



ELSEVIER

Data & Knowledge Engineering 31 (1999) 155–181

**DATA &
KNOWLEDGE
ENGINEERING**

www.elsevier.com/locate/datak

Formal ontology for subject

Christopher A. Welty *, Jessica Jenkins

Computer Science Department, Vassar College, Poughkeepsie, NY 12604-0462, USA

Received 2 June 1999; accepted 3 June 1999

Abstract

Subject-based classification is an important part of information retrieval, and has a long history in libraries, where a subject taxonomy was used to determine the location of books on the shelves. We have been studying the notion of subject itself, in order to determine a formal ontology of subjects for a large-scale digital library card catalog system. Deep analysis reveals a lot of ambiguity regarding the usage of subjects in existing systems and terminology, and we attempt to formalize these notions into a single framework for representing it. © 1999 Elsevier Science B.V. All rights reserved.

Keywords: Subject-based classification; Digital libraries; Description logics; Card catalog system; Formal ontology

1. Introduction

There can be no argument that the wealth of information now available on-line has created numerous problems for information retrieval. One consequence of the information age is the promise of Digital Libraries, with on-line access not just to library catalogs, but to the documents themselves. Furthermore, progress in the text encoding community offers the potential for significant amounts of these documents to exist in some marked-up form (see, for example the TEI [30], the CES [22], the MEP [8], and the WWP [13]).

Having information about the content of a document readily available in a machine-readable format may seem to imply the ability to do document retrieval based on this content information; the community, however, has yet to deliver tools that exploit the markup in a manner that justifies the effort required to create the documents [22]. We have been performing ontological analysis in the domain of library catalog systems as part of a project endeavoring to address this problem.

Many aspects of the library ontology presented here would not seem very novel to anyone experienced with conceptual modeling and ontology building (in fact it is based on Tom Gruber's bibliographic ontology [16]). It is, however, a radical departure from the current paradigm in library catalog systems whose foundation is quite clearly the need to *place books on shelves* in an orderly way. Modern library scientists may find this an over-simplified and out-dated criticism of

* Corresponding author. Tel.: +1-914-437-5992; fax: +1-914-437-7498; e-mail: weltyc@cs.vassar.edu

current library access methods, but we hope to show in this paper that it is in fact still quite relevant.

The most interesting scholarly work that has been ongoing for many years in this effort is a deep and thorough analysis of the ontology of subjects [38]. Subject classification can be viewed from a variety of perspectives ranging from the purely pragmatic to the purely philosophical. The notion of a subject is so deeply ingrained in all aspects of our culture and language that this part of the work has been particularly fruitful and challenging. The bulk of this paper, therefore, will deal with the issues of the representation of subjects, classification by subjects, and subject-enhanced retrieval.

The paper begins with a discussion of the goals of the project, and the goals of the ontology work, then presents the new ontology for library card catalog systems. The remainder of the paper will be devoted to the ontological nature of subjects. Elements of this paper have been published previously in [40].

2. The axiom project

The Axiom project seeks to combine four different aspects of research into a single digital library: text encoding of non-technical (i.e. humanities) documents [8], linguistic text analysis [20], knowledge representation [41], and user interfaces [6].

2.1. Library data

The center of the project is the Model Editions Partnership (MEP), maintainers of a large repository of fully marked-up electronic versions of historical documents from the US Civil War period [8]. The documents include letters, manuscripts, newspapers, and diaries, as well as the more traditional books and magazines. The markup of the electronic documents is quite deep; in addition to structural components (paragraphs, line breaks, font changes, etc.) and basic header tags (title, author, date), the body of the documents include: identification of names, dates, cross-references to other documents, events, etc. The markup is done mostly by hand, and the development of automated tools for assisting these encoders is a topic of current research.

The existence of the data in this structured form opens the door, at least intuitively, to the possibility of dramatically more expressive queries than library users are familiar with today, because it enables retrieval based on the contents of the documents. Web-based keyword searches also permit this, yet they have no (or very little, at best) semantics, and cannot make the simplest differentiations between word sense usage. The markup of names is of particular interest in this regard, because in addition to being tagged as names, the tags also differentiate person names from place names, organization names, etc. In other words, we expect to find markup such as:

```
The city of <name type = city>Washington</name> was named after the first
American
President, <name type = person><firstname>George</firstname>
<lastname>Washington</lastname></name>
```

Clearly this encoded information allows queries to be expressed that differentiate between Washington the place, and Washington the person, a differentiation that keyword search does not allow.

2.2. Project goals

We wish to use the richness of the encoded data to enable multiple modes of retrieval of the documents. We feel that one key ingredient to accomplish this is expanding the “card catalog” representation beyond books to support far more expressive queries, such as “Publications written by people at Vassar College”, or “Works of fiction reviewed by someone at Vassar who is interested in AI”, or even, “Papers on AI published at a conference sponsored by ACM”. The increased expressiveness should serve to enhance the utility of the card catalog for three basic kinds of users.

1. *Users with vague notions about what the object of the search is.* Library users frequently do not know the title or author of the book they are looking for, or even if it exists. For example, someone may be interested in reading about “Formal Ontology”. Such a user has no idea what the relevant titles or authors are. During a search for a publication on that subject, if none were found, the user could generalize the search to, perhaps, “Knowledge Representation”. The key to generalizing (or specializing) a search in this manner is access to the subject taxonomy, which is normally not the case in existing card catalog systems. Furthermore, if the user was aware that interesting research in formal ontologies was taking place at a particular institution, the user might search for “Publications on knowledge representation written by someone at Vassar College”. Again, since we are dealing with a library containing potentially billions of publications, the ability to refine a search will be critical in making the information accessible [39].

2. *Users doing scholarly research.* We are working with two text-encoding groups who are in the business of marking up historical manuscripts and making them available in electronic form. Their primary customers are scholars who do not have direct access to the manuscripts (because of proximity, because the manuscript is fragile, etc.), and having the data in electronic form opens up vast new possibilities. Etymologists (people who study words) would like to make queries such as “what is the date of the first manuscript that uses the word ‘cleave;’” other researchers might ask for “Books by women authors in the 18th century containing mythological characters”. It would be impossible to exaggerate the implications of this technology for scholarly research in the humanities.

3. *Users with vague recollections of the object of the search.* It is not uncommon for a library user to be searching for something they read or heard about at some time, but do not remember exactly what it was. The more information about each publication that can be provided, the better the chance that something the user remembers is actually recorded. For example, “The paper on KR written in the early ‘80s by someone at BBN”, or, “A paper on planning at an AI conference in Florida”. This category was, in fact, the initial motivation for this research [38].

Drawing from these user types, we have developed a new paradigm for library search and retrieval, one that tightly integrates browsing and querying of an information space that represents more than simply books, but the people, places, events, etc., that are useful background information for supporting the retrieval process.

2.3. New ontology for library information systems

Every library card catalog system operates on the basic assumption that the object of a query is a document. When one types “FIND SUBJECT ONTOLOGY” in a library card catalog system, that system understands this to mean, “find *all books* whose subject is ontology”. Our new paradigm does not change the abstract interaction; the result of any search will still be a book (or, more generally, a document). The basic revelation of this new paradigm is the recognition that in a digital library, the object of a *specific query* may not be a document, but a person, place, event, or other kind of object. These other objects, which represent what we term the background knowledge, extend the information space defined by the catalog and add an important step to the normal interaction between a user and the catalog system: browsing. We expect users to engage in an iterative process of querying for information, using the query results as a starting point for browsing, and using their browsing results to guide them in making new queries. This process repeats until the desired documents are found. An example is provided in Section 2.4 below.

This new paradigm requires a representation that allows for the background knowledge to be found, either through queries or browsing, and we have developed a simple ontology for supporting this new style of interaction. We present the ontology here only briefly; a more complete description can be found in [38,41].

Our ontology consists of five principal classes: documents, document modalities, objects, events, and subjects. The objects include people, places, companies, organizations, etc. The events include conferences, meetings, wars, battles, etc. A subset of the taxonomy for this ontology is shown in Fig. 1. The subjects are described later.

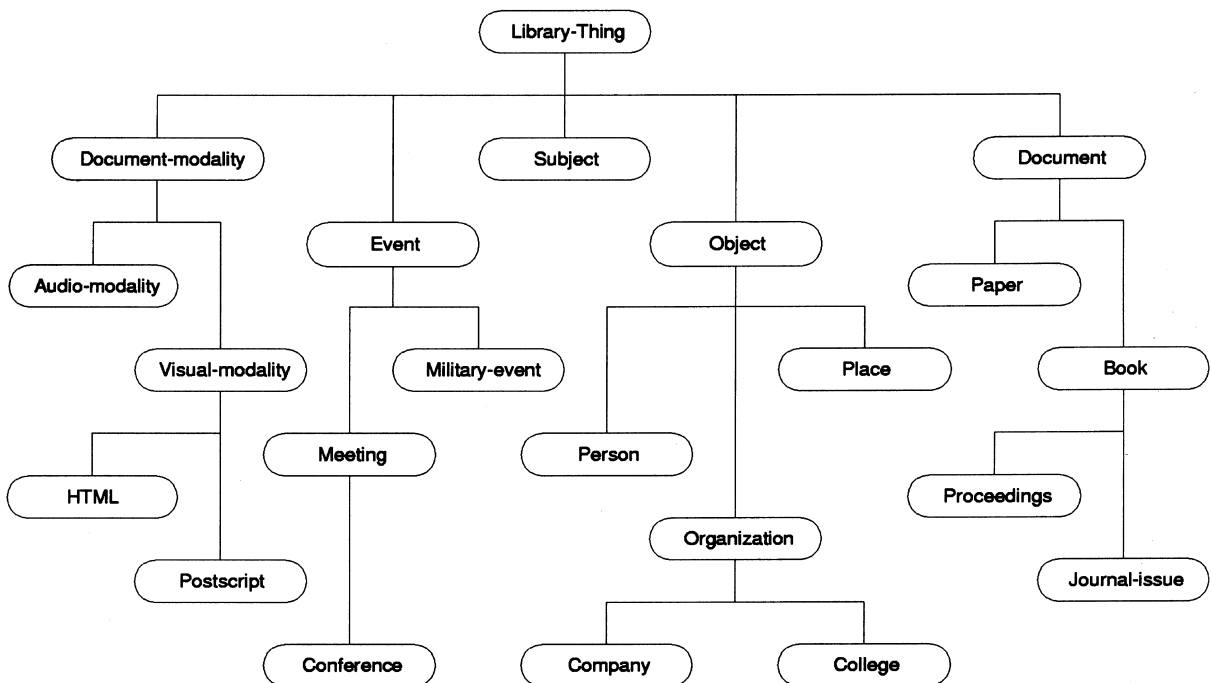


Fig. 1. Some of the key concepts in the electronic card catalog ontology.

