



Project no.507618

DELOS

A Network of Excellence on Digital Libraries

Instrument: Network of Excellence

Thematic Priority: IST-2002-2.3.1.12

Technology-enhanced Learning and Access to Cultural Heritage

DELIVERABLE 5.1.2

“Study to determine the requirements for and usage of extracted knowledge”

Due date of deliverable: 31/12/04

Actual submission date: 14/2/05

Start Date of Project: 01 January 2004

Duration: 48 Months

University of Southampton (UoS, Partner 26)

Revision Final

Project co-funded by the European Commission within the Sixth Framework Programme
(2002-2006)

Dissemination Level: [PU (Public), PP (Programme Participants), RE (Restricted), CO
(Confidential, only for consortium members)]

Study to determine the requirements for and usage of extracted knowledge for bibliometrics, domain analysis, issue tracking and community modelling

Les Carr, Tim Brody, Nicholas Gibbins
School of Electronics and Computer
Science
University of Southampton
Southampton, United Kingdom
{lac,tdb01r,nmg}@ecs.soton.ac.uk

Liz Lyon, Ann Chapman, Michael Day
UKOLN
University of Bath
Bath, United Kingdom
{e.lyon}@ukoln.ac.uk

ABSTRACT

As digital libraries proliferate (together with other forms of online information resources), the challenges facing the DL community are no longer those of basic access and resource discovery by naive textual query. Instead, the appearance of systems that provide a deeper interpretation of the literature based on the bibliographic relationships between authors and the articles that they write, shows a new way to help users of digital libraries to discover, access and understand their holdings. This study examines the potential of added value services that can be constructed from a systematic semantic interpretation of the resources held by a digital library, and makes recommendations for digital library practice that will achieve this goal.

TABLE OF CONTENTS

1	Introduction.....	1
1.1	Scope.....	1
2	Background.....	1
3	Technologies.....	4
3.1	Ontologies and the Semantic Web.....	4
3.2	Knowledge Extraction.....	7
4	Enhancing the Digital Library.....	9
4.1	Bibliographic Management.....	9
4.1.1	Ontologies for Bibliographic Metadata.....	9
4.1.2	Linking with Bibliographic Metadata.....	13
4.1.3	Services for the Distributed Digital Library.....	14
4.1.4	Shared Bibliographic Information.....	19
4.2	Bibliometrics.....	20
4.3	Issue Tracking.....	24
4.4	Community Modelling.....	25
5	Visualisation.....	27
6	Recommendations.....	32
6.1	Enhance bibliographic management with SW technologies.....	32
6.2	Enhance bibliometric measures with community context.....	33
6.3	Enhance community modelling with bibliometric information.....	33
6.4	Develop visualisations of literature and its context.....	33
6.5	Reassess bibliometric measures.....	34
7	Activities by Partners.....	36
7.1	School of Electronics and Computer Science, University of Southampton (UK).....	36
7.1.1	Advanced Knowledge Technologies (AKT).....	36
7.1.2	Open Middleware Infrastructure Institute.....	37
7.2	ETH, Swiss Federal Institute of Technology, Zurich (Switzerland).....	37
7.2.1	Multimedia Information Management.....	37
7.2.2	ISIS - Interactive Similarity Search.....	37
7.2.3	Organization of Individual Information Space.....	38
7.3	FORTH, Crete (Greece).....	39

7.4	Netlab Knowledge Technologies Group, Lund University (Sweden).....	40
7.5	School of Informatics, University of Edinburgh (UK).....	40
7.6	Technical University of Crete (Greece).....	41
7.7	UKOLN, University of Bath (UK).....	41
7.7.1	eBank.....	42
7.7.2	Agentcities.NET.....	42
7.7.3	Resource Discovery Network.....	42
7.8	UNIMI, University of Milan (Italy).....	42
7.9	University for Health Informatics & Technologies, Tyrol (Austria).....	43
8	Activities by other Groups.....	44
8.1	NISO MetaSearch Initiative.....	44
9	References.....	45

1 Introduction

This study has been undertaken as part of the DELOS “Network of Excellence on Digital Libraries” Knowledge Extraction and Semantic Interoperability cluster. The cluster aims to bring together expertise from a number of inter-related fields – knowledge engineering, information management, digital library and Grid computer science – to explore and develop “models, algorithms, methodologies and processes” to enable greater interoperability, new opportunities for knowledge mining, analysis and community building. These recommendations will inform the work of the cluster members as well as the library and research community throughout Europe and beyond.

The growth of complex, large-scale, distributed information systems such as the Semantic Web and the Grid raises important issues which are closely related to those which are felt in more conventional digital library settings [29]. The semantic diversity of a system in which there is no overall control of the use of vocabularies for describing entities and information resources, and in which commitment to any particular vocabulary is essentially voluntary, presents problems both for human users, and also for software agents which may mediate this information for us. An important lesson learned so far is that there is no one-size-fits-all solution for the representation of knowledge in such as system. Knowledge is necessarily contextual, and a description of a work in terms suitable for one cultural heritage domain may not be entirely suitable for another.

This cross-domain understanding is one of the key goals in Tim Berners-Lee’s vision of the Semantic Web [4], but also applies to the nascent Grid. Where the Semantic Web concerns itself with providing an infrastructure for describing distributed information resources in a semantically rich way, the Grid applies a similar infrastructure to distributed computing resources.

The study represents a review of current practice across the cluster members, and related groups, with respect to the requirements and usage of extracted knowledge. The aims of the cluster are closely related to those of the Semantic Web’s efforts. In digital libraries there are many disparate and disconnected sources of data, which if joined would likely provide a greater whole than the sum of its parts. The Semantic Web offers an approach that will allow heterogeneous collections of information – databases and the “Deep Web” – to be mapped together, to allow supra-services to reason over the collection.

1.1 Scope

This study aims to summarise the areas of research the cluster members are working in, how that work fits within the general field of knowledge-based research, the technologies used, and how those technologies can be further used to achieve the aims of the cluster.

2 Background

The development of digital libraries has largely been a response to the changing perceived needs of user communities. Digital library systems aim to satisfy the information requirements of individual users in a way that traditional libraries cannot, and by doing so support the communities of which the users are a part, either implicitly by supporting the tasks of their members, or explicitly by reifying the processes and

workflows of the community as a whole. For example, an open archive in which authors self-archive their scholarly works supports the process by which a research community communicates research results amongst itself.

There have been a number of studies of the typical processes within digital libraries, of which the Open Archive Information System (OAIS) model [24] is a common example. The OAIS model is a conceptual framework for an archival system dedicated to preserving and maintaining access to digital information over the long term. The OAIS Functional Model, shown in Figure 1, describes the flow of information through an open archive, and it is in terms of this model that we shall describe the role that knowledge extraction, issue tracking, bibliometrics and community modelling can play within a digital library, and indicate how this enhances the individual and community processes which the digital library supports.

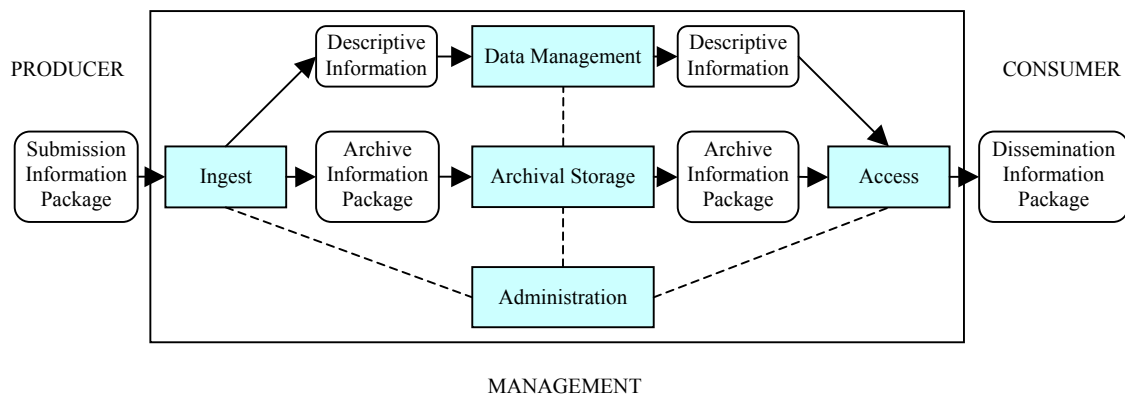


Figure 1 The OAIS Functional Model

The OAIS functions which we primarily address in this report are those which the user community will interact with, namely the Ingest and Access functions. The Ingest function is responsible for transforming a submitted information object into a form suitable for archival, including the generation of a suitable description. The Access function controls the dissemination of information from the archive, both information objects themselves, and the descriptions of those objects. We also consider the Data Management function, which maintains descriptions of information objects, ranging from bibliographic metadata to image thumbnails.

These functions represent a user's visible interface to an archive, and the capabilities provided by them largely determine the degree and nature of the support that the archive provides to the user, and by extension, to the user's community. Enhancements to these functions (for example, support for the automatic creation of descriptive metadata in the Ingest function) may therefore benefit both users and user communities.

In the remainder of this report, we describe the application of knowledge engineering technologies (knowledge extraction, ontologies and the Semantic Web) to the task of enhancing these OAIS functions. We then go on to identify and discuss five areas in which an enhanced open archive could provide additional functionality to the user. Table

1 gives a summary of these technologies and areas of enhancement, categorised in terms of the OAIS functions to which they are most applicable: Ingest, Data Management and Access.

Ingest	Data Management	Access
Knowledge extraction		
	Ontologies and SW	
Bibliographic Metadata		
	Bibliometrics	
	Issue Tracking	
Community Modelling		
		Visualisation

Table 1 Applicability of technologies and applications by OAIS function

3 Technologies

In this section, we describe two technologies from the knowledge engineering community which can be applied to functions within a digital library. Ontologies (and by extension, the Semantic Web) facilitate the semantically rich expression both of descriptive information about an information object, and of the broader context in which that object is situated, while knowledge extraction provides a means to automatically generate descriptive information in the Ingest function.

3.1 Ontologies and the Semantic Web

In Computer Science, an ontology is a formal description of a domain of knowledge¹. Gruber defines an ontology as “a specification of a conceptualisation” [20]: a conceptualisation is an abstract model of some application domain, while a specification is a formal account of that model. Typically, the conceptualisation part of an ontology consists of a set of concepts (things within the domain) and a set of relations that link the concepts in the domain, but may also contain other types of knowledge, from constraints and axioms to procedural (rule-based) knowledge.

Ontologies need not contain relational information, but tend to be known by other names if this is the case. A degenerate ontology which does not contain any relational knowledge other than an IS-A or PART-OF hierarchy is better known as a taxonomy or a meronymy, depending on the sense of the hierarchical relation. Similarly, an ontology which consists only of a set of concept names is a controlled vocabulary. Ontologies therefore form a spectrum from the least expressive controlled vocabularies, to highly expressive ontologies with rules that allow new knowledge to be inferred from that which is already known. This spectrum is summarised in Table 2.

The specification of a conceptualisation is carried out in a *knowledge representation language* with a well-defined semantics, typically some form of mathematical logic. The choice of mathematical logic as a foundation allows the use of software reasoners based on provably sound and complete algorithms for reasoning about the knowledge expressed in a logic-based ontology. Soundness and completeness are terms for properties of the reasoning process; soundness means that a reasoner will deduce no incorrect knowledge, while completeness means that the reasoner will deduce everything that should be deduced – there is a clear analogy with the information retrieval measures of precision and recall.

The artificial intelligence community has extensively studied the issues surrounding the use of formal logic for knowledge representation. Most common knowledge representation languages have some kind of logical foundation, from frame-based systems (from which object-oriented techniques developed – notions of classes, instances and attributes) to network knowledge representations such as semantic networks or conceptual graphs.

¹ The philosophical notion of Ontology as “the science or study of being” is related, however in computer science usage “an ontology” is an engineered artefact rather than a field of study

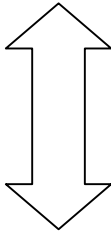
 <p>Less expressive</p> <p>More expressive</p>	Controlled Vocabulary	List of terms
	Taxonomy/ Meronymy	List of terms + hierarchy
	Ontology	Taxonomy + relations
		Taxonomy + relations + constraints
	Taxonomy + relations + rules	

Table 2 From controlled vocabularies to ontologies

The construction of an ontology, particularly the formal, machine-understandable account, allows a system to reason over a set of data, essentially turning a system that is implicitly structured by human understanding into a system that is explicitly structured and therefore understandable by a computer. As an example, Miles-Board describes an ontology for management reporting [32] as follows:

“The domain ontology defines the concepts (for example, Staff, Reports, Projects) and relationships (for example, Staff work on Reports), in order to provide principled and intelligent navigation of the knowledge in the domain.”

The notion of an ontology is central to the Semantic Web [4]. Ontologies are used to structure the knowledge that is published on the Semantic Web, and provide the means for software agents to understand and make use of that knowledge. In terms of the OAIS model, ontologies provide a means for more richly expressing the descriptive information about information objects. A metadata vocabulary with a formal semantics, grounded in a task- and community-based context, enables a greater degree of mediation of information resources to users. If it has a better understanding of the descriptive information, a user’s software agent may be able to better make judgements about which information objects satisfy the user’s requirements.

The key technology of the Semantic Web is the Resource Description Framework (RDF), which is used to express knowledge in terms of an ontology. Unlike HTML, RDF is intended for situations in which information needs to be processed by applications, rather than being only displayed to people [30]. RDF is principally an application-neutral framework for information interchange, with orientation to a particular application domain or domains being provided by one or more ontologies. This has the advantage that application designers can use this common framework and concentrate their efforts on the specifics of the application domain.

For RDF to work, it needs the ability to uniquely reference “things” and “concepts,” e.g. using a URL to reference a Web page. More abstract concepts – a Web page having a creator (authors, editors etc.) – require a more abstract form of URL, a Uniform Resource

Identifier (which can share the same properties of global uniqueness, without requiring centralised control of the identifier space). By explicit definition and strict use of identifiers, RDF allows machines to reason over knowledge, e.g. by always using the same URI to identify the creator relation, wherever that identifier occurs a machine can understand the meaning.

