Topic driven access to scientific handbooks

Caracciolo, C.

CHAPTER 2

A Vision on Access to Electronic Scientific Handbooks

Perhaps the most basic thing that can be said about human memory, after a century of intensive research, is that unless detail is placed into a structured pattern, it is rapidly forgotten.
[Bruner, 1960]

The aim of this chapter is to provide a high-level view of the browsable map that we envision and that underlies the work presented in this thesis. We also provide brief introductions to the core areas and terminology that will play leading roles in this thesis: the semantic approach to the organization of information, Information Retrieval (IR) and digital libraries

This chapter is organized as follows. In Section 2.1 we introduce our proposal for providing access to electronic scientific handbooks, i.e., the idea of a browsable map. Then, in Section 2.2 we provide background on scientific handbooks, and in Section 2.3 on searching and accessing books. In Section 2.4 we discuss various types of semantic structures that inspired the LoLaLi map (introduced in Chapter 3): thesauri and taxonomies, semantic networks and ontologies. In Section 2.5 we present related work on providing focused access to scientific documents by modularization of the text. Finally, in Section 2.6 we discuss the issues presented in the course of this chapter, and their connections to our work in this thesis.

2.1 The Vision

In our vision, a person with some knowledge of the domain at the center of our attention (i.e., the interface between logic and language), but not an expert (typically, a graduate or undergraduate student in the area), should be provided with multiple ways of accessing a handbook in electronic format. The bottom line is that the entire text should be accessible for reading, and the text should be queryable in order to locate words and phrases in it. In addition, the user should be guided through the domain covered by the handbook so that more "vague" information needs can be supported as well. The sort of support we envision is provided by a map of the domain, where each topic in the map leads the user to a specific *excerpt* (in the following, also called *segment* or *passage*) of the text in which the topic is covered.

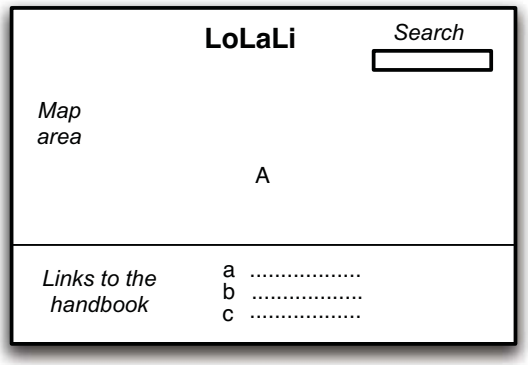
Our idea, then, is to provide the user with an integrated environment, where a browsable map of the domain is provided with two types of links: *internal to the map*, to make explicit (some of) the relations between topics in it, and *external to the map*, connecting the map to the text. The latter type of link connects topics to passages of the book that are internally homogeneous. Such a map is at the same time a tool for users to explore the domain and a tool for retrieving relevant excerpts from the text. In order to ensure maintenance and scalability of the approach, these links should be automatically selected.

What is a good way to create such a map, then? And what is a good way to select links to the text? Although some work has been done on automatically extracting hierarchies from text (see e.g., [Caraballo, 2001, Cederberg and Widdows, 2003, Girju et al., 2003, Roark and Charniak, 1998, Snow et al., 2005]), we assume that a map that is manually created by experts of the domain will be richer, more reliable and of better quality than an automatically created one. The expected downside of this decision is that different authors tend to project their knowledge in different ways, and their opinions about the relations between concepts can vary considerably. Also, the experts who contribute to the map may not coincide with the authors of the text, which, it can be argued, represents an extra source of difficulties.

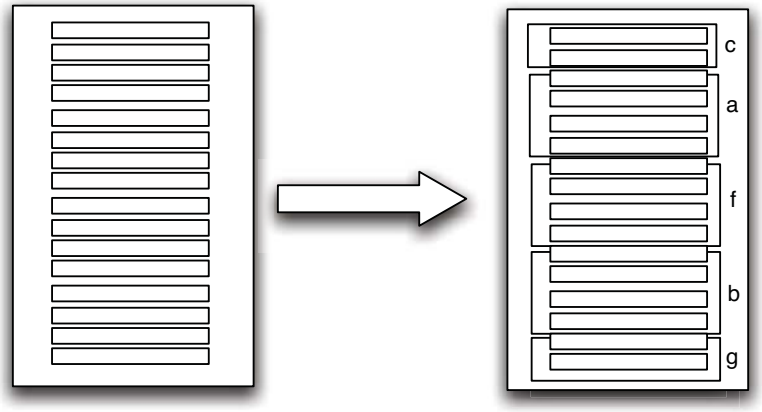
With the help of domain experts, we have organized topics from the domain in a graph[1] where topics are connected by labeled relationships and provided with glosses. The map is organized by means of semantic relationships, both hierarchical and non-hierarchical, that make explicit to the user the relationships between topics in the domain.

