elpub2002 Proceedings — J. A. Carvalho et. al. (Eds.) — VWF Berlin, 2002

Towards Scientific Information Disclosure Through Concept Hierarchies [1]

Caterina Caracciolo[1], Maarten de Rijke[1], and Joost Kircz[2]

[1] ILLC, University of Amsterdam
{caterina, mdr}@science.uva.nl
[2] KRA Publishing
office@kra.nl

Abstract. We report on an ongoing project aimed at providing an exemplary architecture for an electronic dissemination environment for scientific handbooks. We focus on our way of facilitating navigation through and access to electronic handbooks by using a WordNet-like concept hierarchy consisting of synsets that are connected to each other and to external sources by semantic relations for navigational purposes.

1 Introduction

There are many reasons that justify an electronic version of scientific handbooks that cover fairly abstract material such as the *Handbook of Logic and Language* [11] or the *Handbook of Automated Reasoning* [10]. It makes distribution easier and quicker, and readers can be helped considerably when searching for information: even simple keyword searches are more useful than scanning tables of contents or indexes, especially for large handbooks, and tracking down a reference can be as simple as a mouse click. Also, electronic publications are less rigid than their paper counterparts, which facilitates integration with other media types: e.g., computer simulations and visualizations, movies, and tools.

Electronic books facilitate a modular way of reading, as opposed to the linear way of traditional paper books. Indeed, screen, mouse and keyboard constrain the

[1] This research was supported by Elsevier Science Publishers. The second author was supported by grants from the Netherlands Organization for Scientific Research (NWO), under project numbers 612-13-001, 365-20-005, 612.069.006, 612.000.106, and 220-80-001. We thank Anita de Waard and Guus Schreiber for helpful comments and suggestions.

reading (and writing) process and suggest that we should orientate the whole reading environment towards a more modular scenario.[4] proposes a modular structure for articles in experimental sciences, but it is not clear whether this approach can be adapted to handbooks that contain more abstract content. In the *Logic and Language Links* project we aim at defining how an electronic publication should look like: we are especially interested in developing a good hyper-link system, rich enough to account for the complexity of our domain (the interface between Logic and Linguistics), while avoiding disorientation of the reader.

The ideal reader of the envisaged electronic version of Handbook of Logic and Linguistics is not a new learner user, because our hyperlink structure is not meant to provide a learning environment. Moreover, given the size of the concept hierarchy, it is best accessed by using a mixture of browsing and searching, which implies a certain ability in phrasing the information need.

In our approach we use a WordNet-like concept hierarchy to annotate and access the handbook. It consists of synsets (sets of synonyms) that are connected to each other and to external sources by semantic relations. Topic or concept hierarchies are often used for the purpose of navigating through large collections of documents. They are very useful for the organization, display and exploration of large amounts of information. Well-known examples include Yahoo!'s topic hierarchy for exploring the Web [12], and Google's directories [3] (based on the DMOZ open directories initiative [9]).

Moreover, it has been shown that users in a hypertext search task who had hierarchical browsing patterns through the hypertext performed better than users who had sequential browsing paths [7]. Therefore, it is very important that architectures for electronic handbooks allow, or even enforce, such hierarchical patterns: a concept hierarchy is a good way of doing this.

In Section 2 the *Logic & Language* (LoLa) hierarchy is described in some details: Section 2.1 presents the internal structure, Section 2.2 the encoding. The process of populating the hierarchy is addressed in Section 3, where we also outline the roles of editors and author in the process. Section 4 is dedicated to our ongoing work, while some conclusions in Section 5 close the paper.

2 Organization of the Hierarchy

The LoLa concept hierarchy [6] consists of concepts, connected by several semantic relationships. By *concept* we mean every relevant notion or topic in the domain, worth individual discussion. In line with WordNet [2], we make a distinction between words or terms on the one hand, and concepts on the other hand: a concept is denoted by a *synset*, a set of synonymous words (we only use the English nomenclature for our domain). Words are synonymous if they have (more or less) the same

meaning in some settings. For example *first-order logic* is also known as *predicate logic*, *FOL* or *predicate calculus*.

