SWAD-Europe deliverable 12.1.2: Semantic Blogging and Bibliographies - Requirements Specification

Project name:
Semantic Web Advanced Development for Europe (SWAD-Europe)
Project Number:
IST-2001-34732
Workpackage name:
12.1 Open Demonstrators
Workpackage description:
☞http://www.w3.org/2001/sw/Europe/plan/workpackages/live/esw-wp-12.1.html
Deliverable title:
Semantic Blogging and Bibliographies - Requirements Specification
URI:
☞http://www.w3.org/2001/sw/Europe/reports/requirements-demo-1/hp-requirements-specification.html
Authors:
☞Steve Cayzer, HP Laboratories, Bristol, UK
☞Paul Shabajee, Graduate School of Education and ILRT, Bristol, UK
Contributors
Dave Reynolds , Ian Dickinson, HP Laboratories, Bristol, UK
Abstract:
Workpackage 12.1 comprises two demonstrator applications designed to both illustrate the nature of the semantic web and to explore issues involved in developing substantial semantic web applications given the current state of the art.
This document outlines the requirements for the first of these applications; a semantic blogging tool applied to the bibliographic management domain. We start with a brief summary of the domain and our reasons for choosing this particular demonstrator. We outline the context and aims of the project, after which we outline the user requirements and list the associated components, identifying existing resources that are available for inclusion in the demonstrator. The requirements are separated into core functionality, which we expect to deliver, and optional extensions, which will be undertaken if time permits. A series of appendices contain the details of related and background work.
Status:
First release.
Comments on this document are welcome and should be sent to ☞Steve Cayzer or to the ☞public-esw@w3.org list. An archive of this list is available at ☞http://lists.w3.org/Archives/Public/public-esw/

Contents

1 Introduction

This report is part of ☞SWAD-Europe ☞Work package 12.1: Open demonstrators. This workpackage covers the selection and development of two demonstration applications designed to both illustrate the nature of the semantic web and to explore issues involved in developing substantial semantic web applications.

This report forms the requirements specification for the first demonstrator, Semantic Blogging and Bibliographies. The aim of this report is to set out the key criteria that we wish to achieve with this demonstrator. The intention is to provide sufficient framing so that the reader can understand what the demonstrator is expected to do, and why we feel that the capabilities are germane to the SWAD-E agenda. This document is not a design document, and so detailed implementation decisions will not be presented here, although general architectural principles will be discussed where appropriate.

We start by reiterating our notion of semantic blogging for the bibliographic domain. We then set the overall context and aims of the application, discussing the selection process we used in order to draw up a set of requirements for a demonstration vehicle which has sufficient illustrative power, while remaining feasible to implement within the project timescale. Sections 4-6 outline the user requirements and list the associated components, identifying existing resources that are available for inclusion in the demonstrator. The requirements are separated into core functionality, which we expect to deliver, and optional extensions, which will be undertaken if time permits.

The Appendices contain details of work undertaken to support the requirements selection process. The criteria for requirement selection, and details of other use cases considered, are discussed in ☞Appendix A. An extensive survey of related work can be found in ☞Appendix B, while details of current bibliographic standards and software can be found in Appendices ☞D and ☞E. We conducted a short user study to act as a reality check on our assumptions - the results are summarised in ☞Appendix C. .

2 Semantic blogging for bibliographies

In this section we recap some of the key points raised and discussed more fully in our previous report ☞[SWADE_ANALYSIS]. In summary, we aim to take an existing phenomenon (blogging) and semantically enrich it. The new metaphor is grounded by applying it to a concrete domain, bibliographic management.

Web logging, or *blogging* ☞[ESSENTIAL_BLOGGING], is a well known phenomenon that has a number of attractive features. It provides a very low barrier to entry for personal web publishing and yet these personal publications are automatically syndicated and aggregated via centralized servers (e.g. ☞blogger.com) allowing a wide community to access the blogs. Blogs have a simple to understand structure and yet links between blogs and items (so called *blog rolling*) supports the decentralized construction of a rich information network. While we want to extend the blogging metaphor, we also want to preserve its key values, especially its simplicity. We want to build on blogging's proven potential for publishing, syndication & discovery, and community formation.

The notion of semantic blogging builds upon the success and clear network value of blogging by adding additional semantic structure to items shared over the blog channels. This semantic structure has two key effects:

- **Rich Query**: Semantically enriched blog metadata enables new subscription, discovery and navigation behaviours.
- **Rich Structure**: Access to ontological markup enables both richer annotation and sharing of higher level structures (like categorisation schemes), encouraging peer commentary and recommendation activity.

There is some movement in the blogging community to what we call semantic blogging. The Movable Type Trackback functionality ☞[MT_TRACKBACK] allows two way linking between blog items. Some blog commentators envisage the next step, which is attaching semantics to these links [☞LINKING_DANGEROUSLY]. Richer (hierarchical) categories are facilitated by the RSS2.0 standard ☞[RSS2.0]. The Topic Exchange activity ☞[TOPIC_EXCHANGE] uses TrackBack as a step towards the use of shared ontologies. Further details on these and other activities can be found in the appendix on ☞related work, but it is worth emphasising them here. These developments indicate that there is a real need for the capability that we are proposing.

Bibliography management is a large and complex domain, and the appendices contain reviews of relevant ☞tools and ☞standards. Within this domain, there is a need for lightweight tools for small group bibliography management (see the ☞User Study). We feel that this need is an ideal testing ground for the semantic blogging paradigm. It is not our aim to duplicate functionality of the existing bibliographic tools and standards; rather, we seek to integrate our demonstrator with such tools so that

users are enabled to use the additional functionality within the context of their current work practice.

3 Context and Aims

It is not immediately clear why two successful, but distinct, paradigms (blogging and the semantic web) should be brought together. We believe, though, that there are compelling reasons to combine the two. The rich structure and query properties enabled by the semantic web greatly extends the range of blogging behaviours, and allows the power of the metaphor to be applied in hitherto unexplored domains. Bibliographic management is a concrete example of a task that illustrates the benefit from the combined paradigm. Although traditional bibliographic management deals mainly with static categorisations, the needs of a small group collectively exploring a domain exhibit a more dynamic, community based flavour. Here is a task which is characterised by a need to share small items of information with a peer group in a timely, lightweight manner. This information should be easily publishable, easily discoverable and easily navigable. It should be simple to enrich the information with annotation, either at the point of delivery or later. The information should be archived in a commonly understood way for effective post-hoc retrieval. It should be possible to be notified, in a timely way, of new items of interest. We believe that a combination of blogging and semantic web technologies offers an ideal solution to this problem - blogging for low barrier publishing, a simple shared conceptual model, and a mechanism for natural, dynamic community formation - semantic web for rich structure, which enables richer community annotation, and rich query, which enables more powerful discovery and navigation.

It is the aim of this demonstrator to develop a tool that is simple, useful, extensible and illustrative. Simple, because it should be easy to learn and to use. Useful, because it should do something that users actually want, efficiently and reliably. It should be deployable. Extensible, because although we ground the requirements in the bibliographic domain, we expect it to be reusable for other semantic blogging applications. And illustrative, because we wish to incorporate features that demonstrate the advantages of the semantic web approach (semi-structured data, semantics and webness) without losing the key advantages of blogging (low effort publishing, easy subscription and decentralized discovery).

Combining these desiderata, we arrive a key set of capabilities that we wish to illustrate through this demonstrator, and which should therefore be captured by the requirements.

Rich Query -

- **Subscription** Blogging as it stands provides a convenient time-based channel structure. However the channels must be subscribed to independently and the filtering within these channels is limited to free-text search. By using the semantic blogging approach we make it possible to subscribe to categories such as "what's going on in field x". Aggregation need not be bound to a single feed, and can do semantic filtering within these feeds. It is thus possible to define a subscribed category which contains a subset of items from a number of channels.
- **Discovery** This capability means the discovery of new, potentially useful, sources of information. Blogging discovery mechanisms include directories, blogrolling, item-item hyperlinks and informal routes (such as email). The semantic web offers capabilities that allow an even richer discovery capability. For example, bibliographic items can be related via an ontology, and thus a user can query a wide community (and many channels) for papers "in the semantic web category". If items are associated with a consistent (or at least corresponding) identifier, users can also ask for blog entries "about this item".
- **Navigation** Current navigation within blogs is largely chronological, and hence not well suited for non time-based data, such as bibliographic records**.** We want to enable users to browse their own blog (and others') for such data in an intuitive manner (for example, across a user-defined topic hierarchy). In addition, semantic links will allow navigation to other items "related to", "agreeing with" or "disagreeing with" this one. These links are an alternative to explicit citation-type links and thus provide a community based network based around ideas and discussions.

Rich Structure -

- **Shared Ontologies** By exploiting the semantic web ontology layer we are able to not only represent rich topic hierarchies for classifying citations, but also to link and share these topic sets across communities. Thus different communities can use distinct classification schemes and yet the data can be shared across the same infrastructure and potentially the relationships between terms in the different schemes can be explicitly represented to allow cross-community search.
- **Annotation** Annotations need not be limited to free text entries. They can fit into semantically meaningful structures (like ratings, comments and section-specific annotations), allowing more effective discovery, navigation and maintenance of blogged data. They may also be supplied with

context data, such as provenance. Because we index items by URI, we enable such multiple, provenanced annotations to be integrated across a community.

These capabilities need to be set in the context of bibliographic management. So interoperation with existing tools will be explicitly included as part of the requirements.

We seek to build an application which demonstrates these key features and has demonstrable utility. We have chosen bibliography management as a domain, and therefore the overall application should be one that enables a group to manage their collective bibliographic records. We expect to produce a tool which, in general terms, allows a community to effectively manage their bibliographic data, and to harness the power of the group for discovery, recommendation and collective learning. Specifically this means that the demonstrator will exhibit lightweight capture of bibliographic data, rich discovery and navigation mechanisms, useful presentation of the relevant information, and good integration with other tools. In short, an application that is genuinely useful for bibliography management.

4 User Requirements

In this section, we consider the requirements from the standpoint of the user (architectural and technical issues are considered in the next section). We present a use case which captures the core functionality of our demonstrator. It is as simple as possible while still providing useful functionality. It is nevertheless quite feature-rich, and involves a lot of semantic technology. Other features, some of which involve significant research hurdles, can be built onto this framework as extensions. We discuss some of them later.

Core Functionality: Local Group Bibliography Management - This scenario can be summarised by imagining a commonly encountered problem - that of sharing papers and citations with a project group. Existing solutions tend to be ad-hoc and unsatisfactory. For example, email is useful for a speedy, low cost notification of a useful paper. It also allows the shared citation to be annotated at the point of delivery. However, it is often difficult for the recipient to categorize the data appropriately ("I often find that I receive the right paper at the wrong time") and post-hoc, principled retrieval of such received (or even sent) papers is next to impossible. Another method is web pages, on which people can post useful literature with an arbitrary amount of structure. Useful though this is, the publishing process is far from low cost, and the reader is required to understand the publisher's conceptual structure. The reader is also required to 'ping' the website rather than being notified of new papers. Topic portals are another web-based example, but the coverage is often too general and the content not necessarily up to date. A third method is shared bibliographic databases, such as ProCite or EndNote. These formats do allow sharing of bibliographic information between small groups, but there is considerable 'lock-in' to the formats, which are often unwieldy and inflexible. Finally, there is the possibility of managing bibliographic data using the existing blogging infrastructure. We performed some simple, informal trials in our group to identify the main problems and found that blogging is currently an unsuitable environment for the capture of bibliographic content. In particular, it is difficult to organize the blog in such a way that the large numbers of articles can be managed effectively by the publisher, let alone other readers. At a minimum, we need to add structure, in a flexible yet low-cost manner in order to make the metaphor work for this domain.

Note that these are simply our intuitive reflections on the domain. We conducted a short user study to test these intuitions - the results can be found in ☞User Study. This study identified a number of limitations with current approaches and generated a wishlist, which is matched encouragingly well by our requirements.

Illustrative Scenario

Tim is interested in semantic blogging. He does a Google search for relevant papers, and finds some that look interesting. After having read a few, he posts the details on his semantic blog. There are a variety of low cost ways available to do this. For one paper, he chooses a 'copy and paste' importer into which he pastes the BibTex entry (from CiteSeer) and the marked up item is automatically added to his blog entry. He categorises the item, rates it and adds a free text comment. Other, unread, papers are added (to the same category), but not commented on or rated.

For a paper he has not read yet, he wonders if anyone else has. He performs a community query for that paper and receives a summary table with the comments and rating of his peers on that paper. For the paper he has read, he wants to find related papers. He performs a community query for 'papers like this' which generates no hits. He chooses the 'generalize this query option' and this time there are some related papers. Again they are presented in summary form with title, topic and rating (this can be customised) and he follows links to interesting looking papers to examine his peers' comments. Another community query is 'find peer commentary' which finds peers' blog entries linking to this one. Again, the retrieved entries are displayed in summary form.

One of the followed links refers to a paper which looks interesting enough for him to read himself. He accesses the abstract and downloads the PDF. He creates his own blog entry (using a 'blog this' bookmarklet option) which automatically copies all the metadata (title, author etc) created by his peer. It also creates a link between his (new) blog entry and the peer's. He can also, if he wishes, 'bulk import' peers' blog entries from the summary table. He can now add his own comment and rating, and recategorise the item if required. Finally, each bibliographic item also has a list (0 or more) of citation links - papers cited by, and citing, this paper. Tim may follow these links if he wishes to find other interesting bibliographic items, and possibly import them too.

Tim decides to export his blog to ProCite, which he uses for writing papers. He chooses the 'export new items' option which converts all papers added (or modified) since the last export to a ProCite-compatible input file.

Tim is interested in keeping up to date with semantic blogging papers. He does this in a number of ways. Firstly, he enables a tracking feature for his blog entries, which enables him to record all blog entries linking to this one. Secondly, he adds an 'email alert' to his blog entries in the semantic blogging category so that he is notified when anyone links to one of his blog items in this category. Thirdly, he sets up a community alert for the semantic blogging topic, which provides an update on any new community blog entries in this category. He sets up a web page to display a summary of these new semantic blogging entries.

Later, when actually writing a paper, Tim uses his blog (rather than ProCite) because of its superior semantic search capabilities. He browses his bibliography data for papers with a topic (or supertopic) of semantic blogging and again gets a summary table. He can filter this table using other metadata (eg rating) or unstructured data (i.e. free text). He can also augment the table using a community query. Once his is happy he has the right subset of papers, he exports the data to a BibTex file for use with his LaTeX paper.

Requirements List

This scenario illustrates a number of requirements for the core functionality:

- **Painless Import**. Bibliographic items can be generated automatically from custom sources (EndNote, ProCite), from web pages or from copy and paste. At a minimum, copy and paste of BibTex data should be supported.
- **Rich Navigation**. Blogs should be navigable using metadata. At a minimum, a topic should be selected from a topic hierarchy, and a summary table produced showing bibliographic items at (or below, at user discretion) the chosen topic. Such summaries can be filtered easily by metadata or free text search.
- **Assisted markup**. Users should be helped as much as possible to markup items according to the central ontology. For example, much of the metadata can be automatically created on initial import. Categorisation (and recategorisation) into a topic hierarchy should be possible. An intuitive (and customisable) UI to enter other metadata, such as ratings and comments, is essential. A description for the ontology terms should be available on request.
- **Provenance**. Ratings and comments need to be provenanced by the author. However the provenance information need not actually be used here - it is acceptable for different blog entries about the same paper to be listed separately and not integrated.
- **Metadata Visible**. For a single blog entry (or for a summary view), metadata should be visible (this can be customisable). For example: ratings, annotations, comments, author, title, journal and so on.
- **Query** Three types of community query need to be supported here. The first one is a query for a paper, which returns all blog entries that reference a particular paper. The second is a community query for related papers, which returns all blog items about papers *categorised under the same topic as* the source paper. This query should be generalisable - for example, search for papers categorised *under the immediate super-topic*. The third query is a search for any related blog entry. Note that at this stage, the third query is likely to return exactly the same results as the first. It is also achievable using the current TrackBack blogging mechanism. However, this mechanism is a prerequisite of more advanced features, such as semantic networks.
- **Authoritative Identifier**. For community queries to be meaningful, the same paper should be referred to by different people using the same identifier. A simple solution to this is adopted; the use of a common, stable and unique URL. So the implications for the user is that only such sources (eg CiteSeer) will be supported.
- **Access to data**. A blog entry should allow access to the underlying paper. (Indirect access, for example via a CiteSeer URL, will be sufficient).
- **Subscription/Discovery**. Semantic, community alerts such as 'alert me when more semantic blogging articles are added' features will be needed.
- **Common Ontology**. One of the scoping factors for this scenario is that it can be assumed that all the peers use the same ontology. Complex parts of the ontology (eg topic hierarchy) must be

browsable and the ontology should be descriptive (so that people know what the ontology terms mean, and are thus given guidance to apply them correctly). The ontology should include: classes for book chapter, journal article and so on; author; title, journal, volume and issue (for journals); booktitle, chapter, publisher and ISBN (for books); and other appropriate bibliographic fields. The ontology should also include user-supplied metadata such as rating & comment, and classification metadata such as topic.