Taking an example from the RDF Primer, the English-language statement “**http://www.example.org/index.html** has a **creator** whose value is **John Smith**” may be represented in RDF as:

- a subject <http://www.example.org/index.html>
- a predicate <http://purl.org/dc/elements/1.1/creator>
- and an object <http://www.example.org/staffid/85740>

Note that this example uses a URI to refer to John Smith, and also to refer to the creator relation, taken from the Dublin Core vocabulary.

In this way, the knowledge that is obvious to a human (that John Smith wrote the Web page <http://www.example.org/index.html>) can be semantically defined, and therefore re-used within knowledge applications e.g. to locate all of the Web pages authored by John Smith but excluding those pages that were only edited by him (which a simple by-keyword search would find). A more complex application could extract the knowledge that he is an expert in horticulture from John Smith’s home page (also defined in RDF), so a search performed by a school student looking for poem on trees might not return John Smith’s article first (given that John Smith may work with trees, but doesn’t list “subject” poetry as an interest).

The Semantic Web uses languages based on RDF to define the ontologies which are used by RDF to express knowledge. At present, there are two languages which can be used to define the important concepts in an application domain (classes of object, properties which relate objects to each other, constraints which apply to the members of classes). The first of these languages, RDF Schema, is a relatively simple and inexpressive language which is supported by a wide variety of applications and tools, including ontology editors and specialised RDF databases. The second and more recent language is the Web Ontology Language (OWL) [31], which provides a more expressive language for describing the concepts which make up an ontology.

3.2 Knowledge Extraction

Knowledge extraction is the process of extracting structured, contextually-dependant knowledge from existing information, typically unstructured text, in order to enhance the use and reuse of that information. As an example of knowledge extraction, the Artequakt project [1] uses knowledge extraction techniques to extract structured knowledge about artists from unstructured textual accounts of their lives. These structured accounts are then used to segment the unstructured text and assemble composite biographies of the artists from the text fragments and the extracted knowledge.

Given the paragraph below, Artequakt extracts the fact that the person *Rembrandt* was born on the date *15th July 1606*, and that he had attended the University of Leiden:

“Rembrandt Harmenszoon van Rijn was born on July 15, 1606, in Leiden, the Netherlands. His father was a miller who wanted the boy to follow a learned profession, but Rembrandt left the University of Leiden to study painting. His early work was devoted to showing the lines, light and shade, and color of the people he saw about him.”

This process can be seen as semantically marking-up existing unstructured information. The semantics of something are its *meaning* – a word is a name of a person, that word relates to a date and that date is the person’s date of birth. A knowledge extraction tool will be able to parse existing text, flagging up concepts within the document with their semantic meaning.

In the context of the OAIS Functional model, knowledge extraction supports the Ingest function, which receives information from producers and prepares it for storage and management within an archive. A key part of the Ingest function is the separation of descriptive information (metadata, image thumbnails, etc) from the archive information itself. In many existing digital library systems, the generation of metadata is a largely manual process. The goal of knowledge extraction in this context is to automate the generation of the various types of descriptive information from the object submitted to the archive during the Ingest function, ranging from traditional bibliographic metadata to citation and bibliometric information, semantically-enriched marginal annotations, and contextual community information.

An example of a novel use for knowledge extraction in bibliographic management is the extraction of information about the context in which a document was written from its acknowledgements section. Although they may not have made a significant enough contribution to be considered as an author, the people and agencies who support a piece of work have still played an important role; the informal way in which these roles are reported has prevented them from being analysed to the same extent as citations, an oversight which is being addressed by the CiteSeer project [7].

Semantically rich descriptive information is crucial to the support which digital libraries give to the scholarly community. The eBank project [28] attempts to track and support the scholarly knowledge cycle, and makes the observation that research and learning processes are cyclical in nature, in that subsequent outputs from these processes contribute to overall knowledge, because there is a continuous use and reuse of data and

information. In order to support this reuse, it builds upon an environment where original and derived data are described using a metadata description framework (this metadata either being extracted automatically or created by hand) and richly linked.

Another example of a knowledge extraction tool is Amilcare [9], which is an adaptive system that uses machine learning techniques to adapt to new application domains. Amilcare is trained by human-annotated texts in a given domain, which it uses to build a profile of that domain. Once it has been trained, Amilcare can then automatically annotate further documents from the same domain, e.g. picking out the times and speakers from seminar descriptions.

Amilcare, and other tools like it, aim to address one of the main issues surrounding knowledge-based systems, that of the *knowledge acquisition bottleneck*. Simply put, this refers to the difficulty of determining the salient and necessary facts about some application domain when constructing a knowledge-based system such as an expert system. The traditional view that prevailed in the knowledge engineering community in the 1980s was that such knowledge acquisition from a domain expert should be mediated by a specialised knowledge engineer who provides expertise in the task of abstracting raw knowledge and making it suitable for processing by a knowledge-based system [22]. This can be compared with the classification and cataloguing process by which information professionals generate descriptive metadata for information resources in their collection. In both cases, the process of building a suitably structured representation from largely unstructured sources (text, or the contents of an expert's head) is a manual process which is time-consuming, and frequently prone to subjective biases.

By providing automated support for the knowledge acquisition task, Amilcare and other tools like it are likely to be crucial in the widespread uptake of the Semantic Web. The Semantic Web is a development of the World Wide Web which aims to provide an infrastructure for machine-understandable information on the Web. By expressing information in a form which makes the meaning, or semantics, accessible to machines, the goal of the Semantic Web is to create a next-generation Web in which the users and the information on the Web can be mediated by software agents.

The Semantic Web vision relies upon structurally and semantically meaningful information, compared to the current Web that is based on a display-description paradigm with an opaque hypertext system (links do not provide information on the nature of the relationship between pages). Ciravegna [8] points out that manual annotation is arduous and error-prone. Widespread document annotation is unlikely to be something that Web authors undertake out of choice, so information extraction techniques, like those employed by Amilcare, will allow the automatic and semi-automatic annotation of texts necessary for the Semantic Web.

4 Enhancing the Digital Library

In this section, we discuss five areas in which a semantically enriched view of an open archive, together with knowledge extraction techniques, may be used to enhance the functions of a digital library as described by the OAIS model. We consider the development of common protocols and ontologies for bibliographic information to improve interoperability between open archives, bibliometric techniques for describing the patterns of publication, the use of issue and claim tracking to follow the evolution of an idea within a body of literature, the role that the community context of a publication plays in making sense of that publication, and finally the use of visualisation techniques for the effective communication of collection-wide information to users.

These areas are frequently complementary, so an enhancement in one enables more sophisticated behaviour in another. For example, a thorough characterisation of citation types to enable the tracking of claims in scholarly discourse might be used to inform the bibliometric measures which are used to describe publication patterns.

4.1 Bibliographic Management

The main domain to which these knowledge-based technologies are to be applied is that of *bibliographic management*. Although it originated with the description of printed materials such as books and journal issues, it now encompasses a larger space than simply the creation of catalogue entries for physical artefacts. In short, bibliographic data can be described as any data that identifies or describes works of intellectual or artistic creation, regardless of physical form.

4.1.1 Ontologies for Bibliographic Metadata

Bibliographic metadata takes a number of different forms. Standards used range from the complex and expressive, e.g. the family of MARC (Machine Readable Cataloguing) formats used by libraries, to relatively simple vocabularies like Dublin Core, intended for use as an 'interlingua' between more complex systems. Other standards include the XML-based ONIX (Online Information eXchange) format², used by the publishing trade for encoding descriptive information about books, including that typically contained on book covers, e.g. 'blurbs' and quotations from reviews. This standard is currently being extended to cover other resource types, e.g. serials and multimedia.

The Library of Congress originally developed the MARC standard in the 1960s, initially as part of a project exploring the potential uses that libraries could make of a centralised pool of standardised bibliographic data [18]. Development of the format continued in co-operation with the British National Bibliography (BNB), resulting in the development of a standardised structure for the interchange of bibliographic information on magnetic tape (ISO 2709). From the start, national bibliographic agencies implemented the format in different ways, resulting in the proliferation of national MARC formats, further

² Online Information eXchange format (ONIX), developed and maintained by EDItEUR with Book Industry Communication (UK) and the Book Industry Study Group (USA).
<http://www.editeur.org/onix.html>

complicated by the development of variants by bibliographic utilities and library system vendors. Early responses to this proliferation included the development of standardised MARC formats for interchange, e.g. the UNIMARC (Universal MARC) format.³ More recently, there has been an increased focus on format convergence, e.g. the union of USMARC and CANMARC in 1997 to create MARC21,⁴ and on interaction with 'core' formats like the XML-based MODS (Metadata Object Description Schema)⁵ and Dublin Core.

The development of the MARC formats gave a huge impetus to bibliographic record supply and the development of shared cataloguing programmes. The creation of bibliographic data is an expensive activity for libraries, and the existence of a standard exchange format enabled them to obtain records from national bibliographic agencies or bibliographic utilities run by library co-operatives. Some of these co-operatives operate shared cataloguing programmes, maintaining large union catalogues like OCLC's WorldCat⁶ usually only made available to member libraries on a commercial basis.

Other types of union catalogue have been developed to facilitate end-user access to resources. Traditionally, such union catalogues have been based on the physical merging of bibliographic data from multiple databases into a single catalogue. Examples include the California Digital Library's public-access MELVYL system, which combines the holdings of University of California libraries by converting incoming records to a standard format, using sophisticated matching algorithms to merge them into single records with multiple holdings information [12]. The UK Consortium of University Research Libraries (CURL) union catalogue Copac⁷ does an initial check based on matching identifiers or author/title acronyms and data, and merges duplicates identified by this process into a single record for each item [10]. Bibliographic management issues frequently arise in the alignment of different metadata schemes in the creation of union catalogues. For example, the UK Revealweb initiative⁸ is building a national union database of resources available in accessible formats like Braille or audiocassette tape. The databases being combined include records in a range of formats, including UKMARC, modified UKMARC, and non-MARC library system formats; also data stored in Microsoft Access databases. Revealweb is a centralised union catalogue whereby records from different databases have been merged into a single physical database.

Translation between formats is typically based on the development of mapping tables or crosswalks. These are usually manually generated and maintained, and are therefore expensive. The high cost acts to discourage any local variations in the format of descriptive information, even where this might be beneficial to users. Manual translation

³ UNIMARC. <http://www.ifla.org/VI/3/p1996-1/sec-uni.htm>

⁴ MARC21. <http://www.loc.gov/marc/>

⁵ Metadata Object Description Schema (MODS). <http://www.loc.gov/standards/mods/>

⁶ OCLC WorldCat. <http://www.oclc.org/worldcat/>

⁷ Copac. <http://copac.ac.uk/>

⁸ Revealweb. <http://www.revealweb.org.uk/>

is necessary because there is no formal account of the meaning of metadata terms that might assist in the automatic or semi-automatic translation of vocabularies (but c.f. [17]). Most formats also change over time, meaning that mapping tables and conversion utilities need to be regularly updated to take account of these.

An alternative approach to the development of union catalogues is to use distributed search technologies to create virtual union catalogues. Examples of these include the virtual Canadian union catalogue (vCuc),⁹ and the *Karlsruher Virtueller Katalog KVK*,¹⁰ both of which are based on the ANSI/NISO Z39.50 search and retrieve protocol. Virtual union catalogues are much cheaper to implement than the traditional centralised model but in practice there tends to be problems with search and retrieval accuracy. Some problems with Z39.50 were noted by the evaluation report on the AGORA hybrid library project [6]. Firstly, some interoperability problems resulted from the differing extent of implementation of the protocol by library system suppliers, others from the use of different content standards (e.g. cataloguing rules), making the sorting and de-duplication of result sets difficult. Coyle has written, "it appears that the common use of Z39.50 in libraries today is not a distribution of our catalogs, but a kind of harvesting in disparate databases ... we still seem to harbor a somewhat illogical hope that this harvesting will inexplicably yield consistent and accurate results" [11]. Lynch has further argued that the query language that can be supported will be the "lowest common denominator of all the query languages supported by the systems servicing the distributed search" [27]. In comparing Copac with pilot virtual catalogues, a feasibility study for a UK National Union Catalogue [41] concluded, "it was evident that the physical catalogue architecture offered a more reliable, faster and consistent response than any of the virtual systems tested." They also found that the creation and operating costs for both physical and virtual systems were broadly similar, undermining the perception that virtual catalogues would be cheaper to implement.

As was described in Section 3.1, the knowledge engineering notion of an ontology consists both of a vocabulary for describing some application domain, and a formal description of the meaning of that vocabulary in terms which can be understood by a computer. Metadata schemas can be viewed as a type of degenerate ontology, because ontologies are generally more expressive (see Table 2) and allow more subtle distinctions to be made in the definition of terms in a vocabulary. The issue of ontology translation is directly equivalent to metadata schema translation, and has been studied extensively in the knowledge engineering community. Techniques for ontology mapping are used to translate between ontologies, which encourages a heterogeneous environment in which the ontology used by a small community of agents can reflect their true representational requirements, rather than the lowest common denominator for a much larger community.

In addition, the nature of the languages used to define ontologies leads to the construction of modular ontologies, in which multiple viewpoints can be provided of a domain within a general representation framework, with some innate degree of interoperability. A simple example of this is the relationship between basic Dublin Core (the fifteen

⁹ Virtual Canadian union catalogue (vCuc). <http://www.collectionscanada.ca/8/6/index-e.html>

¹⁰ Karlsruhe Virtueller Katalog KVK. <http://www.ubka.uni-karlsruhe.de/hylib/en/kvk.html>

metadata elements), and Qualified Dublin Core. The terms in the latter set extend those in the former in order to define more specific concepts, and in doing so retain a degree of interoperability. We can envisage an environment in which a community may choose to extend a basic common vocabulary with terms which are necessary to fulfil their requirements, and yet still be able to exchange information with other communities who have chosen to commit to the same common vocabulary.

