

SWAD–Europe Deliverable 12.1.1: Semantic web applications – analysis and selection

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 web applications - analysis and selection

URI:

http://www.w3.org/2001/sw/Europe/reports/chosen_demos_rationale_report/hp-applications-selection.html

Authors:

[Dave Reynolds](#), Steve Cayzer, Ian Dickinson, HP Laboratories, Bristol, UK
Paul Shabajee, Graduate School of Education and ILRT, Bristol, UK

Abstract:

This report concerns the selection of two open 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.

We start with a discussion of the nature of the semantic web and in particular what the key aspects of it are that should be brought out by the demonstrators. Then, after a brief summary of the roles that the demonstrators play within the overall SWAD-E project, we look at existing and proposed semantic web applications.

Finally we describe our two chosen demonstrators - semantic blogging for bibliographies and semantic community portals.

Status:

First release.

Comments on this document are welcome and should be sent to [Dave Reynolds](#) 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](#)
 - 2 [The nature of the semantic web](#)
 - 3 [Role of applications within SWAD-E](#)
 - 4 [Criteria for selection](#)
 - 5 [An overview of the application space](#)
 - 6 [Demonstrator 1: semantic blogging and bibliographies](#)
 - 7 [Demonstrator 2: semantic community portals](#)
 - A [References](#)
 - B [Application survey](#)
 - C [Blogging and semantic blogging](#)
 - D [Changes](#)
-

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.

The aim of this report is to select the two specific demonstrators to be developed and provide the rationale behind that choice. We have also tried to put this choice into context. In particular, we offer a picture of what the key features of the semantic web are that the demonstrators should illustrate, together with a survey of many known or proposed semantic web applications.

This report is not intended to specify the detailed functionality or architecture of the chosen applications. Separate deliverables are scheduled to cover these.

We start with a discussion of the nature of the semantic web and in particular what the key aspects of it are that should be brought out by the demonstrators. Then, after a brief summary of the roles that the demonstrators play within the overall SWAD-E project, we look at existing and proposed semantic web applications. We have grouped the applications we are aware of into different categories and in the body of the report we just offer an overview of

these categories - details on the specific applications surveyed is included in [Appendix B](#). We hope to also make this survey available in RDF format.

Finally we describe our two chosen demonstrators and the reasons for selecting them.

2 The nature of the semantic web

Overview - Much has been written about the nature of the semantic web [\[SEMWEB\]](#) [\[SCIAM\]](#) and at first glance the notion is fairly straightforward. The existing world wide web allows anyone to publish human readable web pages that can be connected via hyperlinks. The combination of a common format for marking up such web pages, common access protocols to allow client applications (browsers) to access and view the data and universal hyperlinking, has transformed the way we publish and access information. A simple description of the semantic web is that it is an attempt to do for machine processable data what the world wide web did for human readable documents. Namely to transform information processing by providing a common way that data can be accessed, linked together and understood. To turn the web from a large hyperlinked book into a large interlinked database.

There are several motivations for this.

First there is the view that simply making data available is an end in and of itself and will lead to benefits and applications. Just as the world wide web gave everyone access to information that previously would have been locked away in local file systems or local networks, the semantic web can unlock access to data currently hidden away in databases, freeing that data to be accessed by applications and tools across the globe. In particular the ability to link data from different data sources together allows us to explore data in new ways and discover new relationships and correlations.

Secondly, there is the appeal of automated processing of this information. As long as the data we share across the web is, as now, primarily in natural language there are limits to the ways that our software systems can add value to this information because robust natural language processing is beyond the current state of the art. Current text processing techniques are sufficient to allow us to index and retrieve such documents but not to perform any meaning processing on their contents. We can't, for example, robustly extract the details (artists, locations, dates) from an article on upcoming concerts to check them against our diaries. We can't, other than by using fragile web scraping techniques, aggregate price information and ratings from different product sites to assist with purchases. Making such information available in machine interpretable form means that we can build applications which actively process this information - collect it, analyze it, filter it, correlate it, link it and apply it to the task at hand.

Thirdly, there is the sense in which the semantic web is very much an extension to the current web. The common representation for data allows us to attach semantic information, metadata, to the human readable web - allowing people and machines to work in closer cooperation. This enables applications such as semantic search - search engines that "understand" the difference between computer chips and potato chips, and can therefore present a user with semantically relevant results rather than just syntactic matches.

Technology layers - To achieve this vision the semantic web is built as a series of layers. These are not layers in a strict software architecture sense but levels of functionality. At the base level the semantic web builds on the rest of the web infrastructure - HTTP transport, URIs for naming and location, XML as a common syntactic format. On top of this existing infrastructure the semantic web adds two key pieces:

- A common representation for semi-structured data and metadata, RDF [\[RDF\]](#).
- A common representation for ontologies that enable the terms used by the data layer to be defined and related to each other [\[RDFS\]](#) [\[DAML\]](#) [\[OWL\]](#).