- **Link to other items**. Blog items should link to other blog items in a number of ways. The straight hyperlink is of course possible but links in the other direction also necessary. A "Blog this" feature will copy across metadata relating to the paper, and also create such a bi-directional link. Implicit links (via the topic hierarchy) have been mentioned above.
- **Export**. Bibliographic data should be exported to a variety of useful forms, certainly including ProCite and BibTeX. The export should function should have a choice between 'export new' and 'export all' items.

Exclusions

There are some things explicitly excluded from the core demonstrator, some of which will be discussed as extensions. These considerations should guide the design so that such extensions are not locked out (and where possible are enabled and facilitated by the infrastructure):

- **Disparate ontologies**. Although part of the aims of this demonstrator are to explore disparate, external and extensible ontologies, it is expected that within a small group the peers could be expected to adhere to a common, centralised, predefined ontology. Ontology modifications will be limited to simple extensions and would require no special machinery (text editing of a centralised file would be sufficient).
- **Semantic links**. For the purposes of this scenario, blog items may simply be linked to other blog items. Argumentation networks will not be required here. In addition, explicit, user defined links between the papers will not be supported; the blog links and topic groupings are expected to be sufficient.
- **Diverse content:** A simple solution to providing authoritative identifiers is adopted here; the use of common, stable and unique URLs. Such a solution precludes items from sources other than those with such URLs (eg an agreed online database). The approach taken here is a limitation but one that is necessary in building an feasible demonstrator. Such a limit however not only precludes many papers, it also precludes other useful (and citable) sources such as books, standards, people, companies and personal communications. These will considered below.
- **Authority files** Authority files identify an entity (such as an author, a corporate body or a geographical location) unambiguously. This facility is extremely important for accurate provenancing and annotation, and poses the same problems as for diverse bibliographic content, discussed above. Essentially, it is the problem of generating a stable unique identifier without a universally agreed scheme for doing so.
- **Visual navigation**. The network of bi-directional links is useful, but in order to navigate it effectively, a graphical UI would be helpful. It is considered that within the context of a small group, the network is small enough that although a user would make use of such navigation, the path would not be complex enough to require visualization beyond hyperlinks and simple summary lists.
- **Citation links** The links mentioned above are blog-blog links. However, integrating with an online tool like CiteSeer [☞CITESEER] would allow access to citation links, which join the underlying papers together. For the purposes of this demonstrator, a simple link to the appropriate tool (eg CiteSeer page) will provide suitable access to the citation network.
- **Context dependent markup**. The meaning of an annotation may be non-intuitive. For example, 'rating' is a term which may be ambiguous. Is a user rating the style of a paper or its content? Is the rating from the perspective of a certain class of user (eg "This paper is a great introduction to RDF but not worth reading if you are already an expert")? In the demonstrator, we do not explore the possibilities of context dependency, but ensure that the ontology terms are described (and the descriptions made available to the user) so that such ambiguities can be avoided.
- **Rich annotation**. There are various possibilities that arise with the use of rich metadata. For example, annotation on partial content, such as a comment on 'paragraph 2 of the article' Also, the integration of annotations from various sources (with the use of provenance). Finally, other contextual information associated with annotations (such as 'type' of rating).
- **Reputation**. Community queries may be augmented by the reputation of the source. This aspect, while interesting, will not be explored in the demonstrator.

Extensions - We now discuss three possible extensions to the semantic blogging demonstrator. These are not considered to lie within the scope of the project deliverables, but may nevertheless be implemented as related projects.

Extension 1: Shared Ontology

This extension deals with groups who have the same data but different topic hierarchies. An example might be a group of users who are interested in the same content and have similar, but subtly different ways of categorizing that content. These would include both simple labelling differences and differences at the level of detail. For example, to take the bibliographic domain, some users might have a category 'blogging', which other users call 'web logging'. Some users might be very interested in the topic of blogging, and subcategorise that arena into MovableType, Blogger, Radio Userland and so on. Other users are not that interested and put all blog resources in the same category. Note that these differences are quite subtle and yet present considerable hurdles to interoperability between the members of a community. Essentially, the problem is where a group of users in a loosely defined community want to share a largely consistent conceptual model, but still to allow individual variations on it. The essential capability would be to align two taxonomies (limited here to saying that two nodes are equivalent). This has two consequences. Firstly, it allows a user to view the same concepts but through a set of labels that s/he finds meaningful. Secondly, it enables the reuse of categorisation effort. In the above example, one user might map a category to 'blogging' and hence gain access to all the finer grained categorizations performed by other peers under that concept.

We believe that such an extension provides a compromise position - on the one hand, it enables genuinely useful functionality to a community with different ontologies, yet on the other hand it avoids the scale and complexity of more ambitious projects like APECKS [☞APECKS] which would make it difficult to implement within the project timescale.

Extension 2: Semantic Linking

Attaching semantics to item-item links allows the possibility of navigating an argumentation network, as explored in the ClaiMaker project [☞CLAIMAKER-WEAVE]. Such a possibility is certainly powerful, and yet there is a risk that without appropriate visualisation and navigation tools the capability will simply produce cognitive overload. This extension is therefore limited in scope to three key capabilities. Firstly, an enhanced metadata creation tool that allows users to create semantic links between their blog items and others'. Secondly, an extended query mechanism that allows the user to retrieve 'blog entries that agree with this one' for example. This facility could potentially be made transitive, although there is a danger in assuming that a "someone who agrees with someone I agree with" also agrees with me! Thirdly (and optionally) a visualization tool that enables one to view the activity surrounding a particular bibliographic item (papers that agree, papers that disagree, papers that extend and so on).

This extension is deliberately limited to maintain feasibility. In ClaiMaker, it is not the papers that are linked but the concepts. Thus one paper may give rise to a number of concepts, all of which are nodes in the argumentation network. Such a mechanism is clearly of benefit, but would introduce too much complexity to be considered here. It also raises further issues to address. For example, the concepts have to be identified and categorised so that they can be reused, which presents two difficulties. Firstly, once a non-trivial number of concepts are generated it is difficult to find the right concept to reuse. Secondly, it is difficult to define a concept in a manner both specific enough to be useful and general enough to be reused. In fact, to some extent, even this sophisticated model does not go far enough. Consider, for example, the issue of trust. If I assert that 'paper X contains concept Y' then we need to build a mechanism for someone to dispute this assertion.

Semantic links raise another issue that is currently unexplored. That is - do the links refer to the blog items or to the underlying paper? We have so far blurred the distinction between blog-blog and item-item links. In fact, much of the blog metadata (eg author, title) is more correctly viewed as being attached to the underlying item. Making this distinction explicit allows richer possibilities: for example, "This paper disagrees with that paper" versus "I disagree with what you are saying about this paper". However, such a mechanism might simply be confusing and is thus not considered here.

Extension3: Name by property

In order for two people to talk about the same item, it is necessary that they use a common identifier (or at least that the identifiers can be mapped to one another). The constraint on the naming of items adopted in this core demonstrator provides a simple, workable solution to this problem. Essentially, it uses some socially agreed provider of identifiers (for example CiteSeer) to ensure that when two users reference the same paper, they use the same identifier. Such a solution can provide a significant amount of functionality but more powerful solutions exist. One such solution is for people to take identifiers from different schemes and, where appropriate, link them. One might imagine, for example, an identifier from CiteSeer and an identifier from MEDLINE, referencing the same paper. A user could discover both instances by, for example, performing a pattern matching search on the paper title. Matching papers would be returned in a summary list, and the identical papers linked together. From that point on, as far as the system is concerned, the two papers are the same paper, and the results are available for other

users. A query on 'entries that link to this paper' would return the union of linkages to both papers.

Such a solution, while offering a wider source of items, can be extended further. Using a *name by property* paradigm, users can identify a paper using descriptions such as "The report with an author 'Steve Cayzer' and the title 'SWAD-Europe: Semantic Blogging and Bibliographies - Requirements Specification'". Such an approach would also facilitate versioning of documents. Note that this paradigm does not rely on unique identifiers. It is entirely possible for two users to refer to different papers using the same description (consider "The paper with author 'Ying Ding' and date 2002 and title 'Golden Bullet*' " which matches two papers [☞GOLDEN-BULLET-1, ☞GOLDEN-BULLET-2]). On the other hand, two equivalent papers might have quite different descriptions ("The paper with author 'Boris Omelayenko' in conference 'FLAIRS2002'" [☞GOLDEN-BULLET-2]). But these disparate descriptions can be integrated with a suitable query, and those identical papers linked as before. This time, however, the descriptions would be aggregated and the (more complete) description would then be available for reuse. Similarly, disambiguating metadata can be used to enrich the identifiers of distinct papers with identical descriptions. This will have the side effect of disaggregating the annotations that peers had attached to each of these papers.

An immediate consequence of this is that the core demonstrator should adopt an identification scheme that enables this extension. For example, the unique identifier (eg CiteSeer URL) can be attached to the paper as just another property. Of course, properties which can act as unique identifiers can be marked as such (eg using `InverseFunctionalProperty` from the proposed W3C Web Ontology Language [☞OWL]). Such a mechanism allows the core demonstrator to function as before, while providing a minimal hook for this extension.

Other Ideas

There are a number of other ideas which, although not under active consideration, provide further examples of extensions which it should, at least in principle, be possible to implement over the core demonstrator.

Rich Discovery: The discovery mechanism described above is useful, but it can be made even more powerful. For example, if people annotate their channels then it should be possible to discover "channels about the semantic web" and to perform a search within that restricted domain, formatting the result as an RSS feed. Another example would to generalize a relationship search i.e. "Are there any more blog entries describing this application idea?" - where "application idea" is a concept related to (perhaps indirectly) the underlying paper.

Visualisation: Rich path visualisation affords a powerful way to improve navigation. Visualisation could be of papers, blog items or peers. One possibility would be to present a network view of blog entries (or bibliographic items), connected by (typed) links. Different types of link would be shown in different colours, and added/omitted from the map as the user chooses. Complexity would be managed by limiting the 'window' (path length) from the current blog entry. Another visualisation possibility is a view of the blog organized along 'semantic UI' lines, using an approach similar to the Haystack project [☞HAYSTACK]

Large Group Aggregation: Currently, we expect to support only a small, manageable community. The demonstrator will be built in as scalable a way as possible, but will not be deployed in a large group situation. As the community grows, various infrastructural issues arise. For example: community annotations may require an annotation server; shared access to community ontologies becomes problematic; community editing of ontologies may require a more sophisticated approach such as Kaon [☞KAON] or APECKS [☞APECKS]. In addition, even assuming the stable name problem has been solved, how do we discover blog items 'about' XYZ in a large (potentially worldwide) community. This is a peer-to-peer query issue. Finally, scaling up presents challenges other than the purely technical; the need for different navigation metaphors to avoid cognitive overload has already been mentioned.

Fine Grain and Rich Media Annotation: This extension is particularly intended to explore the annotation on partial content (e.g. a comment on "section 2" of an article). This could be achieved by context data on the comment metadata. A more natural way would be to create a new URI which referred explicitly to the fragment and to comment on that. Of course, such URI's would have to be aggregated together since a user would expect to be able to ask "who has commented on this paper *or any bit of it*". Another possibility is to allow annotations on items such as pictures, audio files and video clips. In both cases, a suitable user interface is required to enter the annotations. This scenario also requires a more sophisticated ontology, to add context (in a similar way to provenance) to disambiguate annotations. For example, is a user rating the resolution quality of a video clip or the illustrative nature of its content?

Content Management: This refers to the need to 'get at' the content underlying each blog item. Essentially, this extension encompasses smoother integration abilities. One example would be increased co-operation with sources such as CiteSeer, so that users are enabled not only to retrieve the full text of the article, but also to browse the citation network in conjunction with the blog network.

Privacy and Reputation Blog entries could be marked with various levels of privacy (personal, community, public) or more flexibly we could implement a role based access control mechanism. The

flip side of this coin is the use of reputation to augment community queries.

Assisted Markup The demonstrator as scoped has lightweight assistance for markup. Richer possibilities exist - for example: visual markup via drag and drop; automatic suggestions where similar objects/data already exists; automated classification of incoming blogs against an existing channel hierarchy; other enrichment of imported data. Another possibility is to 'cluster' blog entries based on link structure.

5 Component Requirements

There are a number of components that could be expected to be built into the core demonstrator. Some of these can be built on existing resources as explained ☞below.

Core Components -

- **Assisted markup tools:** This component will allow the user to easily markup a blog entry with useful metadata. One tool will take an ontology and allow a power-user to build up an *markup template*. Such templates will then be available for other users to apply to their bibliographic data. Widgets will be required to browse certain types of metadata - for example, enumerated lists, typed data and hierarchical browsers. Another tool will be a 'Blog This!' script (implemented as a bookmark) that will take a bibliographic item, copy across the metadata and create a bi-directional link.
- **Ontology:** We will build a suitable ontology. Actually there is more than one ontology - there is an bibliography ontology, an ontology for annotations (comment, description etc), an ontology for topics, an (optional) ontology for semantic links and an ontology for channels (primarily subchannelOf). In addition, the ontologies will be created such that semantic extensions and provenance mechanisms are enabled. As mentioned above, a community editing tool is considered outside the scope of this project.
- **Custom metadata view** We will provide a mechanism for the user to customise the view of metadata (eg summary tables and the like) which will probably be script based. The resulting views will be available as *display templates* which can be used to view a single blog entry, an individual blog, or a query result. Other templates, such as query and markup templates, may also be defined in this way.
- **Navigation module:** We provide a navigation module in order to enable a user to find his or her way around a set of items. These items may be hosted on a single blog (possibly their own) or may be the result of a query. The navigation is primarily through filtering, using *query templates*, and the results will be displayed in a custom metadata view.
- **Query module:** The query module is one of the more complex components. It will primarily offer a community search for blog items of interest. The search can be a metadata based filter: "Find me all blog entries relating to this item", "find me all blog entries marked up with this topic" or other metadata. Note that this query should be generalisable - for example, search for papers categorised *under the immediate supertopic*. It can also be a network based search: "Find me all blog entries linking to this one", or (in the extended case) "find me all papers who agree with this". The results of the query will be available for subscription (eg RSS feed). Provenance data will also be returned (and rendered according to the display template settings).
- **Integration bridges** A variety of mechanisms will be supplied to make it easier to import data (cut and paste of BibTex) and export data (to BibTex and ProCite). Further import/export mechanisms may be added if time permits, according to user need.
- **Infrastructure**: There are a variety of less visible, though necessary, components that act as the 'plumbing' underlying the demonstrator. The communication protocol between peers is a good example (although here we are likely to leverage the existing blogging infrastructure, see below). Another example is a component that allows the propagation of community queries. Although this would ideally be performed in a peer to peer manner, it is not the purpose of this demonstrator to investigate P2P query mechanisms, and so a centralised solution, such as an aggregator, will probably be adopted. A further requirement is the integration of community annotations. For example, if different people comment on the same item (in different blogs) then these comments could usefully be combined. That is, where an explicit equivalence link exists between two bibliographic items, then it would be desirable (though not essential) for comments on one person's blog to be automatically propagated to the other.

Extensions - We now discuss the components necessary to implement the functionality outlined in the ☞extensions

Shared Ontology: In order to implement this functionality, we require a component which would allow a user to take another user's ontology and to mark equivalences between that ontology and their own. We need an extended ontology which defines such relationships. The component is scoped by only

allowing a very constrained ontology (i.e. taxonomy) to be compared, and by restricting the links to be simple equivalence relationships. The query module needs to be enhanced, so that equivalence relationships are followed to return enriched result sets. We also need a mechanism to access the linked ontology (eg by importing a subset of a taxonomy tree, or by providing a link to it). Once such access is provided, the markup, view and navigation modules will need extensions to provide the user with a natural way to access and use the new ontological structure.

Semantic Linking: For this extension, we need a component which allows users to *type* their links (eg agreesWith) between blog entries. An extension will also be needed to the query component (eg "I want blogs that agree with this one, transitively"). It is possible that the extra semantic richness would necessitate an improved metadata viewer, and clearly a network visualisation module would be relevant (though not necessarily essential) here. In any case the navigation module needs to be extended to support the new navigation metaphor.