The connection between the topics and the text is provided by hyperlinks to link targets (Figure 2.1 (a)), which are found in two steps (Figure 2.1 (b) and (c)). First, the text is divided into *passages*, then the passages are matched to the appropriate topics from the map by means of information retrieval techniques. When dividing the text into passages, we are interested in understanding what kind of passage identification

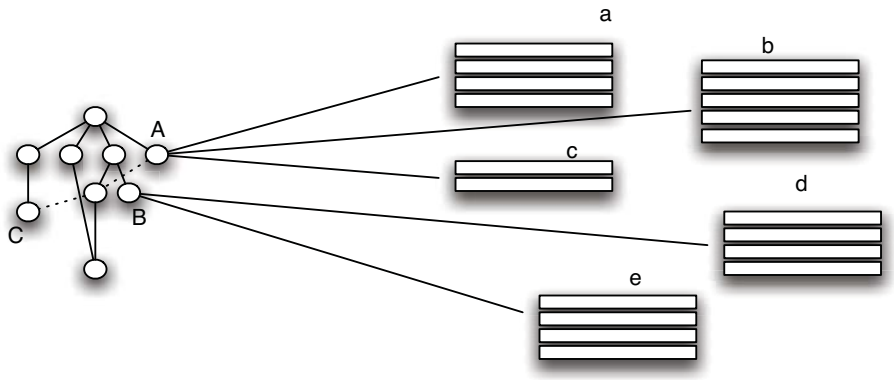
[1] According to standard terminology used to talk about graphs, we will often refer to topics in the LoLaLi map as *nodes*.



(a)



(b)



(c)

Figure 2.1: Graphical representation of the LoLaLi environment and its components. (a) The topics in focus are shown in the center of the map area together with the associated links to the handbook. A search box is always available while navigating. (b) The text is divided into segments. (c) Segments are linked to topics in the map.

technique is most appropriate for the text at hand. But, how can this be evaluated? In order to meet the user's reading needs, we do not only want to evaluate the relevancy of the passage, but also its ability to provide a good *entry point* to the text.[2] This means that the beginning of the text should coincide with the beginning of the relevant text. We tackle these issues by developing and using several evaluation measures. In particular, we interpret the entry point issue in quantitative terms by defining error measures that look at the number of relevant paragraphs missed at the start of the segment (and similarly: the number of irrelevant paragraphs added at the start of the excerpt).

2.2 Handbooks

For most disciplines a time arrives where a need is felt to write a handbook[3] gathering together the knowledge about the domain and passing it to the students. Handbooks can vary in length, organization and style. In the case of comprehensive handbooks dedicated to broad areas, it is not uncommon to have more than one author. An important use of handbooks is as a reference tool, for quick look-up of notions, bibliography checking, often in conjunction with other activities, such as essay writing. These characteristics make the handbook a genre suitable for electronic publication, exploiting the potential of search in a digital environment and facilitating integration with the document writing – and activity more and more directly done on screen. Another reason for exploiting the electronic publication of handbooks and access to them is related to their long publishing cycle, as a result of which new, revised editions become available only after several years, even though the field they cover may develop at a very fast rate. So, an environment for electronic publication of handbooks would be especially valuable, as it would decrease the publication time as well as the time and cost for updates. But in order to have a sound publication environment, we also need effective ways to access such texts, since page browsing is not convenient in electronic documents.

For documents on paper, including handbooks, a number of mechanisms are available for accessing the information contained in them: tables of content, indexes of names, topics, figures, tables and so on. The table of contents refers to the internal organization of the handbook, while an index helps locate relevant topic (figures, tables, names, ...) in the text. Both tables of contents and indexes use pages to locate information and, in the case of indexes, only mention the *important* pages where a *term* appears: they do not mention all terms, nor all occurrences of them, nor any finer mechanism than the page to indicate *where* the term is discussed in the document. We envision a more comprehensive way to retrieve topics in a electronic handbook, one that extends the functionalities of the back-of-the-book index.