Individuals of the non-subject classes have numerous relationships through which they can be connected. A Person can be the *author-of* any document, and the mere existence of such a link implies that the person is also an author. Organizations (such as “AAAI”) can *sponsor* events (such as “AAAI-97”), people can be *members* of organizations (“Lincoln was a *member* of the US Government”), and can *participate* in events (“Sherman *participated* in the burning of Atlanta”). These links or relationships provide the basic structure for browsing the information space. In our interface, when looking at information about an object (e.g. “Lincoln”), clicking on a link (e.g. *member-of*) will move you to the object linked to (e.g. “US Government”).

It should be noted that we do not wish to create an encyclopedia here, only enough information and background material to enhance the on-line browsing process and make the sheer magnitude of digital library information accessible, and useful.

Instances of documents, events, and objects may be classified by subject. For a person, being classified by a subject indicates interest in that subject. For an organization, it means an official association with the subject, not necessarily simply the propagation of the interest of its employees or members (Chris is interested in cars, but Vassar College is not). Subjects should be propagated along certain relationships, but that will be described later.

Finally, we also represent the tags in the markup, and relationships between them from a semantic (not syntactic) basis. We augment the existing tagset with other “virtual tags” (tags that are not present in the markup but can be derived automatically) in the representation. This allows us to process a variety of background information. We have split our background information ontologies, which are fairly simple objects and rules regarding the types of texts that are encoded, into different subject areas as well.

2.4. User scenario

We believe the principal advantage our approach offers is the improvement of searching and browsing for information retrieval. The added knowledge, resulting from both a principled approach to ontology development and the background information that ties the meanings of the tags together, makes it possible to more precisely specify a query.

Imagine, for example, a scholar who is motivated by recent events to research how commonplace it was during the civil war period in American history for government officials to mention government business in their personal correspondence.

The historian would begin by finding a digital library that includes marked-up versions of historical documents, such as the MEP’s Civil War era documents [8]. He would then enter the query, “Find all government officials during the years 1860–1865”, and be presented with a list of all the government officials the system knows about. The system knows about government officials because the markup includes tags that identify government documents, and in one of our historical ontologies we have a rule that says, “The author of a government document is a government official”. There are similar rules for names that appear in tags that signify senders, recipients, signatures, etc. The system can also infer dates of service for these officials from the dates of the documents.

The power here comes from the fact that it is fairly easy to specify such rules, and then capitalize on the data already in the marked-up documents. Thus, rather than enumerating all the government officials during the Civil War, we specify the rules and let the system gather that information for us. We have found an enormous wealth of information in these marked-up

documents that can be exploited in this way, such that the information itself is more accessible. Note that, of course, this information may be incomplete, e.g. if the library does not include documents in which a particular government official is mentioned in the right context (i.e. in an author, recipient, etc., tag), the system will not know that person is a government official.

The user's next step is to pick a person from the list of government officials, and display the information the system knows about that person. This will reveal to the user the kind of information the system keeps about individual people, or in this case about government officials. The user will find links called "author-of, recipient-of, sender-of, member-of", and many others. From these links, this particular historian would check the "author-of" link, and be presented with a list of all the documents this person was the author of. The documents are listed along with their most specialized parent (i.e. the parent concept lowest in the concept taxonomy), so the user will see a list of document titles and document types, for example:

```
PERSON-102: Andrew Johnson
AUTHOR-OF: DOCUMENT-23: Public Address to Baltimore
MEMO-54: Message to Lincoln
LETTER-32: Letter to Mrs. Johnson
```

The user now selects LETTER-32, and sees all the information that the system keeps about such a document, including links like "has-author" and "has-recipient", and all the parent concepts (the list view above reveals only one parent, but most instances will have many parent concepts inherited down through the taxonomy from all its immediate parents). LETTER-32 is an instance of a LETTER, PERSONAL-DOCUMENT, HISTORICAL-DOCUMENT, DOCUMENT, MEP-DOCUMENT, and several other concepts. The user notices the existence of the PERSONAL-DOCUMENT concept, and selects it.

All concepts in the ontology have brief natural language descriptions, so the user will see something like what is below. This information is pruned a bit from the actual output:

```
Concept PERSONAL-DOCUMENT:
Comments: 'A Personal-Document is a document, usually a letter or diary entry, that was considered part of the author's personal life. This concept was originally disjoint with PROFESSIONAL-DOCUMENT, however we have found several exceptions and removed that restriction.'
Parents: DOCUMENT
Ancestors: DOCUMENT OBJECT LIBRARY-THING
```

The user now has learnt all the information he needs to form a more specific query, which is essentially, "Find all the PERSONAL-DOCUMENTs written by GOVERNMENT-OFFICIALs in the years 1860–1865 whose recipients were not GOVERNMENT-OFFICIALs". The result of this query may be quite large, but in fact is precisely what the user is looking for, and finishing the research will involve reading through these documents.

Eventually the scenario above would be augmented by a powerful interface that assists users with the query language. Until that time, and probably even after it, we expect that trained librarians who understand the representation and the query language will assist users with this

system. Note that, of course, modern communication technology, from phones to chat rooms to 3D web spaces, imply that the human assistance does not require proximity, and the goal of making these new digital libraries accessible to the world is still achieved.

There are many example user scenarios that better detail the extraordinary advantages we believe can be gained by applying knowledge representation and ontology techniques to digital libraries of marked-up texts in [41].

3. History of subject classification

Perhaps the most interesting part of the scholarly aspect of this research has been a study of the history of classification, and in particular the use of subject classification in libraries. One could easily spend a lifetime trying to catch up on everything that has been written on the idea of classification. The history of this area, as with all histories, is ripe with colorful characters, personality conflicts, love triangles, deep political intrigue, valiant and selfless acts for the advancement of knowledge, and many other factors which, in addition to being interesting, are also relevant to understanding the meaning of any piece of work. This particular part of the research was always done with clear project goals in mind, and frequently stopped well short of even a substantial survey.

3.1. Etymology

One of the most useful tools in ontology building is a dictionary, and when researching the ontology of common terms, it is often helpful to study the etymology of the terms themselves.

The English word *subject* derives through Latin from Ancient Greek, where it is the combination of the words “to throw” and “under”. The OED lists over 20 definitions of the word, which range from the idea of a vassal or minion (i.e. “a subject of the crown”), to the metaphysical definition of substance, to the “subject matter” of art and science (i.e. “Mona Lisa was the subject of this painting”), to the grammatical role of words in a sentence, to the use in Logic of an object which is predicated, to the theme of a literary composition [31].

The English word *topic* has a similar etymology, deriving also through Latin from Ancient Greek. The original Greek meaning of the word is a place, or a locality (hence the use in “topical ointment”).

Finally, the word *about*, which is normally used to relate books to their subjects, derives also from a spatial etymology meaning “near” or “around” (such as “scars about the face”).

The sense we are interested in for these words is believed to have been first brought into usage by Aristotle, who envisioned topics as “*places* in which a rhetorician might look for suggestions treating his theme” [31]. Subjects were things that were placed under (thrown under) these topics. They would be the very things the rhetorician might be looking for. This implies that subjects are not the same as topics; one is a place and the other is something located at that place.

It is pure conjecture, but nevertheless reasonable to infer, that the very thing Aristotle was thinking of here was a library, and that therefore the importance of organizing knowledge by subject dates back to him.

3.2. Bacon's fields of knowledge

Although undergoing steady development throughout the ages, arguably the next most significant event in the history of classification after its invention by Aristotle was Francis Bacon's attempts to organize and represent all the knowledge gained by human science [3]. Bacon's framework was hierarchical (a "tree of knowledge"), and drew on efforts in individual sciences to classify the natural world. Bacon's work was probably not notable in itself other than being considered the first real attempt to perform this kind of analysis.

A more impressive effort, while clearly based on and inspired by Bacon, was the well-known *Encyclopedia* by Diderot and D'Alembert in the mid eighteenth century [11]. Although built on work done previously, it was the first encyclopedia. In order to organize the knowledge in this work, Diderot (primarily, it is believed) and D'Alembert created a much more detailed and sophisticated tree of knowledge than Bacon's.

3.3. Library subject classification

The *Encyclopedia's* tree of knowledge is significant in many ways. We are interested specifically in two. First, it represents an entirely practical effort. The tree was clearly meant as a contribution to the development of the philosophy of science and the classification of knowledge; however, the effort to devise the tree was guided by the need to organize the *Encyclopedia*. Many deeper philosophical issues were passed over when they were not relevant to the contents of the work itself.