Name by property: We need several components to implement this extension. Firstly, we need a component which allows users to mark two items as equal. The query module needs to be enhanced so that future queries on either of the items will return results germane to the other. Users can also remove equivalence links. Secondly, we need an identifier-generating utility that allows a user to choose the properties used to describe the item (there would be a suitable default, for example; author, title, year). The combination of properties is called an *identifying reference expression* (IRE). Thirdly, a mechanism that detects the equivalence of incoming items (eg from a community query) by comparing IREs (possibly with inference). Such items are then automatically made equivalent just as if they had been marked as such by a user. Fourthly, a component which implements the above approach with respect to *authority files,* for example to identify peers, authors or companies.

Note that this component is scoped down for feasibility and has two limitations. Firstly, items can be identical without having the same IRE. This possibility is mitigated against by having default, automatically generated (and thus hopefully commonly used) IRE patterns, but ultimately users can mark such items as equivalent manually. Secondly, items with the same IRE might actually be different. Users can correct this, partly by removing the equivalence link and partly enriching the IRE appropriately. Note that a side effect of this disambiguation is that the combined annotations would be split; each annotation being returned to its owning item.

6 Existing Resources

There are some resources that are available for re-use or as a base to build on:

- **Blogging Infrastructure** The extant blogging tools offer a powerful and useful base to build on. In particular, tools such as MovableType [☞MT], which offer full control over the blogging environment, provide a rich development environment. One utility of particular interest is the *bookmarklet*, which offers one click blogging of items of interest.
- **Blogging Extensions** There are many useful extensions to blogging in the literature. For example, Trackback [☞MT-TRACKBACK] is a facility that enables citation links to be recorded. What it does is to send a 'ping' (with summary details and a URL) whenever you comment on somebody else's blog entry. The functionality is part of MovableType but is also available as a standalone module [☞TB-STANDALONE]. There are a variety of recent extensions to Trackback (for a summary see [☞TB-SUMMARY]) including *ComeBack* (community annotation 'in place'), *BackTrack* and *MoreLikeThisFromOthers* (following community links). Finally, there are a set of utilities available [☞MT-RDF] that will help in converting MovableType blog entries to, and enriching them with, RDF. More details on these and other utilities are provided in the appendix on ☞related work.
- **Ontologies** There are a variety of useful, well thought out ontologies suitable for our purposes. For bibliographic information we can draw on standards like BibTeX [☞BIBTEX] and MODS [☞MODS] for markup, and ACM [☞ACM] for categorisation (see the appendix on ☞bibliographic standards for more detail). For annotations and semantic links, there are a number of standards, including Annotea Threads [☞ANNOTEA-THREAD], IBIS [☞IBIS-TERMS] and ClaiMaker [☞ CLAIMAKER-SCHEMA]. For channel hierarchies, there are the XML standards RSS2.0 [☞RSS20] and XFML [☞XFML]. We are not making a commitment to any of these standards at this stage but will certainly evaluate them at the earliest opportunity.
- **Ontology Sharing** Divergent ontologies can be reconciled in a number of ways. Trackback can been used to create community topics [☞TOPIC-EXCHANGE] and thus facilitate emergent ontology formation. XFML [☞XFML] offers a low cost way to define and link taxonomies, although as yet there is no RDF serialization of the specification. Nevertheless, the combination of the two is an attractive proposition. Another way to share ontologies is to use orthogonal classification schemes, as captured by the Facet Map concept [☞FACET-MAPS]. Mark Pilgrim [☞THIS-IS-XFML] demonstrates how this idea can enrich blog navigation. Topic maps

[☞TOPIC-MAPS] offer a similar (though richer) approach, for which an open source Java toolkit [☞TM4J] is available.

- **Visualisation tools**: In order to view a network, Apache offer a useful tool [☞APACHE-AGORA]. ClaiMaker [☞CLAIMAKER] also has a online demonstration showing their approach to concept navigation [☞CLAIMAKER-SANDPIT]. And an algorithm is proposed by Noel et al [☞VIS-ALGORITHM] that provides a way to 'untangle' a network of links so that they can be viewed in as uncluttered a way as possible (ie reduce the number of overlapping lines). For an interesting diagram of blogger communities, see [☞ BLOG-TRIBE].
- **Enriched metadata** Citation links between papers can be scraped from CiteSeer [☞CITESEER] pages, or obtained from the Web of Science [☞WEB-OF-SCIENCE], although the latter is subscription only.

Appendices

A Criteria for Requirement Selection

The appendix contains details of the decisions process which led to our final requirements. Our refinement process involved a number of use cases which tested various combinations of our desiderata. These use cases provide helpful context around our final requirements. We provide a brief summary of each use case together with an explanation of how it helped guide our thinking. We conclude this section with a summary table showing how the use cases map to our criteria.

Community Bibliography Management - This use case is where we want to access the bibliographic data of a large (reasonably well defined) community. In this scenario, users would take the semantic blogging approach to bibliography management, aggregating, storing, accessing, navigating and discovering bibliographic items across a loosely defined community. Members of the community are expected to have a similar conceptual model of the domain, although their actual ontologies will probably differ. The underlying data is expected to be, but is not limited to bibliographic data.

Various permutations of this use case covered all of our criteria. However, as formulated, the use case is both too ambitious and too poorly defined to be useful as a design input. Therefore we took a suitable subset of this use case (local group bibliography management) as a core for our demonstrator. We also took the idea of shared (but disparate) conceptualizations as the basis for our first extension - ☞shared ontologies. Further interesting variants will be explored in our second demonstrator, semantic community portals.

Rich Navigation - This use case was inspired by the navigation of an argumentation network, as explored in ClaiMaker [☞CLAIMAKER-WEAVE]. We noted that an argumentation network needs to be supported by a good navigation interface. We also noted that semantic links raise the issue of whether the links refer to the blog items or to the underlying content.

This use case became the basis for our second extension - ☞semantic links.

Semantic Blogging - We drafted a use case as a reminder that we want the demonstrator to be applicable to domains other than bibliography management. Certainly we want our demonstrator to handle content other then bibliographic data. Consideration of this issue highlighted the need for a flexible way of referring to underlying content (the one finessed in the core demonstrator by using CiteSeer URLs).

This use case became the basis for out third extension - ☞name by property.

Rich Discovery - This use case looked at the need for people to make discoveries based on richer channel ontologies or on the semantic content of blog items. Examples would be the discovery of channels "about" the semantic web, or a search for blog entries describing a particular concept.

We decided that this use case was not in scope for the demonstrator, but it is listed as a ☞possible extension.

Visualisation - Visualisation is a good way to improve navigation. We considered use cases which dealt with visualisation of papers, blog items and peers. One possibility would be to present a network view of blog entries (or bibliographic items), connected by (typed) links. Different types of link would be shown in different colours, and added/omitted from the map as the user chooses. Another visualisation possibility is a view of the blog organized along 'semantic UI' lines, using an approach similar to the Haystack project [☞HAYSTACK]

We decided that this use case was not in scope for the demonstrator, but it is listed as a ☞possible extension.

Large Group Aggregation - We discussed the need for the aggregation of blog data from a large group. For example: community annotations may require an annotation server; shared access to community

ontologies becomes problematic; community editing of ontologies may require a more sophisticated approach. In addition, even assuming the stable name problem has been solved, how do we discover blog items 'about' XYZ in a large (potentially worldwide) community. This is a peer-to-peer query issue. Finally, scaling up presents challenges other than the purely technical; the need for different navigation metaphors to avoid cognitive overload has already been mentioned.

Although this need was not built up into a separate use case, it formed the basis for a ☞possible extension.

Knowledge Management - This use case is particularly intended to explore the annotation on partial content (e.g. a comment on "section 2" of an article). Another possibility is to allow annotations on items such as pictures, audio files and video clips.

We decided that this use case was not in scope for the demonstrator, but it is listed as a ☞possible extension.

Document Management - This refers to the need to 'get at' the content underlying each blog item. For example, access to underlying documents, rich media files and potentially even services. This ability takes the blogging demonstrator beyond an annotation exchange mechanism to a content management system.

We decided that this use case was not in scope for the demonstrator, but it is listed as a ☞possible extension.

Privacy and Reputation - Blog entries could be marked with various levels of privacy (personal. community, public) or more flexibly we could implement a role based access control mechanism. The flip side of this coin is the use of reputation to augment community queries.

We decided that this use case was not in scope for the demonstrator, but it is listed as a ☞possible extension.

Assisted Markup - We discussed a variety of possibilities for assisted markup - for example: visual markup via drag and drop; automatic suggestions where similar objects/data already exists; automated classification of incoming blogs against an existing channel hierarchy; other enrichment of imported data. Another possibility is to 'cluster' blog entries based on link structure.

We decided that we would provide basic assistance in the core demonstrator. Richer possibilities are listed as a ☞possible extension.

Use Case Coverage - We present a summary table which shows how the considered use cases map to the functionality required. As we can see, a combination of the core demonstrator and the three optional extensions offer good coverage of all capabilities.

Use Case	RICH QUERY			RICH STRUCTURE	
	Subscription	Discovery	Navigation	Shared Ontologies	Annotation
Community Bibliography Management	Yes	Yes	Yes	Yes	Yes
Core Demonstrator	Yes	Yes	Some	Some	Some
Extension 1: Shared Ontology)				Yes	
Extension 2: Semantic Linking (aka Rich Navigation)			Yes		
Extension 3: Name by Property (aka Semantic Blogging)					Yes
Rich Discovery		Yes		Some	
Visualisation			Yes		
Large Group Aggregation			Some	Yes	Some
Knowledge Management					Yes
Document Management					
Privacy and Reputation					
Assisted Markup					

B Related Work

There is a large body of work related to this demonstrator. It is useful to consider this work within the framework that we have outlined ☞above. So we will start by reviewing current approaches to providing what we have identified as the key capabilities of blogging, including lightweight publishing and community formation. We shall then look at tools which use rich structure for annotation and ontology sharing, or rich query for subscription, discovery and navigation. There have been a number of developments in the blogging community recently, which move the metaphor in the semantic blogging direction, and which we review. Finally, we briefly note some web based methodologies relevant to bibliography management.

Lightweight publishing - One of the attractions about blogging is that it offers a low barrier to publishing content, and hence disseminating information. It is not the only mechanism with this characteristic, and it is instructive to survey other methods and comment on their strengths and weaknesses.

One particularly interesting area is that of collaboratively authored websites. Although many products conforming to the IETF standard ☞[WEBDAV] exist, a simpler, and perhaps more germane, example is provided by Wiki ☞[WIKI]. An extremely simple interface, using some basic formatting conventions (for example, 4 dashes for a horizontal break), allows people to add new content, or change (even delete) existing content. The intended result is that the presented information is a collaborative, emergent, adaptive reflection of a community's thoughts on a subject. There are various communities using the Wiki software, both general and technical. In addition, there are links between the Wiki community and blogging community (for example, a Wiki can be syndicated using RSS). The paradigm works well but there are some disadvantages. For example: it can be difficult to navigate through a Wiki to find the information of interest; information can be accidentally (or maliciously) deleted; the simple editing facilities mean that the presentation is rather monotonous.

Nevertheless, the collaborative, community nature of Wiki is one that we can learn from.

Another approach to low effort publishing is the use of scripting languages to provide authors (and possibly readers) with the ability to change a site's content (e.g. add/remove links).

This facility works well, although the editor is highly constrained in what s/he can do. With sufficient relaxation of these constraints, the paradigm offers similar capabilities to blogging.

Blogging itself of course offers a number of ways to publish information. Usually this is done through a simple web form, but even more lightweight solutions exists, the ultimate probably being to email blog entries. Of course, the email metaphor can also be used by a consumer, and a recent development (see ☞http://www.pipetree.com/qmacro/2003/01/29#nntp) shows how blog entries can be aggregated into virtual newsgroups for browsing through an NNTP client.

Online Communities - The semantic blogging approach encompasses a rich community metaphor. It is important to have some appreciation of how these communities are created, structured, and navigated, as such an appreciation enables us to better support them. For a comprehensive review of online communities see Werry and Mowbray 2000 [☞ONLINE-COMMUNITIES] .

Creation: There are a number of approaches to online community formation. Newsgroups and mailing lists create topic-focused communities, which are a valuable resource, but quite different paradigm to the less constrained, more dynamic blogging communities. Blogging, particularly semantic blogging, also provides a potential escape from the 'one conversation at a time' nature of such communities and allows users to discover, join or leave conversations according to their current interest. Portals offer another form of online community. These can be extremely valuable, but offer an impersonal, centralized organization which may be counterintuitive to an individual user. Online journals have a community of readers, and have the advantage that the information is collated, filtered, summarised and annotated for the benefit of its readers. It is certainly true that a collection of blogs will never have the tight cohesion of a journal, but the semantic blogging metaphor will empower users to tap into a much wider community of annotators and commentators, at a level and in a way that suits them. Personal links within web documents, such as homepages, can also give rise to communities. Web conferencing (eg ☞http://thinkofit.com/webconf/) certainly gives rise to online communities, but in a way only tangentially related to blogging, where the links are more decentralized and serendipitous.

Within the blogging world, communities are formed in a number of ways. The variety of link mechanisms (central aggregators, subscription, blogrolling and informal methods such as email) allow a community to be built up, the nature of which is arguably more flexible and robust than other online communities ☞[BLOG_COMMUNITIES].

Structure: The canonical community structure study is due to Milgram

[☞SMALL-WORLD-MILGRAM], who carried out a real world real world experiment of giving people letters for an unknown destination. The letter had to be passed to somebody who was personally known to the current holder. This experiment suggested a median chain length of 6 (hence "six degrees of separation"), a result which has not always been successfully reproduced. Nevertheless, Tjaden [☞ORACLE-OF-BACON] (who coined the phrase "The Kevin Bacon effect") did show the phenomenon in the movie domain (using IMDB data, where the highest finite Bacon number was 8) and a current experiment (using email) is being conducted by Duncan Watts [☞SMALL-WORLD-WATTS]. Watts also conducted a number of experiments on real world networks (actors, power grid, neural circuit) where he looked at 'small world' parameters. He found that all these networks exhibited typical small world characteristics - namely a large average path length and yet a high degree of clustering. The implications we can take from these studies is that our semantic blogging communities may look quite different (depending on the domain, the people, the content and other factors) yet they may have certain common features (such as small world characteristics). It is instructive, then, to examine the structure of different online communities. A study of communities based on home page analysis of students at Stanford & Yale ☞[FRIENDS] showed each with a slightly different, though clearly structured, nature.
Mazzocchi used Apache's Agora product to map email networks [☞APACHE-AGORA]. Blog networks have been mapped by Ross Mayfield [☞BLOG-TRIBE]. All of these studies are interesting and relevant, but, to the authors' knowledge, there has been to date no quantitative analysis on blogging community structure.

Web content itself can also be clustered (for example web pages with similar content), which has implications for the way that blogged data might be grouped (and associated implications for how best to navigate that data). But we are more interested with the interlinked nature of the blogs. Such links have been used to identify communities [☞SELF-ORGANISE] where each page is linked to more pages within the community than without. Menczer [☞TOPIC-DRIVEN-CRAWLERS] uses linkage patterns as clues to guide a crawler (and to supplement other metrics such as textual or semantic 'distance'). Other techniques include PageRank ("importance") as used in Google, HITS (detection of Hubs), subgraph identification and spreading activation.

Navigation: Hyperlinked data has special properties, as pointed out by Victor [☞HYPERTEXT-THEORY], who suggests that hypertext can be mapped onto an expert's semantic network or knowledge structure. Hence the user can be guided but also taught. Jacobsen [☞HYPERTEXT-LEARNING] also finds that hyperlinks encourage knowledge transfer, and recommends that users should be allowed to modify the navigated network Of course, the concept (but not the term) of a hyperlink is due to Vannevar Bush's insightful 1945 paper "As we may Think" [☞AS-WE-MAY-THINK]. Items can thus be linked to create a new "trail", and Bush boldly suggests that trail creation may itself be a form of authoring, creating new structures which can themselves be shared.

The paper "As we should have thought" [☞AS-WE-SHOULD-HAVE-THOUGHT] extends Bush's metaphor by introducing the term *structural computing*. The essential idea here is that the network of content forms a rich structure, and this (rather than the data) should form the primary design driver for support tools.