[2] The notion of entry point has received a lot of attention in the setting of structured document retrieval; see [Reid et al., 2006].

[3] There is no clear distinction between handbooks and manuals, but manuals are often oriented toward procedural knowledge, while handbooks are usually dedicated to more theoretical issues.

2.3 Searching and Accessing Textual Documents

Traditionally, e.g. in a library or archive, accessing a book has meant retrieving a catalog card that contains a pointer to the "real" document. The catalog card is part of some inventory or catalog, based on *classification*, i.e., a grouping of similar or related things and arrangement of the resulting groups in a helpful sequence. Alternatively, the catalog may be an *index*, i.e., a list of words of a book so that the book they refer to can be easily found. Classifications and indexes can be organized into hierarchies such as classifications systems and thesauri (we look into these in Section 2.4).

The first retrieval experiments were performed on machine readable catalog cards and abstracts in the 1950s: there we can place the birth of computerized Information Retrieval (IR). By storing the catalog in a database system, it is possible to speed up the process of searching for the catalog card, both by searching through the description metadata and the content metadata (manually assigned indexing terms).

With today's availability of full length documents, as opposed to abstracts, keywords, or index terms, comes a more flexible indexing methodology where virtually any word can be an index term, and where virtually any term can in principle be used in query formulation. The possibility of automatically indexing the entire document allows one to retrieve a document not only on the basis of the metadata available, but also on its *content*. Another consequence of the availability of full text documents and indexes is that it is possible to access the content *below* the document level—an area in (textual) IR that has received a lot of attention in recent years (through tasks such as passage retrieval [Salton et al., 1993a], XML retrieval [INEX, 2005], question answering [Maybury, 2004], and entity retrieval [Sayyadian et al., 2004]) and one to which we will return frequently in the thesis.

The main components of an information retrieval process are: the user, an information need expressed as a query, a retrieval model and a document collection. Let us start from the end. A document in a collection can be represented in many ways, by means of its title, or editor (author), or a combination of these or other pieces of information such as abstracts and keywords. In modern information retrieval it is common to represent a document by means of an automatically generated *inverted index*, i.e., a list of the words occurring in the document associated with their position in it. This data structure facilitates fast search of index terms and does not presuppose any predefined classification system nor any knowledge of classification or criteria for manual indexing on the side of the user. The same type of representation is applied both to the document and the query.

Many IR models have been studied and implemented over the last five decades, including logical models, vector space models, probabilistic models and models based on language modeling; see, e.g., [Baeza-Yates and Ribeiro-Neto, 1999, Grossman and Frieder, 2003]. In this thesis we adopt the classic vector space model, where documents and queries are represented as vectors of words, where words can be weighted according to some criteria, such as their frequency, or a combination of term frequency in the document and its frequency in the collection. In a vector space model, the co-

sine of the angle formed by the document vector and the query vector represents the degree of similarity between them. Therefore, documents in the collection are ranked according to the value of the cosine of the angle they form with the query. Since documents may be lengthy and may cover more than one topic at different levels of details, there have been attempts to look inside a document and use *evidence* from parts of it to better assess whether the document is relevant to a query; see [Wilkinson, 1994] for an early example. In the course of this thesis we look at various ways to split a document according to the topics it covers and how to link the resulting excerpts to a map of the domain.

Very early on, the IR community addressed the issue of how to evaluate a system and how to compare systems to one another [Cleverdon, 1970]: over the years a number of measures have been proposed, each capturing different aspects of the retrieval process. The most popular quantitative measures are certainly *precision* and *recall*, where precision looks at the proportion of retrieved documents that are relevant, and recall gives the proportion of relevant documents that are retrieved. Other measures have been designed that take into account the ranking proposed by the system, others are more explicitly user oriented ([Cooper, 1968, de Vries et al., 2004]), and still others have been designed to evaluate specific retrieval tasks. The user plays a central role in a special branch of IR research called Interactive IR [Ingwersen, 1992, Robins, 2000], where the focus is on the interaction of the user with the IR engine: how she expresses information needs into queries, how she evaluates search results, and how she issues new queries. In the view put forward in this thesis, the LoLaLi map supports information seeking activities such as searching and browsing, and the issue of providing focused access inside a handbook is treated as a special retrieval task—for this, we follow the methodologies developed in the IR community. We also propose our own measures in order to quantify aspects that are especially relevant for us and for the reader of the retrieved excerpts; see Chapter 5.