Second, the work of Diderot was well known by an American Francophile, who used it to guide the organization of his own extensive personal library. This American, Thomas Jefferson, was the third president of the United States, and the founder of the US Library of Congress (LOC), whose initial contents were precisely those of Jefferson's library.

This is by no means a claim that the LOC was the first to organize material by subject. Indeed, evidence points to attempts at such arrangements throughout history. The first surviving record of a subject catalog was in 1483, in Melk, Austria [37]. Various other personal and public libraries were organized by subject since then. The turn of the 19th century, which is roughly when the LOC was created, is about the time that subject classification of books began to be an accepted practice in libraries around the world. Until these changes were made, books were cataloged alphabetically by author, by title, or by the first lines of text.

Libraries began to be more dependent on accurate subject classification, and while all early classifiers were students of the philosophy of science, and therefore used the work of Bacon, Diderot, D'Alembert and their scholarly descendants to guide them, by the mid-19th century library classification had become an art of its own, which for practical reasons had to separate itself from philosophy. The main reason for this separation was that the subject catalog was driven by a single goal, the placement of books, whereas the philosophy of science clearly had other desires. "Classification", said John Cutter, developer of the *Expansive Classification* system that was to be the primary influence on the American LOC standard, "is the process of equating subjects with the topic contents of books, rather than with the logically derived elements of the general classification of knowledge" [29].

To further drive a wedge between philosophy and the new library science, many scholars objected to the use of a subject catalog. W. Stanley Jevons believed that such an organization, as

opposed to an author-arranged index, supported students who did not know their literature well enough [24]. He claimed that students should first be educated to know who was doing the important work in a field of inquiry. The well-known philosopher and logician Augustus DeMorgan also strongly objected to the idea of a “classed catalog”, believing it to be too much, “one man’s theory of the subdivision of knowledge, and the chances are against it suiting any other man” [9].

Library classification ensued with its overriding goal still to place books on shelves. Whether intentionally or not, this clearly links back to the etymology credited to Aristotle. Ainsworth Spofford, director of the LOC during the mid-19th century, and a major historical figure in the development of library subject classification, wrote, “Places in the library are subjects themselves” [29].

Further study of classification revealed that the most important part of the process was capturing rules to guide the classifier. The earliest evidence of recorded rules for subject classification is attributed to Adrien Baillet, who developed classification rules for indexing the private library of Chretien-François de Lamoignon in the late 17th century [37]. Accuracy was not considered as important as consistency. Spofford’s system was not even intended to be exploited by library users, but by trained library assistants [29].

Probably the most familiar and extensive work on library subject classification were those of Dewey (the decimal classification system) [10], Hartwig (the Halle Schema) [28] and Cutter. Dewey and Cutter each added two important points to their systems: the ability to add and remove categories, and the use of a relative index. The former was deemed critical as new fields of science and inquiry come and go, and the latter was important because the growth of library collections over time meant they could not guarantee that a certain subject would always be able to fit in a certain place. This gave libraries the ability to move the whole collection without requiring a recomputation (by humans) of the whole index. These two factors, plus the incredible care with which Dewey and Cutter approached the actual contents of their subject hierarchies, are the main reasons why both systems are still in use today, over a hundred years later.

3.4. LOC classification

Library classification is a field, a profession, requiring extensive training and experience. It would not be unreasonable to claim that library scientists have created the first formal ontologies to be used in systems (these systems began as printed indexes, then cards, and now databases). These subject catalogs represent enormous amounts of human labor spread out over two centuries. The accomplishment is nothing short of impressive, representing an essential blend of philosophy, analysis, and practical compromises. We examine here, in slightly more detail, the LOC Classification system (LCC) to gain further insight into the difficulties of representing and using subjects.

To begin with, the LCC is a hierarchical system. Subject categories have sub-categories, and sub-categories may have sub-sub-categories, etc. This hierarchy, perhaps due to its direct correlation with the physical location of books (but perhaps not as discussed below), was considered a mereological hierarchy, “Subcategories are more like parts of their parent categories than subsets” [28].

The design of the system, initially, was not so much meant for retrieval of information, but a way to locate books on similar topics near each other. Again, we have a throwback to the etymology of locality. This organization supported physical browsing better than an author or

title-based approach, something which surely any library user has experienced. If a book's subject and location were, in fact, the same thing, and given that a book cannot physically be in more than one place at a time, it is no surprise that all library classification began with a, "one book, one subject", paradigm.

This created obvious problems, not just for books that contained information on more than one subject, but for books whose single subject seemed to be an amalgamation of two existing categories. Early classifiers noted the difference between intrinsic sub-category relationships, such as, "zoology falls under natural history", or, "crops fall under agriculture", and extrinsic relationships, such as, "politics of agriculture", being under either agriculture, politics, both, or something else.

Eventually, Charles Martel of the LOC devised a strategy for dealing with extrinsic relationships between subject categories, finding that there were a few (politics, history, education), that cropped up far more often in combination with others. His strategy involved the use of *aspects* or *facets* of a subject. Each subject category, and potentially each subdivision thereof, had seven facets: serials, periodicals, dictionaries, etc.; philosophy and theory; history; general works; law and regulations; study and teaching; and the subdivisions. Thus, a book on the history of Chemistry would be classified under the history facet of the chemistry subject. A book on the theory of Physical Chemistry would be classified under the theory facet of the subject Physical Chemistry, itself a category in the subdivision facet of Chemistry.

Other qualities of books that affected their classification were the time of publication, the geographical origin, and the language. Specific sequences were created for works relating to specific persons, and for materials related to particular works (such as books about the works of Shakespeare).

Despite the addition of facets, combination topics still presented a significant problem to classifiers. Although Horne, working in the British National Library, had proposed classification of books under multiple subjects in 1825 [24], by the beginning of the 20th century libraries were still stuck on shelving as the principle factor in book classification, thus every book needed a primary subject class. Martel observed that "countless subject combinations [were] constantly found in new variety", and that transfers of books from one category to another were quite common, as better choices for the primary subject became clear *from use*.

The scheme in use always favored primary classification by the orientation of the author, which while helpful in resolving many classification problems, propagated the secular nature of scholarship. This factor was also clearly heuristic, and could point evidence in entirely the wrong direction at times.

Biographies were handled specially and normally placed in a special area of the history facet of the subject that the person the book was about was associated with. This was not usually done uniformly and depended on the existing structure of the history facet. Some subjects had very limited history facets because they were new.

Literature, fiction and poetry in general, were classified by author, requiring new categories to be created every day.

Letters, collections, and essays were considered the most difficult to classify, and each was done in a manner that seemed appropriate at the time.

Most of these principles are still in practice today. The biggest paradigm shift has been towards retrieval and access and away from locating books on shelves, and the COLON classification

system is probably the most modern approach [1]. The COLON system emphasizes facets more directly, with a richer set of facet categories than the LCC, and was developed by Ranganathan to address the rapid growth of human knowledge, and the rather natural observation that multiple subject headings for a book was the rule, not the exception. Ranganathan did not believe any book should have a primary subject, but eventually relented in this requirement of his system, admitting that the books had to be placed somewhere [34].

3.5. Thesauri

The COLON system also added the notion of *relatedness* in subjects, and this idea along with facets has been adopted by the thesaurus community [7]. One term is related to another if they are normally associated with each other, and one is not more general than the other. For example, “airplanes” is related to “air mail service”. Thesaurus structures do not specifically represent subjects; the goal is to provide relationships between terms used in any setting, but there is significant overlap in the problems of thesaurus representation.

Thesauri typically have four kinds of relationships between terms: *synonym*, *broader*, *narrower*, and *related*. E.g. “plane” is a synonym with “airplane”, broader than “jetliner”, narrower than “vehicle”, and related to “air mail”.

The addition of facets as a relationship between terms is fairly new, and has been driven by the same problems described here: specifically that it is frequently the case that one term may appear *prima facie* to be narrower than another, but further analysis reveals that is not quite right. For example, it is not truly the case that “cloning” is a narrower term than “biology;” in some ways it might be considered broader: with moral, political, religious, economic, and other aspects. With facets, however, we can keep cloning a subtopic of biology and add facets to biology (and cloning) for politics, ethics, etc.

4. Ontological analysis of subjects

In the course of this research, we have investigated the notion of subject hierarchies from many different angles. The main motivation has been the belief that digital libraries imply a profound shift in certain assumptions that have guided library subject classification since its inception:

- Digital documents do not need to be shelved, and can appear in many “places” at once. Modern library classifiers are trained to classify books under as many subjects in the catalog as appropriate. It cannot be denied, however, that the book must have a *primary* classification so that it can be shelved, and that the basic classification scheme itself was designed for the purposes of shelving, not retrieval.
- In the card catalog ontology outlined in Section 2.3 above, the object of a search does not need to be a document. We want to be able to classify people, organizations, conferences, etc. by subject as well. Library classifiers, as mentioned earlier, actually do use information they know about individual authors to help determine the primary subject classification of a book, however this information (about the author) is not recorded. Furthermore, we want to be able to employ rules of inference about the classifications that are made explicitly in order to propagate the classifications to other types of objects.

It should be noted that the purpose of this work is to provide an extensive card catalog system, that supports browsing and retrieval as described in Sections 2.3 and 2.4. Our goal is not to create an encyclopedia, in fact we hope to be able to stop a long way short of Cyc [27], while still providing a significant advance in usefulness over existing card catalog systems.

