the information to an ontology that describes the domain's concepts. In the Web's context it does not seem realistic to represent this semantic layer on a central server, as this model would not reflect the characteristics of presently used, non-centralized networks like the current Web. Therefore we are confronted with a huge number of locally developed and stored ontologies, and we need some kind of integration techniques to connect ontologies developed for the Semantic Web. In this article, we describe our experience with ontology-based e-learning systems and we propose a mechanism to integrate such systems into a Semantic Web context. We concretely present our hyperbook model and show how hyperbooks can be integrated into digital libraries by an ontology mapping procedure.

Reference


Available at:
http://archive-ouverte.unige.ch/unige:46377

Disclaimer: layout of this document may differ from the published version.
Integration of Hyperbooks into the Semantic Web

Gilles Falquet, Luka Nerima and Jean-Claude Ziswiler
Centre universitaire d'informatique CUI – University of Geneva
24, rue du Général Dufour – 1211 Geneva 4 – Switzerland
{Jean-Claude.Ziswiler, Gilles.Falquet, Luka.Nerima} @cui.unige.ch

Abstract
A crucial aspect of the Semantic Web is the capacity to add formalized meanings to information to enable non-human actors to process it. This is usually accomplished by linking the information to an ontology that describes the domain's concepts. In the Web's context it does not seem realistic to represent this semantic layer on a central server, as this model would not reflect the characteristics of presently used, non-centralized networks like the current Web. Therefore we are confronted with a huge number of locally developed and stored ontologies, and we need some kind of integration techniques to connect ontologies developed for the Semantic Web. In this article, we describe our experience with ontology-based e-learning systems and we propose a mechanism to integrate such systems into a Semantic Web context. We concretely present our hyperbook model and show how hyperbooks can be integrated into digital libraries by an ontology mapping procedure.

1 Introduction
Most of the digital libraries available on the Web are collections of documents stored on a document server where users can for instance use full-text search engines. Results are then presented in title lists with annotated, small informational fragments of the (perhaps) most relevant documents. By clicking on these references, users can see the whole text or information, which is stored in the form of a PDF-file or in a similar format. Such systems are simple to build, but have their drawbacks. From the semantic point of view, they are in fact static hypertext systems. There is no semantic representation of the content, and different systems and servers are not interrelated.

Even if most Web sites have a dynamic part, it is normally limited to information which is semantically poor and which can be easily stored in a common database management system. For instance, most e-commerce shops have such functionalities. A virtual bookstore has information about titles, authors, page-numbers and prices, but the content of the books is often not formally represented. Systems including formal concept representation give for instance the possibility of a non-linear reading or the presentation of narratives in multiple forms. Therefore, we propose an approach that allows fragmenting text sources and linking these fragments to domain ontologies. This is the main idea of what we call a hyperbook.

Such a dynamic hyperbook model results in a local, non-centralized and heterogeneous system. Heterogeneity is one of the characteristics of the Web and we anticipate it to be a permanent one. We even believe that one of the major keys to the success of the Internet is the heterogeneity of the systems where all users may create the pages in their own way. When building semantic web applications, we should not restrict authors when they build the content. We furthermore should try to find solutions that involve a great flexibility. But to avoid the risk that users create isolated systems, we propose to assemble many hyperbooks into a digital library. The main task of this process is the integration of different domain ontologies.

The ontology integration problem is one of the most challenging tasks in the Semantic Web. The integration procedure for digital libraries we present uses specific properties of hyperbooks or similar e-learning systems like fragments, documents and personalization elements. But we can imagine that the approach could be extended to other domains, to the extent that there is a good information base or a document repository available.

The paper is organized as follows: Section 2 describes the existing research in the domain of hyperbooks and presents our model. Section 3 includes the step from hyperbooks to a digital library, and section 4 concludes the article with an outlook about future research questions.

2 Hyperbooks
There is presently no consensus on a common virtual document or hyperbook model. Nevertheless, most of the proposed models are comprised of (at least) a domain ontology and a fragment base. These models generally differ on the user interface part, i.e. how to specify the production of user-readable documents (with declarative languages, through specific ontologies, with inference rules, etc.). Crampes and Ranwez [Crampes and Ranwez, 2000] propose two models of virtual documents. Both of them use domain ontologies for indexing informational fragments (the re-
sources). In the first case, a “conceptual backward chaining” strategy can construct reading paths corresponding to the user objectives (described in terms of conceptual graphs). In the second case, a pedagogical ontology defines teaching rules, which guide the assembling of fragments to produce documents with respect to a predefined pedagogical approach. These rules, in particular, force the order of appearance of information in the documents. An inference engine generates documents that satisfy these rules.