2.4 Semantic Structures

Full text indexing is one way of describing the contents of a document. Another way to describe a document (and so to find it among many others) is to use words reflecting its content. Such words are called *keywords*, that could in principle be rare or absent from the document. In order to minimize the chances of mismatch between the keywords the user would assign to the document and the keyword actually used by the (human) indexer, it is common practice to define a set of admitted keywords, usually called a *controlled vocabulary*. For example, indexers can agree, or be compelled, to consistently use the keyword ‘computer science’ even when they would have preferred ‘informatics.’ Controlled vocabularies ensure uniformity of indexing, which should imply high accuracy at retrieval time.

The advantage of keyword indexing by controlled vocabulary is that, assuming that the user is able to find the appropriate keyword to describe her information need, she is

automatically able to find all documents described by that keyword. The disadvantages are that it forces the user to "guess" the appropriate index term, and that the index terms are taken as independent, with no explicit relationship among them. The latter problem is (at least partially) overcome by structures such as taxonomies and thesauri, which enhance a controlled vocabulary with relations.

Taxonomies and Classifications

Taxonomies refer to the classification of objects (or any kind of entity) into categories[4] and in principle the classification could be based on any kind of law or principle. In practice the most common taxonomic relation is: 'A is a kind of B,' whose corresponding relation of exclusion is 'C is a different kind of B' (cf. Section 3.4.1).

A taxonomy is a very versatile structure, used to represent human knowledge (e.g., the tree of knowledge of Raymond Lull, Figure 2.2), regularities found in nature (e.g., the taxonomy of species by Carl Linnaeus (1707–1778)), and subjects for the purpose of book classification (e.g., the Dewey Decimal Classification system (1876)).[5] A taxonomy can also use a parthood relation, in that case it is called a *meronymy*.

Other classification systems, such as the ACM Classification System[6] [ACM, 1964], are specific to a given domain. The ACM classification system (Figure 2.3) is used to index scientific papers in the area of computer science. It is based on a hierarchy (a tree) of terms organized in four levels, of which the 11 nodes at the first level are never used for classification, while the nodes of the levels below are actually used for indexing the articles published by ACM in order to enhance search for papers in the collection of their publications.

Thesauri

In literature, thesauri are dictionaries in which each entry is listed together with its synonyms (e.g., *Roget's International Thesaurus of English Words and Phrases* [Roget and Davidson, 2002]), but the usage of the term thesaurus here, now widespread, dates from the early 1950s in the work of Luhn [1958], who sought ways to automatically create a list of authorized terms for indexing scientific literature. The list was to include a structure of cross-references between families of notions. Usually, though, specialized thesauri are organized by relations forming a hierarchical structure (a tree) among the terms in it. The typical pair of hierarchical relationships in a thesaurus is *broader term* (BT) and *narrower term* (NT), while non-hierarchical relations include

[4] As the ancient Greek etymology suggests: 'class' + 'rule.'

[5] Although there is no agreement on the historical origins of the DDC, Dewey himself openly refers to the Baconian classification [Wiegand, 1998]. According to Bacon, human knowledge depends on three main faculties: Memory (History), Imagination (Poetry/Art) and Reason (Philosophy). Dewey adopts an inverted Baconian system of classification, where Reason takes the classificatory numbers from 1 to 6, Imagination 7 and 8, Memory 9. General works take the number 0.

[6] The first version dates back to 1964, the second version, totally different from the first one, was issued in 1982, then other versions based on that followed. The most recent is from 1998.