4.1. Previous work

This work in subject classification has been ongoing since 1994. The following sections outline other approaches to modeling subjects we have investigated.

4.1.1. Object-based subject taxonomies

An essential ingredient in all systems with subject classification, from Bacon, through Diderot, LCC, Dewey, COLON, and even Yahoo!, is a taxonomy or hierarchy of subjects. This notion seems inexorably tied to subject classification. Our work actually began with the ontology for digital library card catalogs [38] implemented in a description logic [5], and because description logics are particularly good at representing taxonomies, adding a subject taxonomy seemed a trivial and useful additional feature. Five years later, the card catalog ontology has been stable for four years and subjects are an open area of research.

Our main reason to use a subject taxonomy is to *narrow a search*. This has proven particularly effective in web catalogs such as Yahoo!, but it brings up an important point: library subject catalogs, as mentioned in Section 3.3, were developed as a cataloging and shelving method, not for retrieval. Many of the problems we faced, and solutions we experimented with, took a different turn since shelving was not of concern to us.

An initial and recurring obstacle we encountered when trying to add a subject taxonomy to the card-catalog system was in trying to capture the power of description logics to do subsumption over the subject taxonomy. A more complete discussion of description logics can be found in [5], a formal treatment in [2], and a comparison to first-order logic in [4]. A brief overview of the nature of subsumption reasoning in description logics follows, but the reader is strongly encouraged to review the referenced literature.

A description logic *concept* describes a set in such a way that membership can either be stated explicitly or computed. For example, if we define the concept BIRD to be any animal that has feathers, and define an *individual* Harry to be an animal that has feathers, then a description logic will derive that Harry is a member of the set BIRD. Like many representation systems, description logics make a syntactic distinction between concepts and individuals. In addition, unlike many representation systems, description logics are capable of reasoning at the terminological level. For example, if we additionally create a concept called EAGLE to be any animal with feathers and a sharp beak, a description logic will derive that EAGLE is subsumed by BIRD. While taxonomic relationships are a fairly common element of most representation systems, description logics permit describing *why* one concept is more specific than another. Finally, most description logic implementations deal with taxonomies very efficiently by caching all subsumption relationships; as a result, retrieving the individuals of any concept (i.e. the members of any set) is a simple memory access and requires no computation.

Given a representation and reasoning tool like a description logic, creating a subject taxonomy seems *prima facie* simple. We encounter, however, three main problems when trying to integrate a subject taxonomy with our card-catalog ontology:

1. In a description logic, subsumption is only computed for concepts. In order to exploit the existing power of these systems, the objects to be arranged in a taxonomy must be concepts. A subject taxonomy must, therefore, be a taxonomy of concepts as shown in Fig. 2.
2. The only kind of relationship permitted between a concept and an individual is *instance-of* (i.e. member-of). This derives from the interpretation of a concept as a unary FOL predicate, and an individual as an FOL object symbol. The only way to link a particular book to its subject would be by making the book an instance of its subject, as shown in Fig. 3.

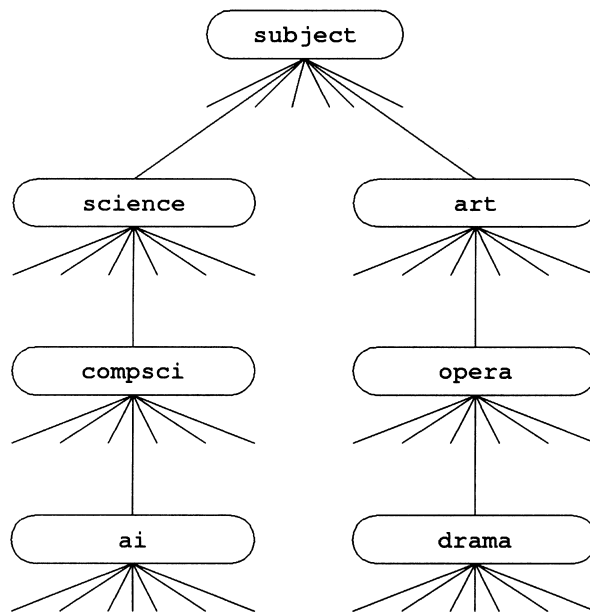


Fig. 2. A concept taxonomy of subjects.

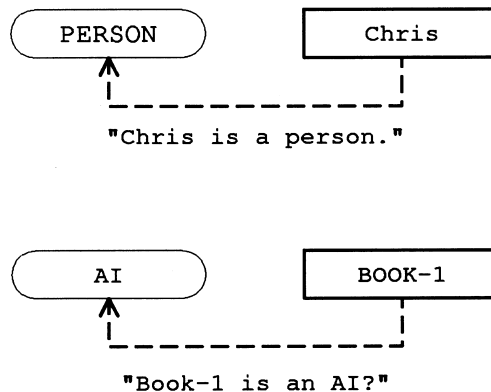


Fig. 3. Individuals can only be instances of concepts, which has an accepted linguistic interpretation.

3. Most library electronic card catalog systems treat subjects as attributes of books, e.g. a book has an author, publisher, etc., and a subject. Role (attribute) values must be individuals, which would require that subjects be represented as individuals, as shown in Fig. 4. This, too, derives from FOL; roles are binary FOL predicates, and a concept (as a unary predicate) cannot itself be predicated, therefore it cannot be the value of a role. From the perspective of our ontology, treating subjects uniformly as individuals would allow other individuals to be subjects. A book review article, as shown for example in Fig. 4, is about the book being reviewed – the subject is another book. This is obviously ignored by existing library systems, and web catalogs like Yahoo! do not treat subjects as attributes, but merely as a taxonomy.

In a subject taxonomy, when we try to classify a book under a particular subject concept, for example in Fig. 3 we represent “AI” as a subject, then in order to identify AI as the subject of a particular book, BOOK-1, we have to make BOOK-1 an instance of AI. This seems simply unnatural from an ontological perspective; no book is an instance of AI. The problem here stems quite simply from an attempt to overload the *instance-of* relationship to mean *has-subject*.

Further analysis reveals what may seem like a trivial solution: concepts in description logics are epistemological primitives that denote sets. What we mean by the concept AI, therefore, is the set of all things whose subject is “AI”. The confusion with calling BOOK-1 an instance of AI is no more than with choice of names: “AI-BOOK” would better reflect the ontological interpretation of this concept.

Taking this conclusion further reveals something about our electronic card catalog ontology (described in Section 2.3). We claimed this ontology was a “radical departure from the current paradigm in library catalog systems”, and now we begin to see why. Our ontology does not deal only with books. It deals with documents in general as well as with people, places, events, organizations, etc. We would like to classify all these types of objects by subject: businesses in AI, people interested in AI, events (such as conferences) on AI, etc. Renaming the concept, then, does not solve the problem unless we create a duplicate subject taxonomy for each thing that can be classified by subject; e.g. as shown in Fig. 5 there would be an entire subject taxonomy for people that included the “AI-Person” concept, and also hierarchies for companies, labs, events (like conferences), etc. In addition, these subject concepts are all connected in some way. The concepts “AI Book” and “AI Person” are related, yet there is no way in general to say “*x* Book” and “*x* Person” and “*x* Company”, etc., are all related because they all have to do with *x* as a subject.

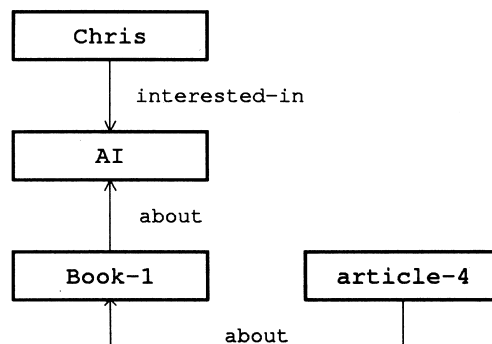


Fig. 4. When subjects are attributes, it becomes possible for objects other than traditional subject categories to be subjects.

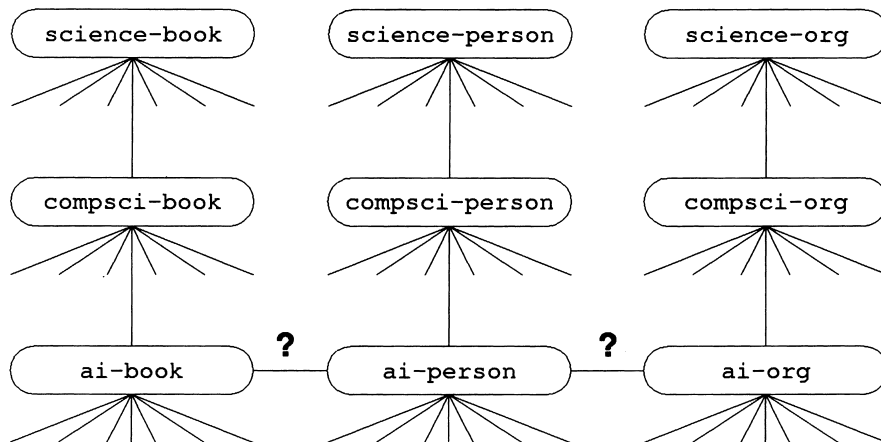


Fig. 5. Alternate naming scheme requiring multiple redundant taxonomies.

Again, this problem does not appear in libraries, or even in Yahoo!, for the simple reason that they are only concerned with books (or web pages). Note further that these problems are not merely created by the choice of representation system. Representing the subject hierarchy as sets and subsets in FOL would have the same drawbacks, unless one chose not to represent sets in the traditional way (as unary predicates).