Modular ontologies also allow agents to minimise their ontological commitments and therefore the cost of commitment. Complex metadata schemas like the MARC format carries high incidental costs; users must be trained to a high degree in order to correctly generate records which conform to such a schema - although this may actually be a consequence of the content (cataloguing) rules in use. While a less detailed ontology may not meet the representational requirements of a community or agent, the adoption of an overly detailed ontology may involve high enough costs to outweigh any benefits that may follow from committing to it. By breaking the representation of an application domain into faceted components, and building layers at different levels of detail, an agent can select only those modules that it considers necessary and disregard the remainder.

Few libraries are now cataloguing all materials they acquire themselves. Some receive bibliographic records along with the physical item from their library supplier. Some acquire records from union or co-operative catalogues, or from bibliographic record suppliers such as the British Library, BDS and BDN. Increasingly in-house cataloguers simply add local data, to meet local needs, and only create records from scratch for those few items for which a record is unavailable elsewhere, though the proportion varies from library to library. This 'create once and re-use' aspect reduces the cost to each user. Customers may be able to specify the level of detail required in the record supplied. Localised ontologies will therefore need to accommodate records from other sources, most likely the source the item originated.

In the 1970s there was a move to short form cataloguing. This was in part prompted by the cost of space on computers, and partly the thesis that users needed far less information. Nowadays, cost of space is not usually an issue, and records contain increasingly detailed data, such as content summaries, tables of content and even images. The reality is that for any item one user may require only minimal details, while another requires as much as possible. The extra detail may be required on content (by a researcher) or on physical attributes (by those with physical or sensory impairment who are restricted in the formats they can access) or on relationship to other items (e.g. digital version of a manuscript, teacher and pupil versions of a textbook, translation of a work originally in another language).

The enhancements that can be made to bibliographic management by using expressive ontologies fall predominantly into the Data Management OAIIS function, because they affect the form of the descriptive information packages. However, they have subsidiary effects at other times during the lifecycle of a metadata record. When open archives are federated, metadata records from other archives may pass through the Ingest function, while a rich language for describing information objects lends itself to the formulation of more expressive and effective queries in the Access function.

4.1.2 Linking with Bibliographic Metadata

There is a strong relationship between bibliographic metadata and the means by which a document may be located; search interfaces used to access information objects typically require that the search expression is expressed using a similar vocabulary to that used for descriptive information. In a Web context, where hypertext links are used to create navigable structures that may be browsed by users, we may use bibliographic metadata as a kind of stored search which is instantiated as a link; traversing the link causes the search expression to be evaluated, returning the document which is described by the metadata.

While the capability of linking is well understood, and widely implemented, being able to reliably and persistently provide links to the location of an item has proved difficult. Links to items by location (e.g. a Web page URL) hard-coded by the author of a document break when the location of the item changes. Broken links degrade the experience of users. When the context of the link may promise an ideal resource, but with insufficient information to actually locate that resource in the event of its location changing, broken links lead to a frustrating experience that may turn the user off the service. Maintaining accurate links by hand is an impossible task as, especially on the Web, the location of pages can often be counted in the days, or entire collections can be moved or simply disappear. In contrast, the use of “bibliographic links” allows for accuracy and persistence of linking within the digital library system, where the linked-to items can be trusted to continue existing (and advertising their presence) within the system. Most importantly the source of the link stores descriptive data about the target, allowing the target to be more easily found in the event of it moving or changing – similar in fashion to a scholarly citation, which has allowed users to follow citations to the cited material for hundreds of years.

SFX and OpenURL is a high-level infrastructure for implementing context-sensitive, dynamic linking to works (e.g. a research article). SFX is based on a two-layer model of a metadata and linking layer. The user traverses these layers by following OpenURLs from the metadata to the linking layer, and then following an absolute URL generated from the OpenURL at the linking layer to the metadata layer.

An OpenURL is an encoding of bibliographic metadata into a URI, appended to the base URL of an *OpenURL resolver* (a Web service that understands OpenURLs). For a journal article this might encode, the journal title, author, volume, and pagination. When a user clicks an OpenURL link, the bibliographic metadata can be used by the *OpenURL resolver* to either redirect the user to the article, or provide a set of links to further information on that article.

SFX introduces two contexts into linking: the source, and the user context. The source, the service which generated the OpenURL, is encoded in the OpenURL using a unique ID. The second context is provided by the OpenURL target. The target may be a machine within the user’s place of work, which knows about the user’s rights of access to Web services.

In order for an *OpenURL resolver* to resolve an OpenURL it needs to have a database of metadata records with which to match the OpenURL against. How the target builds this database is not defined within the SFX infrastructure. This would appear to limit the

potential for the interoperability of SFX services, as a resolver may need to support a large number of protocols to obtain metadata.

The metadata formats currently supported by OpenURL are focused on a set of types of work, e.g. journal articles, conference proceedings, patents, books. The type of work may be difficult to determine, for example a service that automatically adds OpenURLs to citations may find it difficult to determine the type of work being referred to from the reference itself, although it may be possible to identify that a particular term within the citation is a year, and another a volume. In this situation the service would need to pre-resolve the reference metadata (using some other means) to determine the type of the work, and then build the OpenURL using that knowledge. By having to pre-resolve, much of the usefulness of a dynamic linking environment is lost, although OpenURL is still useful as a means for users to navigate services. In highly unstructured situations it may be that a more flexible descriptive framework would be needed in order to build bibliographic links, but which would still be resolvable by an intelligent OpenURL resolver.

Resolution services of this kind support both the Data Management and Access functions in the OAIS model. OpenURLs can be used to express links between documents in the descriptive information as a means to manage citation information internally, or can be exposed to the user (as an Access function) to provide a means to make citations into navigable structures like hypertext links.

4.1.3 Services for the Distributed Digital Library

Enhancing the vocabularies used to express descriptive information is not sufficient by itself. The prevailing view is that the digital library is not a monolithic system, but a distributed federation of heterogeneous archives. The manner in which these archives communicate and the protocols they use to exchange information are a necessary counterpart to the choice of metadata vocabulary.

The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) allows distributed, heterogeneous repositories of documents to be *harvested* to form a single, federated collection. The purpose of the OAI-PMH is to allow repositories to expose, as easily as possible, their collections to service providers. By minimising the barrier to interoperability OAI-PMH aims to achieve widespread adoption, hence establish an environment where services can more easily access material to build collections from.

With such a common protocol in place, we can create services which work across archives. Citebase Search allows users to find research papers stored in open access OAI-compliant archives (see Table 3). Citebase harvests OAI metadata records for papers in these archives, as well as extracting the references from each full-text paper. The association between document records and references is the basis for a classical citation database (similar to Web of Science or Citeseer). Citebase is best viewed as a kind of “Google for the refereed literature”, as it ranks search results based on the number of references to papers or authors. Google combines a search relevance score with a page ranking algorithm calculated from the number of Web links to a page. Citebase supports ranking search results by the absolute number of citations to papers and by search score,

but does not provide a combination of the two. Citebase contains 320,000 full-text eprint records, and 9 million references (of which 2 million are linked to the full-text).

Citebase was developed using data from the JISC/NSF Open Citation Project, which ended December 2002. Citebase has continued to be developed beyond the life of the project and has been integrated with arXiv.org. As part of the final OpCit Project Report a user survey was conducted on Citebase [23]. This was used both to evaluate the outcomes of the project, and to help guide the future direction of Citebase as an ongoing service. The report found that “Citebase can be used simply and reliably for resource discovery. It was shown tasks can be accomplished efficiently with Citebase regardless of the background of the user.”

In the OAIS model, Citebase is primarily an Access function. It provides a Web site that allows users to perform a meta-search (title, author etc.), navigate the literature using linked citations and citation analysis, and to retrieve linked full-texts in Adobe PDF format. Citebase also provides an export of the citation data it collects through its own OAI-PMH interface using the Academic Metadata Format (AMF), a new XML format for scholarly literature. The export also supports Dublin Core (oai_dc), which is essentially the same as the records harvested, and experimental support for the Dublin Core Citation format (dc_citation). The Open Archives Initiative (OAI) Protocol for Metadata Harvesting (PMH) is designed to address the need to expose metadata - titles, authors, abstracts etc. - from research literature archives in a structured form. An XML protocol built on the HTTP standard, OAI-PMH is in effect a CGI interface to databases. Based on 6 commands (or “verbs” in OAI terminology) OAI-PMH allows metadata to be incrementally harvested by service providers (the HTTP client) from data providers (the HTTP server).

Citebase makes use of the OAI-PMH to harvest metadata from e-print repositories. A list of base URL’s (the CGI interface) of repositories is stored, along with the date of the last harvest from the repository. During the daily update Citebase requests any new or changed records from each repository, since the last time Citebase successfully harvested. The date used is the time at the start of the harvest process, otherwise records that may be added during a harvest may be missed in the next update.

arXiv.org	http://arXiv.org/	301754
Biomed Central*	http://www.biomedcentral.com/	14432
University of Southampton*	http://eprints.soton.ac.uk/	2861
Cogprints	http://cogprints.soton.ac.uk/	2052
Research in Computing, Library and Information Science	http://www.rclis.org/	1920
UoS, Electronics and Computer Science	http://eprints.ecs.soton.ac.uk/	1165

Computer Science		
W3 Conference Proceedings	http://wwwconf.ecs.soton.ac.uk/	430

Table 3 Repositories harvested by Citebase Search, with total number of records from each repository.
 *Contains a mixture of free full-text and metadata/publication-only records.

The oai-perl libraries, developed as part of Citebase, provide an abstract interface to OAI. This hides the complexity of PMH flow-control and error-handling. Using the oai-perl libraries for harvesting consists of creating a repository object that contains the base URL and then calling the PMH commands on that object (e.g. ListRecords). The library returns a list of objects that contain the metadata as an XML DOM (Document Object Model), or an error code and message. Only if the harvest completes successfully is the date updated. In the event of an error the harvest is started again at the time of the next update.

As changes have been made to the OAI-PMH the oai-perl libraries have been updated, requiring minimal changes to the main code base. The libraries also allow Citebase to harvest from version 1.0, 1.1 or 2.0 PMH repositories (silently converting syntax and responses to the most recent version).

As a developmental service it is often the case that Citebase's database has to be rebuilt, e.g. when author name parsing code is changed. As repositories may potentially contain many millions of metadata records the Celestial cache was written to harvest records from OAI repositories and re-expose those records to OAI services, in effect an OAI cache. Celestial harvests records from multiple repositories at very high speeds, using multiple processes to harvest from repositories simultaneously. It is designed to allow the export of PMH records as fast as possible, trading storage requirements for speed. This allows Citebase, as a harvester from Celestial, to rapidly rebuild its database from the source material without loading the source archives and usually faster than the source repositories OAI implementations allow.

As Celestial uses the same oai-perl libraries as Citebase, it can harvest from any version of the PMH. All harvested repositories are silently converted to the latest version (2.0), and re-exposed as version 2.0. Apart from the protocol syntax itself, the only other change required is to change the namespace of the mandated metadata oai_dc format (Dublin Core).

The libraries also contain the ability to fix errors that repositories may have. Often OAI repositories import data without converting and checking that characters are in UTF-8 (a way of encoding non-English characters). A single bad character could prevent an entire repository being harvested, therefore the oai-perl libraries attempt to replace the location of unparseable characters with character that will parse ("?"), rather than generating an error and giving up. To avoid the overhead of parsing XML when exporting the metadata records, Celestial stores the XML as raw data in a database.

Citebase's reference parsing either parses semantically structured or unstructured documents. Structured documents are LaTeX or XML from respectively arXiv.org and biomedcentral.com. Any other format falls under unstructured.

The first method used by Citebase (under the Open Citation Project) to extract references from documents was by parsing the ‘bibitem’ entries from LaTeX documents, as written by authors. These bibitem mark-ups (“/bibitem”) enclose a free-text string containing a citation. Depending on the citation style it may contain a reference to more than one real-world article. The Citebase parsing code adds another mark-up around each Bibitem, runs LaTeX over the document, then using the custom mark-up extracts the unstructured references. Processing the LaTeX source is necessary to expand any macros that the author may have included within the references (which would otherwise require writing a custom LaTeX processor!). Documents from biomedcentral.com are parsed for ‘bibl’ structures, which contain fully structured citations and require no further processing.

Unstructured documents are where there is no semantic mark-up within the document to indicate what the text means (e.g. is this an author’s name?). Postscript, Adobe PDF, and HTML are all unstructured text (with varying degrees of typographical structure). To parse the references from these Citebase converts them to plain text, then passes the text to code that attempts to find the references. The first step to extracting the references from the plain text is to locate the reference section or bibliography. Citebase does this by locating a title containing the word “reference”, “bibliography” or “notes”. A title is a line of text by itself, preceded by a number (e.g. “20. References”), or capitalised (e.g. “REFERENCES”). If the text is in two columns it is de-columnised by finding the modal distance from the left margin where a large space occurs (a check is made by finding if this distance is approximately 50% of the mean line length), and splitting each line around that point. A number of rules are then checked against the text body below the references title to separate out individual references, and to find the end of the reference section (figures and acknowledgements may be at the end of the document). The simplest reference style to parse is numbered, either “1.” or “[1]”. If the references are unnumbered they may be separated by whitespace, or by the publication year (e.g. “authors ... (1993) ...”).

Parsing references from documents is generally more successful the closer to the original version the parsing can be done. While processing Latex files seemingly requires an infinite number of style libraries (Latex macros), parsing the Latex will avoid any errors that may be introduced by getting back from a presentation format (e.g. PDF) to a usable format. It is ironic that PDF – a format designed to preserve the layout of a document – makes it nearly impossible to get back to the original document. The nature of PDF makes it difficult to reliably extract the text of the document, as PDF does not store the thread of the document. E.g. the end of one column is not associated with the beginning of the next section of text.

The reference parsing facility of Citebase is a set of cases evolved over time that provide a good coverage of the literature Citebase is required to process. The documents that can not be parsed may not be convertible (e.g. no current support for Microsoft Word format), or convert badly (e.g. conversion of postscript format to plain text often results in garbage, as postscript may store only pictographic representations of characters, rather than the letter). Once in plain-text the ability to parse the references is dependent on the format and style used by the author. With a seemingly unlimited number of possibilities some references will be unavailable (or just incomprehensible!).