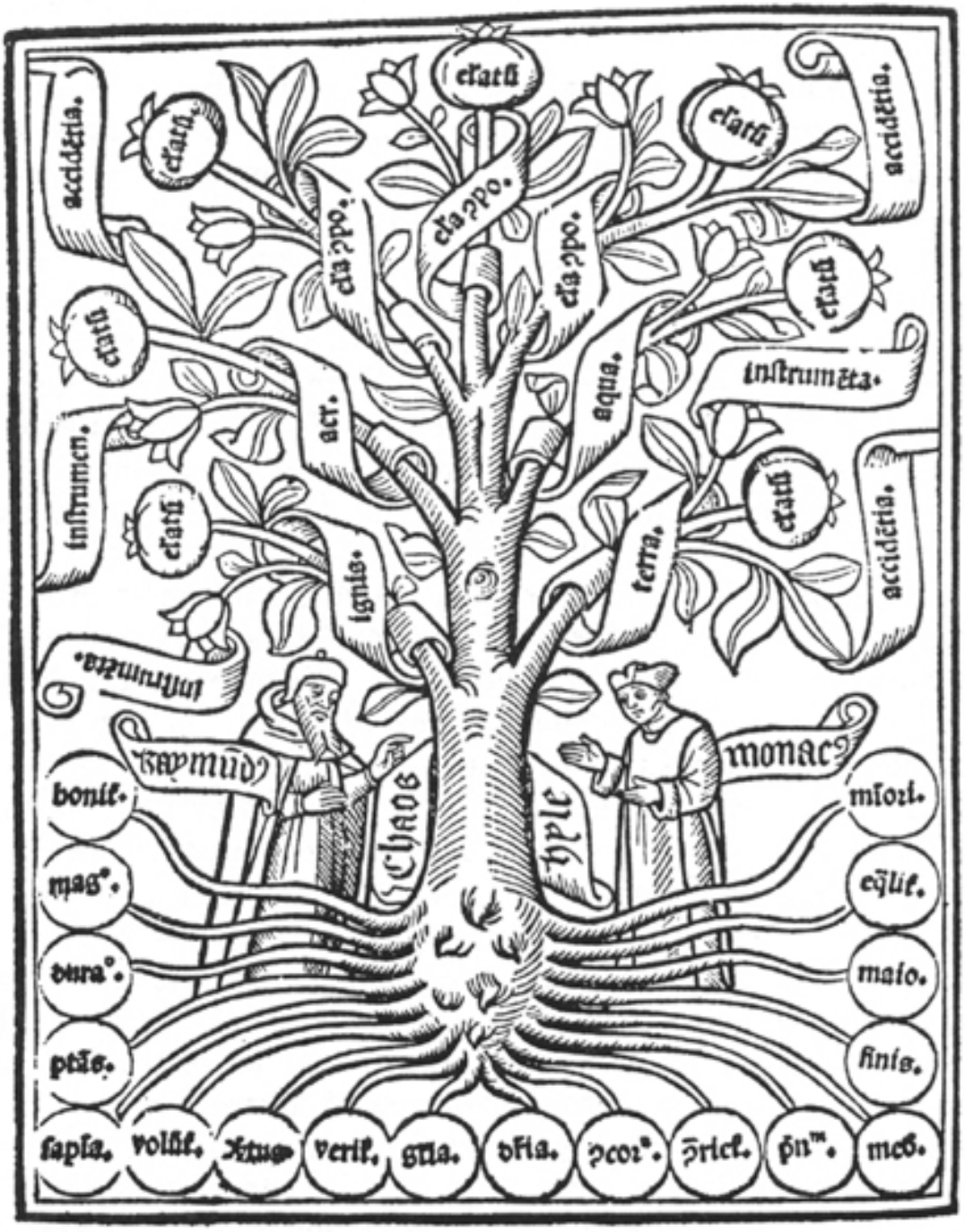


Figure 2.2: A medieval representation of the Arbor Scientiae (tree of knowledge) by Ramon Lull.

used for (UF, which marks a *preferred term* or *descriptor* to be used for indexing instead of *non-preferred terms* or *non-descriptors*; the inverse relation is *use* (USE)) and

The ACM Computing Classification System (1998)