4.1.2. Concept-individual pairs

After encountering these initial obstacles, the next approach we experimented with was to have every subject represented by a concept-individual pair (for those versed in description logics, note that this is not a “meta-individual” [5]). At first glance, this seems to solve our problems: objects are no longer instances of their subjects, and the individual part of a represented subject can be predicated. It creates a problem, however, because the subject information is not passed down the hierarchy as one would expect.

We would expect, for example, a query for “Books about Computer Science” to return a list of all books about computer science, and also any books about any of the subjects below computer science in the taxonomy, including books about AI. In order to achieve this, either objects must be restricted to be about only one subject (which is not desirable), or the description logic would require the SOME operator [33] (which many do not provide due to its complexity). With the SOME operator, this query would be:

```
(AND BOOK (SOME ABOUT CompSci))
```

The problem would then become that, while some of the subjects of an object would appear as role values, the “inherited” ones would not. In the example shown in Fig. 6, AI-IND and PHYSICS-IND are the fillers for the *about* role on Book-1, however COMPSCI-IND is not, even though Computer Science should be considered a subject. Queries, therefore, would need to take this into account and not use the role values at all, which puts us back to using the concepts and treating individuals as instances of their subjects.

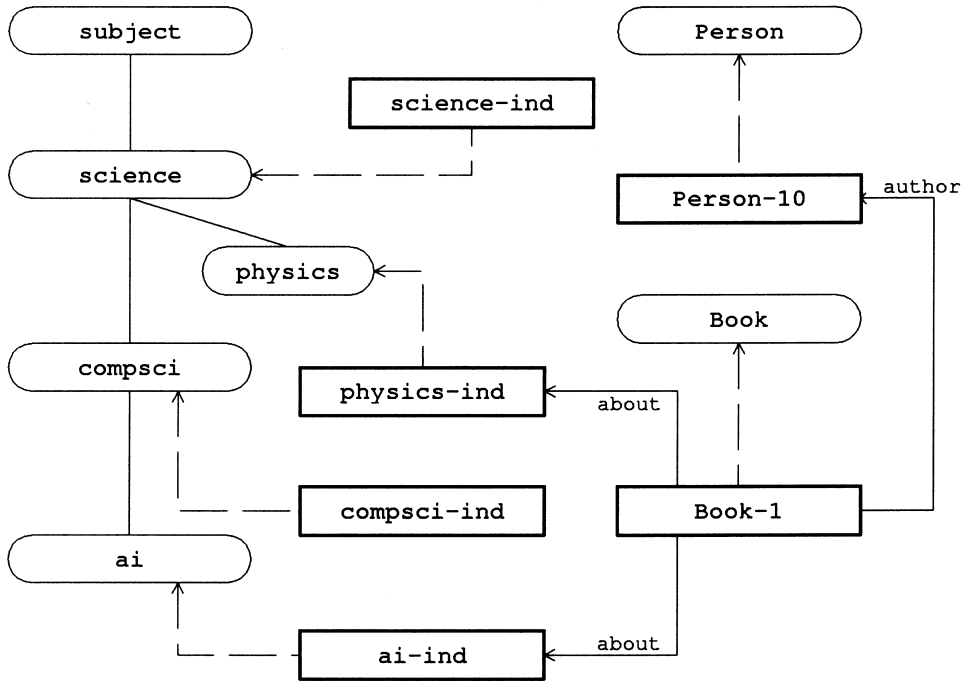


Fig. 6. Representing subjects as a concept-individual pair does not cause the expected inheritance of more general categories.

Another problem with this approach is that it is not entirely clear what things are. The instance parts of the subjects (e.g. AI-IND) are instances of SUBJECT and so perhaps they are the subjects themselves, but what are the concept parts? Are they simply taxonomic placeholders or do they have some meaning beyond that?

This representation would be quite a bit less efficient than the first approach in focusing a search on objects in a certain subject area. Despite these drawbacks, this is a possible solution, given the SOME operator.

4.1.3. Subject-based instantiation

Another alternative is to alter the syntax and semantics of description logics to include a special relationship between individuals and their subject concepts, as shown in Fig. 7. Again, description logics define only one link between individuals and concepts, the *instance-of* link. This

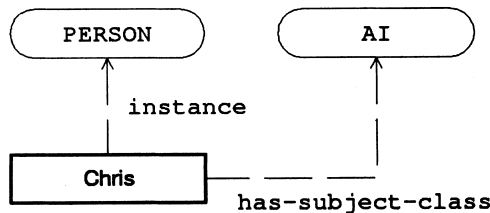


Fig. 7. Creating a special syntax for subject classification.

link is part of the syntax of the language, and has nothing to do with roles, which can only link individuals to each other. A special *has-subject-class* link would require fairly extensive modifications to the language, since operationally it would behave precisely as the *instance-of* link with respect to subsumption, yet would need its own record keeping and so forth within the implementation.

This approach lacks the ability to deal with the third problem above (treating subjects as values for the *about* role), and like the others pushes a lot of interpretation onto the user interface. We are not currently considering it due to the complexity of the implementation changes required. Other groups may be trying this approach [25].

4.1.4. Subject things

The next solution we considered, and actually are still using due to its simplicity (see [42]), returns to the basic idea of the first approach (Section 4.1.1). Recall the observation that what we really mean by a particular subject concept, e.g. “AI”, is *the set of all things whose subject is AI*, be they books, web-pages, people, organizations, etc. The problem, then, may be no more than one of choosing the proper name. While “set-of-all-things-whose-subject-is-AI” is operationally a bit cumbersome, “AI-THING” seems to fit the bill.

In this approach, a book on AI becomes an instance of both the concept BOOK and AI-THING, a person interested in AI becomes an instance of both PERSON and AI-THING.

Through experimentation, we found this approach also allows the succinct expression of rules for propagating subject information. For example, it makes sense to say, “A person who writes a book about AI is interested in AI”. It is difficult to express this rule using the previous techniques (see [40]).

This solves the first and second problems by slightly altering the way subjects are considered and pushing any interpretation onto the user interface. It does not deal with the third problem of individuals as subjects at all. We cannot make a book, or a person, the subject of anything.

4.1.5. Subjects as instances

We then considered making a rather major shift in our thinking by focusing on how to solve the problem of allowing individuals to be subjects as well. There are numerous reasons to do this, such as a book review article whose subject is a book, or a biography whose subject is a person. If our ontology is to be consistent, then if it is possible for a subject to be represented as an individual, all subjects should be represented as individuals. If subjects are individuals, does this mean that they exist, the same way that a book exists?

These considerations led us to initially conclude that, in fact, the subject of a book does exist and that all subjects should be individuals. We think, in this approach, of subjects as individuals that correspond not to concepts in a taxonomy as in the Concept-Individual Pairs approach, but to each item to be classified by subject. That is, for each individual that has some sort of subject classification there is an individual that represents the subject of that item.

We then represent a hierarchy of subject *types* as concepts. In previous approaches, these concepts represented the subjects themselves; if a book was about “AI”, then the book would be an instance of the AI subject concept. In this new ontology, the concepts represent categories of subjects, and the individuals represent a thing that the book is about. This has a similarity to the etymology for subject and topic discussed in Section 3.1; the concepts in the hierarchy are what Aristotle called topics and the individuals are the subjects, which are “thrown under” topic

headings. The objects (books, people, etc.) themselves not classified directly by subject, but have an *about* (or *interested-in*) link to their subjects. Any object can be a subject, as well.

A simple example of classification in this manner is shown in Fig. 8. In this example, Person-1 is Ernest Hemingway, Book-2 is *The Old Man and the Sea*, Book-1 represents a biography of Ernest Hemingway, and Person-2 is the author of the biography. Each book has an associated subject that it is about. The subject is always an individual, and will be an instance of all the concepts that represent topics under which the subject falls. In the case when a book or (as in this example) a person is the subject, the object that represents that book or person is the value of the *about* role. This condition is sufficient to make the object an instance of the concept SUBJECT; in other words, if x is about y , then y is a subject. This inference is shown as a dotted line in the figure.

What first prevented us from representing subjects as individuals was that we wanted to exploit the power of description logics to deal with our subject taxonomy automatically. One of the principal reasons for representing the subject hierarchy, as mentioned in Section 4.1.1, was to enable the user to *narrow a search* based on the subject. It seemed that representing subjects as individuals was mutually exclusive with maintaining a taxonomy of subject concepts. This ontology seems to account for that in an elegant way, and seems to address all the problems of the previous approaches.

This is a very different approach than any of the others we tried, however, and led to an interesting issue: is the subject of each object unique? If so, then how do we represent the fact that there are many books about the same person? If not, how do we represent the fact that a book may be about more than one person? A related issue is, when a book is about more than one thing, does it have more than one subject?