The roadmap for the semantic web also offers a vision of future layers [\[SEMWEB LAYERS\]](#) to support richer knowledge representation and to provide a trust infrastructure so that the results inferred from semantic web data can be traced back to the assumptions that lead to them. However, our aim in these demonstrators is to illustrate the semantic web as defined so far - other work packages will explore these other layers, especially the important trust issues.

Features - There are three core aspects to the semantic web which we feel are critical to capture and illustrate in the open demonstrations.

Data representation

The foundation of the semantic web is a common format, RDF, to represent data. This format is designed to be suited to representing semi-structured data and metadata. Data is broken down into conjunctions of individual assertions in the form of subject/predicate/object triples. Each of the components (other than simple literals) is a web URI and thus has a defined place in the global namespace. This allows many sorts of data (property values of objects, relationships between objects, value annotations) to be represented uniformly and allows data from multiple locations to be combined without accidental clashing of property names or structure mismatches.

Semantics

The aspiration of the semantic web is to be able to express *meaning*. It is the second layer of the semantic web - the schema and ontology layer - that begins to do this. It enables the properties and types used in the data layer to be related to each other. To say, for example, whether two terms are distinct, or equivalent or whether one term is a subset of another. This capability allows a data source to expose its conceptual model explicitly in machine processable form thus allowing a software agent accessing it to make decisions on how the data can be processed and what the semantic relationship is between data from different sources. Ontologies do not provide an absolute way of conveying semantics. They allow classes and properties be related to other known classes and properties thus allowing the meaning of new terms to derived from

combinations of other known and "understood" terms. However, the semantic web does not require some standardized global upper ontology to function. Like the web, it remains decentralized so that data sources are free to mix and match terms from different ontologies - so long as two entities share a common ontology they can communicate.

Webness

There is nothing new about either semi-structured data representation or explicit representation of conceptual models through ontologies. The critical innovation of the semantic web does is put both of these concepts into a web framework. This is manifested in deceptively simple ways such as the use of URIs to provide a global namespace for both entities and concepts (properties, types). However, the impact is substantial. An agent accessing a data source now has a means for discovering the ontology associated with that data source. Ontologies can be developed in a decentralized way to suit particular needs, but the terms defined in different ontologies can be related and combined to enable transformation of data from one domain to another.

Each of these features is important but it is the combination of all three that forms the fundamental nature of the semantic web and our demonstrators should ideally illustrate that combination. This is a critical point to emphasize - there is a difference between applications which happen to use parts of the semantic web stack and applications which serve to demonstrate the vision of the semantic web itself. For example the Mozilla browser [\[Mozilla\]](#) uses RDF internally to represent the structure of mail messages, web links and so forth. This is a great use of RDF but it lacks the use of deeper semantics or the webness to be an illustration of the full semantic web vision. Similarly, there have many applications of ontology technology (and indeed richer knowledge representations) over the years which do not themselves illustrate their role in connecting data representations across the web.

3 Role of applications within SWAD-E

Before we look at example applications we should clarify what the role of the application work is within the SWAD-E project. We see two different classes of role - communication and investigation.

Communication - A key role for the demonstrators is to illustrate the nature and value of the semantic web, to enable both users and developers to understand the potential benefits. It is primarily to meet this requirement that we prefer that our applications attempt to illustrate the semantic web concept as a whole and avoid concentrating too strongly on, for example, just the ontology aspects.

As well as communicating value and potential, the demonstrations should also communicate practicality and feasibility. A core aim of the SWAD-E project is to ensure enough of the tools and understanding are in place to allow practical development of serious semantic applications. The demonstrators should show that existing tools and techniques are indeed sufficiently mature to support such applications or give specific guidance on any current limitations.

Investigation and analysis - The second role of the demonstrators is to test the current capabilities and limitation of existing semantic web standards and toolkits. By pushing the technical boundaries, the demonstrator work should generate advice for current developers on practical limitations, important feedback to toolkit developers on key requirements for future tools and guidance for future standards development.

The semantic web is an ambitious vision and fully realizing it will require substantial research - not simply engineering development. While the aim of the demonstrators is not to tackle such research issues head on, they do have an important role in probing the boundaries of these hard problems to determine which issues can be worked around and which are the critical ones that should be targeted by future needs-driven research investments.