- A. General Literature
  - A.0 GENERAL
    - *Biographies/autobiographies*
    - *Conference proceedings*
    - *General literary works (e.g., fiction, plays)*
  - A.1 INTRODUCTORY AND SURVEY
  - A.2 REFERENCE (e.g., dictionaries, encyclopedias, glossaries)
  - A.m MISCELLANEOUS
- B. Hardware
  - B.0 GENERAL
  - B.1 CONTROL STRUCTURES AND MICROPROGRAMMING (D.3.2)
    - B.1.0 General
    - B.1.1 Control Design Styles
      - *Hardwired control* [**]
      - *Microprogrammed logic arrays* [**]
      - *Writable control store* [**]
    - B.1.2 Control Structure Performance Analysis and Design Aids
      - *Automatic synthesis* [**]
      - *Formal models* [**]
      - *Simulation* [**]
    - B.1.3 Control Structure Reliability, Testing, and Fault-Tolerance [**] (B.8)
      - *Diagnostics* [**]
      - *Error-checking* [**]
      - *Redundant design* [**]
      - *Test generation* [**]
    - B.1.4 Microprogram Design Aids (D.2.2, D.2.4, D.3.2, D.3.4)
      - *Firmware engineering* [**]
      - *Languages and compilers*
      - *Machine-independent microcode generation* [**]
      - *Optimization*
      - *Verification* [**]
    - B.1.5 Microcode Applications
      - *Direct data manipulation* [**]
      - *Firmware support of operating systems/instruction sets* [**]
      - *Instruction set interpretation*
      - *Peripheral control* [**]
      - *Special-purpose* [**]
    - B.1.m Miscellaneous
  - B.2 ARITHMETIC AND LOGIC STRUCTURES
    - B.2.0 General

Figure 2.3: A fragment of the ACM classification schema (1998).

related term (RT) [ISO:2788, 1986]. Sometimes the relation *equivalent term* is also used, to express a relation of synonymy. For the sake of uniformity, only descriptors are used for indexing. Monolingual thesauri are usually a strict tree (each term has only one broader term), but multilingual thesauri often allow multiple parenthood.

Thesauri are widely used for indexing and cataloging in library and information sciences, especially when dealing with restricted domains where a high degree of detail is required. A wealth of organizations have developed their own thesauri, in areas as diverse as engineering [Aitchison, 1970] and art [Peterson, 1994], up to the point that from the 1960s throughout the 1980s thesauri became *the* tool for document indexing and occupied a leading role in the area usually called knowledge management. In fact, both standards for the creation of thesauri, i.e. for monolingual and multilingual, maintained by the International Organization for Standardization (ISO) date back to the 1980s [ISO:2788, 1986, ISO:MLT, 1985]. At the same time, the organization also approved a standard for indexing [ISO:IND, 1985].

Thesauri have also been used in IR since the very beginning of the field for the so-called "thesaurus approach" to information retrieval [Joyce and Needham, 1958, Salton, 1968]. The idea was to combine classification with indexing by grouping words together in "notional families." These families of grouped keywords are considered to define a classification. By using a thesaurus provided with relations of BT/NT one could also make an inclusive search, meaning the possibility of automatically propagating the assignment of keywords by means of the hierarchical structure of the thesaurus. The expected advantages are "easier" indexing for the indexer (because there are more indexing terms to choose from) and "easier" retrieval for the user (since indexing terms are grouped together and organized in a structure and there are more chances for the retrieval to be exhaustive). In IR, thesauri have also been used for the automatic enhancement of queries (an application usually called query expansion), a method that has proven to be useful when queries are short (2–3 words) [Voorhees, 1994].

Semantic Networks and WordNet

Recently, thesauri and thesauri-like structures have enjoyed a renewed interest in the area of the Semantic Web [Berners-Lee et al., 2001]. Before touching on the Semantic Web below, we introduce Semantic Networks and WordNet, historical antecedents of the semantic web.

A *semantic network* is a graphic notation for representing knowledge in patterns of interconnected nodes and arcs [Sowa, 1991]. In practice, it is a directed graph consisting of vertices representing concepts and edges representing relations between concepts. Semantic networks can be more complex than thesauri in that they often admit multiple parenthood, and can contain virtually any relationships (e.g., 'instrument for,' 'mother of,' 'affected by,' ...), but usually concepts are organized according to levels of generality based on taxonomic distinctions. Also, properties and relations of the nodes in the network are usually expressed in some formal language.[7]

[7]The origin of semantic networks can be traced back to Charles C. Peirce (1839–1914), who devel-

Semantic networks have been intensely studied and developed between the 1960s and 1990s within the framework of computer science, artificial intelligence, knowledge representation and logic, where emphasis was put on the *reasoning* capability of the system. The most common type of reasoning implemented was the inheritance mechanism (relations that hold for all concepts of a given type are inherited through the hierarchy).

