# Topic driven access to scientific handbooks

Caracciolo, C.

CHAPTER 7

# Conclusions

In this thesis we have explored the possibility of providing topic driven access to scientific handbooks by means of a domain map that is automatically linked to appropriate segments inside the underlying handbook texts. For our case studies we made use of the *Handbook of Logic and Language* [van Benthem and ter Meulen, 1997]. Our approach is described in detail in Chapter 2, and the core of our work is presented in Chapters 3, 4, 5 and **??**.

In order to facilitate the exposition of our conclusions, we group the Research Questions presented in Section 1.1 into two groups, concerning the map on the one hand and linking on the other (Sections 7.1 and Section 7.2 respectively). Next, we discuss the resulting "bigger picture" (Section 7.3), and ideas for future work (Section 7.4).

## 7.1 How to Organize and Visualize a Domain Map

We started our work by asking: "What requirements should we impose on a map that is to be used for human browsing and as a skeleton to provide focused access to the text?" (Research Question 1). Focusing on non-expert end users, our answer was to impose three requirements on the LoLaLi map (Section 3.1): inclusion of relevant topics from the domain, informativity to the audience addressed, and a low risk of information overload. The trade-off between informativity and information overload, together with the observation we made during the user studies, led us to define the set of relations used within the map that we presented in Chapter 3. In the course of that chapter we also presented examples to suggest that more refined, formal, or specific relations could be used, but we argued that in that case we would actually be addressing a different

type of end user.

Next, we asked: "How do we present the map to readers of a handbook in such a way that we ensure broad coverage of the domain (with detailed information per topic), while making sure that users do not get lost?" (Research Question 2). We proposed a user interface tailored to address an audience of non-expert users (Chapter 4). Our user interface was developed to be available from the Internet, therefore it does not require special software to be installed by the user. It shows the main features of the map while hiding the complexity of the structure, and it integrates browsing and searching. The usability of the proposed user interface was tested during the user studies we conducted (Chapter 4). The proposed interface was found to be clear and usable by both groups of end users on which it was tested (undergraduate and masters students), although masters students showed a better grasp of the relations and a stronger preference than undergraduate students for a combination of searching and browsing.

During the user studies we also aimed at gathering evidence about how our intended users perceive an environment that should support a variety of information seeking behaviour and support focused access to text. The user studies we conducted, one of the very few of that type, confirmed the usability of the interface and helped in refining the set of relations presented in Chapter 3. Moreover, they showed that the model of the domain we adopted is easily grasped. We found that a user's background especially affects browsing activities. Therefore, the user's background should be carefully considered when illustrating the semantics of the relationships used, as misunderstandings are likely despite the apparent intuitiveness of the relations. Finally, the studies suggested that the appreciation of the possibility of inspecting a text from the map is independent of the user background.

As the map was intended to be an aid for human navigation, an ingredient of a bottom-up data-driven approach to information, as well as a starting point to experiment with our ideas about focused information access, it was not endowed with formal semantics. We went for a map formally under-specified, deciding for a trade-off between the possibility of imposing constraints enforceable at the level of consistency checking and a more user-friendly approach to modeling. The approach we adopted is sensible given the purpose of our work. If the possibility of automatic consistency checking should be considered for later work, the level of formality should still be kept minimal given our intended users.

The major bottleneck we found during the making of the LoLaLi map was the lack of tools to organize the authoring and editorial processes. We needed methodologies and tools to organize a workflow involving *editors* and *authors* of the map, where editors are subject experts, but untrained knowledge modelers, and authors are mainly subject experts (e.g., linguists, logicians, and computer scientists). Under this view, editors are in charge of the general planning of the map, while authors are responsible for its actual population. A person may play both roles at once, but the roles are logically distinct. Besides common editorial activities, such as deciding on terminological conventions (e.g., topic names, templates for glosses, etc.) and ensuring uniformity of style, editors should ensure balanced development of the map, and share duties among

subject experts. The former activity consists in sketching the highest level topics in the map, and setting its goals, purposes, and the level of detail (granularity) required. The latter activity consists in selecting and defining "areas" to assign to authors and validating their contributions.