The nature of this structure is examined more closely by Dalgaard [☞SCHOLARLY-ARCHIVE], who uses the terms intertexts (eg citations), paratexts (eg introductions) and metatexts (eg reviews & annotations). According to Dalgaard, we navigate in a scholarly archive using these forms of 2nd order textuality. Indeed journals can be seen as a metatextual overlay, and often different overlays (or gateways) onto the same concepts are appropriate. The implications of these studies are threefold. Firstly, semantic blogging communities will need support tools that explicitly support navigation through semantic networks. Secondly, different types of metadata 'overlays' will be used by different people and at different times, depending on the user's need. Thirdly, the users need to be empowered to augment or enrich the content (annotation) rather than being passive consumers (of portals).

Shum and colleagues, from the KMI institute at the Open University, take some ideas from structural computing to form the basis for (what will become) ClaiMaker [☞CLAIMAKER-SCHOLARLY]. In particular, the importance of detecting (contrasting) perspectives, the need for varying levels of granularity, the acceptance (even encouragement!) of conflicting statements, and query ("what if I pose this hypothesis? What's the evidence for/against?")

. ClaiMaker has considerable similarities with our semantic blogging demonstrator, and provides the inspiration behind our ☞semantic linking extension.

For bibliographies, *citation indexing* is an old idea [☞CITATION-INDEXING]. There are well

known implementations of it, such as the Web of Science [☞ WEB-OF-SCIENCE] and CiteSeer [☞CITESEER]. The important thing here is that we can navigate in two directions, both from a paper to its citations and from a paper to the papers that cite it (these of course imply navigation forwards and backwards in time).

Thus within the context of the web, citation indexing can be though of as bi-directional hyperlinks.

Annotation - There are a variety of approaches to community annotation. Essentially these approaches comprise a set of tools for marking up content and sharing this markup. The most pertinent example is probably Annotea [☞ANNOTEA] which is a semantic web approach to marking up web pages. Users are allowed to make comments on arbitrary web pages using an *annotation server,* which manages the annotations (annotation servers can also be found in other domains, such as bioinformatics [☞DISTRIBUTED-ANNOTATION]). XPointer expressions [☞XPOINTER] are used to allow annotation on a particular section of the document. Annotations have metadata associated with them - for example authors/dates, annotation type (explanation, comment etc). Some relationships between annotations (ie various types of reply) are also allowed to support threaded discussions.

Another example of an annotations editor is CREAM [☞CREAM], from Steffen Staab. The difference here is that annotations are intended to be made while authoring the web page. In accordance with RDF principles, URIs are used as values - these URIs are harvested (using an RDF crawler) from the web at large, thus encouraging consistency across communities. Other annotation tools are listed at [☞ANNOTATION_TOOLS].

Once annotations are captured, users need some ability to discover, query and navigate these annotations. Annotea provides the *Algae* query language, navigation is the responsibility of client applications such as Amaya [☞AMAYA]. An obvious approach to discovery is the use of concentrators, a role fulfilled by annotation servers and an approach adopted in this demonstrator.

Another useful capability is automatic markup, such as that provided by COHSE [☞COHSE], a tool created at the University of Manchester. This is a vocabulary tool, where terms in the text are mapped to some ontology (essentially a concepts hierarchy). From this, resources (eg shops) related to the concept (ie word) can be retrieved. Klarity [☞KLARITY] also offers automated markup using Dublin Core tags.

Useful markup requires a well thought out ontology for such markup. There has been considerable activity in the bibliographic domain around markup standards, this is reviewed in a ☞separate appendix. For annotations and semantic links, there are a number of standards, including Annotea Threads [☞ANNOTEA-THREAD], IBIS [☞IBIS-TERMS] and ClaiMaker [☞ CLAIMAKER-SCHEMA].

For topic hierarchies, there are two considerations. Firstly, there is the basic structure of the hierarchy (various uses of subclass, narrows, subtopic etc, all of which may have different semantics). The RSS2.0 [☞RSS20] standard has a method (introduced in RSS 0.92) to classify a blog item according to some external ontology. For example, in order to classify an item into the DMOZ [☞DMOZ] hierarchy, one might add the metadata:

```
<category domain="http://www.dmoz.org">
Business/Industries/Publishing/Publishers/Nonfiction/</category>
```

An alternative approach is used by XFML [☞XFML], which specifies the structure needed to both define such a hierarchy and to link to other defined hierarchies. Such defined hierarchies include DMOZ (as mentioned), Yahoo! and so on. For topic hierarchies in the bibliographic domain, see the ☞appendix on bibliographic standards.

Ontology Sharing - Community ontology sharing has been approached by a number of groups, tackling a variety of issues associated with this idea.

One approach is to simply adopt and share a previously defined, centralised ontology. Within the web community, DMOZ [☞DMOZ] is popular as a topic ontology. This is a system where users can enter web pages under a rich topic hierarchy, and the hierarchy itself is under user control. Other centralised ontologies more relevant to the bibliographic domain are reviewed in a ☞separate appendix.

At the other end of the scale is approaches that allow a community to collectively define, in an gradual manner, a suitable ontology. Kaon [☞KAON] and APECKS [☞APECKS] are both large scale engineering projects concerned with community management of a shared ontology. The take a variety of approaches to solve the issue of shared ontology management. These systems offer rich functionality, but also illustrate that it is not really feasible to tackle this issue head on in a short project.

We are interested in more restricted paradigms, where a community has a rather looser collaboration, either on a smaller ontology or with each person making fairly minor changes to a localised portion of the ontology. The key intuition here is that the community ontology sharing should make mark up easier. One such approach is the use of *archivelets* [☞KEPLER], which is part of the Open Archives Initiative (OAI;

☞http://www.openarchives.org/). This is an approach to harvesting publication data from multiple, distributed providers, all of whom subscribe to a certain basic metadata provision (eg Dublin Core tags). The approach is interesting because they recognise the need for low cost publishing (the archivelet is a simple publishing tool that offers assisted markup) and for diverse repositories, all of which may have different organisational strategies (over and above the core OAI requirements). However, there are differences. In particular, our ideas about community annotation and emergent ontologies do not seem central to KEPLER's *modus operandi*.

An intriguing possibility is the use of the blogging infrastructure to build up community topics. The internet topic exchange [☞ TOPIC-EXCHANGE] offers a facility (powered by TrackBack [☞MT-TRACKBACK]) which enables a user to nominate a topic and for others to link their blog entries to that topic. In effect, this is building an emergent, community based ontology, though with only a basic structure, and no topic hierarchy. An orthogonal approach is that of XFML [☞XFML], which offers a low cost way to define and link taxonomies. Topic maps [☞TOPIC-MAPS] offer a similar (though richer) approach, for which an open source Java toolkit [☞TM4J] is available. Finally, classification schemes can be combined as separate 'facets', as advocated by the Facet Map concept [☞FACET-MAPS]. Some bloggers have used this approach to provide richer navigation for their blog [☞THIS-IS-XFML].

Rich Query - We aim to enrich the current blog mechanisms for discovering new content. Our assertions is that adding semantic capabilities to discovery will provide the user with a much more powerful and useful mechanism. Indeed, such an effect has been found in other domains. For example, Kalfoglou et al [☞MY-PLANET] describes a web agent in which simple profiles are used to select news articles of interest. Such a mechanism is partially available already, by performing keyword based searches on RSS aggregators and returning the result as an RSS feed. The use of an ontology augments such simple keyword searching. Even simple metadata based searches (author, title etc) can be surprisingly powerful. Inference based searching (for example, articles on projects related to a particular research area) offers correspondingly greater flexibility. Ontology-based document structure matches can be specified as templates (for example, "X visited Y" where X and Y are typed nodes). Recent work in such ontology based agents has shown promising results [☞FINDUR, ☞ONTOSEEK]..

The semantic blogging paradigm retains some of the attractive features of other discovery mechanisms. The informal recommendation nature of email can be supported through targeted RSS feeds from peers. The more definitive nature of portals can be captured in part by a composition of blogs from domain experts. The timeliness of bulletins can be mirrored by community feeds.

An important area is that of information visualization. In order to navigate a semantic network, the user will need visual and intuitive tools in order to make sense of the structure. There are significant challenges here, not all of which are directly related to the semantic web. For example, it is not obvious how one should present a network so as to reduce overlap and improve clarity. Various algorithms exist which tackle this problem, including one by Noel et al ☞[VIS_ALGORITHM] and one by Mazzocchi which is used in Apache's Agora product [☞APACHE-AGORA]. Such techniques are directly applicable to visualization of blog items.

There are a variety of software products that offer visualization tools for concept mapping. For example, customised shapes in Visio [☞http://www.microsoft.com/office/visio/default.asp] allow support for mind mapping (including bibliographic management), peer mapping and process mapping. Atlas [☞http://www.atlasti.de/] allows the user to associate selected text with a concept.. Nudist (now NVivo) [☞http://www.qsr.com.au/products/nvivo.html] has hierarchical concepts. MindManager [☞http://www.mindjet.com/] and TheBrain [☞http://www.thebrain.com] are concept mapping tools. Within the bibliographic domain, the ClaiMaker project has an online demonstrator showing their approach to concept navigation [☞CLAIMAKER-SANDPIT]. Finally, Ideagraph [☞IDEAGRAPH] is a recent (and open source) semantic web approach to concept mapping, which enables import from a number of sources, including web logs.

Current 'Semantic Blogging' Developments - Where appropriate, we want to build on and re-use tools, frameworks and ideas from the blogging community. Blogging tools are constantly being extended, and some of these extensions are related to this project.

An essential first step towards semantic blogging is the incorporation of richer metadata. In fact, the blogging community has already proposed some extensions to the core metadata format. The Weblog metadata initiative [☞WMDI] suggests various tags (predominantly Dublin Core based) to mark up blogs and blog entries. The Ol' Daily weblog [☞OLDAILY] has a 'Research' facility, where readers may search for other articles by the same author, belonging to the same category or published in the same journal. RSS 2.0 [☞RSS20] has a notion of category taxonomies, while XFML [☞XFML] is an XML format for defining topic hierarchies. Matt Biddulph has written a set of utilities [☞MT-RDF] that will help in converting MovableType blog entries to, and enriching them with, RDF.

MovableType pioneered the TrackBack mechanism [☞MT-TRACKBACK], which enables citation links to be recorded. What it does is to send a 'ping' (with summary details and a URL) whenever you comment on somebody else's blog entry. The functionality is also available as a standalone perl or

Python module, or by using XML-RPC [☞TB-STANDALONE]. Using TrackBack functionality, many blogs provide a list of trackbacks (or citations) for each blog item. Further, trackback can be enabled for an entire category, so that one can associate a blog entry with that category. Over time, the category becomes a collection of items pertinent to that community topic. A good example of this sort of emergent ontology formation can be seen at the Internet Topic Exchange [☞ TOPIC-EXCHANGE].

There are a variety of recent extensions to Trackback [☞TB-EXTENSIONS]. In order to explain this, it is helpful to think of *citing* articles, which send trackback pings to *cited* articles. *PostIt* allows full details of the citing article to be sent back with a TrackBack ping. This means that a reader of the cited article does not have to follow the TrackBack link and thus can read the citing article in its entirety 'in place'. *ComeBack* is a related utility that allows readers of a cited article to post comments about a citing article. Again, they do not have to leave the aggregator site to do this, and the comments are automatically forwarded on to the citing blog. A similar feature can be implemented in RSS2.0 using the comments tag [☞http://backend.userland.com/rss#comments].

Other extensions [☞TB-EXTENSIONS] allow access to the extended network of peer commentary. Trackback threading enables a threaded display of a TrackBack discussion. *BackTrack* is where the cited blog sends the citing blog a list of all other citing articles so that these can be listed on the citing blog. *MoreLikeThisFromOthers* is an ability for the cited blog to search through the citing blog for related articles. It does this by searching through the citing blog's RSS feed for blog items in the same category as the citing article.

Web Based Bibliography Management - In this section we review some of the ideas, approaches and systems relevant to web based bibliography management. A separate appendix details ☞concrete tools and applications in this domain.

The primary source of bibliographic references on the web for many people is CiteSeer [☞CITESEER]. The CiteSeer vision has been developed in a series of papers [☞CITESEER_ARTICLES]. In the original paper, the authors outlined their plan to:

- **Acquire:** (initially Postscript only) papers (using search engines and simple heuristics, eg "+publication")
- **Parse/extract:** features, eg header, abstract, citations, word frequencies from such papers
- **Browse/query**: navigate through citations, group documents by similarity, retrieve further details on demand (e.g. BibTeX). Document similarity may be judged through content (TFIDF, string matching) or semantic feature matching (e.g. co-citations).

CiteSeer has since been updated by adding a user profile, which can specify interesting features (keyword based, "documents like this", "documents from this URL"). There is also a notion of citation *contexts.*

There are a number of other systems that provide similar functionality to CiteSeer. In the commercial world, Web Of Science [☞WEB-OF-SCIENCE] provides a subscription-based access to a large database, with citation linking and rich searching. In the academic world, several prototype systems are in development. Bibliography on the web [☞http://www.cs.huji.ac.il/~derel/Projects/bow/] allows users to rate, index, annotate and link papers- so called *RIAL* publishing . In particular, the user is assisted (by machine learning techniques) to classify the document into a deep, rich hierarchy. The papers can be linked to each other (although there is no notion of semantic, i.e. different types of, link). The current prototype looks promising, although there is not a large amount of data on it. Other relevant projects include the hypertext BIB project [☞http://theory.lcs.mit.edu/~dmjones/hbp/] and WEBBIB [☞http://www.cs.wpi.edu/~webbib/] which perform a similar, though more limited, service to CiteSeer.

The usefulness of a hyperlinked archive is underlined by the fact that in some domains, people are prepared to put in a great deal of work to produce such an asset. Metzl [☞CATALOG-OF-CATALOGS] describes a manually created, hyperlinked archive for law documents. Classification (into subject taxonomies) is another useful feature of many archives. Classification-based navigation is provided by online portals like LANL (the ePrint archive at ☞http://xxx.lanl.gov/) and NCSTRL (☞http://www.ncstrl.org/). The Computing Research Repository (CORR; ☞http://arxiv.org/archive/cs/intro.html) is a node on both the LANL and NCSTRL networks. However, in both cases, the user navigates by simple queries (eg title and author); intertextual navigation (i.e. navigation by citations) is not supported.

Finally, personal bibliographic management tools (e.g. ProCite, EndNote and others in ☞review) provide certain integration with web based search. These tools are popular (or, more accurately, widely used) within the technical community. However they suffer a number of limitations, as explained in the original framing document and explored in the ☞user study. An extensive review of such tools is provided in a ☞separate appendix.

A project quite closely related to our work is ClaiMaker [☞CLAIMAKER], whose original name was ScholOnto. In this original framing [☞CLAIMAKER-SCHOLARLY], the authors were interested in building an ontology based digital library server, which supported scholarly discourse. Hence, the idea was to build up networks of contestable claims. Enrico Motta's OCML language [Operational Conceptual Modelling Language; ☞http://kmi.open.ac.uk/projects/ocml/] is used as the schema specification

language; this is a standard logic notation, which supports subclass/property inferencing. A key idea is the emergence of perspectives (documents supporting ideas of a certain type). The key issue is the management of complexity, and the finding of the appropriate concepts to mark up. There are two tools to support this activity - a tool for annotation of a new document, and a tool for browsing concepts & relationships (ie schema and instances).

ClaiMaker's work is interesting, and the online demonstration [☞CLAIMAKER-SANDPIT] is impressive. The visualisation is intuitive yet powerful, although with increasing complexity a user might be aided by alternative visualisation techniques. Currently, there seems to be no way to provenance claims (ie Paper X supports idea Y) although later instantiations seem to support formulations like "Person X stated...". The ClaiMaker team have clearly tried to balance richness with complexity. As a result, the schema [☞CLAIMAKER-SCHEMA] is not that complex, and it should not take long for a user to become familiar with it. However, the instances are complex, and it is possible that users will struggle to locate the right concepts to use, with the attendant problems of duplication and redundancy. In addition although the schema seems reasonable, it may not match everyone's ideas about how scholarly discourse should happen. Although the schema could be simply extended for use in other domains [☞CLAIMAKER-WEAVE], this solution will not necessarily suit all eventualities. But the problem in the large is a general ontology mapping problem for which one should not perhaps expect ClaiMaker to have a general answer.

C User Study:

Background and Aims:

This small and informal study was part of the background research activity for the development of the document '☞ SWAD-Europe: Semantic Blogging and Bibliographies - Requirements Specification'.

The aim of the study was to help gain preliminary insight and overview of the use of bibliographic data and associated software and/or management of bibliographic data by individuals and small groups. The study was conducted in parallel with two other activities; a short literature review (mostly Web-based) and evaluation of existing bibliographic software systems (see ☞Appendix D). The interview data thus augmented and clarified (or not) our findings from other activities and more broadly provided a 'test our intuitions' with regard to many aspects of potential design requirements.