WordNet [Miller, 1995], usually defined as a "a lexical database of English," is perhaps the best-known example of a semantic network.[8] WordNet distinguishes nouns, adjectives, verbs and adverbs and uses different hierarchical relations for each group. The basic non-hierarchical relation is the lexical relation *synonymy*: each node in the graph consists of a set of terms, a *synset*, that are synonyms in a certain context. Any pair of synsets in the graph is connected by one of the following relations: hypernymy/hyponymy, coordinate terms (i.e., siblings under the same hypernym), holonymy/meronymy (i.e., has part/part-of). Each synset is also endowed with a gloss to explain the exact meaning of the terms in it, added because the presence of synonyms in the same synset does not always allow one to disambiguate the term.

The distinctive feature of WordNet is that it does not use preferred terms (as opposed to classical thesauri where preferred terms ensure uniformity of indexing), because it groups together partial synonyms (words that are interchangeable in some context) in synsets. A synset, as a whole, provides a notion of *meaning* of a word. This structure makes WordNet a *sort of* dictionary (as glosses aim at giving the *sense* of a synset, not the definition of a single term), and also a sort of ontology (see below).

Beside its importance as a psycholinguistic experiment, WordNet has had great success in computational linguistics and in information retrieval applications. WordNet is also widely used for enriching ontologies and thesauri for indexing. In computational linguistics applications WordNet is especially used for word sense disambiguation, in IR as a tool for query expansion [Voorhees, 1994]. Other work has concentrated on the study of conceptual distances among words for purposes such as opinion detection in political speeches [Kamps and Marx, 2002]. Currently, dozens of WordNet projects are being carried out around the world, including MultiWordNet [MULTIWN, 2005], Indi and Marathi WordNet [IWN, 2005], EuroWordNet [EUWN, 2005], and Cornetto [Vossen et al., 2007].

As we will see in Chapter 3, the LoLaLi map has adopted some of WordNet's features, such as a gloss attached to each topic and a synset-like way of grouping of terms.

---

oped en *existential graph* as a graphical system of logic.

[8] WordNet was created in 1985, and is still actively maintained at the Cognitive Science Laboratory of Princeton University under the direction of psychologist George A. Miller, to test the hypothesis that the human memory groups words together according to their function in the language, or linguistic category. Miller's interest in human memory traces back to the 1950s, when he wrote his seminal paper: "The Magical Number Seven, Plus or Minus Two" [Miller, 1956].

Ontologies and the Semantic Web

The latest incarnation of semantic structures, usually called *ontologies*, has become a crucial ingredient of the Semantic Web [Berners-Lee et al., 2001, van Harmelen and Antoniou, 2004, W3C, 2006]. The idea behind the Semantic Web is that by associating metadata with objects on the Web they can be better retrieved, it is possible to achieve interoperability and exchange of data, and to perform automatic inference (as opposed to inference commonly performed by human agents reading text in natural language). In such a scenario, ontologies are crucial because they can be used to formalize the semantics of the metadata used to annotate objects in the Web.

In practice, there is a loose usage of the term ontology. It can indicate different structures, with big variations in specificity and complexity, such as controlled vocabularies, glossaries, thesauri, term hierarchies, strict subclass hierarchies, frames and value restrictions (see [McGuinness, 2002] for a short survey of them). However, a popular definition is given by Gruber [1993, page 5], according to whom an ontology is "a specification of a conceptualization." This definition corresponds to the view adopted in computer science, where an ontology consists of a set of concepts connected by relations describing a domain of interest [Vickery, 1997]. Nowadays, though, the tendency is to consider ontologies to be structures that are at least as complex as a taxonomy[9] and encoded in a formal language such as RDFS [RDFS, 2004] or OWL [OWL, 2004] that allows some kind of inference.

The key role of ontologies for the Semantic Web has fostered a wealth of research: on standardization of formal languages to encode ontologies, on tools to edit, manage and visualize them, on methods to reuse ontologies and share them, on software interoperability through metadata, and more. Although a deeper discussion about the Semantic Web is beyond the scope of this thesis, we looked at the area for technologies and tools that could help us in our effort to build and maintain the LoLaLi map; see Chapter 3 for details, where we also discuss the extent to which the level of formality often required by semantic web-based solutions is appropriate for our domain and tasks.