The activity of planning requires that some modeling language be available (similar to UML for semantic structures or ontologies) to support the modeling process and enable unambiguous communication of it. The activity of sharing duties implies the possibility of: defining fragments (modules) to assign to subject experts, checking for consistency of the new module with the rest of the map, and keeping previous versions in order to restore them when needed (versioning system).

At the time when we actively worked on the development of the LoLaLi map we found very little support from existing tools and methodologies, and this represented a real bottleneck for our work. Meanwhile, the need for this type of support has been widely acknowledged in the Semantic Web research community and a wealth of studies is currently in progress, including, for example, the two European funded projects Knowledge Web[1] and Networked Ontology[2] (NeOn), about ontologies on the web and ontologies life cycle respectively.

Concerning the life cycle of the map, we stress the importance of suitable functionalities for visualization, modularization, searching and integrity checking. All the functionalities mentioned so far should be integrated in the editorial tools, and should be adapted according to the role played by the user and to the task she has to perform.

We found that for editors it is most important to manage the global view of the map, while for authors the local view is the one most commonly used, as authors need to be able to visualize and modify individual pieces of information attached to topics. Authors may also need to visualize some (selected) branches in great detail, while only sketching other branches to provide some context (i.e., a variation of the focus-context principle). Zooming facilities should enable zooming into subgraphs and single nodes: when zooming in, the author should be able to select also a textual view, for example with paths from the root node indicated in a linear way, handy for printing. Members of the editorial board will need a global, graphical view of the complete map, for all those tasks related to the overall development of the map. In such a visualization, the entire structure would be represented as a graph where each node is visualized with a minimum set of pieces of information (for example, the title only).

## 7.2 Linking the Map to the Handbook

Our Research Question 3 was: "What are suitable targets in the handbook to establish focused links from topics in our browsable map?" Since our aim was to provide *focused*, topic-driven access to the handbook, we concentrated on passage retrieval techniques and compared two types of structural segmentation (by paragraphs and by

[1]URL: http://knowledgeweb.semanticweb.org/.
[2]URL: http://www.neon-project.org/.

sections) and two algorithms for semantic segmentation (TextTiling and C99) against a gold standard we created (Chapter 5). We highlighted and discussed the issue of building a corpus of annotation segments, illustrating how the writing style affects the task. We also discussed measures used for topic segmentation and found that given the nature of the text, precision and recall on topic breaks are the best, although crude, measures. We found that when only the text is considered, the best link targets for topics in the LoLaLi map are segments created by applying TextTiling and structural segmentation by single paragraphs. Actually, segmentation by paragraphs performs best in terms of recall (obviously), but TextTiling behaves (slightly) better when considering both precision and recall.

Given the target links we found, how can the ones relevant to the topics in the map be identified? (Research Question 4) We ran a second set of experiments (Chapter **??**) in which we used the segments obtained in the previous experiment as a collection of documents to link to topics in the map. We used two different weighting schemas (OKAPI and tf.idf) to rank segments with respect to their similarity to a given topic. In order to evaluate these experiments we built a manually annotated reference corpus, and we used several measures for evaluation. Among them, we used three novel measures that we introduced: C-precision, to give account of the amount of relevant matter in a link target; the early onset error, to measure the amount of non-relevant content placed at the beginning of a link target; and the late onset error, to measure the amount of relevant content missed by the beginning of the link target. The two error measures give a measure of the quality of the entry point of link targets. The three measures are complementary, in that they look at different aspects of a link targets. For example, when applying C-precision one looks at the entire set of relevant paragraphs in a segment, while when measuring the onset errors one only considers the beginning of a segment and, in case of the late onset error, the paragraphs preceding its beginning (the relevant paragraphs "missed").