The InterBook adaptive hyperbook project [Brusilovsky et al., 1998a] is based on two models: the domain model and the student model. The domain model is represented by a network whose nodes correspond to domain concepts and links to their relationships. In fact, the domain model corresponds to what is named ontology in many research papers. The student model describes the student knowledge as well as the student learning goals, both expressed in terms of the concepts of the domain model. A glossary is used to describe the navigational paths in an InterBook hyperbook. Several books can be integrated in a bookshelf. They are connected by sharing the same set of domain concepts, thus avoiding the ontology integration problem.

Iksal and Garlatti [Iksal et al., 2001] propose a comprehensive and detailed model of virtual documents. It is based on four ontologies for modeling the domain, the metadata, the applications and the user. These ontologies allow a fully declarative approach to document composition.

Another approach is the Scholarly Ontologies (Scho3Onto) project [Buckingham Shum et al., 2000]. It defines a digital library server that supports multiple and possible conflicting interpretations of a research document. Essential parts are the formalized encoding of interpretations of a scholarly document and the compilation to create semantic hypertexts. Objects (concepts or data) are connected by multi-typed links defining the properties of the objects. Reading a document is defined as going through a set of claims. Claims are the basics of formalized interpretations, which are considered as two connected objects and a (optional) typed link. The result is a set of semantic annotations about the document’s contributions like citations to other key literature. As we will see, this approach is similar to our hyperbook model in the way in which links are typed.

[Dicheva et al., 2004] present a framework for standards-based, ontology-aware course libraries and an environment for building, maintaining, and using such libraries. The aim is to provide a system built of (existing) learning resources to assist students. Dicheva et al. propose a topic-map based system that allows authors to describe the metadata of their learning material according to Dublin Core (DCMI) or IEEE LOM. The model is close to our approach in the way relationships were represented. Each relationship has a type and authors can use a pool of predefined relationship types.

Bocconi [Bocconi, 2003] describes a hypertext generation system to automatically select and compose scholarly hypermedia. The presented content is generated through a domain ontology containing the concepts and their relations and a discourse ontology containing different roles and narrative units describing different genres. The discourse ontology holds a very detailed and highly formalized description of the points of interest that a user can have about a domain. By integrating hyperbooks into digital libraries, we should also consider the question how to integrate elements stored apart of the domain ontology like essentials for hypertext personalization purposes. For this task, the above-mentioned model seems to be very interesting and might be a good base.

In general, the hypertext personalization problem [Brusilovsky, 1998b] seems essential for developing hyperbook systems in the domain of e-learning. In [Wu et al., 2001], the authors propose a model of adaptive hypertext which includes a domain model, a user model and adaptation rules. The domain model is a semantic network consisting of domain concepts and relations between concepts. This model serves essentially to define adaptation rules, depending, for instance, on the concepts known or appropriated by the user. We have included in a similar way some adaptive mechanisms, such as points of view, into our system [Falquet et al., 2004].

A new research field has emerged in the last years that concentrates on the concept of personalizable virtual documents [Crampes, 1999; Garlatti and Iksal, 2001]. Personalizable virtual documents are defined as sets of elements (often called fragments) associated with filtering, organization and assembling mechanisms. According to a user profile or user intensions, these mechanisms will produce different documents adapted to the user needs. The model we will present is based on this approach.

Our hyperbook model is build upon a fragment repository, a domain ontology, and an interface specification [Falquet and Ziswiler, 2003]. The fragments and the ontology including their interconnecting links form the structural part of the hyperbook (Figure 1).

![Hyperbook structure diagram](image_url)

**Figure 1:** The hyperbook structure

The basic informational contents of the hyperbook are made of reusable fragments. These fragments can be small texts, or even illustrations or programming code, but we want to avoid that authors write large fragments. They have to divide the content sources, like documents, and to place the created fragments around concepts. Fragments are connected by structural links, for instance from fragments to sub-fragments. These typed links indicate the roles they play in a group of fragments (compound fragments). For instance
an exercise could be made up of a question fragment, one or more answer fragments, and a discussion.

The semantic structure is described by a domain ontology. It is intended to hold a formal representation of the domain’s concepts and used for indexing or qualifying the fragments.

By establishing typed links between fragments and concepts, the information content stored in the fragments is referenced by the concepts of the domain ontology. This shows relations between the fragments and the concepts, but also puts the fragments into a context or a semantic environment. Typical link types are Definition, Property, Example, Illustration, Exercise, Instance, or Reference. Link types should be predefined, so that authors of a hyperbook can choose from a limited number of unambiguous defined link types.