These questions led us to drive the topic hierarchy deeper into more fine-grained areas. For example, if we take the position that a subject need not be unique for each object, then it is clear

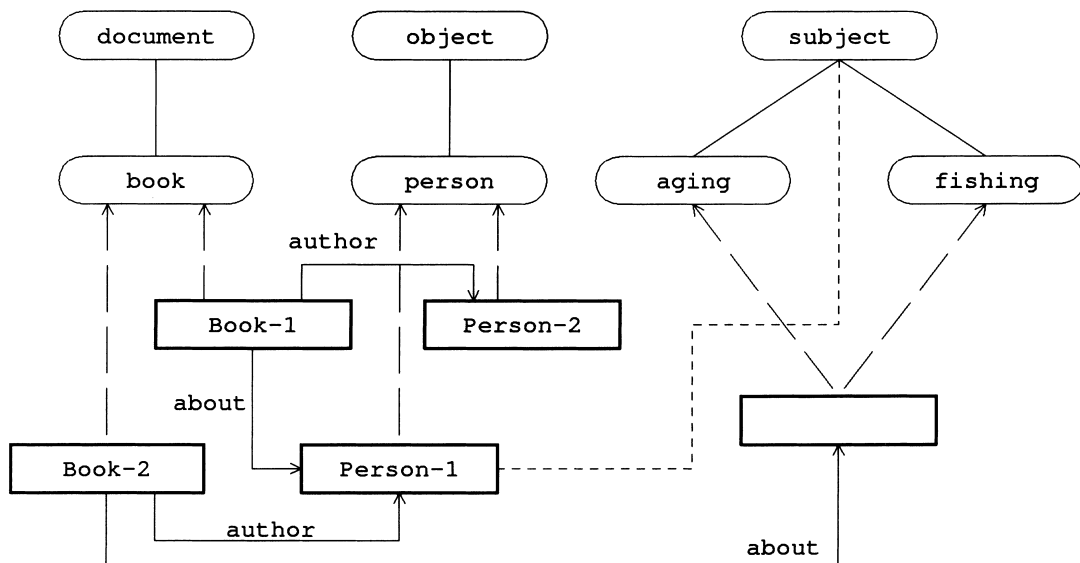


Fig. 8. Creating a unique subject for each document, and classifying the subject under a taxonomy of topics, which allows for other individuals to be used as subjects.

that all books about Hemingway have the same value for their *about* role. It is not clear what happens in the case of multiple books about AI: do all such books have the same subject, or is each one different? For this issue, we believed perhaps it was a question of granularity; it is probably not the case that two books about AI are about the same thing (as giving them the same value for the *about* role would imply), but if we had more specific topic concepts representing sub-fields of AI, perhaps it would make more sense to share subjects among multiple objects.

In pursuing this course, we encountered numerous problems as our topic hierarchy pushed down into more fined-grained levels. We found that the representation becomes incredibly messy, and the taxonomy loses cohesion. Is, e.g., “ontology” a sub-field of “knowledge representation” and “philosophy?” Is “formal ontology” a sub-field of “ontology”, or vice-versa? Are “ontologies for part-whole relations” really “ontology” or are they “epistemology”, or both, or neither? Our experience has been that the LCC, Dewey, and Yahoo! subject hierarchies are so coarsely grained for good reason: it is incredibly difficult to push these structures past a certain point.

4.2. Other views

At the time of the Formal Ontology in Information Systems (FOIS) meeting in Trento, we were experimenting with the ontology in the form described in the previous section (see [40]). At that fruitful meeting, and afterward through email, we benefited from exchanges with other experienced ontology researchers. Some of their informal observations are quoted below.

4.2.1. Subjects as social artifacts

In the course of attempting to draft a taxonomic index of AI subjects for AAI, Jon Doyle commented:

It became clear to me that the main organizing principle for indices, at least for most people, is sociological. That is, one structures the index not to reflect conceptual relations, but to reflect things like what populations of people like to work together, what do they think of as the current main topics of interest, etc. [12].

This is similar to the sentiments of DeMorgan [9], and also sheds light on the problems we were having in pushing the topic hierarchy down into more specific areas. In addition, as a consequence of the social nature of topic organization, the stability of a particular section of the hierarchy seems to be directly tied to the field’s maturity.

David Israel had a similar intuition:

[Subject] phrases like ‘universal algebra,’ ‘high-energy particle physics,’ ‘vertebrate zoology,’ have the logic of mass nouns – like ‘mud’ – not count nouns or sortals. Note also they’re ambiguous as between the topic/subject/field and activities engaged in the topic/subject/field (E.G., ‘He been doing a lot of mathematics lately.’), and that activity words are also mass-noun-like (E.g. ‘He is been doing a lot of walking lately.’). Put all [this] together and what you get is: [subjects] are rooted in the ontology of activities [23].

Existing libraries have recognized the social aspect of subject categories in a limited way. Charles Martel, mentioned in Section 3.4 as the developer of the rules for the LCC, acknowl-

edged, “When in doubt for a primary classification, we often classify books based on what the author does...” [29] One motivation for Ranganathan’s COLON system was apparently the need to represent the changing nature of subjects over time [34]. Recent research in citation and co-citation indices also acknowledge that a “field” of study is defined by the activities of the people who are in the field.

4.2.2. *Subjects as qualities*

Another important aspect of the ontology in Section 4.1.5 is the claim that the subject of each book exists. Nicola Guarino has done seminal ontology research into the nature of object *qualities*, such as color, size, etc., and observed:

From the philosophical point of view, the idea [of a subject as an extant object] is similar to the choice of admitting individual qualities, i.e. instances of determinables like color, size, and, in this case, genre and subject, which are unique for each particular object. This has been recently proposed in [17]. However, the problem in this case is that subject and genre do not seem to be true determinables, in the sense that they rather express a ‘full-shaped’ relationship between books and areas of knowledge. The problem is similar to that of choosing between attributes and relations in an ER diagram: is subject an attribute of a book, or a relationship between two independent entities, namely the book and something else?” [18].

Most implemented systems that do deal with subjects have chosen to make subjects attributes (in an ER sense) instead of relations, however our ontology makes *about* a full relation, and this is similar to the claim that the color of each rose exists. In the latter case, the color of ROSE-1 (a particular rose) is COLOR-1 (a particular color), which is an instance of the quality RED-COLOR. The problem with this analogy is that a subject instance (SUBJECT-1, an instance of CHEMISTRY) is not clearly a determinable quality the way COLOR-1 is.

4.2.3. *Subjects in CYC*

Outside of the library community the only other group we found that is investigating the formalization of subject classification is at Cycorp. Cyc ontologies contain axioms for deriving subtopic relations in order to answer queries. We should note again that our purposes are clearly different; while Cyc seeks to provide access to the *information content* of documents, we seek to provide access only to the documents. Going back to our scenario in Section 2.4, our system provides the information to help a user *find the documents* that may contain answers to the question, “Was it common for government officials during the American Civil War to mention government business in personal correspondence?” Cyc, on the other hand, would seek to answer the question.

This may seem like a dramatic difference in ontology goals, however it is actually just one of granularity. Our claim that this is a card catalog for digital libraries, and that we are incorporating SGML (or XML) markup into our ontology in order to be able to deal with document parts, implies that we may have to consider classifying these parts of documents. For example, a book with several chapters, each of which is a separate paper treating a subtly different topic, should be represented as the aggregation of the different chapters each with its own subject. If that is the case, it might be desirable (or even necessary) to represent sections of each chapter with unique subjects, or sub-sections, or paragraphs, or even sentences.

We are endeavoring to avoid extending the subject representation to such a level, but avoiding this may not solve any problems. Cyc, obviously, does not avoid this issue, and Fritz Lehmann has some experience with axiomatizing the derivation of subtopic from parts of a document:

Subtopic is like part, but the subtopic relation is a very specialized kind of part relation, if it is a part relation at all. Subtopic is of at least two kinds, task-related and general. A topic may be a relevant subtopic given a certain task; or, there may be a general subtopic as in the Dewey Decimal System, LC, COLON, Bliss classifications, thesauri, etc. The latter kind is also task/purpose related, but it assumes ‘general purposes’. If you’re fixing a sink, certain subtopics of ‘sink’ are relevant, like tools for fixing it, codes governing the repair of sinks, methods for repair, etc. Others are irrelevant, like when the sink was invented, decorators’ ideas on sinks, whether there are sinks in Kamchatka, etc. But for a scholar, these might be quite relevant. Subtopic is clearly faceted, ‘18th Century Japanese Lead Statues’ has time, space, substance, process facets, i.e. it can be specialized along several different dimensions or facets.... Subtopic is not subclass. The Shoe–Sandal link is legitimately both, but Shoe–Heel, Shoe–Shoemaking, Shoe–RussianShoeMarket, etc. obviously are not. Subtopic may be derived, with care, from certain sub-class, instance-of, and other links, but this can get subtle. Relevant-subtopic may not be transitive. If you’re interested in ‘The Mating Habits of Modern Americans’ as a topic, you probably aren’t looking for ‘Sex Acts of Texas Serial Killers’, even though it is logically subsumed. [26].

Although there are several important points here, the critical one is that *subtopic may not be transitive*. It was this observation that helped us to escape the traditional view of subject hierarchies and formulate the new view of subjects presented below.

4.3. Further discussion

Based on the work cited in the previous sections, these are the main points of interest regarding an ontology of subjects:

1. Subject classification is inherently hierarchical. This hierarchy may be taxonomic, mereological, or something else. The etymology of the word *vis*. Aristotle is mereotopological (topics are a “place” where subjects are found). The sub-topic relationship may not be transitive.
2. Subject hierarchies are not conceptual. They derive from “folklore” more than from any specific theory of knowledge. At the higher levels, we may reach some form of agreement regarding the organization of e.g. the sciences, but this seems to be dependent on the maturity of a field. The more specific category breakdowns (is “formal ontology” a sub-topic of “knowledge-representation”?) are personal, and may even vary depending on an individual’s needs.
3. Subjects are not just for books. They can be used to classify people, events, organizations, etc. In these cases, the association with a subject is related more to the activities that are performed than to any specific content. It should be reasonable, however, to classify a person who writes a book “about” ontology as someone “interested in” ontology. One of the problems here is that once the activity of writing the book has ended, it may very well be the case that the person is no longer validly “interested in” the topic (most Ph.D. students would admit to this condition after finishing their theses). Therefore any true theory of propagating subject must take this into account.