Methodology and Analysis

The interviews were confidential, informal and semi-structured, lasting between 20mins and one hour. The participants were asked four questions of the form:

1. how have you or do you use bibliographic data?
2. how did/do you capture, manage the data and 'publish' the data e.g. create bibliographies?
3. if you used software, what software was it and a) what did you like about it? and b) what could have been improved?
4. do you have any 'wishlist' functions or capabilities that an ideal bibliographic system should have?

The participants largely directed the discussion, focusing on the particular aspects of the questions that were of interest to them and to which their experiences were relevant. Notes were taken by the researcher and collated, under three broad headings; 'data capture', 'management, manipulation and sharing' and 'publishing', a fourth category 'wish list' was used to capture the 'wish list' items detailed by participants. These were then used in conjunction with findings from the literature and software review, to produce composite findings, see ☞Appendix D.

Participants

The five participants were recruited via internal e-mail and personal contact. Thus they included academic researchers, a university lecturer and software/systems engineers. All of the participants were highly computer literate, and use computer software, systems and on-line services to support their work on a day to day basis.

The very small size of this sample, the relatively narrow range of contexts and high levels of technology experience along with the informal nature of the methodology, mean that the ***findings cannot and should not be generalised*** to the wider community of users. However the aim of the study is to gain initial insight/overview and augment data from the literature and systems review, the findings should thus be seen and used in that context. Further interviews and systems testing with a wider range of users is planned as part of the design and implementation phases.

Findings

This section provides a **brief summary** of the key findings of the interview study - under headings of the questions. ☞Appendix D (Review of Personal Bibliographic Systems) contains a more extensive and integrated set of findings from the three strands of research i.e. literature review, bibliographic software review and this interview study.

Reasons for Creating and Using Personal and Small Group Bibliographies:

By far the major use of the software was to create specific bibliographies for specific papers or other publications. Only one of the participants used the software to hold (the majority) of their bibliographic data on a day to day basis. One participant was using the software as part of a bigger project to collate data from a large number of people.

Capture, Management, Sharing and Publication of the data:

Practice varied greatly across the participants. One participant captured and kept track of their data using a paper/cardfile based system, another originally used an Access database and now uses a simple Microsoft Word file, others via EndNote and Reference Manager.

Where participants (now or in the past) had pressing reason(s) to keep close track of references and the associated publications e.g. active research projects or studying for a Ph.D. they had developed highly sophisticated processes, part of which utilizing software (i.e. Access, ProCite and EndNote), but it was clear in at least three cases that the software was only a part of a larger process, including filing systems, vocabularies, selection processes, continually capturing and managing up dated information, etc.

However where participants did not have pressing reason(s) to keep close track of references e.g. simply keeping up to date with interesting papers for general professional interest, this was not the case and maintenance of their databases and filing systems tended to slip and in one case reduce in complexity to a simple formatted Word file.

There were many specific details about the actual **locating and capture of the data,** see ☞Appendix D for more detail. The **primary common issues** with software are that:

- it is time consuming to enter data (e.g. entering the data, even into a forms based interface takes too long, especially with the need to format the entries 'correctly' e.g. personal names)
- it is difficult to ensure accuracy and consistency of terms (e.g. spelling and format of author names)
- it is often difficult to identify all necessary data from or about a resource.

The **sources of data** were varied inc. the physical papers and books themselves, electronic papers/journals (e.g. PDF files, HTML), on-line/CD databases inc. 'abstracting' and 'table of contents' services, online library systems, references from colleagues, references taken from other publications, news items, e-mail lists, etc.

With respect to **publishing the data**, the participants all use any software primarily to produce bibliographies for particular 'projects' and publications. Where they had used commercial software for this, they were in general happy with the flexibility that the system that they had used. The bespoke Access and Word solutions were not so effective in this regard.

When **sharing or moving data** from one format to another, the participants were less satisfied with the commercial systems, while the commercial systems had built in flexibility for the output for bibliographies, they did not when exporting to other bibliographic management data format e.g. BibTeX (see ☞Appendix D) or other competing products. This was found to be frustrating. One participant had to create custom scripts (i.e. computer programs) to perform a conversion themselves.

In general the users of the commercial **software** were happy with the data **maintenance/management** facilities once the data had been entered. The interfaces generally allowed for effective retrieval and updating of records. The primary difficulty identified was in **creating and maintaining indexing terms**, keyword facilities were helpful but in one case the participant found it frustrating that they had to look up or remember the available terms.

Use of Software:

All the participants had used software to capture, manage and publish bibliographic data. We were lucky with the sample that a variety of software had been used. These included commercial products; ProCite, EndNote, Reference Manager (see ☞Appendix D for detail of these systems) and generic applications adapted to that purpose; Microsoft Access and Word.

There were a number of generic issues (these were raised in the previous section) and many specific issues about individual packages (e.g. one participant noted that integration with a word processor was problematic because the package was very processor hungry when updating the data links; another noted that backing up was very problematic in another). See ☞Appendix D for more specific detail, and integrated findings with the literature and software review data.

Wishlist:

Participants were asked if they had any 'wishlist' functions or capabilities that an ideal bibliographic system should have. Below is a list ordered under rough headings.

Capture:

- Auto extraction of data from text citations and references in papers, for example the Windows clip board, including BibTeX format.
- Auto capture of bibliographic data from a paper e.g. where the papers were from a common journal and thus had a common format - import filter for whole papers.
- Help in 'getting into' a subject area e.g. the common problem of knowing what to ask and what vocabulary to use in a new field. Possibly have a system to cross search other peoples' data and say that 'people who have indexed this paper have used the following terms' or 'expert y has used the following terms to index this paper'.
- Links to external databases to confirm spellings of names, journal titles etc...
- Import data from 'table of contents' services data inc. abstracts.
- Copy some/all details from one record to another e.g. when adding lots of papers by the same author of same journal
- Capture references at the end of an electronic publication
- Auto or semi-auto detection of document type (e.g. book, journal...)
- Notification of additions to selected internal and external databases

Management & Augmentation:

- More effective classification tools inc. more structured vocabularies
- Access and update [the] data online, from any machine via a browser
- Hierarchical keyword structures, "so that they are easier to navigate and conceptualise"
- Richer annotation possibilities other than 'notes' fields and better searching of them.
- Semi-automated indexing or keywording from the text of an on-line document
- Visualization of aspects of data e.g. timelines for papers under a given keyword, citation relationships, or identification of 'camps' of researchers with particular points of view...
- More flexible merging of databases.
- Semi-automatic updating of documents that have used a record [e.g. in a bibliography] when the record changes - i.e. the system should remember when and which files have used the individual records and notify the user when changes are made
- More effective backup

Sharing:

- More extensive export facilities
- See or be notified, when a record was added or edited and by who
- Ability to select records and e-mail selection from within system
- Ability to search other people's bibliographic databases e.g. known expert/specialist or researcher in field, in particular found via publications e.g. person x published paper on y, therefore they might have a bibliographic database.

Publication:

- Output to any bibliographic software
- Ability to integrate (export to and import from) and output to other applications e.g. graphics, mind mapping, database or visualization program, e.g. export data to a mind mapping program so that main branches were authors, next level publication year and then title
- Easy selection of records for inclusion in a bibliography e.g. tick boxes next to record on screen, multiple select (using ctrl, left mouse button), shopping cart type approach
- XML output and input

While these are not necessarily extensive or generalisable to the wider community of users and potential users of such systems, they do give some insight into the perceived needs of users and the significant potential for improvements in design, and functionality of the existing software and services.

Conclusions

This small scale and informal study has provided some initial insight into the use of bibliographic data management systems and software. The ranges of experience and needs, even in this small and fairly narrow sample was high, however there are clearly common needs and issues.

The findings will be used with caution, due to the small scale and informal nature of the study. However they provide significant insight and when combined with data from the literature and software review and software, they give a broad and importantly, cross-referenced assessment, of how existing systems are used and what users might want from the prototype developed as part of the present project.

D Review of Personal Bibliographic Systems

Introduction

The personal and small group management and publication of bibliographic information is a ubiquitous problem for academics, researchers, students, writers and more broadly authors of many kinds. Many systems have been developed by individuals, groups and organizations to meet their own specific requirements, some these utilise software, that supports the capture, management, sharing and publishing bibliographic information.

This appendix aims to provide a brief overview of key issues in the context of the creation of bibliographies, these are:

- the kind of motivations behind the personal and small group use of bibliographic management systems - be they paper or electronic or a hybrid
- the functionality required by users and offered by the various computer based systems
- the software systems available
- relevant standards

As part of this work we reviewed 20 (see below) bibliographic software products and services, reviewed on-line literature and conducted a small interview study of users of such systems (see ☞Appendix C). The interview study consisted of 5 interviews, with individuals from academic and commercial backgrounds, they were asked about their use of bibliographic data, systems including software that they have used to capture, manage and publish the data, along with any positive or negative feedback about the systems. Finally they were asked about 'wish list' items, that they would include in any ideal system.

Reasons for Creating and Using Personal and Small Group Bibliographies

> ***bibliography***: *"a list of the books and articles that have been used by someone when writing a particular book or article" Cambridge Dictionaries Online (☞http://dictionary.cambridge.org/)*

The definition above indicates that the term *bibliography* relates to a particular list of books or articles. However in broader definition might cover any type of 'work' or 'resource' and may simply be a list of works held for any purpose.

There are a large number of specific reasons for capturing, managing, sharing and publishing bibliographic data - for example: creation of bibliographies; as part of a personal knowledge management toolset; organizing a collection of books or papers; group annotation or indexing of resources; locating resources previously used. Although no specific research in this area has been identified as part of this study, it seems likely (based on the functionality of software systems and small scale interview survey), that the dominant uses are to produce bibliographies related to specific topics. Specific examples of the need to create a bibliography include: student essays, grant applications, curricula vitae, book chapters, project reports, sharing references with colleagues, and reading lists. In general these lists must be formatted in specific manner depending on the standards of organization and/or publication.

Needs, Functionality and Affordances

In order for a user or a group of users to do any of the things listed in the previous section, it is necessary for any system to facilitate a set of core activities, these can be broken down in a number of ways, the following is a fairly generic set:

1. **assisted capture** of necessary bibliographic (and other e.g. notes) data
2. **management and manipulation** of the data once captured, for example: editing of records; locating of previously entered data; searching for subsets of data for the creation of a bibliography; ability to categorise and annotate the records.
3. **publishing and sharing** of the data in an appropriate bibliographic format.

There will be specific needs related to each of these three categories in any particular context. However if we focus on the examples related to the creation of bibliographies, these include:

Assisted Capture

Users are likely to obtain data from a range of sources - extracting the necessary bibliographic data from these can be problematic. For example: it is time consuming; it is difficult to ensure accuracy and consistency of terms (e.g. spelling and format of author names); it is often difficult to identify all necessary data from or about a resource. Thus functionality such as the ability to automate or

semi-automate the identification and capture of data are helpful.

Management and Manipulation

Once the data is captured it generally needs to be managed and sometimes augmented. The requirements with respect to the creation of bibliographies are less extensive than other potential applications (e.g. knowledge management). However there are a great many functional requirements even in the relatively simple context. Some specific examples of tasks identified during our interview (see ☞Appendix C) survey include:

- *users may wish to add extra keywords - requiring tools to assist in consistent key-wording e.g. retrieve all records indexed with a particular keyword and index/sub-divide them using finer grain terms*
- *editing records to ensure consistency (e.g. UK or US spellings)*
- *adding bibliographic data as it becomes available*
- *augmenting notes and annotations about a work following further reading.*
- *the desire to understand relationships between works, e.g. by drawing timelines or mindmaps from the data*

Fundamental to these and others is the ability to retrieve records efficiently and effectively. Coupled with this is the ability to annotate, organise and in some cases visualise the data.

Publishing and Sharing

'Publishing' of the bibliographic data takes a number of forms. In practice the most basic case is the creation of a simple formatted text list (bibliography) of selected records drawn from the system. The format/style (see below) and media (e.g. paper, word processor file, HTML) of the list, will depend on the particular context. More complex examples include, the automatic embedding of references and citations within a word processor document (cite while you write, type system, see below).

The sharing of bibliographic data with individual colleagues or within teams, is a common practice, e.g. passing on references to colleagues for their use in their report or small teams of researchers keeping a common word processor or bibliographic database file.

Sharing data is used for a number of purposes, those identified as part of our study (see ☞Appendix C) included:

- *prevent duplication of effort in finding references*
- *as part of authoring process e.g. academic papers*
- *flagging up newly found interesting documents to members of a team*
- *share thoughts about research papers or publications (e.g. via notes fields)*
- *ensure common format for data captured*
- *helping members of a team keep track of new developments/research in a research field*

Standards

There are a variety of aspects of bibliographic data capture, management and publishing systems that require standards to be set, if systems and data are to be interoperable. These include, the pieces of data to be captured, the naming of the database fields, the organization of the data, the rules of when and how to use a particular indexing [or keyword] terms, the syntax of the data within fields, and the storage formats. ☞Appendix E reviews some of the more extensive library related standards.

While the library standards are largely 'overkill' with respect to personal and small group management of data to produce a bibliography, many of the basic needs behind the requirements remain the same. From ☞Appendix E:

1. *what information should be captured about the 'publication' i.e. cataloguing data*
2. *the structure of the record*
3. *detailed rules or guidelines for how to deal with specific cataloguing situations/issues e.g. what to do when there are two different formats of an authors name used within one publication. Broadly these also include the use of authority lists/files as definitive authorities over for example the spelling of a place or personal name.*
4. *how the subject or content of the 'publication' should be described i.e. how the publication should be indexed.*
5. *the specific syntax (e.g. use of punctuation) for encoding of the record*
6. *The specific transfer protocols for transferring the data between locations.*

In these areas, the formal library standards are well defined. However at a personal and small group level, they are largely poorly defined or there are many competing 'standards'. Examples include citation styles and interchange formats; these will be discussed in turn.

Citation Styles

There are a very large number of citation styles i.e. how citations are written to acknowledge works used or referred to in a document, and how the references should be formatted, e.g. one of the most commonly used is the Harvard style in which the: author and year of publication is written (i.e. cited) in the text (e.g. J. Smith (2002)) of a publication. In the case of a book, the reference contains the name(s) of the author(s), editor(s) or the institution responsible for writing the book, date of publication (in brackets), title and subtitle (if any) should be underlined or highlighted or in italics, but must be consistent throughout the bibliography, series and individual volume number (if any), edition (if not the first), place of publication (if known) and publisher. The references are place in alphabetical order of the family name of the main author. Other types of publication will have different data e.g. in the case of a journal the journal volume and number.

The other basic citation method is the 'numeric system' (Vancouver) style, in which references are cited by a number in the text which is then used to label and order the references in the bibliography.

There are generic, high level, international standards, for example:

- ISO International Standards ISO 690:1987 (Information and documentation, Bibliographic references Content, form and structure, see ☞http://www.nlc-bnc.ca/iso/tc46sc9/standard/690-1e.htm)
- ISO International Standards ISO 690-2: (Information and documentation, Bibliographic references Part 2: Electronic documents or parts thereof, see ☞http://www.nlc-bnc.ca/iso/tc46sc9/standard/690-2e.htm)
- British Standards Institution standards e.g. BS 1629: 1989. References to published materials. BS 5605: 1990. Recommendations for citing and referencing published material (see ☞http://bsonline.techindex.co.uk/)

However with regard to specifics of the syntax and format of citation and references traditionally different disciplines have tended to use different styles e.g.

- ***Humanities**: MLA (Modern Language Association) - MLA Handbook for Writers of Research Papers, 5th ed. MLA, New York, 1999) & The Chicago Manual of Style,*
- ***Scientific**: APA (American Psychological Association) - Publication Manual of the American Psychological Association, 5th ed. APA, Washington: 2001 & CBE (Council of Biology Editors) - Scientific Style and Format, 6th ed. Council of Biology Editors, 1994)*
- ***History**: Turabian - A Manual for Writers of Term Papers, Theses, and Dissertations 6th edition, Kate Turabian's, University of Chicago Press, Chicago: 1996. and Chicago (as above)*

In addition, many journals, publishers, governments, corporate bodies and other organizations define their own 'house style'.

In many ways these style rules are analogous to the International Standard Bibliographic Description, ISBD (☞http://www.ifla.org/VII/s13/pubs/isbd.htm) and AACR2 (Anglo American Cataloguing Rules ☞http://www.nlc-bnc.ca/jsc/) standards for larger systems, discussed in ☞Appendix E, in that they provide similar guidelines (e.g. data to be captured and order, syntax etc.), but for a fewer pieces of data (fields/elements).