The fragment-ontology structure allows representing different elements of an e-learning system. Domain knowledge can be represented in different, sometimes contradictory ways according to pedagogical or narrative purposes. Besides, we can annotate concepts with examples, illustrations, exercises and solutions. Even a discussion is possible (topic and message fragments connected through about and replyto links), and authors can give their opinion about a subject (arguments, opinions).

This hyperbook structure enables to generate different hypertext views. As the number of concepts in the ontology is generally much smaller than the number of information fragments, the user can browse the ontology and then go down to the connected fragments. Fig. 2 shows an extract of a hyperbook that we have used in a computer science course this year. Around a concept, we found links to other concepts (some of them are even automatically generated by link inference techniques) and annotated fragments like comments and examples.

This model seems more adequate to e-learning purposes than approaches that are based on large ontology repositories, sometimes also on several types of ontologies (top-level ontology, application ontology). Our experience has shown that people are more used to building small text fragments and annotating them with the most important concepts of the domain than creating complex ontology structures. The result will be small, but expressive domain ontologies. Of course, a more formal representation of the content might have advantages for the ontology integration process. In the next section, we show how we integrate such small ontologies by using the fragment repositories.

\begin{figure}
\centering
\includegraphics[width=\textwidth]{figure2.png}
\caption{An extract of a virtual hyperbook}
\end{figure}

\section{From Hyperbook to Digital Library by Ontology Alignment}

As all the hyperbook models presented above are based on ontologies, their integration plays a major role in the domain of virtual books and a fortiori in the domain of virtual libraries. If we suppose that each virtual book has its own domain ontology, we need an integration to create a semantically coherent virtual library. It is important to note here that it is not very realistic to suppose that all the books will be linked to the same (global) ontology, because either such an ontology does not currently exist or even if it exists, it contains only stable and well-established concepts and will not have the desired level of specialization or diversity. Thus, it will not be convenient for books on new and advanced topics.

Our ontology integration approach consists in preserving the concepts and the links of the initial ontologies and establishing relations of equivalence between elements of the origin ontologies. This approach seems to be much more flexible than a complete fusion of the hyperbook’s ontologies. In the domain of e-learning with non-centralized, semantic information systems, we have to consider that content is under a permanent evolution process. New parts of documents will continuously be added to the Digital Library.

We are focusing on work that is based on alignment techniques. This means bringing two ontologies into mutual agreement by extending the ontology with drawing links between concepts [Klein, 2001]. As subcategory, mapping ontologies means relating similar concepts or relations from different ontologies to each other by an equivalent relation. In both cases, the existing ontologies will persist.

Some heuristics for semi-automatic ontology integration can be found in different approaches. They are based either on identifying structural or naming similarities as can be found in the SMART algorithm [Noy and Musen, 1999] (now integrated in the interactive ontology merging tool PROMPT) or on machine learning techniques, used for instance in the GLUE system [Doan et al., 2002]. Other approaches like OBSERVER [Mena et al., 2000] use semantic interrelations (limited to the analysis of taxonomic relations) to specify mappings between (not synonymous) concepts of
two ontologies. It defines lower and upper bounds for the precision and recall of ontology-crossing queries that are based on manually defined subsumption relations.

As a conclusion to this introduction to ontology integration, we will remember the following points:

- Ontology integration is essential for building Digital Libraries or semantic web applications, because it does not seem realistic to assume that there is one common ontology that covers all the domains and subjects. Using local domain ontologies allow users to express concepts in a higher explicitness.

- Automatic ontology mapping for building Digital Libraries is a multi-level task. We should take advantage of the larger information base of the hyperbooks (taxonomic relations between concepts, but also other attributes of the concepts and especially instances and annotated fragments or documents). Including all this information will allow automatic procedure, contrary to other applications in the Semantic Web where automatic ontology mapping seems more difficult.

3.1 Ontology Mapping by measuring the semantic similarity

For mapping ontologies, we propose to calculate semantic similarities between the concepts. We have noticed that in the existing literature, ontology mapping is a crucial assumption for measuring similarity. We will not consider these approaches, mainly because most of them carry out ontology mapping manually and just focus on how to represent them. This shows that developing algorithms for ontology integration is a crucial, but also a critical task, and might serve not only to research in the domain of semantic similarity measurement, but also to the whole Semantic Web community.