4. Subjects can be faceted. All library classification systems support faceted subject categories, because the “politics of physics” may not be of interest to someone looking for books about politics. Note, however, that the origins of subject facets are in the need to bring books with multiple topics under a single subject heading. With this requirement removed, do facets simply become membership in multiple categories?
5. Subjects can be real things. If the subject is considered to be “that which a book is about,” then clearly a book can be about Science and a book can be about Hemingway. The former case is the usual subject classification, however the latter case becomes complicated by the library ontology presented in Section 2.3 – how do we then relate Hemingway the author to Hemingway the subject. Are they the same thing?
6. Subjects may or may not exist as individual qualities. Using a representation like that presented in Section 4.1.5, does a book with multiple subjects imply the existence of multiple qualities, or is it a single quality that is classified under multiple categories?
7. Subjects can be propagated. It is useful to conclude, as mentioned above, that a person who writes a book “about” ontology is a person who is “interested in” ontology. In addition to issues of changing information, how far does the propagation carry? If a person writes a review of a book about ontology, is the review about ontology? This is clearly not the case for people: a book about Hemingway may not include his interest in fishing. Is this simply because books have limited topics? What about an encyclopedia?
8. General Topics have a different semantics. The “general works” facet provided by library subject classifications (potentially) on any subject category, is supposed to account for textbooks that survey a field, and compendiums that are about a group of sub-topics. For example, imagine classifying the Proceedings of IJCAI-97. The book itself contains articles in almost every discipline of Artificial Intelligence. It would normally seem reasonable to classify this under the heading “Artificial Intelligence” (and so it is). If a person is searching for “books about neural networks” and “neural networks” is a sub-topic of AI, do we exclude the IJCAI-97 proceedings since it is in a more general topic, even though it contains articles that are relevant? Perhaps for this case the answer is simply that cataloging the book and not the articles is the wrong granularity, but what about a textbook on AI that discusses neural nets?

We have tried to address as many of these points as possible in the new approach presented below.

5. A place for subjects

After careful study, taking into account the deep history and the many problems we encountered with subject classification in this research, we propose throwing away the traditional notion of subjects as taxonomic and returning to Aristotle’s definition of topic as a “place”. In a digital library, not only can a book be in many places at once, but those places can have unlimited size, unlimited dimensionality, and can come and go dynamically as people create new places based on those that exist. Furthermore, there is no reason why these spaces could not be tailored to suit the needs of a particular individual or group.

5.1. Relations

In this multi-dimensional space, topics (as places) can be related to each other by all the standard mereotopological relationships: *contains*, *overlaps*, *borders*, *near*, and *far*. In a purely

relational representation, all these relations except *contains* are reflexive, however a more robust representation could include the degree of overlap (which is not reflexive) and the distance for near and far. This information could be determined dynamically over time through active usage data, and could be personalized, which is similar to the ideas presented in recommender systems [35]. Each time a user travels from one place to another those places will be made closer to each other. Frequently traveled pathways would become “worn” over time, like a footpath through grass, leading others to go the same way.

The topics that are normally agreed to be taxonomic (e.g. “Physics is a sub-topic of Science”) are actually contained, and *contains* is known to be transitive. The areas that cause difficulty and confusion in the traditional taxonomic view represent overlaps and bordering topics, explaining the cases where sub-topic is not transitive. The *related-to* relation is actually two topics that overlap, border, or are near each other.

5.2. Topical organization

People and other objects when interested in a topic would also be “located” in those places, from the perspective of the catalog, and of course can be in many places at once. When a book is placed under a topic, the author will be “dragged along” to that place as well. Furthermore, as people’s interests change, they could move from place to place, although this aspect would need to be considered carefully. It may still be useful to associate a person who once worked in an area with that area, even if she no longer is really “there”.

Note that while it is easy to imagine such a catalog with a visual interface that represents different objects (e.g. books, people, etc.) with different avatars, we are not proposing a “chat room”. The presence of a “person” in a particular “place” (topic), would be a fairly static object about which a user could gather information. This is merely a different way of viewing and thinking about the catalog.

Facets of topics would become sub-areas or overlaps, or standard ways of arranging places. Books on the “history of chemistry”, for example, could be in both the history place and the chemistry place, a place that represents the overlap between the two, or some designated region of the chemistry topic.

When topics are considered places, the neighboring regions can be organized such that objects which may also be relevant to a search are located nearby (i.e. topically). General works, then, could always be positioned close by, but still on the periphery, of all the sub-topics they deal with.

Note that the underlying representation of topics as places is very similar to the approach presented in Section 4.1.5. However, some changes to the reasoning will need to be made to incorporate the mereotopological relationships more accurately. We have replaced the notion of class membership with the notion of location, an idea that is gaining stronger and stronger support in the ontology community [36].

5.3. Linguistic evidence

A feature of this ontology is the degree to which it agrees with natural language. Objects that deal with a certain topic will be *about the topic*, and here both meanings of the word “about” apply. Very specific topics will be treated as narrow spaces, and more general ones will be larger,

with *subjects* being arranged *topically*. Again, both meanings of “topical” apply. Note also how this view captures the multiple meaning of “area”, as in a “topic area” and an “area of space”. The words “field”, and “domain”, also have the same dual meanings that become unified with this ontology.

We have attempted to be careful about using the word “subject” and “topic” slightly differently here. In fact, the previous paragraph is the first to use the word “subject” within this section. Like our ontology in Section 4.1.5, we treat subjects and topics differently. At this time, we believe that where topics define *regions* of space, subjects may be *points* within those regions. The subject of a book is its precise location within the topic space. A book considered to have multiple topics will be located in multiple places.

5.4. Formal ontology in a description logic

In our initial experiments with this ontology, we have chosen to represent topics as individuals that potentially have any of the mereotopological relations defined in Section 5.1. We have specifically not used coordinates to define strict spatial regions, because we believe our regions are n -dimensional, with n unknown.

Our basic card-catalog ontology (discussed in Section 2.3) introduces types such as PERSON, EVENT, DOCUMENT, etc. We augment this ontology with several new concepts for the purposes of subject classification using our mereotopological approach: TOPIC and SUBJECT.

Individuals of TOPIC may be related to each other by the relations *sub-topic* (which corresponds to *contains*; inverse: *sub-topic-of*), *overlaps* (inverse: *overlaps*), *borders* (inverse: *borders*, this relation is currently unused in our experiments), *near* (inverse: *near*), and *far* (inverse: *far*). Individuals of SUBJECT may be related to individuals of TOPIC by the *located-under* relation (inverse: *contains-subjects*), as well as by *near* and *far*. Individuals from the basic ontology can be related to individuals of SUBJECT by the *location* (inverse: *location-of*) relation.

The ontology also includes several axioms that are useful for deriving specific relationships between individuals:

- If a SUBJECT is *located-under* multiple TOPICs, then the TOPICs overlap.
 $\forall s, t1, t2 \text{ located-under}(s, t1) \vee \text{located-under}(s, t2) \rightarrow \text{overlap}(t1, t2)$
- Individuals are *about* the TOPICs their SUBJECTs are *located-under*.
 $\forall x, s, t \text{ location}(x, s) \vee \text{located-under}(s, t) \rightarrow \text{about}(x, t)$
- *far* is transitive over *located-under*
 $\forall s, t1, t2 \text{ located-under}(s, t1) \vee \text{far}(t1, t2) \rightarrow \text{far}(s, t2)$
- *contains-subjects* is transitive over *subtopic*
 $\forall s, t1, t2 \text{ subtopic}(t1, t2) \vee \text{contains-subject}(t1, s) \rightarrow \text{contains-subject}(t2, s)$
- *about* is sufficient to be a TOPIC
 $\forall x \exists y \text{ about}(x, y) \rightarrow \text{topic}(y)$

Note that *near* and *overlaps* are not transitive over *located-under*. It is possible for an individual of SUBJECT to be *located-under* a TOPIC that is *near* (or *overlaps*) another TOPIC, but the SUBJECT is not *near* (or *overlaps*) the second TOPIC. In other words, TOPIC regions may be large enough for two locations within the region to be *far* from each other.

Fig. 9 shows a simple example of the usage of our ontology.

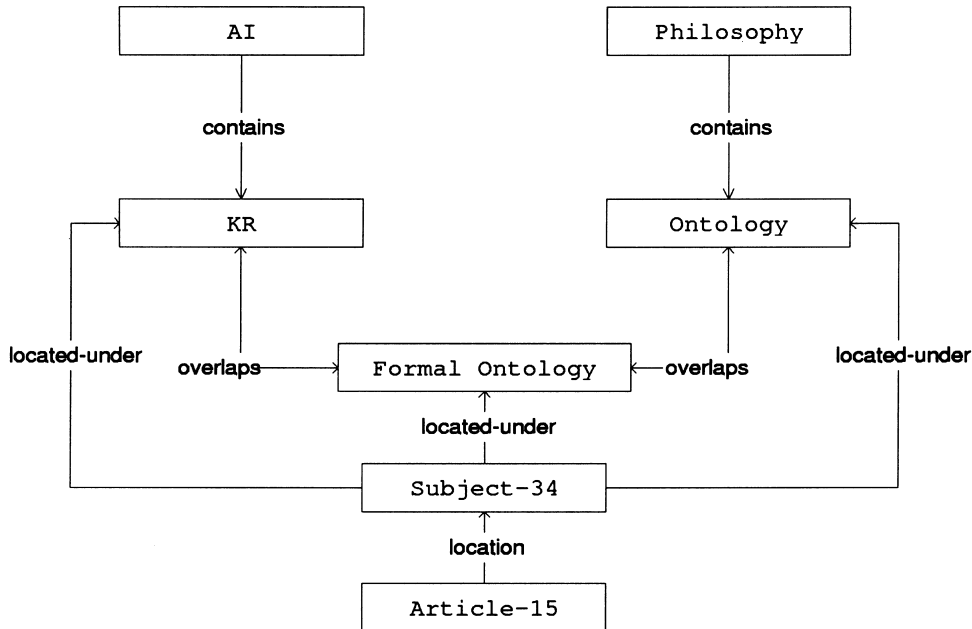


Fig. 9. An article about Formal Ontology that lies in the intersection between KR and Ontology. Although inferred relationships are not shown, one result of this classification of Article-15 will be that KR and Ontology overlap.