With successful extraction of the references from the full text, Citebase can parse individual references into their components: authors, title, year of publication, serial, identifiers, etc. The citation components can then be used to locate the full-text record of the cited article. This creates links between citing and cited articles, which creates a citation database.

Citations are intended as unambiguous references to a real-world thing. They do this through providing a meta-description of the cited thing. In most cases this is a bibliographic reference to a journal article, e.g. consisting of author, year of publication, journal and issue, and page reference. In a database it is convenient to store numbers, as it makes it quick to locate matches. A bibliographic reference can be reduced to 3 numbers (year, issue and page number) with an author and/or journal name to avoid false numeric-only matches. This, coupled with identifiers provided by authors, allows references to be linked to the cited article (where that cited article also occurs in Citebase).

Where static linking of references has failed within the document collection harvested by Citebase, some experiments with dynamic linking have been added. The first of these was to link to the publisher's site where a journal title is recognised (again through sets of regular expression rules). If an author cites an article in *Physical Review B*, Citebase will generate a link to the American Physical Society's Web site using their link service. This link service uses URLs constructed using a journal identifier (in this case "PRB"), volume and page number. If a user has a subscription to the APS they can follow the links from Citebase to the publisher's full-text. Similar links are generated for other APS journals, as well as the journals Nature and Science. Unfortunately most publishers do not provide a convenient mechanism for structured linking to their sites, e.g. *Nuclear Physics B* as part of ScienceDirect (published by Elsevier)¹¹.

Citebase now has support for OpenURL resolution and acting as an OpenURL *source*. As a source Citebase adds OpenURL ('O') links to all references and citations. OpenURL links encode the bibliographic data contained in a reference (e.g. author, journal, year) as a URL. The base of the URL is an OpenURL resolver – a server that can search a database of citations to map that bibliographic data to an appropriate copy of the target article for the user. As a resolver (or OpenURL target) Citebase attempts to find matching articles using bibliographic data supplied by an OpenURL link, performing a search similar to the existing static reference-linking. If that search fails a number of links are provided to perform a more generic search in Citebase, which may generate more than one match. In future – if the article isn't available from Citebase - this OpenURL jump-off page may provide links to other services that support OpenURL, or have knowledge about user-context in order to provide user-specific links e.g. to services that the user has

¹¹ E.g. Pierre Le Doussala and Kay Jörg Wiese (2004) "Derivation of the functional renormalization group β -function at order $1/N$ for manifolds pinned by disorder", Nuclear Physics B:

http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6TVC-4D97JMF-1&_user=126770&_handle=B-WA-A-W-Z-MsSAYVW-UUW-AAUCYBCCYY-AAUBVAZBYY-YAVVVWDYV-Z-U&_fmt=summary&_coverDate=11%2F29%2F2004&_rdoc=2&_orig=browse&_srch=%23toc%235531%232004%23992989996%23537365!&_cdi=5531&view=c&_acct=C000010399&_version=1&_urlVersion=0&_userid=126770&md5=5e03577ad864a834b2e5971cf0f519d4

subscribed to. In this way all reference links could be driven through the OpenURL resolver, allowing the service to dynamically change with the user's context.

As mentioned earlier, Citebase exports its citation data as fully-formed OpenURLs (pointing to Citebase's resolver). This would allow an OpenURL resolver service to harvest metadata records from Citebase in order to pre-resolve references to Citebase-indexed papers, where available. There is a difficulty with building truly dynamic linking systems in that the resolving service has to first query a target before it knows it can provide a successful link to that target. This is further complicated by there possibly being multiple targets, with differing amounts of access granted to the user being serviced by the resolver. This problem can be mitigated by the resolver pre-harvesting all possible links from targets, thus when a user clicks an OpenURL link the resolver can search for all possible links, in real-time, in its own database.

An alternative use for the OpenURL OAI export from Citebase would be to build a CrossRef-like service for resolving bibliographic references to unique identifiers. As it currently stands the OpenURLs exported by Citebase could be used to resolve references to OAI identifiers (which are guaranteed unique only within the context of the repository), or specifically to arXiv.org identifiers which are globally unique and persistent (which use the arXiv.org identifier in the OAI identifier). In this way any repository could be linked to by OpenURL export/resolution.

The insertion of OpenURL-style citation links can be retroactively applied to all documents, not just to new documents during the authoring process. The ParaCite system builds on OAI-compliant systems such as ePrints, and provides the means to build links between an online document and the documents which it cites in a simple way.

ParaCite couples a reference parser (that extracts from a plain string bibliographic terms e.g. author, year etc) with search engines that contain research material. As a first step ParaCite parses references entered by a user through a Web interface into an internally stored OpenURL. The user can then search for this citation in a number of bibliographic search services, depending on the type of the reference, in order to quickly scan across multiple services to find the text of a cited article.

OAI-like services enable the distribution of the OAI Data Management function, and support the development of more advanced services such as ParaCite and CiteBase which support the Access function.

4.1.4 Shared Bibliographic Information

In the same way that OAI-compliant systems can share bibliographic metadata at an archive level, there are services which enable the sharing of citations and bibliographic metadata at a personal level. These can be viewed as a logical progression of annotation services, which allow users to create and share marginal annotations on web pages and other documents. An example of such a shared annotation service is the Annotea system from the World Wide Web Consortium; users use an enhanced web browser which can talk to a specialised annotation server. When the browser loads a web page, it consults the server to see if there are any annotations which can be applied to the page. Similarly, when the user creates an annotation (typically by highlighting a phrase within the

document and typing a textual annotation), the browser informs the server of the annotation so that it can be accessed by other users.

One example of a citation sharing service is Bibshare, which provides extensions for common tools such as Microsoft Word that enable users to search Web-based bibliographic servers for citation information to insert into the bibliography of the documents that they are authoring. Bibshare provides a search engine which handles the federation of search across multiple heterogeneous bibliography servers, and supports OAI-PMH-style services in addition to more Web- and Web Service-based services.

CiteULike is a similar service which mediates access to bibliographic information for a variety of sources, including PubMed, ScienceDirect, Wiley InterScience and JSTOR. These services enhance the OAI-PMH Access function, by providing users with the means to reuse existing sources of bibliographic and descriptive information.

4.2 Bibliometrics

Bibliometrics uses quantitative analysis and statistics to describe patterns of publication within a given field or body of literature [36]. Ronald Rousseau [38] traces the start of bibliometrics to 1913 with the discovery of what was later called Zipf's law¹². Zipf's law is based on the frequency analysis of word-occurrence, but similar patterns are found in other areas. Studies [40,37] have found Zipf's law applies to the citation impact of journal articles (where articles are rank-ordered by the total number of citations to them). There is some debate as to whether adherence to Zipf's law reflects a deeper meaning, or is simply a function of the nature of the data. Regardless, the consequence for citation impact is that the more citations a paper has, the more likely it is that those papers will receive further citations.

More recently, bibliometrics has gained importance due to the use of statistical analysis of research material for the purposes of evaluation. Gene Garfield's¹³ Science Citation Index (first developed as a printed index in the 1950's) has grown to be used not only as a navigational and discovery tool for science (by following citations created by authors), but also as a quantitative measure (by counting citations) of a work's importance, and, by proxy, the importance of the journal the work was published in, the importance of the authors that wrote it and their institutions with which the authors are affiliated¹⁴. A new study by the Times Higher Education Supplement (THES) ranks universities globally, using in-part citation impact scores from the Science Citation Index¹⁵.

¹² P_n similar to $1/n^a$, where P_n is the frequency of occurrence of the n th ranked item and a is close to 1. <http://www.nist.gov/dads/HTML/zipfslaw.html>

¹³ Gene Garfield's home page: <http://www.garfield.library.upenn.edu/>

¹⁴ "Performance Mandate of the Swiss Federal Council" points to bibliographic studies as evidence for the success of the ETH ("[ETH is among] the 50 best of the 5,000 universities worldwide") http://www.sl.ethz.ch/docs/oeff/la/leistungsauftrag_e.pdf

¹⁵ THES "World Rankings": <http://www.thes.co.uk/worldrankings/>. The THES world ranking of higher education institutions ranks by peer-opinion, number of foreign students, staff to student ratios and citation impact. The citation impact score of research staff is determined using ISI's science citation index.

Bibliometrics relies on structured data about its subject. Most commonly, this is a citation database: an index of published texts, authors, publishing body and references from those texts linked to the cited texts. The Science Citation Index (by ISI Thomson) covers around 7000 “core” journals in the sciences, along with separate databases for the humanities and conference proceedings. While the ISI database provides structured citation data, information is missing on other scholarly outputs and publications not covered by those databases.

Current citation databases rely upon a combination of automated and manual systems to extract bibliographic data from online and printed documents. CiteSeer¹⁶ [13] is an “autonomous citation agent” that crawls the Web for research articles, parsing them for bibliographic data (titles, authors, abstracts) and references to other works. References are further parsed in order to link them to the cited documents, hence creating a citation database. CiteSeer also provides a number of extended bibliography services which attempt to fill in gaps in a paper’s bibliography by suggesting similar or related documents. These services function as a kind of recommender system, choosing similar documents on the basis of textual similarities or co-citations. Services of this kind fit in the OAIS Access function, and provide new ways in which users can explore the literature.

In a similar fashion, Citebase¹⁷ autonomously parses and links research articles, but utilises existing collections of documents. Citebase uses the Open Archives Initiative Protocol for Metadata Harvesting¹⁸ [26] (OAI-PMH) to retrieve metadata records from research archives that support the OAI-PMH. The newly released Scholar¹⁹ service by Google works in a similar fashion to CiteSeer, but in addition trawls publisher’s sites through special access agreements. This allows Google to index both research material available on the Web, and material available only through the subscriptions, licensing or pay-per-view. Both Scholar and CiteSeer track citations to research materials that have been cited, but are not available to the indexer’s agents, e.g. Google tracks citations to books (of which very few are available online, and few of those are freely available to be indexed by Web crawlers). In addition to these freely available Web tools, most large publishers maintain their own citation indices, which typically cover only their own journals. Elsevier have extended their in-house collection of journals with citation data from other publishers to create Scopus²⁰, which will cover 14,000 journal titles.

CiteSeer et al have been sufficiently successful at automating citation linking (augmented with human-correction) to provide useful services. An important factor in the

¹⁶ CiteSeer at PSU <http://citeseer.ist.psu.edu/> (<http://www.citeseer.com/>)

¹⁷ Citebase Search at the University of Southampton <http://citebase.eprints.org/>

¹⁸ Open Archives Initiative <http://www.openarchives.org/>. The OAI-PMH allows *service providers* to harvest (download) metadata records from *repositories*. A service provider might, for example, harvest Dublin Core records from repositories and provide a Web search service. The OAI-PMH is based on XML and the Web.

¹⁹ Google Scholar: <http://scholar.google.com/>

²⁰ Elsevier Scopus <http://www.info.scopus.com/>

development of these services is the increasing amount of material available either because it is “born-digital”, or from the digitization of older material. “Content is King” is no truer than when it is applied to digital library systems. As more and more content becomes available to the citation indices, so their usefulness will increase. Given the dominance of Google in the Web search field, and increasingly as the first port of call for researchers, it isn’t difficult to foresee the monoculture of Google forcing material either to be made available to its crawlers, or simply disappear from current research use (which is not to say all material will be free to access, but that it’s “Web or dead”).

Bibliometrics based upon citation data – explicitly linking document identifiers using bibliographic references – relies upon extracting the bibliographic references and knowing about the cited item. The more data that is available, the easier it is to automate bibliographic linking; an omniscient system can significantly reduce the difficulty of the problem by using vocabularies of known entities, e.g. author names, publication titles etc. While the problem could be solved by authors providing defined bibliographic metadata for every paper and reference, it is likely that automated systems will continue to be the most common method of reference linking, simply because there is currently no mass-market authoring tool that produces structured (semantically marked-up) documents. Unique identification systems that support reference linking, i.e. the Digital Object Identifier, rely upon the cited item having an identifier, the author knowing that identifier and it being accurately written into an article’s bibliography.

While automated systems have shown success at building citation databases, they cannot solve the problem of resolving names to authors. Duplication of names (authors who share the same name, discipline and sometimes institution) defeats the need to provide accurate by-author publication lists and analyses. While every project and organisation that works with names shares this same problem, there is no widely implemented system for uniquely identifying authors, editors, contributors etc. that allows document repositories to attach to an author name a globally unique identifier. Without that data being captured at the location documents are stored, services that wish to identify authors either have to accept a high level of inaccuracy, or respectively mark-up document collections with author-identification, either through allowing authors to perform that role (by creation of an author record and then claiming documents to being their own) or through large numbers of editors.

Research institutions are increasingly building systems for recording the research output of their faculty and students. The Theses Alive! project²¹ [25] aims to support the implementation of electronic repositories for theses and dissertations in UK institutions by producing a software system which can be used and built on by those institutions. As well as increasing the exposure of the institution’s research output, these institutional repositories allow the capture of structured data about the material deposited. The eBank project [28] uses specialised institutional repositories to store original data, allowing secondary information sources (published articles etc.) to link back to the source data.

Bibliometric techniques are also being applied to the Web at large, typically using the link structure as a substitute for bibliographic citations. The canonical example of this is

²¹ Theses Alive! (Edinburgh University Library) <http://www.thesesalive.ac.uk/>

the PageRank algorithm [34] used to rank web pages that are retrieved by the Google search engine. This identifies pages as hubs (those with a large number of outgoing links) or authorities (those with a large number of incoming links), and assigns a high score to pages which are linked to by other pages with a high score.

Bibliometrics span the OAIS Data Management and Access functions; they can be viewed as an enhancement of existing descriptive metadata, but also enable the provision of advanced services that give users more sophisticated access to the literature.