[Weinstein and Birmingham, 1999] present an overview about semantic similarity measures. They have identified three groups:

- The first category are filter functions and are presented as inexpensive, universally applicable, and appropriate for identifying initial sets of candidate recommendations. A first subcategory concerns measures based on path distance, but they are labelled fragile due to their sensitivity to the degree of detail in the ontological structure. When classifying new concepts, the measures can change although the compared concepts have not changed. Other filter measures leverage the assumption of measures based on path distance, that local concepts in differentiated ontologies inherit. They use inheritance links to identify concepts and the definitions of both the source and target concepts.

- The matching-based functions build and evaluate (maximal) one-to-one correspondence between elements of concept definitions represented as graphs. They enable analysis of similarities and differences between the concepts.

- The probabilistic functions require domain-specific knowledge of the joint distribution of primitives. To model the joint distribution, Bayesian Networks are necessary.

The identified groups cross a spectrum with respect to the degree of required knowledge specification. Filter functions do not involve role semantics. The matching-based functions use knowledge of roles considered independently, and the probabilistic functions exploit knowledge of the interactions among roles.

3.2 Similarity measure in Digital Libraries

For the integration of hyperbooks with their domain ontologies into a Digital Library, we focus on techniques that try to establish one-to-one correspondences between ontologies of two hyperbooks. We first carry out the computation of the semantic similarity to decide if two concepts belong to the same semantic field or not. Using string similarity measurement approaches might be the first and simplest task for this computation. However, considering only the syntax of the concepts in question is too basic. In the domain of Digital Libraries, we can first assume that there are not many words with the same spelling in the different ontologies, and second that there exists terms with the same spelling, but which differ semantically (polysemy). Figure 3 shows extracts from four sample-ontologies, each speaking about Football. As we can see, the word "Football" does not always stand for the same meaning.

Using the WordNet ontology or another available top-level ontology for word disambiguation and resolving the meanings of polysemous words might be a solution. But in these kinds of ontologies, we can't usually find specific concepts, as only more general terms are represented. Looking at the ontologies in Figure 3, we can assume that there are at least some words in the upper-class level that can be found in the WordNet ontology. This would allow introducing some links between the different ontologies. But as our experience with e-learning systems has shown, authors do not normally model the upper level part of their ontologies. Another question is also whether the time needed to calculate word disambiguation with WordNet is reasonable.

![Figure 3: String similarity measurement with word disambiguation by WordNet](image-url)
The domain ontology of a Digital Library is not just a terminological ontology where the collection of concepts are organized by a partial order. It is an axiomatized ontology whose concepts are distinguished by different kinds of relations to other concepts. We have mentioned above the importance of link typing in a hyperbook structure, and we can now take advantage of this formalized representation for the ontology integration process. Thus, we need a similarity measure approach that considers more elements than just terminological relations. For this, we propose to use an extension of the technique of [Rodríguez and Egenhofer, 2003]. The approach is based on entity classes that are groups of equivalent or very similar words. The similarity between two entity classes is the weighted sum of three measurements: similarity of the terms (set of synonyms), similarity of the semantic neighbourhood (set of concepts close to the entity class in the graph) and similarity of the attributes (characteristics and properties, so called distinguished features like set of values). The similarity function takes into account the depth of the entity classes relative to their respective ontologies. The approach considers also cognitive properties of similarity by introducing asymmetric measurement of semantic similarity. The authors argue that according to people's judgement, the base and the target concept can have different roles. For instance, the perceived similarity from a concept to its super-concept is greater than the perceived similarity from the super-concept to the concept. Finally, comparing not only sets of synonyms, but involving also distinguished features into the similarity measures means introducing grades of similarity. The result is no longer a simple binary expression (same or different word), it allows to detect also similar words.

As we have mentioned, most of the known techniques for semantic similarity measurement need a primarily integrated ontology. As we want to invert the process (integrating ontologies by first determining the semantic similarity measurement), we profit of the above-described approach because it can process the measures in not yet related ontologies. This is done by simply establishing a relation to an imaginary and more general entity class "anything" from the two root-concepts of the ontologies in question. This allows to calculate of the distance from two entity classes to the immediate super-class even if there is no "natural" common super-class.

3.3 Discussion

A first problem of the algorithm of Rodríguez and Egenhofer that is to mention is the fact that the assigned weights of each specification component depend on the characteristics of the ontologies. For instance, the authors mention that, when many polysemous terms occur within the ontologies, the synonym set component might not be a good indication of similarity. In consequence, the relative weight assigned to this component should be reduced. But it seems critical when we have to analyze the occurrence of polysemy first. If we want to decide how the weights have to be assigned, we have to look at the three components in detail.