We found that structural segmentations by sections, and the topic segmentation performed by C99 do not constitute good options, while structural segmentation by paragraphs, and the topic segmentation performed by TextTiling do. Tuning of parameters did not affect the results. We conclude that TextTiling and segmentation by paragraphs, and possibly a combination of them, are appropriate techniques for this task.

## 7.3 The Bigger Picture

In this thesis we proposed that a domain map endowed with links to a text would provide focused access to the text while enabling information seeking activities. We learned that a domain map can be an informative tool, and the typed relations used in it are at the same time its strength and its weakness. If, on the one hand, they could lead non-expert users to find their way through the text and the domain, on the other hand, their understanding relies on previous knowledge of the user, which may lead

to difficult interpretations. From the user studies we learned that some features of the interface need to be improved and that certain types of users can profit from the LoLaLi environment better than others.

Our conclusion is that a map of topics can be the basis for an electronic environment where a variety of information seeking behaviours are supported and topic driven access to the text is enabled. However, in our opinion the grand vision of a comprehensive domain map should be abandoned in favor of smaller maps that can either be manually built or (semi-)automatically extracted from the text. They would specifically provide access to the text and represent the fragment of domain treated in the text according to the view of the author(s). These smaller maps could then be connected to one another allowing users to navigate (i.e., browse and search) a variety of related yet independent resources in a unified way. The reasons for preferring document-based maps over domain maps are both practical and theoretical. Since in that case consistency across the entire map would not be an issue, it would be possible to access documents embodying different views of the domain in a more flexible way. Also, in case the maps were manually built, they would require fewer resources also in terms of time needed for the development. To a certain degree the folksonomy (and social tagging) phenomenon that has emerged over the past few years may be viewed as a partial instantiation of this view.

The main challenge related to the extraction of candidate links for the map was related to identifying the limits of the segments. Our work concentrated on converting a text originally intended to be linearly read, and the map was written by different people than the authors of the text—and a few years later. These factors quite naturally lead us to consider an alternative scenario where text and map are built at the same time, so that one reflects the other and vice versa. In this way the map can become an aid for writing and a general publishing model, and the problems related to the conversion of an existing text would be avoided altogether. Harmsze [2000], who took a very different research approach than ours, arrived at similar conclusions (cf. Section 2.5). We do not share Harmsze's confidence, though, that such an approach would be a viable solution to providing focused access to scientific information, nor that this is the direction that electronic publishing will (or should) take. In that scenario, in fact, when writing a new text, an author should take into account the relations "allowed" in the map. The set of relations that we took into consideration in LoLaLi should necessarily be enlarged, which would stress the tension between informativeness and information overload (both for the end user and the authors of the map) that we consider fundamental. Also previous work in the hypertext community [Baron et al., 1996, Conklin, 1987, DeRose, 1989, Smolensky et al., 1987, Trigg, 1983, Trigg and Weiser, 1986] suggests that this issue should be treated with a great deal of care. In fact, the many typologies of links proposed run the risk of overloading the reader with a large number of very detailed (and often ad hoc) definitions and restrictions, between parts of the same text, or between different texts, or between text and the annotations one may want to attach to the text (as an extreme case, consider Trigg [1983], who defines about 80 link types). If, on the other hand, the author were free to decide what relation to use and

include, then the risk is to have personal semantics, somehow like in folksonomies, but without the advantages of a community effort.

These considerations make us think that the role of semantic structures like our map, or more refined ones, as an aid to writing new handbooks is limited (Research Question 5) in domains or disciplines where a standards body is absent. In our view, handbooks will continue to be written "linearly" (i.e., without heavy constraints on using sets of pre-defined relationships between topics) because this form allows the author the level of flexibility that is required to express complex and articulated thought. However, this does not imply that there is no room for improvement in the way writing tools are shaped or in the functionalities they offer. Technologies like wikis have already—and dramatically—changed our understanding of collaborative writing and rapid publishing [Miller, 2005]. Experiments with the so-called "semantic wiki" are also looking at the possibility of leveraging the search functionalities available in common wikis by defining relations between pages [SemWiki, 2007]. In particular, authoring and editorial environments will leave behind the model of electronic typewriters to be more and more plugged into the Internet and connected with other available resources. Although not of practical use for the *writing* of new handbooks, we believe semantic structures can have interesting applications for *disseminating* and *accessing* information. For this reason, we emphasize the importance of exploring tools to (semi-)automatically extract maps of documents (as opposed to maps of domains), highlight topic breaks in long documents, and, in general, refine search facilities to bring the user "below" the document level.