2.5 A Modular Approach to Focused Access

The work of Harmsze [2000] (see also [Kircz and Harmsze, 2000] and [Kircz, 1998]) relates to the model we hinted at in Section 2.1, in that she proposes a modular model for the electronic publishing of scientific articles. Her idea is that in an electronic environment, scientific information may be communicated more efficiently and effectively if it is presented as a network of articles with a modular structure rather than as a set of linear, essay-type articles. The model would provide focused access to information and enable re-use, better retrieval and clarity.

[9] In fact, the word ontology is often used as a synonym for taxonomy, although an ontology can use a broader set of relations than the subclass relation.

Harmsze's work is grounded on the analysis of the argumentative structure made by Grootendorst et al. [1996], and on the basis of that she provides a multi-dimensional classification of modules typical of certain types of scientific articles (e.g., hypothesis, data, experiments), and of their aggregation. She also provides a multi-dimensional classification of links: both of those that aggregate simple modules together, and of those that connect aggregated modules to one another. An article with a modular structure then consists of a coherent collection of explicitly linked modules, representing a coherent network of related conceptual information units within the larger network of published information. This structure allows the user to take into account the role that concepts of interest should play within a document, as well as the relations between specified concepts.

Harmsze defines requirements for relations and modules, and applies the model to the modularization of two articles. Such conversion turns out to be possible, although very difficult, because of the difficulty of extracting modules from originally paper-based sections. In particular she found that: (1) some pieces of information were inherently difficult to characterize unambiguously, which led to overlap of different modules; (2) even when the characterization of different strands was unambiguous, some information of different types was so closely interwoven that it was difficult to disentangle it; (3) in order to obtain a complete module, it was necessary in some cases to add extra information not provided in the original version. Her study concluded that scientific articles written for different media need a different structure. Therefore, she developed extended guidelines for writing modular articles from scratch.

Contrary to Harmsze, who started from the documents and concentrated on modules within the texts (articles on experimental sciences) or among them (e.g., a set of articles resulting from the same project), we start from the map, focus on the modeling of a map of the domain and then address the issue of providing an automatic way to link excerpts from the document to the map. In this sense, our approach is more general than the one described in [Harmsze, 2000], but some overlap is evident, such as in the types of relationships found. For example, the relation `mathematical result` (in the LoLaLi map, cf. Chapter 3) is reminiscent of the module 'results' in Harmsze's approach. Despite these similarities, though, Harmsze focuses more on argumentative or organizational relations than on domain relations. For example, whereas Harmsze considers relations (among others) such as 'Elucidation relations,' 'Clarification relations,' and 'Explanation relations,' we concentrate on relations such as 'part of,' and 'mathematical results.'

2.6 Discussion

The vision presented in Section 2.1 aims at enabling focused access to information, while supporting information seeking activities that include searching, browsing and exploration of a domain map with explicit relations between topics. This approach should support different information needs, ranging from vague to precise. In the

course of this chapter we have presented our vision on accessing electronic handbooks, and we connected it to the areas in which our approach is grounded.

We acknowledge the importance of semantic structures to provide an overview of a subject area, and add to it the possibility of searching at various levels: the map level, the *Handbook* and the links connecting them. In the course of this thesis we concentrate on the making of the map, on the selection of links to connect to it; more details about the search on the map can be found in [van Hage, 2004].

Research in Information Retrieval started with search on catalog cards and ended up generalizing the concept of indexing to the point that inverted indexes are now the data structure of choice in the area. Full indexing allows one to search on virtually any word in a document, but it relies on the ability of the user to issue the "right" query, and may force her to issue different queries to capture documents written according to different styles and vocabularies. Structures like thesauri work on the idea that optimal search and retrieval can be achieved by looking at the content of the document, or its *meaning*. We called these structures *semantic* so as to emphasize their ambition to represent the meaning of words and the meaning of documents. We stressed the fact that classification and grouping is very intuitive and was used very early on to represent documents and subject areas. In the context of electronic search and reading environments, though, systems solely based on this approach suffer several major limitations, including the ability of the indexer to capture *all* meanings of a document, and the need for the user to have some knowledge of the classification system adopted in order to "guess" the right classification or keywords.

Against this background, the LoLaLi map we present in the next chapter was inspired by the structures we presented in this chapter (especially WordNet) but with the main purposes of providing a map of the domain and a bridge to the document. As will become clear from Chapter 3, we took from WordNet the use of synsets and some relations. From the Semantic Web area, we took some of the tools we used to implement the LoLaLi map.