4.3 Issue Tracking

Issue tracking is a management term for tracking the progress of work in relation to a business issue. In [32], Miles-Board describes a *management reporting system* (MRS) that allows the production of status reports during the progress of projects. This necessitates the use of issue tracking: bringing together all of the aspects of a problem, tracking it through to its conclusion, and making that information available in future. In addition to issue tracking by itself, the embedding of the issue tracking process within a user's workflow is an important subsidiary consideration. MRS uses elements of open hypermedia technologies to augment an author's existing tools and environment in order to support the assisted creation of documents such as status reports, using (and reusing) knowledge from diverse sources.

A different view of issue tracking plays down the document-based aspects described above, and adopts a more purely behavioural. In this, issue tracking occurs only in the context of workflows, in which the actions carried out by users are informed by the progress of an issue. An example of this approach has been carried out within the I-X project [43], which provides a suite of tools that maintain an intelligent 'to-do' list for a group of users. These tools use ontologies for modelling the activities, issues and constraints that are shared by a group of users, and provide support for the collaboration both of human and of software agents towards the resolution of those issues.

A related task to issue tracking is that of claim tracking in scholarly discourse, which also questions the nature of citation. Citation is rarely, if ever, a neutral activity. When the author of a document cites another document, they usually do so for a specific reason; perhaps the cited document supports the author's argument, or perhaps they are refuting a claim made by the other document. A collection-level view of these citations allows us to build a network of claims and counter-claims which we may use to follow the progression and development of an idea or issue.

Citation typing has been discussed by several communities. The hypertext community, which sees citations as a particular type of hypertext link, has investigated other types of link; the seminal contribution here is the taxonomy of link types defined by Randall Trigg [44]. A more recent project examining the nature of citation in the context of claim tracking is the Scholonto project [45], which has developed an ontology for describing argumentation and supporting software for sense making and claim tracking in a collaborative environment. In addition to describing the support/refutation relation, the Scholonto ontology also allows citations to be described in other ways: causal relationships, similarity and contrast, taxonomy and meronymy, and problem-solving.

The addition of issue tracking considerations to a digital library could enhance both the OAIS Data Management and Access functions. At one level, issue tracking can be seen as another type of rich descriptive metadata which attempts to capture the implicit meaning of the relationships between documents that are reified as citations. Issue tracking also provides a navigable structure which users can use to explore the literature, either by browsing the network of citations and claims, or by formulating queries which use information about claims to specify abstract relationships within the content of documents (e.g. "show me the documents which support the claim made in this document, and which were not written by one of the author's common collaborators").

4.4 Community Modelling

Community modelling can be taken to mean one of two things, either the identification of the implicit communities which are engaged in some joint task or activity, or the identification of the joint task or activity in which an explicit community is engaged. These notions are jointly known as communities of practice [46], and have been the subject of some interest in the business community as a means for identifying and harnessing the implicit knowledge within organisations (knowledge that is known to a select few in an organisation, but which is not known by the organisation as a whole).

A community of practice is characterised by three concepts: the joint activity undertaken by its members, the style of the mutual interaction between members that binds it as an entity, and the pool of communal resources that the community creates and may draw upon. The key notion is that these communities are participatory, rather than enforced; membership of an individual in the community is at the discretion of the individual, rather than of some third party. In this respect, communities of practice are self-organising systems.

The modelling of these communities, and the links between their members, shares many similarities with the modelling of scholarly literature, and similar techniques have been applied to both. In [15], Flake et al discuss techniques for identifying Web-based communities through analysis of hyperlinks. Work on communities of practice is coming to prominence within the Semantic Web community, where it is seen as a possible vehicle for the implementation and deployment of mechanisms for expressing and communicating trust in knowledge presented on the Semantic Web.

Friend of a Friend²² (FOAF) provides a vocabulary for describing the kind of information that is found on people's home pages in a machine-understandable fashion, e.g. "My name is", "I am interested in" and "You can see me in this picture". This allows queries to be made over communities of people, e.g. "Show me pictures of people who are interested in Marilyn Manson who live near me." FOAF, more fully described by Dumbill [14], is a domain-specific vocabulary to support the social interactions of humans within the general Web. It isn't necessary for FOAF to be an ontology for the entire Web, as in the Semantic Web different communities with domain-specific vocabularies can be mapped together, to create a greater whole.

Alani et al [2] describe the Ontological Network Analysis (ONA) technique for discovering potential communities of practice by analysing Semantic Web knowledge that has been expressed using an ontology. The aim of this work is to examine the formal, explicit relations that exist between people (e.g. A has authored a paper with C, and B has authored a paper with C, but no direct relation exists between A and B) with a view to inferring the informal, implicit relations (e.g. A shares interests with B). Individuals that are linked by such informal relations can be transformed into a possible community of practice by ranking them according to the strength of the relations and discarding those which fall below a given threshold.

²² Friend of a Friend (FOAF), <http://www.foaf-project.org/>

This technique has been demonstrated to be effective at identifying the implicit community that surrounds an individual (the community of practice whose common interests are those of the given individual) given a knowledge base containing information about the publications and group affiliations of the individuals. Alani et al note a number of potential applications that have some bearing on the digital library domain.

The first application is to recommender systems. A recommender system takes a user profile (interests, documents viewed, publication history, and so on) and generates a list of documents which may be of interest to the user, as a kind of personalised current awareness service. ONA (and other community of practice techniques) could be used to reduce the burden on the users to explicitly construct their profiles by building an initial profile based on their community of practice. In this sense, ONA could be used to augment the Ingest function by providing support for user profile generation, or to enhance the Access function by providing a context in which documents are retrieved that could better align the result set with the user's information needs.

The second application applies ONA to the problem of *coreference*. When taking the union of multiple databases, the manner in which entities are identified may cause problems if two of the component sources choose different identifiers for the same object. Due to the philosophy of its design (and the design of the Web itself) the Semantic Web suffers from this issue of coreference; there is not global authority which is responsible for minting new URIs, and there is also no unique name assumption. It is not possible to determine if two distinct URIs refer to the same object by inspection of the URIs (the Semantic Web treats URIs as opaque identifiers), so when presented with two different URIs, a Semantic Web agent should by default deduce that the URIs refer to two different objects.

ONA addresses the coreference issue by behaving as a similarity measure between individuals; in addition to traditional similarity measures such as string edit distance on the names of objects, it also provides a means for comparing people by comparing the communities that surround them. When used alongside the traditional methods, ONA could provide supporting evidence that two people are the same (same/similar name, similar community of practice, etc). This is of great use when maintaining the descriptive information about information objects; where possible, the metadata for objects should seek to minimise redundancy by not creating new instances of people, journals, conferences, etc wherever possible. An enhanced Data Management function would therefore seek to use ONA to clean up the descriptive information passed to it from the Ingest function.

5 Visualisation

The final area in which we will consider possible enhancements to digital libraries is that of visualisation. In conventional digital libraries, users typically interact with one information object at a time, even though they might require a more high-level view which spans several objects. However, communicating the required aspects of a large and dynamic information space such as that represented by a community-based open archive without overwhelming the user in a deluge of data is a difficult problem. The rise of bibliometrics can be seen as one way of abstracting collection-wide knowledge in order to transform it into a form better suited to human abilities.

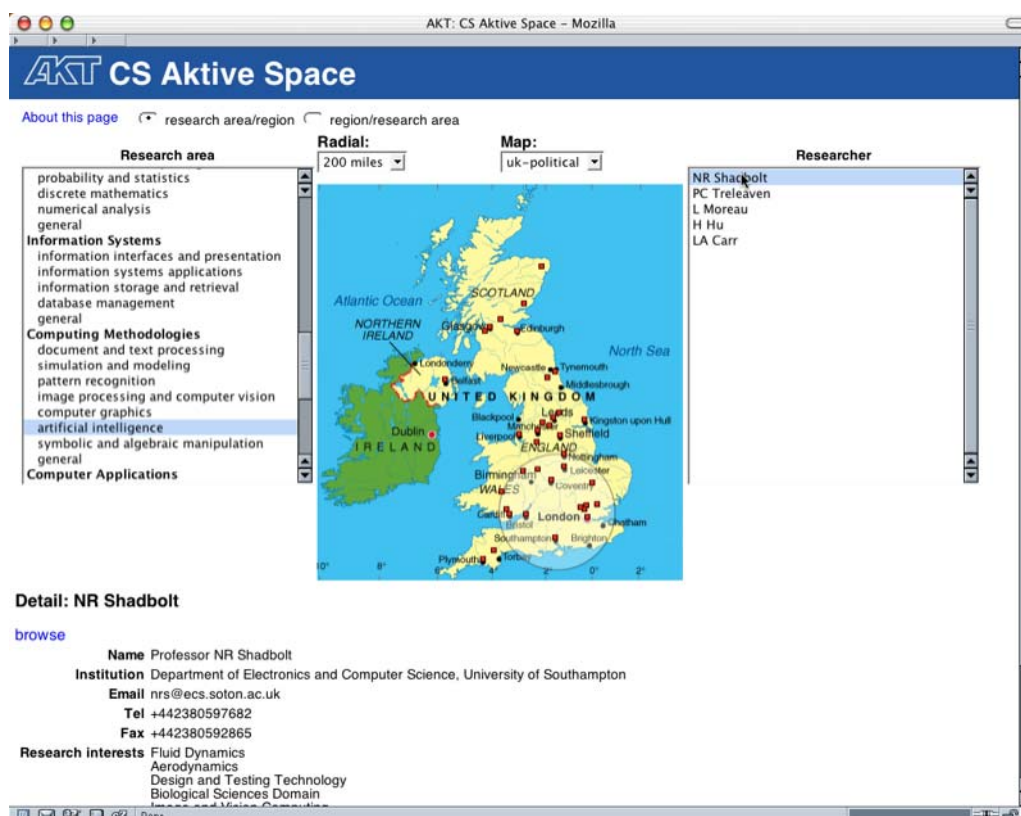


Figure 2 The CS AKTive Space explorer interface

The Advanced Knowledge Technologies (AKT) project has produced a number of systems which use Semantic Web technologies as an infrastructure for building tools to visualise and explore scholarly literature and its context. The first of these, CS AKTive Space [39], aims to represent a large ontological space in a meaningful fashion, and contains information describing the state of computer science research in the UK, including a directory of active researchers, information about the funding supporting research, a snapshot of the most significant research outputs (from the 2001 Research Assessment Exercise, a UK government initiative to determine the effectiveness of research funding). This information is presented in such a way that a user can explore it in several ways, for example by geographical region or by research specialisation, in order to gain a gestalt view of the domain: what research is being conducted where and

by whom. Various measures are used to judge the reputation and impact of people and institutions, including the value of grant funding.

Figure 2 shows the explorer interface of CS AKTive Space. Across the top are three panels which can be used to progressively narrow the search by selecting items: a list of research areas taken from the Association for Computing Machinery (ACM) Computer Science Classification Scheme, a map of the UK which can be used to express geographical constraints, and a list of researchers. Each of these panels informs its successors, so that a selection can be seen to restrict the available choices, and the order of the panels can be changed to allow the user to explore the space in a different manner. At the bottom of the window is a detail view which provides contextual information on the most recently selected item. At present, this is showing information on the selected researcher NR Shadbolt, including details of his publications and project affiliations. At any time, the user may drill down and examine the raw data which was used to generate this view by clicking on the “browse” link.

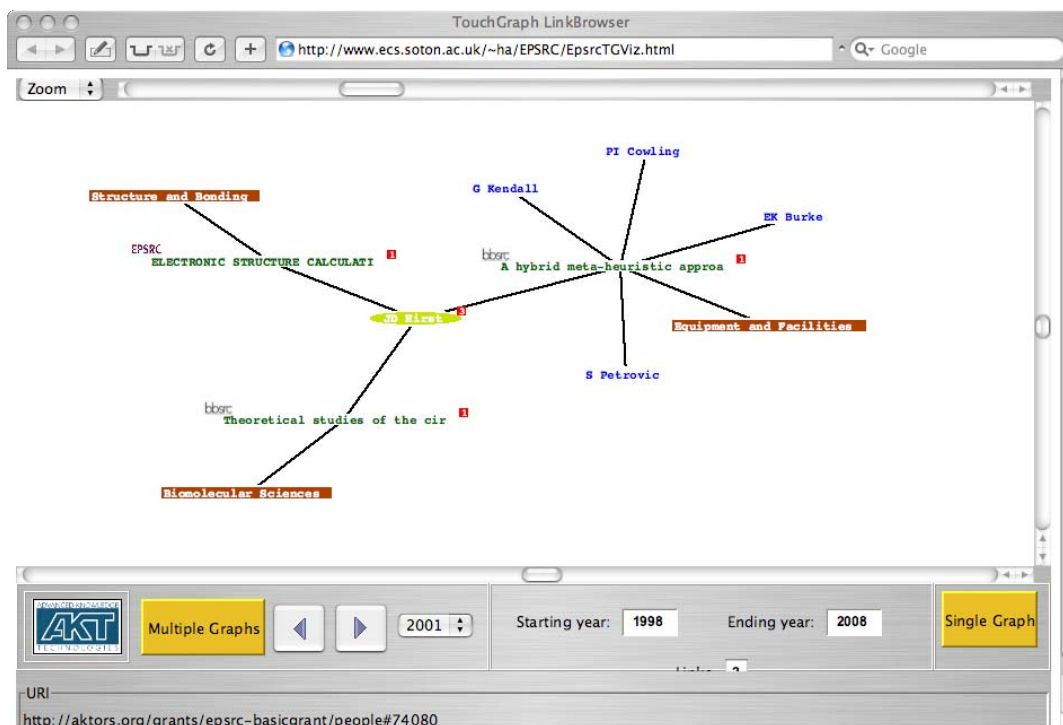


Figure 3 Examining collaboration and funding networks in AKTive EPSRC²³

Further development of this work in collaboration with the EPSRC²³ produced a different interface, shown in Figure 3. This window shows the network surrounding an individual who receives funding from multiple research funding agencies (BBSRC²⁴ and MRC²⁵), along with his co-investigators. Graph-based visualisations like this provide a different

²³ Engineering and Physical Sciences Research Council (EPSRC), <http://www.epsrc.ac.uk/>

²⁴ Biotechnology and Biological Sciences Research Council (BBSRC), <http://www.bbsrc.ac.uk/>

²⁵ Medical Research Council (MRC), <http://www.mrc.ac.uk/>

method for examining the context in which research outputs are produced. While a full graph of an entire community would be too complex to easily visualise, techniques like this allow a user to select specific areas for examination, possibly restricted to particular time periods.