First, comparing different synonym sets by a word matching process is a very basic level of similarity measure. When checking the number of common and different words in the synonym sets, there is a gradual similarity measure, but polysemy will not be detected.

Second, the semantic neighbourhood is determined by analyzing related super-concepts of an entity class. The links are labelled with types like "is-a" (hyponymy) or "part-whole" (meronymy). The authors use path distance (with a preliminary defined upper-bound) to define the semantic neighbourhood of an entity class. In a hyperbook, we can assume that the semantic neighbourhood is a priori of "good quality" for similarity measures. The problem is how we should determine the boundary of this distance.

But the most important problems can be found in the third component, introducing distinguished feature. Although when they provide a more formal description of a domain, Rodríguez and Egenhofer have mentioned in their evaluation section that feature matching alone is insufficient for detecting the most similar entity classes as many entity classes share common features or have a common super-class from which they inherit common features. Even in combination with the other two components, the recall is lower for finding the most equivalent concepts than without using distinguished features.

The approach detects equivalences between concepts of two ontologies, but by using distinguished features, it also returns similar concepts. This is an important fact. In contrary to the ontologies that Rodríguez and Egenhofer have used for the evaluation of their approach (top-level ontologies like WordNet), we are primarily confronted with small domain ontologies. This means that we can't be sure that there are many equivalent concepts. To get a stronger mapping between the two ontologies, we need an approach that can at least detect some similar concepts.

At this moment, it seems important to remark that this approach is adapted to hyperbooks and Digital Libraries and most probably can't be used in other Semantic Web applications as such. A high degree of domain formalization and a strong explicitness of concept representation are important assumptions.

Before we explain how we use an extended concept of distinguished features to improve semantic similarity, we want to mention another unresolved problem of this approach. In fact, there probably is no single root in each of the ontology to compare. But our model presented in section 2 explicitly allows a flat hierarchical ontology structure, where a single root might not appear very often. Such a hyperbook model might be much closer to e-learning purposes. So, the question remains how to build a common virtual root if relevant domain ontology is empty. We could try applying the approach of Rodríguez iteratively, but applying the process to an even smaller part of the ontology might represent the best solution.

The above-mentioned approach of [Doan et al., 2002] is particularly interesting because they deal with taxonomical relations and instances to establish links between two ontologies. Contrary to Rodriguez, they don't establish an arti-
ficial top-level root concept to measure the semantic similarity, they primarily match the instances with machine-learning technique to be able to define the semantic similarity of two concepts. Then, they define distributions over the concepts to determine the final semantic similarity by a user-supported approach. Finally, they determine which of the derived relations will be considered into the ontology mapping according to given domain constraints and relaxation labelling (taking into account the neighbourhood of a concept).

3.4 Improvements

In our case of virtual documents, we make use of additional information to evaluate the similarity between concepts. We also consider the annotated fragments and we take advantage of the fact that links between fragments and concepts are typed. As we have seen above, the fragments were annotated according to a predefined list of link types. If two concepts $A$ and $B$ are bound by links of the same type $t$ to sets of fragments $t(A)$ and $t(B)$ respectively, the “documentary” similarity between $t(A)$ and $t(B)$ will be taken into account in the definition of the similarity between $A$ and $B$. To define the similarity between $t(A)$ and $t(B)$, we will use a traditional technique of information retrieval (for instance, the cosine between the $t$-idfs vectors representing the documents in the space of terms [Salton, 1989]). Then, we define the similarity between $t(A)$ and $t(B)$ based on the similarities between documents (for example by taking the maximum similarity found between all the fragments of $t(A)$ and $t(B)$).

The similarities obtained for all types of links will then be added up to the similarity measure computed at the conceptual level.

It is important to remark that link typing is crucial here. Indeed, the comparison makes sense only if the compared fragments play the same role with respect to a concept. If, for instance, fragment $a$ is an example of concept $A$ whereas $b$ is a counterexample of $B$, a strong similarity between $a$ and $b$ does not imply a strong similarity between $A$ and $B$, on the contrary.

4 Future Research Questions and Conclusion

In this article, we have presented our hyperbook model, which is based on domain ontologies and fragment repositories and an integration process into a Digital Library, which is an extended version of the approach of Rodríguez and Egenhofer. We have applied their algorithm to Domain ontologies by involved typed relations between fragments and the domain ontology. In future research, we will work on the integration process, especially on the validation of matched links between concepts of the domain ontology, but also on the integration of hypertext personalization elements like points of view. This also seems an interesting approach for validating the established relations because points of view help to describe the concept of an ontology more precisely.

References