Information Interchange Formats

There are no ubiquitously used standards to encode personal and small group bibliographic data i.e. there is no equivalent of MARC (see ☞Appendix E) in the library world. In general the commercial software packages specifically designed for individuals and small groups use proprietary formats for internal storage and in many cases export options are very restrictive e.g. EndNote (probably the dominant product for personal level data in UK Higher Education) only has txt, rtf and HTML formats as options under the 'export' menu.

In some cases it is possible to use the tools to output different styles to create output 'styles' that match any other text based format. e.g. in the case of EndNote (see table below) there is a fairly comprehensive style template language. Base level formats such as comma delimited text, can be imported as standard by nearly all of the packages reviewed.

In contrast, the majority of such systems can import a great many formats (including competitor's formats) and as in some cases advanced users can create their own 'import filters' (i.e. parsing rules to extract the necessary data from the source format). In general for example MARC records can be imported and the data parsed to extract only the required fields to match the internal data format, having done the necessary mapping between data standards.

The Dublin Core [☞DC] Metadata Initiative has a citation working group (☞http://dublincore.org/groups/citation/) which is working on refinements and encoding of bibliographic data. Dublin Core unqualified or qualified can provide a base level metadata standard. There are also various 'cross walks' (mappings) between formats e.g. MARC to Dublin Core (☞http://www.loc.gov/marc/marc2dc.html). However as yet Dublin Core has not become widely used as

the basis of a data exchange system for commercial systems reviewed.

Other metadata and encoding related initiatives include the TEI (Text Encoding Initiative - ☞http://www.tei-c.org/), EAD (Encoded Archival Description - ☞http://www.loc.gov/ead/) and DocBook (☞http://www.docbook.org/). These all have elements that deal with bibliographic data, however again as these are designed for different and more complex application areas they are not used in the systems reviewed here.

BibTeX [☞BIBTEX] is one of the more widely used (and older) formats that has been used and well supported with publishing tools. BibTeX files are text files with appropriate encoding - when printing a document using a LaTeX processor, the document contains markup that instructs the processor to import and format the records appropriately. This facility means that the BibTeX file can act as a data exchange format. And for example CiteSeer [☞CITESEER], a widely used digital document library, provides a BibTeX entry (as plain text) which can be cut & pasted into a BibTeX file. BibTeX formatted data can be imported into some of the commercial systems.

Software

The UK, academic market is dominated by ISI ResearchSoft (☞http://www.isiresearchsoft.com/) who produce the EndNote, ProCite and Reference Manager products, all of which are widely used in Higher Education institutions. However there are a large number of pieces of software that are designed specifically to facilitate the personal and small group capture, management, sharing and publishing of bibliographic data and other generic applications that are used to do the same (e.g. spreadsheets, word processors, and database programs). The list below details the majority of the more developed products and services, identified while researching this report.

The phrases in quotes marks are taken from the cited Web site and where there are no comments from the authors, this means that the product had the basic functionality (see end of this section), only 5 of the products have been tried/tested in any depth, other data is taken from Web based research.

Archiva	(☞http://www.legend2000.com/archiva/arc_index.asp?arcMenu=Archiva) - "an advanced reference management system with integrated thesis processor."
AskSam	(☞http://www.asksam.com/)- a generic data management system that is very flexible and facilitates the capture, management and publishing of structured, semi-structured and unstructured data. Thus just as with generic structured database systems, it can be adapted to work with bibliographic data.
Bibliographix	(☞http://www.bibliographix.com/) marketed as a publication planner as well as a bibliographic management tool, it provides an 'ideas manager' in which ideas can be indexed using a "short thesaurus" hence keeping and linking related ideas. It also enables network access to group bibliographic databases.
Biblioscape	(☞http://www.biblioscape.com/) There are different levels of product. The standard edition has all the basic functionality and in addition allows relationships to be defined between references, e.g. "Supportive", "Contradict" (it calls this cross linking). There is a freeware version (☞http://www.biblioscape.com/biblioexpress.htm)
Bookends	(☞http://www.sonnysoftware.com/) Is a Mac based product and is "a full-featured and cost-effective bibliography/reference and information management system for students and professionals"
Citation	(☞http://www.citationonline.net/) "Citation is a powerful and easy to use bibliographic database system and notes organizer for research writing."
Database Software (Personal)	Generic database systems such as Microsoft Access that individuals or groups use to create small scale applications to manage their bibliographic data. Other examples include: Cardbox: ☞http://www.cardbox.co.uk/ and Idealist (☞http://www.bekon.com/)
EndNote	(☞http://www.endnote.com/) EndNote is from is from ISI ResearchSoft and seems to be the dominant product for personal bibliographic data management in UK Higher Education. It has all the basic features, including very extensive and customisable import and export filters, and a 'cite while you write' facility which integrates with MS Word and Word Perfect. It provides means of importing images and other files. It can use Reference Web Poster (☞http://www.endnote.com/rwprod.htm) which enables the Web

	publishing (on their server) of EndNote bibliographies.
Spreadsheet Software	Spreadsheet software is can be used to store bibliographic data as a flat field database - the columns being the data fields. These generally providing sorting of the table, fields can be added as required and more sophisticated versions (e.g. MS Excel - ☞http://www.microsoft.com/office/excel/) provide filtering by columns. However there is no easy way to produce actual bibliographies.
Library Master	(☞http://www.balboa-software.com/) "Library Master automatically formats the bibliography, footnotes and citations for your paper, thesis or book in numerous bibliographic styles. It makes it easy to organize research notes and project records."
Ibidem/Nota Bene	(☞http://www.notabene.com/brochure/ibidem.html) "Store bibliographic information in the simple database format, and Ibidem will generate your bibliographic references for you."
Papyrus	(☞http://www.researchsoftwaredesign.com/) Has DOS (version 7) that runs under Windows, and a Mac version (version 8). In addition to basic functionality this provides linking between records e.g. relationships such as "Reviews" or "Refutes" and keywords e.g. "Synonyms", "Supercategory/Subcategory,". Allows the embedding of images. It claims to allow import of references form 'anywhere' including from existing bibliographies in word processor format, using "artificial intelligence techniques in reading your source file, alerting you to potential problems"
PowerRef for Windows	(☞http://www.cheminnovation.com/powerref.html) provides the majority of basic features along with high levels of flexibility in document type, adding user defined fields also allows attaching of graphics to records.
ProCite	(☞http://www.procite.com/) ProCite is also from ISI ResearchSoft and provides comprehensive basic functionality plus a network version with access to a single file, multiple read but only one person can write at one time. It also captures the URL a Web pages title information and stores in the reference collection, and then text from the page can be pasted directly into the ProCite record - although one interviewee noted problems using this facility.
Pybliographer	(☞http://canvas.gnome.org:65348/pybliographer/) is a Linux based product, it has a basic level of functionality. Licensing is based on a GNU GENERAL PUBLIC LICENSE i.e. "Everyone is permitted to copy and distribute verbatim copies of this license document, but changing it is not allowed." (☞http://www.gnu.org/copyleft/gpl.html)
Reference Manager	(☞http://www.refman.com/) Also from ISI ResearchSoft, again with comprehensive basic functionality along with full multi-user networking i.e. multiple uses can both read and write to the shared database.
RefWorks	(☞http://www.refworks.com/) This was the more extensive of the two Internet-based services that we found. It is designed specifically for group as well as individual use.
Scholar's Aid	(☞http://www.scholarsaid.com/) "Scholar's Aid 2000 is a program package that includes a bibliographical data manager called Library and a notes/information manager called Notes", it seems to have a good range of basic functionality. It also claims to export to XML format.
Scribe SA	(☞http://www.scribesa.com/) Is a online service designed to capture data via a forms interface - it has different version for different output formats i.e. APA, MLA or ISO 690.
WebCrimson	(☞http://www.webcrimson.com/) WebCrimson is essentially an online web publishing service. It uses a server based content management system, which has a large number of pre-defined data fields, to generate pages from templates, the user can choose from a variety of predefined template and edit them or create their own. They then add their data to the database via a forms interface and the system generates the pages. One of the pre-defined (customisable) templates is a bibliography template, which (with some tweaking) produces

Web-based bibliographies.

There are a number of other products (e.g. InMagic: ☞http://www.inmagic.com/) and systems (e.g. ADLIB: ☞http://www.uk.adlibsoft.com/) that are designed for a larger scale of usage.

As noted above there are very many products - commercial, open-source and developed by individuals - of which this list is a relatively small sample. However, we believe that the major products have been covered. We are not aware of any publicly available data regarding market share of products and/or actual usage of the systems. Anecdotal evidence in the UK points towards Endnote being a dominant product in Higher Education.

In general the systems have similar **basic functionality** e.g. template based data entry, simple search facility, import from Z39.50 servers, default and customisable output bibliography templates (see below), in many cases integration with word processors, etc. Some also provide support capture of electronic document and/or ideas management/capture along with other functionality (see below) to provide an integrated workspace for researchers, author, students and other potential users.

Key Features

From the above review and our interview study we have complied a list of 'features' that the various systems provide (with respect to bibliographic data management, we have not included below other features such as ideas management). These have been grouped into three sections; data collection/entry, data management/manipulation & sharing, publication/export. However in many cases features cross these areas or choices in one impact on another; we have tried to make these clear in the text.

Before going into detail about particular features it is worth discussing issues related to platform/operating systems. In the majority of cases the systems above are pieces of software installed on individual machines (as opposed to Web-based products, such as RefWorks) The majority of the packages above are Windows products, those which are Mac based or have Mac versions include Bookends, EndNote, PAPYRUS Version 8.0 and ProCite. The only Linux product that we found was Pybliographer.

The feature list below aims to provide a list of the many of the key features (i.e. it is not comprehensive) that are available in the systems reviewed. Some are very common; where features may only be available in a small number of products, this has been noted.

Data Collection/Entry

Under this heading we include all aspects of the capture of bibliographic data from any source.

- **Pre-formatted publication types**: different types of publication require that different data is captured e.g. a journal is likely to require 'journal title' and 'volume number', which are not relevant to other types of publication. Most systems provided a number of standard types along with the ability to add others and in some cases edit the standard version e.g. to add a URL to a book reference.
- **Specifically designed forms based input:** to facilitate the data entry e.g. most systems order the fields with author(s), title, year and other key elements at the beginning of the form - the majority do not allow customisation of this ordering e.g. one participant in our survey felt that it was annoying to have to go to the end of a form to enter the URL.
- **Predictive and auto-completion:** Some systems monitor text as it is entered into a field, and if there is a matching value in the database already it fills in the field, the user can accept the value or keep typing. This can increase the speed of entry considerably, however where many values are similar e.g. journal names, this is less valuable, some may find it annoying.
- **Indexing/Key-wording records**: Many systems provide means of applying keywords to documents to allow the categorising /grouping of records. In general these were un-structured, one product enabled users to link keywords via relationships e.g. e.g. "Synonyms,", "Supercategory", "Subcategory,"
- **Authority Files/Drop Down options**: One participant in our interview study had created their own database using Microsoft Access, they had created separate tables for authors, journal names and publishers (but not keyword) and added dropdown option lists on their main data entry form to make data entry faster and more consistent. While this is done to some degree by the predictive completion of text (see above) the authority file is a more robust system. None of the systems we actually tested had this 'drop down' type facility.
- **Spell checking:** related to authority files is the facility to check spellings of all the text in a record, this was provided in a many of the products.
- **Import filters:** these allow the user to import bibliographic data held in file formats different to that of native system, e.g. BibTeX, or data from competing products or CD Rom based databases. Some products provide a simple parsing language to allow user defined filters.
- **Ability to search and retrieve records from remote Internet accessible bibliographic database services:** This is via the Z39.50 protocol. Many systems provide a set of standard 'connection files' for commonly used services, these are pre-configured to access these. In general these can be edited and new ones created as necessary - however this requires some work and

generally some technical expertise.

- **Automatic field completion for Web pages**: One product had the facility to gather data directly from a Web browser and so auto-complete some of the relevant fields and support capture of text from the page into relevant fields.
- **Document capture**: Electronic documents themselves (as opposed to simply the bibliographic details) can be captured e.g. PDF and HTML files. These can then be used to do more advanced searching e.g. full text searching.
- **Attaching image objects to a record:** Some products allowed images to be attached to the record and stored in the internal database.

Data Management/Manipulation & Sharing

This includes any processing, management, manipulation or sharing taking place between capture and publishing, including sharing of data.

- **Searching/Filtering of records:** Probably the most basic requirement, most systems provide tools to search the data using any field and in some cases multiple fields. Results can generally be exported to an external bibliography.
- **Customized/sorted views**: The data can be sorted in to different orders e.g. by author or title and in some cases a table (spreadsheet) view. In some cases it is possible to define which fields to view.
- **Defining relationships between records**: A small number of the products offered the facility to define relationships between records of the form "Supportive", "Contradict". These can then be traversed to explore the records and select them for exporting into a bibliography. Other systems allow references to be grouped, which is similar in functional terms to keywording and filtering on specific keywords.
- **Hyperlinking to URLs**: Many products provided URL fields which provide a means of direct linking to the source document.
- **Network and Multiple User Access:** A number of the products are designed to be used by groups. These generally use a server based database or a single database on a network share. These take care of locking the records to prevent simultaneous editing. Two of those reviewed were Web based services in which the data was held on an external server with logon from any Web browser. Of course many other products can be used for non-simultaneous access to a database file on a network share.
- **Duplicate detection**: One product provides duplicate detection, hence assisting in maintaining the database, others provide tools that could facilitate this.
- **Global Search and Replace**: Some products provide a means of changing a piece of text across all records e.g. a name spelt consistently incorrectly.
- **Saving and Backing up:** Many systems had database back ends and so individual edits are automatically saved, however in many cases actual backing up of the whole database, requires technical knowledge e.g. where the files and which require backing up.
- **Merging of database:** In many cases a product allowed the use of multiple databases, some provide tools to assist in merging these, however there may be some problems with dealing with duplicate and near duplicate records.

Publication/Export

- **Creation of Bibliographies in commonly used styles:** This is probably the most basic publishing feature. The majority of products allowed customisation of the templates or filters using custom languages or integrated tools.
- **Export filters:** More generic than bibliography publishing, is the ability to export the database or a selected sub-set of the database to an external format. Most products offered very limited export facilities, compared to their import features. This may be to try and 'tie' customers into their product. The most basic format seems to be standard comma delimited, with a custom ordering of fields.
- **Integration with standard word processors:** Many products provide integration with word processors (e.g. MS Word and WordPerfect). In most cases this takes the form of the ability to place a mark (automatically or not) denoting a citation in the document. The program then either automatically adds a reference to the references section at the end of the document, or the document is processed later to do so.
- **Web posting/publishing of bibliography and/or database:** A very common facility was to output bibliographies or the data in HTML or to an online and searchable database.
- **Word processor templates for manuscripts of common journals**: Using Microsoft Word Macros one product leads the user through the creation of the document with relevant formatting, text styles and citation styles.

Usability

We are unaware of any usability studies with respect to the bibliographic software reviewed. However from our limited review of systems and comments gained during our interview study, it seems likely that there are a number of usability barriers to the easy and continued use of these systems in general and specific problems with individual products.

Central to the problems we have identified seem to be ease of data entry, all participants in our (albeit small) interview study indicated that the time consuming and often 'fiddly' nature of data entry was problematic. Other key issues relate to customization (e.g. difficulty in creating customized styles), integration with other packages (e.g. technical bugs) and the time consuming nature of simple day to day maintenance. The bottom line for all bar one of our interviewees was that they tended to use the software for personal use only when they need to write papers or other documents with bibliographies, rather than as a general repository for bibliographic data.

It would be interesting and useful to conduct, or find existing data about, usability studies in this area.