The semantic relationships linking synsets come in two kinds: ones that are *internal* to the concept hierarchy (Section 2.1), and ones that link the concepts to *external* resources (Section 4.4).

2.1 Internal Architecture

Concepts in the hierarchy are annotated with a gloss; for instance, *the study of language meaning* is a gloss for *semantics*. Moreover, they come with a longer description, provided by the authors of the concept especially for the LoLa hierarchy.

The hierarchy consists of a TOP concept, under which 4 main branches find a place: *computer science*, *mathematics*, *linguistics* and *philosophy*. Concepts are related to other concept(s) by one of the following relations:

1. is a kind of: *epistemic logic* is a related subtopic of *modal logic*;
2. is part of: *metaphysics* is a part of *philosophy*;
3. technical notion: *operator* is a notion in *mathematical logic*;
4. mathematical result: *Goedel's incompleteness theorem* is a mathematical result (theorem) of *logic* and *mathematical logic*;
5. computational tool: *SPASS* is a computational tool for *first-order logic* (it is a first-order resolution-based theorem prover);
6. historical view: the concept *Frege on quantifiers* gives an historical view of the concept *quantifiers*.

Such relations are called *vertical relations*, because they form the backbone of the hierarchy. The type of vertical relation is now indicated to the user using an abbreviation placed beside the name of the parent concept. More refined and compact visualizations (e.g., colors or icons) will be tested by means of usability tests with an appropriate sample of users.

The above set of relations is currently undergoing a detailed analysis to make sure that they provide a reasonable coverage of important semantic and cognitive connections between concepts.[3]

The concept hierarchy is not a strict tree, as long as multiple parenthood is allowed: for example the concept *logic* has the concepts *mathematics* and *computer science* as parents. In fact, this is properly a graph structure.

[3] In particular, we are currently studying the distinction between the notion of *is a*, interpreted a the set theoretical notion of subclass (for example *modal logic* is a kind of *logic*, where modal logic stands for a family of logics using modal operators), and the notion of *instance* (using the above example, *M* is a instance of modal logics, i.e. a particular axiomatization of a modal system.)

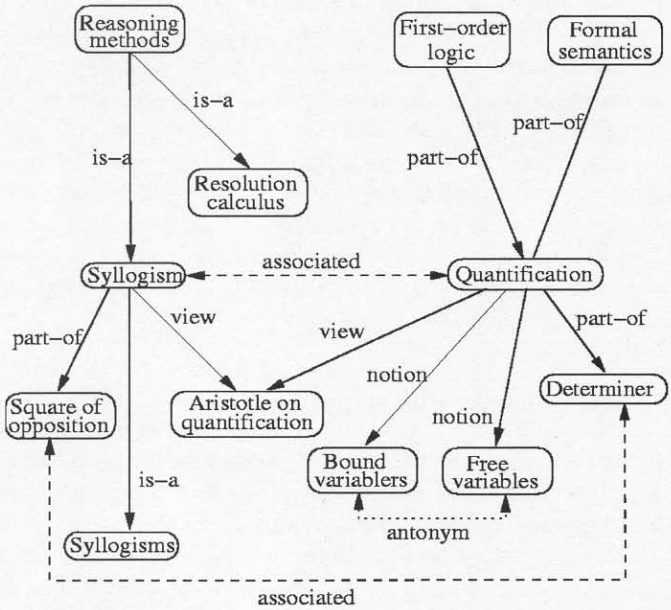

Fig. 1. *A graphical representation of a fragment of the LoLa concept hierarchy.*

In graph theory, cycles are sequences of connections between nodes of the graph that start and end up in the same node. In our setting, we do not allow cycles consisting of only vertical relations (i.e., of the kind mentioned above), since they disorientate readers. An example of such a cycle is the following: *logic, model theory, logic.* Nevertheless, cycles are allowed when they also involve (possibly only) horizontal relations (e.g., *updating, belief revision, knowledge representation, updating*, where updating and belief revision are subconcepts of knowledge representation and are associated to each other).