Figure 4 shows a third visualisation project from AKT. This interface integrates numerical data, in this example a graph of crude oil production for Iran and Iraq since 1973, with access to the relevant literature. The rear window, which is displaying the graph, has a sidebar on the right which enables the user to initiate a search for related documents about crude oil or petroleum. The user can refine this search by selecting extra terms which frequently occur with the existing search terms from the drill-down list on the right. Selecting a search brings up a results window containing a ranked list of related documents, shown in the front window.

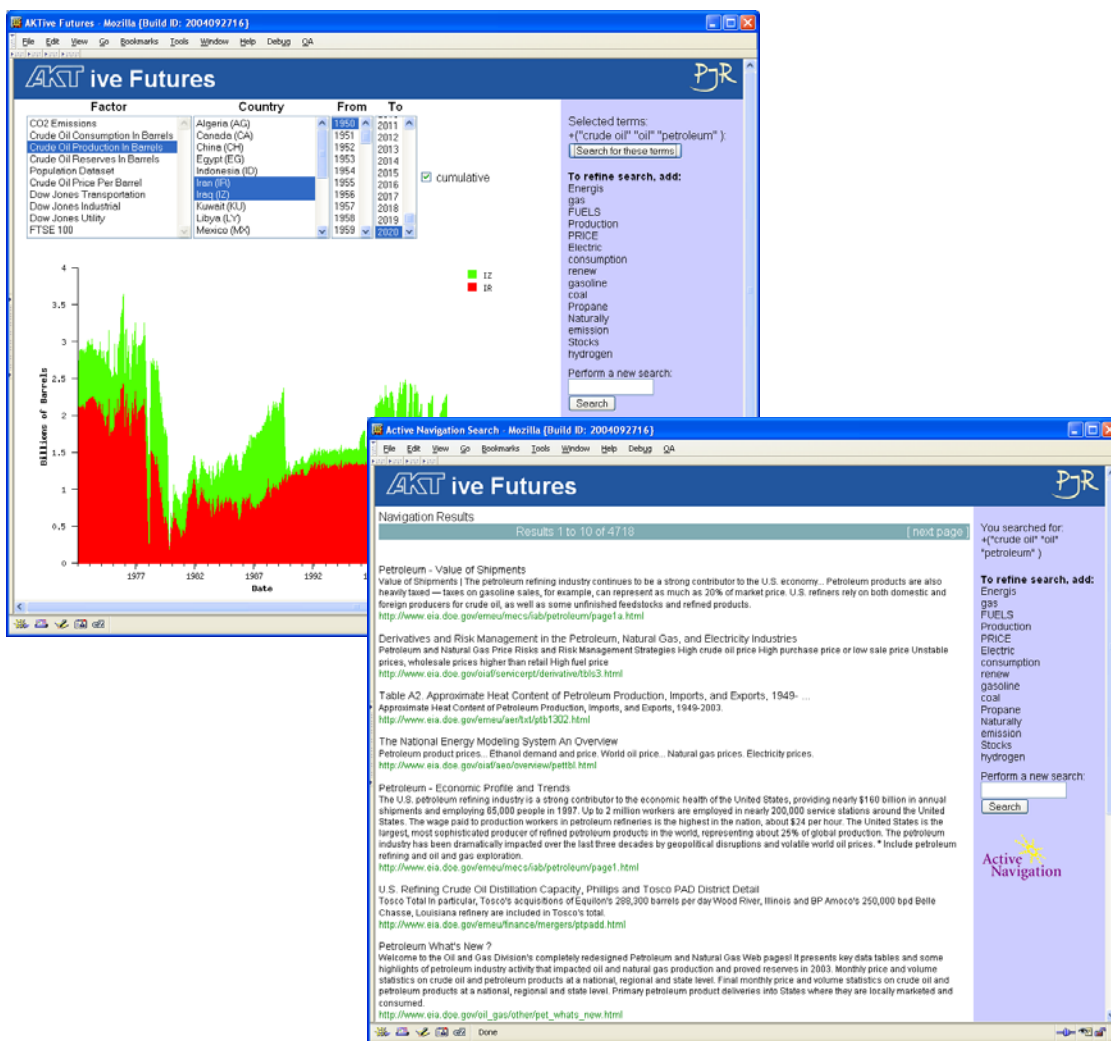


Figure 4 Linking numerical data to the relevant literature in AKTive Futures

The DOPE project [42], developed by the Free University of Amsterdam in collaboration with Elsevier, uses an ontology to structure information about pharmacology in order to

describe and visualise a collection of papers (DOPE stands for Drug Ontology Project for Elsevier).

Figure 5 shows the graphical browser developed as part of this work. The frame on the left of the window contains the hierarchy of terms in the ontology that are co-occurrent with the focussed term (in this case, acetylsalicylic acid, or aspirin). The user can select any number of these terms in order to see how they are distributed across the document set retrieved from the focussed term. The matching documents are displayed in the top right frame as overlapping clusters, where each cluster corresponds to a selected co-occurrent term. In the example, the user has selected the terms “warfarin”, “mortality”, “practice guidelines” and “blood clot lysis”, and the overlap between these terms is clear. Each blob within a cluster corresponds to a document; the colour of the blob represents the type of document – full article, review article, abstract and so on. In the lower right frame are listed the documents in the currently selected cluster, in this case those that relate to blood clot lysis.

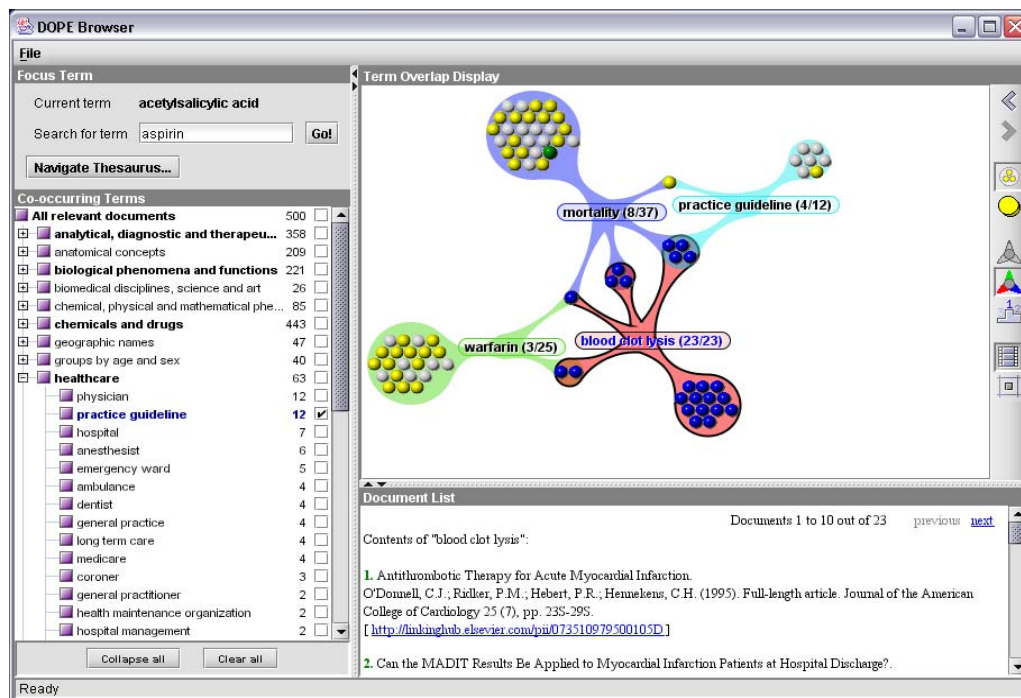


Figure 5 The DOPE browser

In OAI terms, these visualisations are an enhancement of the Access function, and provide novel ways for users to explore the contents of open archives. These visualisations are not intended to be an exhaustive list of what can be done to present new views on digital libraries to users, but an indication of the possible – DELOS WP4 is looking in greater detail at information visualisation systems.

The Work Package 4 cluster²⁶ of the DELOS project will survey current user interfaces to Digital Libraries, including visualisation tools. WP4 is looking at the entire lifecycle of

²⁶ DELOS Work Package 4, <http://www.delos.info/WP4.html>

the development of user services, with the aim of developing a theoretical framework to support the development of DL systems using the expertise of the cluster members. In the first of these activities the cluster will systematically study user requirements in relation to the ongoing development within the DELOS project. In particular this will involve the characterisation of DL users including those with special needs. This could lead to adapting visualisation systems to e.g. facilitate users with impaired vision. The development stage of WP4 will include re-assessing information filtering and retrieval, in the light of user requirements. The cluster will look at extending existing visualisation systems, particularly effectiveness, expressiveness and interactivity.

6 Recommendations

Following our examination in the previous sections of the current state of digital library technology, specifically relating to the featured areas of enhancement, we make the following recommendations. These recommendations identify areas which we believe will be fruitful avenues for future research, and suggest practical steps which could be taken, including technology choices.

6.1 Enhance bibliographic management with SW technologies

As described in this report, Semantic Web technologies offer improvements in open archive interoperability by facilitating translations between metadata vocabularies, and by providing a machine understandable foundation that enables agent mediation.

There are existing standards for bibliographic management which use Semantic Web technologies, chiefly the Dublin Core metadata vocabulary. However, Dublin Core's simplicity means that it is too semantically impoverished for many applications, which has led to a proliferation of domain-specific extensions. Current practice with respect to Dublin Core is varied, not least because Dublin Core imposes minimal constraints on its usage by design. While the Dublin Core Metadata Initiative is making progress on the creation of a body of best practice, there remain notable holes which encourage a diversity that works against interoperability.

The contents of the `dc:creator` field is largely unspecified, for example, and there is no agreement on how to represent the decomposition of authors names into name components. The common practice for author names is to use a set of cataloguing rules such as AACR2 to write the names in a canonical form, but this does not split the names into components that can be easily manipulated by Semantic Web agents. Similar issues exist regarding the representation of canonical forms for journal names, or the characterisation of information objects as journals, journal papers, books and so on.

The development of enhanced metadata vocabularies is essential for all of the future research areas discussed in this report. Richer semantics enables richer uses, supporting more sophisticated, accurate and powerful tools; a common criticism levelled at Web search engines is that they do not allow domain-specific queries, e.g. searching by chemical structure. A formally defined meaning for the relations between entities is necessary for improvements in bibliometric calculation. Bibliometrics is study of THE literature (a study of all the papers), so there is a need for a common semantically enriched view.

We recommend that DELOS should encourage the application of Semantic Web technologies to bibliographic management, both by the definition of richer modular ontologies for bibliographic information that extend Dublin Core, and by the development of tools to allow authors to provide semantic annotations. This work should take place within the context of the open archive community, and use the existing Web infrastructure where appropriate.

6.2 Enhance bibliometric measures with community context

Existing bibliometric measures treat citations as neutral references, and disregard the intent of those citations; by considering this intent, bibliometrics can distinguish between favourable and unfavourable citations. Citation intent can be considered to be a reflection of the relationships that exist within the community in which the citation is made. Bibliometrics can be further enhanced if they examine this community, looking at the affiliations of authors, sources of research funding, and the patterns of co-citation. Similarly, impact factors can be applied to things other than journals; people, projects, institutions, sectors and even countries can be considered to have citation factors.

These use cases can be accomplished by modelling the community context of a publication in a principled fashion, and using this to enhance our view of existing bibliometric measures.

We recommend that DELOS encourage the development of bibliometric techniques that reflect the context in which a document is published, the intent of a citation, and the broader community that surrounds a publication.

6.3 Enhance community modelling with bibliometric information

As a counterpart to our previous recommendation, bibliometrics can be used to improve our understanding of the community context in which papers are published. Patterns of publication and the relationships between papers are a reflection of the implicit communities in which authors participate. Services such as expert finders, star finders (which locate rapidly rising researchers), and community of practice search are all improved by using our knowledge of the literature to inform our understanding of the interactions between individuals.

We recommend that DELOS support the investigation of techniques for using bibliometric information to improve the study of the interactions between researchers within communities.

6.4 Develop visualisations of literature and its context

Search is not the only way in which users interact with information. Browsing and exploration are valid alternative interaction metaphors, as anyone who has walked the shelves of a library instead of using the catalogue should recognise. If a user's information requirement is to get an overview of a particular literature, than to obtain a specific resource from that literature, browsing is a better approach than search.

Browsing can be informed both by explicit relationships from a ontologically-motivated description of documents and the context in which they are published (the relationship between two papers whose authors work on the same project, for example) and by implicit relationships that result from further exploration of a paper's context (for example, the active bibliographies that are assembled by CiteSeer based on various measures of similarity between documents).

An expanded view of documents which includes not only a representation of a paper, its authors and the concepts within, but also the relationship between works and people implied by co-citation, both provides an alternative to search, and can be used to enhance

search (visualisation and clustering of search results as an alternative to ranking, as demonstrated by DOPE). Or documents could be broken down into components, using the concepts within the document and relationships outside, to be re-assembled in differing forms depending upon the user, e.g. through automatic semantic presentations [34, 33].

Appropriate visualisations can greatly facilitate exploration, browsing and other forms of sensemaking, but it should be noted that visualisation is necessarily both task- and domain-specific. A visualisation tool which is aimed at an author who is trying to discover new developments in his chosen speciality would not be appropriate for a user who is trying to assess the impact of research across a discipline.

Similarly, different communities have different processes and common practices. Electronic preprints are considered essential in the high-energy physics community (not least due to the success of ArXiv), but are rarely cited in the social science community. The difference in the distribution of citation rates between disciplines (for example, between the pure sciences and engineering) also affects the requirements of a visualisation; what is considered as signal in one might be discounted as extraneous noise in the other.

We recommend that DELOS should investigate tools that allow end users to visualise the contexts in which documents are published in order to explore the literature at a higher level of abstraction.