Wish Lists

This 'wish list' of features are compiled from the interviewees in our small scale study, and so can only be indicative of such items that might come from a larger survey. It should be noted that in some cases the features are available in some products, however they were not in the system(s) used by the interviewees. These are ordered under rough headings:

Capture:

- Auto extraction of data from text citations and references in papers, for example the Windows clip board, including BibTeX format.
- Auto capture of bibliographic data from a paper e.g. where the papers were from a common journal and thus had a common format - import filter for whole papers.
- Help in 'getting into' a subject area e.g. the common problem of knowing what to ask and what vocabulary to use in a new field. Possibly have a system to cross search other peoples' data and say that 'people who have indexed this paper have used the following terms' or 'expert y has used the following terms to index this paper'.
- Links to external databases to confirm spellings of names, journal titles etc.
- Import data from 'table of contents' services data inc. abstracts.
- Copy some/all details from one record to another e.g. when adding lots of papers by the same author of same journal
- Capture references at the end of an electronic publication
- Auto or semi-auto detection of document type (e.g. book, journal, etc)
- Notification of additions to selected internal and external databases

Management & Augmentation:

- More effective classification tools inc. more structured vocabularies
- Access and update [the] data online, from any machine via a browser
- Hierarchal keyword structures, "so that they are easier to navigate and conceptualise"
- Richer annotation possibilities other than 'notes' fields and better searching of them.
- Semi-automated indexing or keywording from the text of an on-line document
- Visualization of aspects of data e.g. timelines for papers under a given keyword, citation relationships, or identification of 'camps' of researchers with particular points of view...
- More flexible merging of databases.
- Semi-automatic up dating of documents that have used a record [e.g. in a bibliography] when the record changes - i.e. the system should remember when and which files have used the individual records and notify the user when changes are made
- More effective backup

Sharing:

- More extensive export facilities
- See or be notified, when a record was added or edited and by who
- Ability to select records and e-mail selection from within system
- Ability to search other people's bibliographic databases e.g. known expert/specialist or researcher in field, in particular found via publications e.g. person x published paper on y, therefore they might have a bibliographic database.

Publication:

- Output to any bibliographic software
- Ability to integrate (export to and import from) and output to other applications e.g. graphics, mind mapping, database or visualization program, e.g. export data to a mind mapping program so that main branches were authors, next level publication year and then title
- Easy selection of records for inclusion in a bibliography e.g. tick boxes next to record on screen, multiple select (using ctrl, left mouse button), shopping cart type approach
- XML output and input

While these are necessarily very extensive or generalisable to the wider community of users and potential users of such systems, they do give some insight into the perceived needs of users and the significant potential for improvements in design, and functionality of the existing software and services.

Sources of Reference - These sources were the main Web sites from which generic background information was gained as part of our on-line literature review. In the majority of cases information about products and services came from the publishing organisations' Web sites. Other citations are referenced directly in the text.

Evans, Peter. (2002) A review of 3 major Personal Bibliographic Management tools. Available: ☞http://www.biblio-tech.com/html/pbms.html.

Information Systems and Technology University of Waterloo. (2000) Which Personal Bibliographic Management Package Should I Choose? Available: ☞http://ist.uwaterloo.ca/ew/biblio/which.html.

Kent, T. (2002) Bibliographic Software. The UK Online User Group. Available: ☞http://www.ukolug.org.uk/links/biblio.htm.

Maggie. Shapland. (1999) Evaluation of Reference Management Software on NT. 2001. University of Bristol. Available: ☞http://www.cse.bris.ac.uk/~ccmjs/rmeval99.htm.

Memorial University of Newfoundland Libraries, Bibliographic Control Services (2002) Technical standards for electronic bibliographic data/metadata. Available: ☞http://www.mun.ca/library/cat/standards.htm.

Morton, D. (2001) Personal Bibliography Software. Available: ☞http://library.uwaterloo.ca/~dhmorton/dnh4.html.

Online Computer Library Center. (2003) OCLC Online Computer Library Center Homepage. Available: ☞http://www.oclc.org/home/.

E Overview of Major Library Focused Bibliographic and Related Standards

Introduction

The extensive and necessarily complex bibliographic standards used by the international library communities are likely to be over complex and in some cases seem irrelevant (e.g. classification schemes to place a book on unique place on a shelf), in the present context of applying blogging and Semantic Web approaches to personal and small group bibliographic data management.

However many of the concepts that underpin the Library systems and their standards are highly relevant. Indeed concepts such as the use of authority files, consistent description of resources and encoding of records to facilitate interoperability are at the heart of Semantic Web approaches. Many other concepts may be simplified or adapted such as cataloguing rules, copy cataloguing and the theories and practices of knowledge representation that underlying subject cataloguing and indexing schemes.

This appendix aims to provide a very short and simplified overview of relevant concepts and the various types of standard related to bibliographic data and their roles, along with a review of the dominant standards and likely future developments.

Background

It is important to understand that current bibliographic standards have evolved (primarily) in the context of physical libraries over a very long period (100+ years), to meet the requirements of librarians and library users. Probably the most fundamental requirements are to catalogue items (generally books or periodicals). The cataloging originally used cards in a physical catalogue and was used to assist the librarian and library user to locate items (of which there is generally only one copy) on physical shelves, that were relevant to their query (information need).

The standards related to bibliographic data are particularly numerous, and their inter-relationships complex. This is because bibliographic data management is a ubiquitous problem, for which many individuals, groups and organisations have historically developed solutions to meet their own specific needs. For example, with respect to physical media: traditional libraries have tended to manage physical books and periodicals, and their needs are very different to that of a picture library, museum or archive

of historic manuscripts. Similarly, the subject indexing needs of a generic public library and a specialist research library on behaviour of primates will be very different.

Understanding this background provides necessary context to, what might otherwise appear over complex and even bizarre, in an age of electronic multimedia 'documents' and ubiquitous use of computer database systems, along with associated search engine technologies. Further problems in understanding the standards relate to differences in the use of terms. For example, in the context of library cataloguing, 'descriptive cataloguing' refers to describing the object itself (and not the subject or content of the object), whilst in contemporary usage, 'descriptive metadata' refers to data that describes the object or relating to what the object contains or is about.

Bibliographic Records

Overview: The need to exchange consistent and accurate bibliographic records between libraries and in particular in a 'machine readable' form, has led to the creation of standards for the creation of Bibliographic Records (data). In generic terms these standards define 6 things:

> *1) What information should be captured about the 'publication' i.e. cataloguing data.*
> *2) the structure of the record.*
> *3) detailed rules or guidelines for how to deal with specific cataloguing situations/issues e.g. what to do when there are two different formats of an authors name used within one publication. Broadly these also include the use of authority lists/files as definitive authorities over for example the spelling of a place or personal name.*
> *4) how the subject or content of the 'publication' should be described i.e. how the publication should be indexed.*
> *5) the specific syntax (e.g. use of punctuation) for encoding of the record.*
> *6) The specific transfer protocols for transferring the data between locations.*

Clearly some of these can be independent, for example, given a set of data to be captured, they could be encoded in many ways and an encoding might cater for a large number of possible indexing systems. However they are intimately inter-related - for example, it is not possible to define a specific encoding without some knowledge of the types of information and structure of the data. Thus standards can be seen to form inter-related sets that are mutually dependent.

The remainder of this appendix will explore the major standards and their practical uses in a library, digital library or archive context. These broadly illustrate the key concepts.

The **major international standards** which deal with issues 1), 2) and some aspects of 5) above are based around a general framework, the International Standard Bibliographic Description, ISBD(G) (Background to ISBD ☞http://www.ifla.org/VII/s13/pubs/isbd.htm) which provides recommendations for:

- *What information should be given inc. what detail is required*
- *What order the information should be given in*
- *How punctuation should be used to distinguish between elements of the description.*

> *There are variations of ISBD for many types of material (see http://www.ifla.org/VI/3/nd1/isbdlist.htm) e.g.*
>
> - *ISBD (CM): International Standard Bibliographic Description for Cartographic Materials*
> - *ISBD(CR): International Standard Bibliographic Description for Serials and Other Continuing Resources*
> - *ISBD (M): International Standard Bibliographic Description for Monographic Publications*

However, the *Functional Requirements of Bibliographic Records* (FRBR pronounced "fur ber") report by IFLA (The International Federation of Library Associations and Institutions) produced in 1998, describes a new and more generic approach to organising and specifying the components of a bibliographic record. Fundamentally, this defines bibliographic entities (i.e. *Work*, *Expression*, *Manifestation*, Item*)* and their relationships (e.g. a *Work* is realized through an *Expression*). IFLA have initiated a full-scale review of IFLA's "family of ISBDs" to ensure conformity between the provisions of the ISBDs and those of FRBR. This may necessitate changes in the other relate standards e.g. MARC (☞http://www.loc.gov/marc/marc-functional-analysis/home.html) and AACR2R (☞http://www.nlc-bnc.ca/jsc/frbr1.html)

The major standard for **Cataloguing Rules** are the AACR2R (Anglo American Cataloguing Rules

☞http://www.nlc-bnc.ca/jsc/), these provide rules for description of 'publications' or physical media and are based on ISBD. They allow for three levels of description (full, core, or minimal), each implementing institution can opt for the appropriate level.

Information Interchange Formats (Data Structure Standards): Standards such as ISBD and AACR2 Standards for the content and cataloguing rules. This data needs encoding in a common machine-readable format. The dominant format in the Library community is MARC (MAchine-Readable Cataloging). "It provides the mechanism by which computers exchange, use and interpret bibliographic information and its data elements make up the foundation of most library catalogues used today." (☞http://www.loc.gov/marc/faq.html#1).

MARC records are divided up into fields (☞http://www.loc.gov/marc/bibliographic/ecbdlist.html) that correspond to AACR2 areas and would be called an element in metadata terms. Fields are grouped in 100s e.g. 2XX fields correspond to Title and statement of responsibility. Each specific field corresponds to an element of the record e.g. 'field 250' for edition statement.

There are various localised flavours of MARC e.g. USMARC (developed by the Library of Congress) and UKMARC (developed by the British Library), to meet the specific practices of a particular national bibliography. However, MARC21 is the current standard and is being adopted widely; the British Library are implementing it for 2004.

There are XML implementations of MARC (☞http://www.loc.gov/marc/marcxml.html), in particular MARCXML (☞http://www.loc.gov/standards/marcxml//), for which there are XML Schema, DTDs and tools to assist with conversion - e.g. a MARC XML to RDF Encoded Simple Dublin Core Stylesheet (☞http://www.loc.gov/standards/marcxml//).

MODS [☞MODS] is a subset of MARC 21. "As an XML schema it is intended to be able to carry selected data from existing MARC 21 records as well as to enable the creation of original resource description records. It includes a subset of MARC fields and uses language-based tags rather than numeric ones, in some cases regrouping elements from the MARC 21 bibliographic format. This schema is currently in draft status and is being referred to as the "Metadata Object Description Schema (MODS)". MODS has many potential uses and seems to be seen as a more usable system for a number of applications.

ISBD (and at present FRBR) can be thought of as acting for the foundation of MARC and AACR2 in that they have both been amended to conform to the ISBD standards.

Subject Headings

There are many standard 'subject heading' systems that is, systems that are used to assign subject headings describing the content of the object, being catalogued. Subject headings are described as providing additional **access points** [i.e. a name or a term that can be used to retrieve the bibliographic information from a card catalogue or an online catalogue]. Examples include authors name, title of the book, and subject heading.

Subject headings are used to place cards in card catalogues or entries in an index, of which there may be many. Whereas subject classifications (see below) are used to place actual items (e.g. books).

Subject headings are generally used in conjunction with other types of **authority files** (see below) that are authoritative forms of headings according to international, national or locally specified criteria. In general only 3 or 4 subject headings are added to records, this derives from the expense of the original system of printing, filing and maintaining the additional cards.

There are important distinction between so-called *pre-* and *post-coordinate* systems.

> *In pre-coordinate systems terms are arranged in a 'logical' order, assuming that there will be one preferred sequence which will be appropriate for arranging the terms or in the case of the need to place a physical object, item.*
>
> *Post-coordinate systems do not try to impose an order on the sequence. There is generally an independent retrieval system that allows terms to be combined (coordinated) at the time of retrieval.*

Pre-coordinate systems place the onus on assigning full combined subject descriptions at the time of cataloguing, while post-coordinate systems allow the combination after cataloguing – as in modern Boolean key worded searches on the Web. Pre-coordinate systems are very uncommon in relation to the Web and many users of modern indexing systems may find the use of pre-coordinate systems problematic.

By far the most widely used is the LCSH - Library of Congress Subject Headings (a pre-coordinate system) others include:

Sears List of Subject Headings (☞http://www.hwwilson.com/print/searslst.htm)
PRECIS (Preserved Context Indexing System) - (see Chan 1994 pp244-254)

Subject heading systems all have various tools to assist cataloguers - for example paper or written

manuals of rules or guidelines, scope notes which clarify how terms should be used and may draw attention to distinctions between terms, references e.g. USE and UF (Use For), SA (See Also) in the LCSH and pointers to BT (broader term) and NT (narrower terms)

Other subject/domain specific subject headings (or thesauri) include: e.g. MeSH the Medical Subject Heading (☞http://www.nlm.nih.gov/mesh/meshhome.html) developed by the National Library of Medicine (NLM) in the USA and Zoological Record Thesaurus by developed by Biosis (☞http://www.biosis.org.uk/products_services/zoorecord.html)

Subject Classifications

Subject Classifications differ from Subject Headings in that they are designed primarily to provide a means of identifying an appropriate location on shelves for a book, e.g. books with the same 'main' subject area grouped together on shelves. However the place of any particular book in any particular library may be different, as different libraries will try to organise their shelves to meet the needs of their users.

The main standards

LCC – Library of Congress Classification (☞http://www.loc.gov/catdir/cpso/lcco/lcco.html)

DDC - Dewey Decimal Classification (☞http://www.oclc.org/dewey/)

UDC - Universal Decimal Classification (☞http://www.udcc.org/)

As with Subject Headings there are a number of subject specific systems e.g. the National Library of Medicine (NLM) Classification associated with the MeSH subject headings, and the ACM Computing Classification System.

Many other systems exist and are in relatively common use (see Chan 1994 for more examples). Other standards include subject classification standards e.g. MARC can use LCC and DDC.

Authority Control

Authority control is central to effective and consistent cataloguing and indexing in library contexts. Authority control is a means by which 'access points' (e.g. personal names, geographic location, corporate names, titles and subject headings) can be represented in a consistent manner. In general authority files are divided conceptually into two 1) subject authority (how to use general subject/topic terms e.g. LCSH) and 2) name authority files (how to refer to **unique** entities e.g. NACO, see below). Authority files (now generally machine readable) are held by libraries or accessed from an external body to provide a means of defining a uniform heading for the data thus ensuring that all works are catalogued appropriately under for example a particular author. Then cataloguing rules (e.g. AACR2R see above) provide the means of implementing the heading in the record.

There are various national and international authority control initiatives, e.g. NACO (the name authority component of the Program for Cooperative Cataloging - ☞http://www.loc.gov/catdir/pcc/naco.html). The Library of Congress authority files are widely used and are now online (☞http://authorities.loc.gov/) and many tools e.g. online or CD products are also provided to assist with authority implementation e.g. the Library of Congress Cataloguer's Desktop (☞http://lcweb.loc.gov/cds/cdroms1.html#desktop)

Copy Cataloguing

Copy cataloguing is an efficient means of cataloguing in which library records are imported from an external sources (e.g. OCLC) and are then amended as necessary to ensure they comply with any local/internal library standards. This clearly also provides a means of increasing consistency across libraries and reducing error propagation as cataloguing and indexing takes place.

Data retrieval and transfer:

In order to transfer the data, MARC and similar encoding standards (see above) provide a means of encoding in a common machine-readable format, these can then be transferred between computer systems. However where users wish to retrieve specific records from a large collection across a network it is necessary to provide a standard protocol for the initial querying and delivery of the data. ☞Z39.50 (☞http://www.loc.gov/z3950/agency/) is the mostly widely used standard within the library community and is used in the vast majority of systems.

Conclusions:

Extensive and comprehensive standards exist in the library domain for all aspects of bibliographic record, data capture, management and sharing - these standards help provide effective interoperability between very large systems used by very large numbers (often millions) of users. In the context of the current project (personal and small group systems) these are in general, significantly over specified and complex to be appropriate. However the principles underlying these standards provide a basis for defining the necessary components of standards, workflow, and protocols for smaller scale systems, especially in the context of interoperability, which is an issue not widely dealt with by existing personal bibliographic software - see appendix D.

Specific lessons include:

1. The value of a generic 'framework' for the necessary elements of the bibliographic record describing the object/publication and the content/subject of the object, e.g. International Standard Bibliographic Description (ISBD) and *Functional Requirements of Bibliographic Records* (FRBR)
2. In the interests of consistency data within the record, the need for some form of cataloging 'rules' which define how the elements of the record should be competed e.g. Anglo American Cataloguing Rules (AACR2R)
3. The need for ubiquitous form of encoding of the record data to provide a means of large scale interoperable transfer of data (e.g. MARC) and a protocol for searching remote and distributed systems (e.g. Z39.50).
4. The requirement for, and complexities of, providing effective yet comprehensive vocabularies for describing the content/subject of the objects/publication. These vocabularies provide a basis for the retrieval of the records and hence access to the actual object/publication for users. Specific issues relate to the need to cater for both general and specialist user groups.
5. The use of authority control to provide a means of ensuring unique identification of names (e.g. authors, organisations, publishers and geographical locations) and in the case of descriptive vocabularies their consistent use.
6. The value of work practices (e.g. copy cataloging) that have evolved to make cataloging and indexing more consistent and efficient.