The horizontal relations mentioned above are non-hierarchical relations, mainly used for navigational purposes. They include the following:

1. **Sibling**: all concepts having the same parent(s). Informal experiments indicate that readers find it useful to know what the siblings of a given concept are. Provided that the siblings are listed in some meaningful order, they prevent the 'lost in space' problem. Siblings are automatically computed and presented to the reader with a flag indicating what kind of relation they have with the parent.
2. **Other meanings**: all concepts having the same title, but with a different gloss. For example both *computer science* and *mathematics* have *logic* as subconcept,

with the following gloss: "A system of calculus or reasoning"; while *logic* under *philosophy* has the gloss: "The branch of philosophy that analyze inference". This relation is automatically computed, too.

3. **Associated concepts**: concepts sharing some properties or somehow analogous to each other. For instance, *finite state machine* is similar in this sense to *regular language* (check belief revision-updating or the example from jan). This relation is provided with a short explanation of the reason of similarity.
4. **Antonymy**: as in the case of *completeness* and *incompleteness*. Learning the antonym of a concept not only teaches us more about the meaning of the antonym, but also about the concept itself [8]. Like the similarity relation, antonym comes with a short explanation.

2.2 Encoding the Concept Hierarchy

Each concept is given a unique identifier and is represented as an XML document in which the following information is stored: references to the parent concept(s), type of relation with the parent concept, gloss, reference to the extensive description, and references to the associated and antonym concepts. Moreover, all elements in the XML tree are given an unique identifier to be further addressable. Descriptions are stored separately because they are typically written in LaTeX and can also contain non textual objects, like images. Users do not access the XML base but a static set of HTML documents, searchable and browsable, generated from the XML base at regular intervals.

Despite the existence of more sophisticated languages to represent hierarchical structures, as RDFschema, we decided to stick to XML, a less expressive but more consolidated language.

3 Populating the Hierarchy

The *Logic & Language* concept hierarchy is currently being populated by hand. The focus of our work is on creating the hierarchical structure, complete with glosses and internal relations; extensive descriptions have mostly been left out at this stage, while external links will be addressed in a later stage.

The current version of the hierarchy is populated with close to 500 concepts, provided by the LoLa group at ILLC. Domain experts from the University of Amsterdam and the University of Utrecht are now being involved in the process of building the hierarchy as large-scale community building effort. Our plan is to double the size of the concept hierarchy by involving about ten groups from the two universities.

In this scenario a slightly unusual notion of editorship takes shape. Editors are responsible for subparts of the hierarchy: they contact the author/s, assign

them specific concepts and/or parts of the hierarchy to create, and check their contributions. In particular, editors are supposed to avoid unbalanced growth of the hierarchy, and check that the classification of the relations between concepts be coherently applied.

Being an author of the LoLa CH means being author of a concept, e.g., to provide a title, a gloss, a description and all the relations that link the concept to the rest of the hierarchy. An author can also revise existing concepts to update its content (gloss and description) or modify/add new relations.

The relationship author-editors in this context is characterized by a longer contact that in usual paper publications. In fact, since the notion of "finished" publication is obsolete in electronic publications, authors are likely to be called for revision of their contribution on a regular basis. Good examples of this change in style and requirement are the Encyclopedia Britannica On-line [1] and the Living Reviews in Relativity [5].

Documentation giving guidelines about how to contribute in the concept hierarchy is available. In order to facilitate this process we have also developed a web based authoring environment that allows authors to add new concepts.