## 7.4 Directions for Future Work

The work presented in this thesis shows that the issue of providing topic driven access to scientific handbooks involves a variety of ingredients. The following directions for future work resulted from our work:

- improvement of the user interface for hierarchical structures;
- more user studies to test the proposed interface and understand the information seeking behaviour of end users in complex electronic environments;
- query-based topic segmentation techniques that are to be used on the fly and that also take into account the structure of the map; and
- evaluation of link targets, and their visualization for the purpose of navigation.

Moreover, the authors and editors of the map should be provided with appropriate tools and methodologies for development and maintenance. Finally, as mentioned in the previous section, we suggest investing in methods for the automatic extraction of maps and hierarchies of topics directly from the text.

As hierarchical structures like the LoLaLi map and ontologies are increasingly being used in a variety of domains, it is essential that appropriate user interfaces be devised for end users (i.e., people with no experience with ontologies or knowledge representation). The user interface we proposed is a good starting point that needs to be refined and improved to accommodate the feedback we collected during the user studies. Next, it is recommendable that more advanced user studies be conducted to confirm the changes made. Such a second run of user studies will also be the occasion to check how the revised organization of the map (presented in Chapter 3) is taken by real users.

As for the segmentation of text to produce link targets, we suggest experimenting with integrating text segmentation into the retrieval algorithms (as opposed to having the segmentation phase as a pre-processing). The segmentation would then be performed based on the topic to which the segment should be linked. Another approach that we consider worth investigating is performing structural segmentation based on paragraphs and aggregating paragraphs at a later stage, during or after the information ranking algorithm has been applied. Next, we need to better understand how to integrate structural information from the map into the segmentation and retrieval algorithm. Finally, the issue of evaluating link targets for the purpose of focused access deserves further study. What we did was to concentrate on the quality of a segment measured in terms of the amount of relevant content (C-precision), and in terms of *where* this relevant content is placed in the segment. We encourage further work on including into the evaluation measures a dimension of "readability" of the segments. Related to this issue is the issue of how to visualize link targets. Our observation about the level of presupposition of the text at hand suggests that the notion of what a link target for focused access is cannot be separated from the way the link is going to be visualized, navigated and read by the end user.

In the course of Section 7.1 we discussed at length the bottleneck we found during the making of the LoLaLi map. That discussion gives direction to future work needed in order to enable authors and editors to develop and maintain a map of topics, in terms of both tools and methodology. Now we add to that discussion the convenience of modularization for assigning subgraphs to authors for development: in order to do this, methods to define, select and manage *modules* from the map are needed.[3] It is also important to be able to keep control of the module during its entire life cycle, therefore a versioning system and the possibility of recovering older versions are also imperative. We also suggest that search facilities for authors of the map be improved and integrated in the environment for editing. In particular, it is worth investigating how different notions of similarity among topics can be integrated in the search facility. In fact, depending on the task at hand, two topics can be defined as being similar if they have exactly the same name, if they occupy the same position in the map, if they have

[3]It can be useful to distinguish between *modules* and *partitions* of ontologies, where the difference is whether modules overlap or not. This issue is gaining increasing attention and a number of algorithms have been put forward, see for example [dAquin et al., 2006, Noy and Musen, 2004, Seidenberg and Rector, 2006, Stuckenschmidt and Klein, 2004].

similar glosses or link to the same parts of the handbook, if they have been authored or edited by the same person, and so on.