These and other lessons drawn from this review have guided the production of the document '☞SWAD-Europe: Semantic Blogging and Bibliographies - Requirements Specification' and will continue to provide guidance for ongoing, more detailed requirements and systems specification processes.

Sources of Reference

These sources have been used for generic background information in this section. Other, more specific, sources are referenced directly in the text.

Rowley, J. and Farrow, J. (2000) Organizing Knowledge: An Introduction to Managing Access to Information, 3rd Edition, Ashgate Publishing Ltd, Aldershot, UK.

Burke, M. A. (1999) Organization of Multimedia Resources: Principles and Practice of Information Retrieval, Gower, Aldershot, UK.

Deegan, M. and Tanner, S. (2002) Digital Futures: Strategies for the Information Age, Library Assocation Publishing, London.

IFLA Study Group on the Functional Requirements of Bibliographic Records. (1998) "Functional Requirements of Bibliographic Records: final report." München: K. G. Saur. Available online at http://www.ifla.org/VII/s13/frbr/frbr.pdf. (Downloaded 29 August 2002.)

Hagler, Ronald (1997) The Bibliographic Record and Information Technology, Third Edition, American Library Association, London.

Chan, Lois Mai (1994) Cataloging and Classification: an Introduction, Second Edition, McGraw-Hill, London.

F References

[ACM]

The ACM Computing Classification System [1998 Version]
☞http://www.acm.org/class/1998/

[APACHE-AGORA]

Apache Agora: A community visualisation toolkit by Stefano Mazzocchi 2002
☞http://cvs.apache.org/~stefano/agora/

[APECKS]

APECKS: a Tool to Support Living Ontologies Jenifer Tennison & Nigel R. Shadbolt in Proceedings of KAW'98 (Eleventh Workshop on Knowledge Acquisition, Modeling and

Management) Voyager Inn, Banff, Alberta, Canada
☞http://ksi.cpsc.ucalgary.ca/KAW/KAW98/tennison/

[AMAYA]
Amaya - the W3C editor/browser
☞http://www.w3.org/Amaya/

[ANNOTATION-TOOLS]
Semantic Web - Annotation and Authoring
☞http://annotation.semanticweb.org/ontomat.html
Also, see [☞SMORE]

[ANNOTEA]
☞Annotea: a system for creating and publishing shareable annotations of Web documents. (latest draft Dec 2002)
☞http://www.w3.org/2002/12/AnnoteaProtocol-20021219

[ANNOTEA-THREAD]
Reply protocol in Annotea
Document: ☞http://www.w3.org/2001/Annotea/User/Protocol.html#ReplyProtocol
Schema: ☞http://www.w3.org/2001/03/thread

[AS-WE-MAY-THINK]
Bush, V. As we may think. The Atlantic Monthly, 176, 1 (1945), 101-108.

[AS-WE-SHOULD-HAVE-THOUGHT]
Nuernberg, P.J., Leggett, J.J. and Schneider, E.R. As We Should Have Thought. In Proceedings of Hypertext '97: 8th ACM Conference on Hypertext, Southampton, 1997, pp. 96-101
☞http://journals.ecs.soton.ac.uk/~lac/ht97/pdfs/nuern.pdf

[BibTeX]
LaTeX: A Document Preparation System by Leslie Lamport, 1986, Addison-Wesley.
BibTeXing (☞btxdoc.tex), by Oren Patashnik, February 1988, (BibTeX distribution).
☞http://www.ecst.csuchico.edu/~jacobsd/bib/formats/bibtex.html

[BLOG_COMMUNITIES]
Weblog Communities an essay by Jeremy Bowers 2001
☞http://www.jerf.org/writings/weblogCommunities/

[BLOG-TRIBE]
Blog Tribe Social Network Mapping, a weblog article by Ross Mayfield January 2003
☞http://radio.weblogs.com/0114726/2003/01/02.html#a176

[CATALOG-OF-CATALOGS]
J. F. Metzl, Searching for the catalog of catalogs". In Books, Bricks & Bytes: Libraries in the Twenty-First Century, S. R. Graubard and P. LeClerc (eds.), pp. 147-160, Transaction Publishers, New Brunswick, NJ, 1999

[CITATION-INDEXING]
Eugene Gaxfield. Citation indexing: Its theory and application in science, technology, and humanities. Wiley, New York, 1979. ISBN 089495024X.

[CITESEER]
CiteSeer Publications Research Index (NEC Research Institute)
☞http://citeseer.nj.nec.com/cs

[CITESEER-ARTICLES]
K. Bollacker, S. Lawrence, and C. Lee Giles. Citeseer: An autonomous web agent for automatic retrieval and identification of interesting publications. In Proceedings of the Second International Conference on Autonomous Agents, pages 116-113, New Yor k, 1998. ACM Press.
K. Bollacker, S. Lawrence, and C. Lee Giles. A system for automatic personalized tracking of scientific literature on the web. In Proceedings of the Fourth A CM Conference on Digital Libraries, pages 105-113, New York, 1999. ACM Press.

[CLAIMAKER]
ClaiMaker: a tool from the ScholOnto project
☞http://claimaker.open.ac.uk/

[CLAIMAKER-SANDPIT]
ClaiMaker sandpit, an online demonstration of concept visualisation
☞http://claimaker.open.ac.uk/Sandpit/

[CLAIMAKER-SCHEMA]
ClaiMaker RDFS schema
☞http://kmi.open.ac.uk/projects/scholonto/resources/Scholonto2.rdfs

[CLAIMAKER-SCHOLARLY]
Buckingham Shum, S, Domingue, J and Motta.E. (2000) Scholarly Discourse as Computable Structure. Proceedings of SC2: Second International Workshop on Structural Computing, San Antonio, Texas, June, 2000. Springer-Verlag: Lecture Notes in Computer Sc ience.
Buckingham Shum, S., Motta, E., and Domingue, J., Representing Scholarly Claims in Internet Digital Libraries: A Knowledge Modeling Approach, in Proc. 3 d European Conference on

Research and Advanced Technology for Digital Libraries (Paris, France, September 22-24, 1999), Springer Verlag.

[CLAIMAKER-WEAVE]
ClaiMaker: Weaving a Semantic Web of Research Papers Gangmin Li, Victoria Uren, Enrico Motta, Simon Buckingham Shum, John Domingue presented at International Semantic Web Conference (ISWC) 2002, June 9-12th, 2002 Sardinia, Italia
☞http://kmi.open.ac.uk/projects/scholonto/docs/ClaiMaker-ISWC2002.pdf

[COHSE]
The Conceptual Open Hypermedia Project
☞http://cohse.semanticweb.org/

[CREAM]
S. Handschuh and S. Staab. Authoring and annotation of web pages in CREAM. In The Eleventh International World Wide Web Conference (WWW2002), Honolulu, Hawaii, USA 7-11 May, 2002.

[DC]
The Dublin Core Metadata Initiative
☞http://dublincore.org/groups/citation/

[DISTRIBUTED-ANNOTATION]
RD Dowell, RM Jokerst, A Day 2001 The Distributed Annotation System *BMC:Bioinformatics* 2:7, 2001

[DMOZ]
DMOZ - The open directory project
☞http://dmoz.org/

[ESSENTIAL_BLOGGING]
Essential Blogging: Selecting and Using Weblog Tools by Cory Doctorow, Rael Dornfest, J. Scott Johnson, Shelley Powers, Benjamin Trott, Mena G. Trott, 2002 O'Reilly

[FACET-MAPS]
FacetMap - the home for faceted classification
☞http://facetmap.com/

[FINDUR]
D.L. McGuinness; Ontological Issues for Knowledge-Enhanced Search; Proceedings of Formal Ontology in Information Systems, June 1998. Also in Frontiers in Artificial Intelligence and Applications, IOS-Press, Washington, DC, 1998.

[FRIENDS]
Friends and Neighbors on the Web Lada A. Adamic and Eytan Adar
☞http://www.hpl.hp.com/shl/papers/web10/

[GOLDEN-BULLET-1]
GoldenBullet: Automated Classification of Product Data in E-commerce Y. Ding, M. Korotkiy, B. Omelayenko, V. Kartseva, V. Zykov, M. Klein, E. Schulten, and D. Fensel BIS-2002: 5th International Conference on Business Information Systems, Pozna, Poland, April 24-25, 2002.

[GOLDEN-BULLET-2]
GoldenBullet in a Nutshell Y. Ding, M. Korotkiy, B. Omelayenko, V. Kartseva, V. Zykov, M. Klein, E. Schulten, and D. Fensel FLAIRS-2002: The 15th International FLAIRS Conference, Beachside Resort and Conference Center, Pensacola Beach, Florida, May 14-16, 2002.

[HAYSTACK]
The MIT Haystack project
☞http://haystack.lcs.mit.edu/
Haystack: A Platform for Creating, Organizing and Visualizing Information Using RDF by David Huynh David Karger Dennis Quan 2002 ☞http://haystack.lcs.mit.edu/literature

[HYPERTEXT-LEARNING]
Jacobson, M. J., & Spiro, R. J. (1993). *Hypertext learning environments, cognitive flexibility, and the transfer of complex knowledge: An empirical investigation.* Urbana-Champaign: University of Illinois, Center for the Study of Reading.

[HYPERTEXT-THEORY]
Hypertext Theory and User Support: A Brief Review of the Literature by Stephen Victor
☞http://citeseer.nj.nec.com/291537.html

[IDEAGRAPH]
Ideagraph application by Danny Ayers
☞http://ideagraph.net/

[IBIS-TERMS]
Issue-Based Information Systems (IBIS) Terms, Danny Ayers October 2002
☞http://ideagraph.net/xmlns/ibis/

[KAON]
Kaon ontology management environment
☞http://kaon.semanticweb.org/

[KEPLER]
☞Kepler - An OAI Data/Service Provider for the Individual *by* Xiaoming Liu & Kurt Maly & Mohammad Zubair ☞D-Lib Magazine 2001 7(4) (ReLIS:doi:dlibma:y:2001:v:7:i:4:p:2)

[KLARITY]
Klarity - automatic concept-based categorisation
☞http://www.klarity.com.au/

[LINKING_DANGEROUSLY]
The Year of Linking Dangerously a weblog article by Shelley Powers 2003
☞http://weblog.burningbird.net/fires/000796.htm

[MODS]
Metadata Object Description Schema
☞http://www.loc.gov/standards/mods/mods-overview.html

[MT]
MovableType personal publishing system
☞http://www.movabletype.org/

[MT-RDF]
Selection of utilities to add RDF to MT blog entries, written by Matt Biddulph
☞conversion template, ☞n3 parser, ☞href detector

[MT_TRACKBACK]
MoveableType Trackback utility
Specification: ☞http://www.movabletype.org/docs/mttrackback.html

[MY-PLANET]
Kalfoglou, Y., Domingue, J, Motta.E., Vargas-Vera, M. and Buckingham Shum, S, (2001) myPlanet: an ontology-driven Web-based personalised news service. Proceedings of the IJCAI'01 workshop on Ontologies and Information Sharing.

[OLDAILY]
Ol'Daily: The weblog of Stephen Downes
☞http://www.downes.ca/news/OLDaily.htm

[ONLINE-COMMUNITIES]
Online Communities: Commerce, Community Action, and the Virtual University by Chris Werry (Editor), Miranda Mowbray (Editor), Hewlett-Packard
Prentice Hall 2000

[ONTOSEEK]
N. Guarino, C. Masolo, and G. Vetere. OntoSeek: Content-Based Access to the Web. IEEE Intelligent Systems, 14(3):70--80, May 1999.

[ORACLE-OF-BACON]
The Oracle of Bacon at Virginia
☞http://www.cs.virginia.edu/oracle/

[OWL]
Web Ontology Language (OWL) Reference Version 1.0 (W3C Working Draft 12 November 2002)
☞http://www.w3.org/TR/owl-ref/

[PROCITE]
ProCite Information Toolbox (ISI ResearchSoft)
☞http://www.procite.com/

[SCHOLARLY-ARCHIVE]
Dalgaard, Rune. Hypertext and the scholarly archive: intertexts, paratexts and metatexts at work. In Proceedings of the 12th ACM conference on hypertext and hypermedia, Aarhus, 2001, pp. 175 - 184.

[SEMWEB]
W3C Semantic Web activity
☞http://www.w3.org/2001/sw/

[SELF-ORGANIZE]
Self-Organization of the Web and Identification of Communities by Gary William Flake and Steve Lawrence and C. Lee Giles and Frans Coetzee
IEEE Computer 35(3): 66--71 2002 ☞http://citeseer.nj.nec.com/flake02selforganization.html

[SMALL-WORLD-MILGRAM]
S. Milgram. ``The small world problem.'' Psychology Today 2, pp. 60-67,, 1967.

[SMALL-WORLD-WATTS]
D. J. Watts. ``Small Worlds: The Dynamics of Networks Between Order and Randomness.'' Princeton University Press, 1999.
D. J. Watts and S. H. Strogatz.``Collective dynamics of `small-world' networks.'' Nature 393, 440-442 (1998).
Small World Research Project: ☞http://smallworld.sociology.columbia.edu/index.html

[RDF]

☞*Resource Description Framework (RDF) Model and Syntax Specification*, O. Lassila and R. Swick, Editors. World Wide Web Consortium. 22 February 1999. This version is http://www.w3.org/TR/1999/REC-rdf-syntax-19990222. The ☞latest version of RDF M&S is available at http://www.w3.org/TR/REC-rdf-syntax.

[RSS2.0]
RSS 2.0 Standard
☞http://backend.userland.com/rss

[SMORE]
Kalyanpur, Aditya, Bijan Parsia, James Hendler, Jennifer Golbeck, "SMORE - Semantic Markup, Ontology, and RDF Editor." Submitted to WWW2003, 2002.

[SWADE_ANALYSIS]
Semantic Web Applications - Analysis and Selection HP SWADE Report 2002
☞
http://www.w3.org/2001/sw/Europe/reports/chosen_demos_rationale_report/hp-applications-selection.html

[TB-EXTENSIONS]
PostIt by David Raynes
☞http://www.rayners.org/archives/000150.php
ComeBack by David Raynes
☞http://www.rayners.org/archives/000140.php
Trackback threading by Ben & Mena Trott☞
http://www.movabletype.org/news/2002_08.shtml#000568
BackTrack by Shelley Powers
☞http://weblog.burningbird.net/fires/000838.htm
MoreLikeThisFromOthers by Ben & Mena Trott
☞http://www.sixapart.com/log/2002/12/more_like_this_.shtml

[TB-SUMMARY]
Summary of Trackback developments by Ben Hammersley
☞http://www.benhammersley.com/archives/003862.html

[TB-STANDALONE]
Trackback Module (in Perl) by Timothy Appnel
☞http://www.mplode.com/tima/archives/000191.html
Trackback module (in Python) by Mark Paschal
☞http://markpasc.org/code/tbpy/
Pingback fucntionality, using XML-RPC, by Stuart Langridge and Ian Hickson.
☞http://www.hixie.ch/specs/pingback/pingback

[THIS-IS-XFML]
This is XFML, a weblog article by Mark Pilgrim Dec 2002
☞http://diveintomark.org/archives/2002/12/03/this_is_xfml.html

[TM4J]
Topic Maps for Java: Open source Java toolkit for Topic Maps
☞http://tm4j.org/

[TOPIC-DRIVEN-CRAWLERS]
F Menczer, G Pant, and P Srinivasan. Topic-driven crawlers: Machine learning issues. ACM TOIT 2002. ☞http://dollar.biz.uiowa.edu/fil/Papers/TOIT.pdf

[TOPIC_EXCHANGE]
Internet Topic Exchange activity
☞http://topicexchange.com/

[TOPIC-MAPS]
Topic Maps activity
☞http://www.topicmaps.org/

[VIS-ALGORITHM]
Visualization of Document Co-Citation Counts Steven Noel, C.-H. Henry Chu, Vijay Raghavan presented at 6th International Conference on Information Visualization, London, England, July 2002

[WEBDAV]
WebDav standard for collaboratively authoring web standards
☞http://www.webdav.org/

[WEB-OF-SCIENCE]
ISI Web of Science
☞http://www.isinet.com/isi/products/citation/wos/index.html

[WIKI]
WikiWikiWeb
☞http://c2.com/cgi/wiki?WikiWikiWeb

[WMDI]
The Weblog metadata initiative

☞http://www.wmdi.org/

[XFML]

eXchangeable Faceted Metadata Language - core specification

☞http://www.xfml.org/spec/1.0.html

[XPOINTER]

The XML Pointer Language

☞http://www.w3.org/TR/xptr/