4 Ongoing work

4.1 Author Environment

We are setting up an appropriate environment to facilitate the work of contributing to the hierarchy. It is important for the author to have a graphical map of the entire hierarchy, with the possibility of isolating only the fragment assigned to him or her. Moreover, it would be of great advantage to be able to simulate at runtime the result of accommodating new concepts in the existing hierarchy. We are currently working on a visualization of the graph, while the preview of the author's contributions is left for a later stage.

4.2 Editorial Environment

As the population of the hierarchy grows, the control and integrity checking of its content becomes more and more difficult. Therefore, a set of utilities to automatically check for integrity constraints is under development, to support the editors and the administrators of the system. Finding cycles, repetitions, cross references and inconsistent relations, these are all operations in the scope of the integrity checking tools.

4.3 User Environment

A module for sophisticated searching of the hierarchy at the user end is under development. The search facility is a crucial feature for users, since access through

browsing is not suitable for significantly large graphs. The user will be able to search the hierarchy using a structured search that allows for queries like "has title...", "has gloss..." or "has sibling called...". Besides that, a generic, i.e., unstructured search (string matching), will also be possible.

4.4 Linking the Hierarchy to the Handbook

In addition to the internal links, our concept hierarchy will also accommodate *external* links in the sense that they are between concepts and targets outside the hierarchy. We distinguish between *handbook* links (to information in the handbook but outside the concept hierarchy), and *web* links (to information sources on the web). Here we focus on the former.

The target of a handbook link can be of different levels of granularity (a part, a chapter, a subsection, a definition, etc.). Ideally, concepts higher in the hierarchy refer to larger fragments in the handbook, while lower concepts refer to smaller parts. However, as the handbook chapters are written by different authors, resulting in a different structuring and writing style for every chapter, this is hard to achieve.

Handbook links come with meta-data describing crucial information about the publication linked (e.g., author, editor, publisher), enriched with an indication of the link type (e.g., definition, theorem, discussed-in, example, counterexample, ...).

In a earlier stage of the project, we have experimented with automatically generating hypertext links from concepts in the hierarchy to (electronic versions of) chapters in the original Handbook of Logic and Language. As the documents to be retrieved, we took pages of the original handbook; while arbitrary, this choice was forced upon us by the diversity of the writing styles of the contributing authors. For the queries we explored several possibilities (term, term plus description, term and description plus additional weights on the term). We plan to use the current (richer) hierarchy to run more refined experiments and concentrate on the segmentation of the text with respect to the topic treated, and classification of the topic itself.

5 Conclusion

We have reported on ongoing work aimed at providing an exemplary architecture for an electronic dissemination environment for scientific handbooks. We focused on facilitating navigation through and access to electronic handbooks by means of a WordNet-like concept hierarchy consisting of synsets connected to each other and to external sources by various semantic relations. We also reported on the state of the project, and outlined current and future developments.

References

1. Encyclopedia Britannica. http://www.britannica.com/.
2. C. Fellbaum, editor. WordNet, an Electronic Lexical Database, MIT Press, 1998.
3. Google, http://www.google.com/.
4. F. Harmsze, A Modular Structure for Scientific Articles in an Electronic Environment, PhD thesis, Universiteit van Amsterdam, 2000.
5. Living Reviews in Relativity, http://www.livingreviews.org/.
6. Logic and Language Concept Hierarchy (prototype), http://benedictus.wins.uva.nl/LoLaLi/alpha/.
7. J. E. McEneaney, Visualizing and assessing navigation in hypertext. Proc.10th ACM Conference on Hypertext and Hypermedia, pages 61-70, 1999.
8. V. L. Muehleisen, Antonymy and Semantic Range in English. PhD thesis, Northwestern University, 1997.
9. DMOZ Open Directory Project, http://dmoz.org/.
10. J. Robinson and A. Voronkov, Handbook of Automated Reasoning. Elsevier, 2001.
11. J. van Benthem and A. ter Meulen, Handbook of Logic and Language. Elsevier, 1997.
12. Yahoo!, http://www.yahoo.com/.