6. Conclusion

We have presented a perspective on a digital library project which seeks to incorporate some of the potential advantages implied by digital libraries into an intelligent card catalog system. Requiring the ability to classify not only documents, but people, objects, places, and events, by subject, our system will support a new paradigm of intelligent browsing.

Central to this system is a proper representation of subjects – not only an accurate hierarchy of topic categories, but a thorough understanding of what a topic and a subject is. We have reviewed a great deal of history of subject classification in philosophy, library science, and the recent attempts at building an ontology for subjects.

Inspired by the origins of the word itself, we proposed and briefly explored the view of topics in a digital library as multi-dimensional spaces. The objects in these spaces have no limits on the number of places they may appear in, and the spaces have no limit on the other spaces they may border, contain, be contained in, or overlap with. In addition to these formal mereotopological relationships, we also proposed using empirical data to represent the degree of overlap between topics, when appropriate.

7. For further reading

[14], [15], [19], [21], [32].

Acknowledgements

I would like to thank Nicola Guarino for forcing me to write this paper, and David Israel, Fritz Lehmann, and Jon Doyle for their comments. Derek Gaasch, Anthony Schorr, and especially Tim McGraw contributed to the work over the several years since it began, and probably at various points Nancy Ide, Ron Brachman, and Alex Borgida had an influence on some of these ideas.

References

- [1] R. Aluri, A. Kemp, J. Boll, Subject analysis in online catalogs, *Libraries unlimited*.
- [2] F. Baader, H. Bürkert, J. Heinson, B. Hollunder, J. Müller, B. Nebel, W. Nutt, H. Profitlich, Terminological knowledge representation: A proposal for a terminological logic, DFKI Technical Memo TM-90-04, May 1991.
- [3] F. Bacon, in: W. Wright (Ed.), *The Advancement of Learning*, Clarendon Press, Oxford, 1873.
- [4] A. Borgida, On the relative expressiveness of description logics and predicate logics, *The Artificial Intelligence Journal*, to appear.
- [5] R. Brachman, D. McGuinness, P. Patel-Schneider, A. Borgida, L. Resnick, Living with CLASSIC: When and how to use a KL-ONE-like language, *Principles of Semantic Networks*, Morgan Kaufman, Los Altos, 1991, pp. 401–456.
- [6] F. Bruneseaux, B. Evelyne, R. Laurent, A user-oriented linguistic resource server: the silfide project, *Actes DRH 97, Digital Resources in the Humanities*, Oxford University Press, Oxford, pp. 303–306.
- [7] H. Chen, K. Lynch, K. Basu, D. Ng, Generating, integrating and activating thesauri for concept-based document retrieval, *IEEE Expert*. 8 (2) (1993) 25–34.
- [8] D. Chesnutt, The Model Editions Partnership, *D-Lib Magazine*, November 1995. Available at <http://www.dlib.org/>.
- [9] A. DeMorgan, On classification, *Philosophical Magazine*, Third Series, 26 (1845) 522.
- [10] M. Dewey, *Dewey Decimal Classification and Relative Index*, 16th ed., Lake Placid Club, NY Forest Press, 1958.
- [11] D. Diderot, *The Encyclopedia*, Trans. by S. Gendzier, Harper & Row, New York, 1967.
- [12] J. Doyle, personal communication, 1998.
- [13] J. Flanders, The Brown University Womens Writers Project. Available at <http://www.wwp.brown.edu/>.
- [14] A. Gordon, E. Domeshek, *Deja Vu: a knowledge-rich interface for retrieval in digital libraries*, in: *Proceedings of the 1998 Intelligent User Interfaces Conference (IUI-98)*, ACM Press, New York, January 1998.
- [15] T. Gruber, A translation approach to portable ontology specifications, *Knowledge Acquisition* 5 (2) (1993) 199–220.
- [16] T. Gruber, Introduction to the Bibliographic Data Ontology. Available as <http://www-ksl.stanford.edu/knowledge-sharing/ontologies/html/bibliographic-data.text.html>.
- [17] N. Guarino, Semantic matching: formal ontological distinctions for information organization, extraction, and integration, in: M.T. Paziienza (Ed.), *Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology*, Springer, Berlin, pp. 139–170.
- [18] N. Guarino, personal communication, 1998.
- [19] S. Harie, N. Ide, J. Le Maitre, E. Murisasco, J. Véronis, SgmlQL – An SGML Query Language, in: *Proceedings of SGML'96*, p. 127.
- [20] N. Ide, J. Véronis, Modelling lexical information, in: H. Susan, N. Ide (Eds.), *Research in Humanities Computing* 4, Oxford University Press, Oxford, pp. 193–206.
- [21] N. Ide, T. McGraw, C. Welty, Representing TEI documents in the CLASSIC knowledge representation system, in: *Proceedings of the 10th Workshop of the Text-Encoding Initiative*, November 1997.
- [22] N. Ide, Corpus encoding standard, First International Language Resources and Evaluation Conference (LREC), Granada, Spain, proceedings to appear. CES Documentation and DTDs available at <http://www.cs.vassar.edu/CES/>.
- [23] D. Israel, personal communication, 1998.
- [24] W.S. Jevons, *The Principles of Science: A Treatise on Logic and Scientific Method*, Macmillan, London, 1874.
- [25] P. Lambrix, N. Shamehri, N. Wallöf, Dwebic: an intelligent search engine based on default description logics, in: *Proceedings of DL-97, The International Workshop on Description Logics*, September 1997.
- [26] F. Lehmann, personal communication, 1998.
- [27] D. Lenat, M. Shepherd, D. Pratt, K. Pittman, R. Guha, Cyc: towards programs with common sense, *Communications of the ACM* 33 (8) (1990) 30–49.
- [28] J. Metcalfe, *Subject Classifying and Indexing of Libraries and Literature*, Scarecrow Press, New York, 1959.
- [29] F. Miksa, *The development of classification at the library of congress*, Occasional Papers of the Graduate School of Library and Information Science, University of Illinois Press, Champaign, IL, August 1984.

- [30] E. Mylonas, A. Renear (Eds.), *Journal of Computers in the Humanities: Special Issue on the Tenth Anniversary of the Text-Encoding Initiative*, Kluwer Academic Publishers, Dodrecht, 33(1–2).
- [31] *The Unabridged Oxford English Dictionary*, 1989.
- [32] P. Padgham, *The Description Logic Home Page*. Available at <http://www.dl.kr.org/dl>.
- [33] P. Patel-Schneider, B. Swartout, *Description Logic Knowledge Representation System Specification*, From the KRSS Group of the ARPA Knowledge Sharing Effort, November 1993. Available at <http://www-db.research.bell-labs.com/user/pfps/krss-spec.ps>.
- [34] M. Sajita, *Colon classification*, 7th ed., *A Practical Introduction*, Ess Ess Publications, New Delhi, 1989.
- [35] L. Terveen, *Collaborative Recommender Systems for the Web*, *SIGART Bulletin*, 9(3–4), ACM, Winter, 1998.
- [36] A. Varzi, *Basic problems of mereotopology*, in: N. Guarino (Ed.), *Formal Ontology in Information Systems*, IOS Press, Amsterdam, 1998.
- [37] M. Verner, *Adrien Baillet and his rules for an alphabetical subject catalog*, *Library Quarterly* 38 (3) (1968) 217–230.
- [38] C. Welty, *Knowledge representation for intelligent information retrieval*, in: *Proceedings of the CAIA-94 Workshop on Intelligent Access to Digital Libraries*, March 1994.
- [39] C. Welty, *Intelligent assistance for navigating the web*, in: *Proceedings of the 1996 Florida AI Research Symposium*, May 1996.
- [40] C. Welty, *The ontological nature of subject taxonomies*, in: N. Guarino (Ed.), *Formal Ontology in Information Systems*. IOS Press *Frontiers in Artificial Intelligence and Applications Series*, June 1998.
- [41] C. Welty, N. Ide, *Using the right tools: enhancing retrieval from marked-up documents*, *Journal of Computers and the Humanities*, Spring, Kluwer, 33 (1–2) 1999.
- [42] C. Welty, *The Untangle Home Page*. Available at <http://untangle.cs.vassar.edu/>, 1998.

Christopher Welty is an Assistant Professor at Vassar College, Poughkeepsie, NY, USA, on leave at LADSEB-CNR, Padua, Italy. He received a Ph.D. from Rensselaer Polytechnic Institute in 1995, and is editor of *intelligence Magazine*, published by ACM, as well as an ACM Distinguished Lecturer. His main research interests are formal ontologies for digital libraries and for automating software engineering.

Jessica Jenkins is an undergraduate student at Vassar College, Poughkeepsie, NY, USA, concentrating in Artificial Intelligence. She spent the summer of 1999 working at Cycorp, Austin, Texas, USA, and is doing her senior research work in ontology development and application.