6.5 Reassess bibliometric measures

Journal Impact factors are frequently treated as a necessary evil that are required solely because there is incomplete access to the scientific literature. They provide an abstract measure of a paper's importance based on where it is published, the assumption being that highly-cited papers are published in high-impact journals, and that high-impact journals only publish highly-cited papers (as if this could be determined before publication!) By their nature as aggregate measures that abstract the importance of literature, and their origins in meaning-neutral citations, impact factors can present a false view of the world. Are all papers in Nature or Science equally highly cited? Should a paper which is only cited in order to refute its claims (the original Fleischmann and Pons paper on cold fusion, for example) be considered to be an important paper?

Our recommendations for the reassessment of bibliometric measures are twofold. Firstly, based on concepts from issue and claim tracking, use an explicit representation of the rhetorical relationship between cited and citing works to inform bibliometrics. Considering the broader research lifecycle beyond the scope of individual papers allows us to assign measures of importance to ideas rather than just their manifestation in the literature.

Secondly, expand the relationships abstracted by bibliometrics beyond citations. The greater context in which a paper is published can be used to discover patterns in the literature, from cycles of mutually favourable citations, to the emergence of new disciplines (and therefore new communities) that might need extra attention in order to grow to maturity.

There is an important issue associated with these recommendations that must be addressed before the recommendations can be acted upon. The purpose of a citation is something that can only really be determined by the author who creates the citation, rather than by an information specialist who takes a neutral view of the literature. The elicitation of the purpose behind a citation from the author is likely to present a variant on the knowledge acquisition bottleneck; knowledge extraction techniques may be able to assist by automating or part-automating this process.

Institutional Repositories, and by proxy the digital library, could capture the context that a paper is published in, and the greater role that its authors have in scholarly discourse. IRs capture the research output of the institution, which as well as the prestige output (journals, books, or monographs depending upon the research field), can include other measures of esteem, such as editorships, committee chairs and so on. Given this additional knowledge citations from e.g. the editor of a prestigious journal could be given greater weight than citations from a paper authored by a research student. This community modelling (which falls under all OAIS functions) may be a potential means to infer meaning onto citations, and better inform the use of citations as a measure of scholarly impact.

7 Activities by Partners

The DELOS WP5 cluster members represent a broad spectrum of experience and expertise within the digital libraries community. Within the broader aims of DELOS it is necessary to promote collaboration, to gain a greater cross-over of ideas and the establishment of common technical infrastructures to maximise the benefit to European and worldwide DL research. In this section we highlight a number of projects undertaken by the cluster members within the last few years.

7.1 School of Electronics and Computer Science, University of Southampton (UK)

<http://www.ecs.soton.ac.uk/>

7.1.1 Advanced Knowledge Technologies (AKT)

<http://www.aktors.org/>

The Advanced Knowledge Technologies project started in 2000 with the aim of developing tools and technologies for managing the complete lifecycle of knowledge, from initial acquisition to use, reuse and eventual disposal. The stance taken is that knowledge is distinct from information; information is structured data, where knowledge is information that applied to the context of a particular task. This task-orientation requires that not only must the objective meaning of a fragment of knowledge be understood, but also its relevance to the activities or processes of the entity which seeks to make use of it. The AKT Project has developed a number of technologies to support the conceptual extraction and re-use of knowledge. Underpinning many of these tools is 3store [21], a scalable open source knowledge repository, or triplestore, which has been used in a variety of application domains, and which provides a high-level storage and query facility based on RDF.

Cabral, L., J. Domingue, et al. (2004). Approaches to Semantic Web Services: An Overview and Comparisons. Proceedings First European Semantic Web Symposium (ESWS2004), Heraklion, Crete, Greece.

The next Web generation promises to deliver Semantic Web Services (SWS); services that are self-described and amenable to automated discovery, composition and invocation. A prerequisite to this, however, is the emergence and evolution of the Semantic Web, which provides the infrastructure for the semantic interoperability of Web Services. Web Services will be augmented with rich formal descriptions of their capabilities, such that they can be utilized by applications or other services without human assistance or highly constrained agreements on interfaces or protocols. Thus, Semantic Web Services have the potential to change the way knowledge and business services are consumed and provided on the Web. In this paper, we survey the state of the art of current enabling technologies for Semantic Web Services. In addition, we characterize the infrastructure of Semantic Web Services along three orthogonal

dimensions: activities, architecture and service ontology. Further, we examine and contrast three current approaches to SWS according to the proposed dimensions.

7.1.2 Open Middleware Infrastructure Institute

<http://www.omii.ac.uk/>

The Open Middleware Infrastructure Institute is an institute of the University Of Southampton, located in the School of Electronics and Computer Science.

Our vision for the OMII is for it to become the source for reliable, interoperable and open-source Grid middleware, ensuring the continued success of Grid-enabled e-Science in the UK.”

7.2 ETH, Swiss Federal Institute of Technology, Zurich (Switzerland)

http://www.ethz.ch/index_EN

Prof. Hans-Jörg Schek schek@inf.ethz.ch

<http://www-dbs.inf.ethz.ch/externalprojects/index.html>

7.2.1 Multimedia Information Management

“Multimedia information systems consist of many specialized components such as databases, object repositories, special image servers, feature extractors, and indexing components. ISIS, our Interactive Similarity Search engine, builds on top of OSIRIS that provides a framework to implement, call and combine services. In this context, ISIS consists of a number of core services to store, analyze and index multimedia documents. These services run in a large cluster (with more than 100 nodes) which is maintained and observed by the underlying OSIRIS system. Simple transactional processes for insertion, similarity search, and bulk load can run in parallel and the subtasks are "optimally" and reliably assigned to the components by the OSIRIS system as shown in Figure 4. At any point in time, a new component can be added to the cluster in order to improve response times. Interactive similarity retrieval is based on the VA-File, a simple but efficient approximation of the inherently high- dimensional feature vectors. In order to improve the retrieval effectiveness, we support complex similarity queries consisting of several reference images, several feature types, textual attributes and predicates. In combination with relevance feedback, our similarity search system provides a convenient interface for effective queries, as exemplified in Figure 5. We further apply these techniques to organize, manage, and present the individual information spaces of users in a more natural and efficient way.”

7.2.2 ISIS - Interactive Similarity Search

“Similarity search in multimedia databases is a difficult and expensive task. Not only is the extraction of features to describe documents rather time consuming, but also is searching for relevant documents a costly operation. Within ISIS, our research aim is

merely on supporting efficient search operations over the most effective document descriptors. ISIS further targets the entire similarity search process including query refinement steps with sophisticated relevance feedback methods. ISIS builds on top of OSIRIS, our Open Service Infrastructure for Reliable and Integrated process Support. While OSIRIS provides a framework to develop, distribute and combine services, ISIS provides specialized services and processes to implement a similarity search engine.

In the past, we successfully implemented a huge image database with around 370,000 images (CHARIOT) and covering the most effective color and texture features. The interactive search times stem from the underlying vector approximation file (VA-File) which supports fast nearest neighbor retrieval in high-dimensional feature sets. Moreover, the VA-File is capable of combining features at run-time and to query the index data with several reference objects at once. This makes it perfectly suitable to implement sophisticated relevance feedback methods that require such complex similarity searches.

Recent developments within ISIS are region-based image retrieval and combined text-based and content-based queries over multimedia objects (images, audio, video). The former kind of queries is supported by a filter-and-refine search algorithm over a rather expensive but effective dissimilarity measure (it involves the solution of an Assignment Problem). For the later query type, we first developed techniques to extract text features from web pages for embedded objects. In contrast to conventional approaches, our method assigns text blocks according to their visual closeness in the layout rather than based on the distance between embedding and text block in the HTML source code. For query evaluation we now investigate an optimal way to integrate the textual query evaluation within the content-based index structures.

Another interesting aspect is how to integrate relevance feedback techniques into the search algorithms. In the literature, a large number of feedback techniques have been proposed in the past. But not all of them are meaningful as they are too expensive and disallow for interactive search scenario. We aim at the development of good relevance feedback technique that leads to queries for which efficient search algorithms are available.”

7.2.3 Organization of Individual Information Space

“While the technical methods for storing and retrieving multimedia information have improved steadily over the last years, the user access interface and the organization of documents have almost remained the same. For instance, most file systems are hierarchically organized and the user is responsible for maintaining the hierarchy and for storing the documents at the "right place". However, it is difficult to define a proper hierarchy at the beginning, and, once the hierarchy is set up, it takes a lot of effort to evolve or change the hierarchy to newer demands. Furthermore, hierarchical organizations often do not fit to all the users needs ("the document I changed last week" or "the image I sent to my parents").

In our vision, the system maintains and stores documents without any interaction with the user. It further provides a model of the individual information spaces which are not restricted to any technical restrictions (i.e. hierarchy in file system) and which can be freely adapted according to the users needs. In this model, the basic access primitive is

"query by association". For instance, one can access documents related to a certain topic, to a certain person, or to arbitrary viewpoints. Furthermore, persons, groups of persons, and organisations have a virtual counterpart that denotes their "personality", or the context of the user. Due to relationships between personalities, a user can gain access to information exchanged between personalities, can discover new information that was relevant to one of his or her related personalities, and the system can accommodate the user's personality.

As a first step towards this vision, we are currently investigating sophisticated methods to visualize and present documents. In recent projects, we developed methods to derive thumbnail previews of web documents, and to layout these documents in 2D and 3D according to their mutual relationships. An interesting aspect thereby is the similarity between two documents. Obviously, if two documents are similar to each other, they have to be placed next to each other on the screen. Further, we are seeking for methods that provide users with powerful tools to enter, change and adapt queries. For instance, similarity scores depend on a number of features and on relationships between documents. However, the influence of all these basic similarity assessments depend on the current information need of the user. But often the user is not able to map his/her information need to the technical level of features and relationships (so-called semantic gap)."

7.3 FORTH, Crete (Greece)

<http://www.forth.gr/>

Martin Doerr <martin@ics.forth.gr>

http://www.ics.forth.gr/isl/people/people_individual.jsp?Person_ID=2

http://www.ics.forth.gr/isl/publications/by_author.html#martin

Martin Doerr (2003) "The CIDOC CRM – an Ontological Approach to Semantic Interoperability of Metadata," AI Magazine, Volume 24, Number 3

“Abstract: This paper presents the methodology that has been successfully employed over the past 7 years by an interdisciplinary team to create the CIDOC Conceptual Reference Model (CRM), a high-level ontology to enable information integration for cultural heritage data and their correlation with library and archive information. The CIDOC CRM is now in the process to become an ISO standard. This paper justifies in detail the methodology and design by functional requirements and gives examples of its contents. The CIDOC CRM analyses the common conceptualizations behind data and metadata structures to support data transformation, mediation and merging. It is argued that such ontologies are property-centric, in contrast to terminological systems, and should be built with different methodologies. It is demonstrated that ontological and epistemological arguments are equally important for an effective design, in particular when dealing with knowledge from the past in any domain. It is assumed that the presented methodology and the upper level of the ontology are applicable in a far wider domain.”

7.4 Netlab Knowledge Technologies Group, Lund University (Sweden)

<http://www.lub.lu.se/knowtech/>

<http://netlab.lub.lu.se/>

Traugott Kock <Traugott.Koch@lub.lu.se> <http://www.lub.lu.se/netlab/staff/koch.html>

<http://netlab.lub.lu.se/Projects-current.html>

“The European Schools Treasury Browser: The objective of the ETB is to build a Web educational resource Metadata Networking and Quality Processing infrastructure for schools in Europe. This infrastructure aims to link together existing national repositories, encourage new publication, and provide a reliable level of quality and structure. The proposal aims to build a simple yet effective distributed "Schoolnet Information Space". The project will enable and encourage trans-cultural and trans-national co-operation and communication and will enable individuals (students, teachers, administrators, parents) and workgroups to produce, handle, retrieve and communicate information in the languages of their choice, and to combine information resources from different regions and countries, and of different levels.”

“Renardus: Renardus is an academic subject gateway service, co-ordinated by European information gateway initiatives. The Renardus partner gateways cover about 64000 predominantly digital web-based resources from within most areas of academic interest, mainly written in English. Renardus allows you to find Internet resources selected according to quality criteria and carefully described by Subject Gateways from several European countries. You discover the individual resources and collections by searching and browsing these descriptions (metadata), not the full text of the resources themselves. A special feature of Renardus is the option to "Browse by Subject" through hierarchical trees of topics and subsequently to jump to one or several related subcollections of the contributing Subject gateways.”

7.5 School of Informatics, University of Edinburgh (UK)

<http://www.inf.ed.ac.uk/>

AKT, e-Science Centre

<http://www.inf.ed.ac.uk/publications/report/>

List, T, Fisher, R (2004) “CVML An XML-based Computer Vision Markup Language,” To appear in the Proceedings of the International Conference on Pattern Recognition 2004 (ICPR) August 23-26, Cambridge, UK

“We propose an XML-based Computer Vision Markup Language for use in Cognitive Vision, to enable separate research groups to collaborate with each other as well as making their research results more available to other areas of science and industry, without having to reveal any proprietary ideas, algorithms or even software. The Computer Vision Markup Language can communicate any type and amount of information, making unavailable functionality accessible to anyone. In this paper we introduce the language and describe how we have implemented it in a very large

cognitive vision project. We provide a free open source library for working with this language, which can easily be implemented into existing code providing seamless network communication abilities and multi-platform support. Last we describe the future of CVML and how it might evolve to include other areas of research.”

7.6 Technical University of Crete (Greece)

<http://www.music.tuc.gr/Research/Projects.htm>

The activities of MUSIC/TUC include research, development, training and technology transfer in the area of multimedia information systems. The staff's research interests include Multimedia Information Systems, Very Large Data Bases, Multimedia Communication Systems, Collaborative Environments, Information Retrieval, Human-Computer Interaction, Electronic Commerce, Tourism and Cultural Systems and Applications.

For this reason the laboratory maintains strong links with other universities, research institutes and high technology companies, all over the world, and actively participates (or has participated) in numerous EU research and development projects (IST, ESPRIT, ACTS, RACE, AIM, DELTA, LINGUA, INCO, STRIDE, SPA etc.).

A second activity is to train graduate and undergraduate students of the Technical University of Crete in advanced technology related to the area of Information Systems. Many members of MUSIC/TUC are also associated with the Technical University of Crete and university students have easy access to the advanced research facilities of MUSIC/TUC and to the experience of its personnel.

A third area of MUSIC/TUC activities consists of technology transfer and collaboration with leading Greek and European companies. MUSIC/TUC has already established strong links with the leading Greek forces in the area of communications and computer technology. These links are maintained through joint participation in EU and National (competitive) projects.

7.7 UKOLN, University of Bath (UK)

<http://www.ukoln.ac.uk/>

UKOLN is a centre of expertise in digital information management, providing advice and services to the library, information, education and cultural heritage communities by:

- Influencing policy and informing practice
- Promoting community-building and consensus-making by actively raising awareness
- Advancing knowledge through research and development
- Building innovative systems and services based on Web technologies
- Acting as an agent for knowledge transfer

7.7.1 eBank

eBank UK is a JISC-funded project which is a part of the Semantic Grid Programme. The project is being led by UKOLN in partnership with the Combechem project at the University of Southampton and the PSIGate Physical Sciences Information Gateway at the University of Manchester. This new initiative is set in the context of the JISC Information Environment, JISC funded development supporting end-users to discover, access, use and publish resources as part of their teaching, learning and research activities. The eBank UK pilot service will demonstrate linking of research data with other derived information.

<http://www.ukoln.ac.uk/projects/ebank-uk/>

7.7.2 Agentcities.NET

The aim of this project (finished February 2003) was to investigate the use of an ontology server in an interoperable agent network (agentcities.NET). The server provides a publication environment for the disclosure of metadata vocabularies and customised application-specific profiles of these vocabularies. The metadata vocabularies (also known as schemas or metadata element sets) may be regarded as simple forms of ontologies. In this registry environment, individual terms as well as whole vocabularies can be investigated for adaptations, local usages and relationships with other vocabularies. The project builds on previous work within the SCHEMAS and MEG registry projects and on our involvement in the Dublin Core Metadata Initiative's registry activity.

<http://www.ukoln.ac.uk/metadata/agentcities/>

7.7.3 Resource Discovery Network

<http://www.rdn.ac.uk/>

The Resource Discovery Network (RDN) is a UK-based internet portal that provides integrated access to the subject-specific portals (subject gateways) that were developed during the JISC-funded eLib programme.

7.8 UNIMI, University of Milan (Italy)

<http://www.unimi.it/engl/>

<http://dakwe.dico.unimi.it/>

Our research activities cross the areas of Artificial Intelligence, Database Systems, and Mobile Computing. Reasoning techniques and well-founded logical approaches are applied to data and knowledge management. A theoretical line of research investigates time related aspects in data and knowledge management. A more applicative line of research investigates the application of knowledge-based techniques to different problems in mobile computing.

More specifically, current research includes:

- Web Engineering and Mobile computing (Adaptive internet services for mobile devices, location-based services, context-based services, advanced distributed bookmark management, ...)
- Time granularity in database systems, knowledge representation and reasoning
- Computer Security: Temporal Access Control Models, Logical Approaches to Policy Specification and Management, Release Control
- Temporal Knowledge Representation

7.9 University for Health Informatics & Technologies, Tyrol (Austria)

http://www.umat.at/index_e.cfm

8 Activities by other Groups

8.1 NISO MetaSearch Initiative

The distribution of digital libraries and the rise of technologies for federated search across these distributed archives require that metasearch services must be able to access a wide variety of heterogeneous sources. These services may have access requirements that range from open standards such as Z39.50, OAI and Web Services, to proprietary APIs or Web interfaces designed primarily for people. The absence of widely supported standards and best practices has a negative impact on the effectiveness of metasearch environments, for both the content providers and for the end-users.

The NISO Metasearch Initiative addresses this issue by supporting the development and evolution of common technologies and practices to enable more effective and responsive services, which deliver enhanced content to end-users while protecting the IP of content providers.

9 References

1. Alani, H. et al (2003) "Automatic Ontology-Based Knowledge Extraction from Web Documents", *IEEE Intelligent Systems* 18(1), 14-21
<http://eprints.ecs.soton.ac.uk/7396/>
2. Alani, H., Dasmahapatra, S., O'Hara K. and N. Shadbolt (2003) "Identifying Communities of Practice through Ontology Network Analysis". *IEEE Intelligent Systems* 18(2), 18-25.
3. Andrew, T. (2003) "Trends in Self-Posting of Research Material Online by Academic Staff", *Ariadne* 37
<http://www.ariadne.ac.uk/issue37/andrew/>
4. Berners-Lee, T., Hendler J. and O. Lassila (2001) "The Semantic Web", *Scientific American*.
5. Bollacker, K., Lawrence, S., Giles, C.L. (1998) "CiteSeer: an autonomous Web agent for automatic retrieval and identification of interesting publications" In *Proceedings of the Second International Conference on Autonomous Agents*, 116-123
<http://www.neci.nec.com/~lawrence/papers/cs-aa98/cs-aa98.pdf>
6. Brophy, P. and Butters, G. (2000) "AGORA: Evaluation Report"
<http://hosted.ukoln.ac.uk/agora/documents/eval-mar2000.doc>
7. Butler, D. (2004) "Acknowledgements hit the limelight", *News@Nature*,
doi:10.1038/news041213-3
8. Ciravegna, F. and Y. Wilks (2003) "Designing Adaptive Information Extraction Designing Adaptive Information Extraction" In S. Handschuh and S. Staab (eds). *Annotation for the Semantic Web*. IOS Press
<http://eprints.aktors.org/archive/00000314/01/AmilcareAnnotation.pdf>
9. Ciravegna, F., A. Dingli, Y. Wilks and D. Petrelli (2002) "Timely and Non-Intrusive Active Document Annotation via Adaptive Information Extraction", in *Proceedings of the ECAI workshop on Semantic Authoring, Annotation & Knowledge Markup (SAAKM02)*, held in conjunction with the 15th European Conference on Artificial Intelligence, Lyon, France
10. Cousins, S. (1999) "Virtual OPACs versus union database: two models of union catalogue provision." *Electronic Library*, 17(2), 97-103.
11. Coyle, K. (2000) "The virtual union catalog: a comparative study." *D-Lib Magazine*, 6(3). <http://www.dlib.org/dlib/march00/coyle/03coyle.html>
12. Coyle, K. (2004) "The virtual union catalog." In: A. Lass and R. E. Quandt, eds., *Union catalogs at the crossroad*, pp. 51-66. Hamburg University Press, Hamburg.
13. Cronin, B. (2001) "Bibliometrics and beyond: some thoughts on web-based citation analysis", *Journal of Information Science* 27(1), 1-7
<http://www.ingentaconnect.com/searching/Expand?pub=infobike://sage/jis/2001/00000027/00000001/art00001>

14. Dumbill, E. (2002) "The Friend-of-a-Friend vocabulary can make it easier to manage online communities," *IBM developerWorks*, 1st June 2002
<http://www-106.ibm.com/developerworks/xml/library/x-foaf.html>
15. Flake, G., Lawrence, S. and C.L. Giles (2000). "Efficient Identification of Web Communities", *Proceedings of the Sixth ACM SIGKDD International Conference, Boston, MA, USA, 20-23 August 2000*, 150-160.
16. Garfield, E. (1998) "The use of Journal Impact Factors and citation analysis for evaluation of science", *Cell Separation, Hematology and Journal Citation Analysis*, Rikshospitalet, Oslo April 17th 1998
http://www.garfield.library.upenn.edu/papers/eval_of_science_oslo.html
17. Godby, C. J., D. Smith and E. Childress (2003). "Two paths to interoperable metadata." 2003 Dublin Core Conference (DC-2003): Supporting Communities of Discourse and Practice - Metadata Research and Applications, September 28 - October 2, 2003, Seattle, Washington, USA. Information Institute of Syracuse, Syracuse, N.Y. <http://purl.oclc.org/dc2003/03godby.pdf> Also at:
<http://www.oclc.org/research/publications/archive/2003/godby-dc2003.pdf>
18. Gredley, E. and A. Hopkinson (1990) Exchanging bibliographic data: MARC and other international formats. Library Association Publishing, London.
19. Griffin, S. (2002) "Fourth DELOS Workshop: Evaluation of Digital Libraries: Testbeds, Measurements, and Metrics", *Hungarian Academy of Sciences*, Budapest, Hungary 6-7 June 2002
http://www.dli2.nsf.gov/internationalprojects/working_group_reports/evaluation.html
20. Gruber, T.R. (1993) "A translation approach to portable ontologies", *Knowledge Acquisition* 5(2):199-220.
21. Harris, S. and N. Gibbins (2003) "3store: Efficient Bulk RDF Storage", In *Proceedings of the First International Workshop on Practical and Scalable Semantic Web Systems*, Sanibel Island, Florida, USA
<http://eprints.aktors.org/273/>
22. Hayes-Roth, F., Waterman D.A. and D.B. Lenat (1983) "Building Expert Systems", Addison-Wesley.
23. Hitchcock, S. Woukeu, A. Brody, T., Carr, L., Hall W. and S. Harnad (2003) "Evaluating Citebase, an open access Web-based citation-ranked search and impact discovery service." Technical Report ECSTR-IAM03-005, Electronics and Computer Science, University of Southampton.
<http://eprints.ecs.soton.ac.uk/8204/>
24. International Standards Organisation, "Reference Model for an Open Archival Information System (OAIS)", (2003), ISO 14721:2003.
25. Jones, R. (2004) "DSpace vs. ETD-db: Choosing software to manage electronic theses and dissertations", *Ariadne* 38
<http://www.ariadne.ac.uk/issue38/jones/>

26. Lagoze, C. and H. van de Sompel (2001) "The Open Archives Initiative: Building a low-barrier interoperability framework." In *Proceedings of the First ACM/IEEE Joint Conference on Digital Libraries*, Roanoke, VA.
<http://www.cs.odu.edu/~pothen/Courses/CS791/oai-jcdl.pdf>
27. Lynch, C. (1997) "Building the infrastructure of resource sharing: union catalogs, distributed search, and cross database linkage." *Library Trends*, 45(3), 448-461.
28. Lyon, L. (2003) "eBank UK: Building the links between research data, scholarly communication and learning", *Ariadne* 36
<http://www.ariadne.ac.uk/issue36/lyon/>
29. Lyon, L. (2004) "Knowledge Extraction and Semantic Interoperability", *Delos Newsletter 1*
<http://www.delos.info/newsletter/issue1/cluster-reports/#5>
30. Manola F. and E. Miller (Eds.) (2004) "RDF Primer".
<http://www.w3.org/TR/rdf-primer/>
31. McGuinness, D.L. and F. van Harmelen (2004) "OWL Web Ontology Language Overview".
<http://www.w3.org/TR/owl-features/>
32. Miles-Board, T. (2003) "Supporting Management Reporting: A Writable Web Case Study", In *Proceedings of The Twelfth International World Wide Web Conference (WWW2003)*, 234-243, Budapest, Hungary
33. Millard, D. E. et al (2003) "Hyperdoc: An Adaptive Narrative System for Dynamic Multimedia Presentations," Technical Report ECSTR-IAM02-006
<http://eprints.ecs.soton.ac.uk/7279/>
34. van Ossenbruggen, J. et al (2001) "Towards Second and Third Generation Web-Based Multimedia," WWW10, Hong Kong
<http://doi.acm.org/10.1145/371920.372143>
35. Page, L., Brin, S., Motwani, R. and T. Winograd (1999) "The PageRank Citation Ranking: Bringing Order to the Web".
<http://newdbpubs.stanford.edu:8090/pub/1999-66>
36. Palmquist, R. "Bibliometrics"
<http://www.gslis.utexas.edu/~palmquis/courses/biblio.html>
37. Redner, S. (1998) "How Popular is Your Paper? An Empirical Study of the Citation Distribution", *European Physics Journal B4* 131-134
<http://arxiv.org/abs/cond-mat/9804163>
38. Rousseau, R. "Timeline of bibliometrics"
http://users.pandora.be/ronald.rousseau/html/timeline_of_bibliometrics.html
39. Shadbolt, N.R., N. Gibbins, H. Glaser, S. Harris and m.c. schraefel (2004) "CS AKTive Space or how we stopped worrying and learned to love the Semantic Web", *IEEE Intelligent Systems* 19(3), 41-47.

40. Silagadze, Z.K. (1997) "Citations and the Zipf-Mandelbrot's law", *Complex Systems* 11, 487-499
<http://arxiv.org/abs/physics/9901035>
41. Stubley, P. Rob Bull and Tony Kidd (2001) "Feasibility study for a national union catalogue," JISC, RLSP, British Library's Co-operation and Partnership Programme
<http://www.uknuc.shef.ac.uk/NUCrep.pdf>
42. Stuckenschmidt, H. , F, van Harmelen, A. de Waard, T. Scerri, R. Bhogal, J. van Buel, I. Crowlesmith, Ch. Fluit, A. Kampman, J. Broekstra and E. van Mulligen (2004) "Exploring Large Document Repositories with RDF Technology: The DOPE Project", *IEEE Intelligent Systems* 19(3), 34-40.
43. Tate, A., Dalton J. and J. Stader (2002) I-P² - Intelligent Process Panels to Support Coalition Operations, *Proceedings of the Second International Conference on Knowledge Systems for Coalition Operations (KSCO-2002), Toulouse, France, 23-24 April 2002.*
<http://i-x.info/documents/2002/2002-ksco-ip2.pdf>
44. Trigg, R. (1983) "A Network-Based Approach to Text Handling for the Online Scientific Community", PhD thesis, University of Maryland, University of Maryland Technical Report, TR-1346
45. Uren, V., S. Buckingham Shum, G. Li, J. Domingue and E. Motta (2003) "Scholarly Publishing and Argument in Hyperspace", In *Proceedings of The Twelfth International World Wide Web Conference (WWW2003)*, Budapest, Hungary.
46. Wenger, E. (1999) "Communities of Practice: Learning, Meaning and Identity", *Cambridge University Press.*