In particular, in exploring potential applications we found that the issues raised by the semantic web vision of many decentralized data sources with separate and evolving ontologies seem particularly important. Such issues include:

- encouraging ontology convergence rather than explosive divergence, perhaps by facilitating ontology reuse
- efficient data source combination (query routing aspects of distributed query, ontology transformations needed to query divergent sources) over high latency networks
- coping with conflicts and inconsistencies in both the data and the merged ontologies - the semantic web will include many different and divergent conceptualizations of the world, it should be possible to discover and reason with these disagreements

As well as investigating the technical issues of semantic web applications the demonstrators should also give some insights into the social and economic issues of uptake. The semantic web is, by definition, a *network effect* technology [\[NETWORK EFFECT\]](#). Its value depends on its deployment and vice versa. Such issues include the question of how to seed community ontologies (too top down and they are rigid and slow to develop, too bottom up and you have too much divergence for the network effect to kick in) and how to encourage adoption of common access and linking approaches in the absence of processing standards.

There is clearly a conflict between these two requirements. To meet the needs of communication, publicity and advice to current developers then the demonstrators should be modest, low risk affairs picked for their ease of comprehension. To begin to probe the research boundaries and give useful feedback on the limitations of current technologies and guidance for future investment the demonstrators should be ambitious and deliberately touch on some of the research issues noted above.

We address this conflict in two ways. Firstly, by choosing one demonstrator which we believe to be in the low risk, easy uptake category and one which is more risky in involving more serious issues of multiple ontologies. Secondly, in both cases we chose a broad area with a modest initial core so that each demonstrator should be

expandable in many directions to explore a variety of research and engineering issues when appropriate.

There is also a risk that our emphasis on practical illustrations of feasibility will lead to demonstrators that fall short of the hyped expectations that are being generated about the semantic web. We regard this as an entirely acceptable risk - communication not evangelism is our aim.

4 Criteria for selection

Given this view of the key features of the semantic web and the role of the SWAD-E open demonstrators we can then summarize our selection criteria as follows:

- illustrate the overall semantic web vision, good mix of semi-structured data, webness and deeper semantics;
- a good chance of impact and uptake across a large enough community to become a live application and not simply an artificial demonstration;
- the semantic web aspects of the demonstrators should be visible and apparent to the end users and not hidden behind the scenes;
- a balance, across the pair, of low risk, low cost of entry and higher risk exploration of the technical boundaries;
- basic feasibility, a meaningful core application must be constructable within the 9 man month bounds, this feasibility includes aspects of availability of data as well as technical feasibility and also suggests that the bulk of the work should be on the semantic web core and not for example on user interface issues;
- where the application requires significant community support (in terms of user community or markup of data) we must have reason to believe that the community can be motivated to collaborate on the project;
- the application and domain area should match the interests and expertise of the builders and of their parent organization.

5 An overview of the application space

To get a deeper understanding the the space of possible semantic web applications we conducted an initial survey of applications including current known or completed projects and prior suggestions and proposals. We don't claim that this survey is comprehensive but we were able to find enough examples (approximately 60) to give a reasonable picture. Summary information on these applications (brief descriptions, links, status) have been capture in RDF format and a subset of this information (translated to HTML) is included in [Appendix B](#).

We then conducted an initial informal clustering exercise which led us to identify some 11 categories of types of application. A description and discussion of each category is included below. This categorization is imperfect - the categories overlap somewhat, applications appear in multiple categories, and some category pairs could be merged without a great loss information. However, this level of structure has been very useful to us. It gives enough detail to provide a good overview of the different ways in which the community believes the semantic web can be applied; while at the same time it is rather more succinct than the raw application data and helps one to see the wood for the trees.

In addition to classifying the applications into these type categories we have also looked at other dimensions of classification, in particular in terms of the domain to which the applications are applied. This is valuable in understanding the breadth of current semantic web explorations though is not a primary criterion for us in choosing a demonstrator - we are fairly agnostic about the information domain itself.

One interesting alternative classification dimension that could be explored further in future analysis is that of information lifecycle phase. Some of our application categories emphasize the use of metadata for management of information in the creation and storage phases of the lifecycle, others during the discovery and selection phases, or others still the application and delivery phases. Traditionally the metadata used in the early lifecycle phases is hidden away in the internals of the data format or the particular content management system used. The semantic web approach to making such metadata externally visible might enable this data to be reused in later phases of the lifecycle. This "end to end" use of semantic metadata could be a powerful motivator for the semantic web and further work in developing an information lifecycle model that explores this aspect could well be fruitful.

Semantic web application categories

1. Data integration

In this class of applications we use the semantic web as a way of exporting data from multiple datasources to allow integration and cross-source queries. Several categories of applications involve some data integration but the essence of this class is that the data itself is seen as having substantial value and simply "freeing" the data and providing cross database query is a value in its own right. Thus the user of the application may be simply and explicitly issuing queries to the merged data sources or viewing the information.

Examples

- [B2B web service mediation](#)
- [Catalogue integration](#)
- [Database integration example](#)
- [Financial Portals](#)
- [Gene Ontology](#)
- [Mozilla](#)

- [eScience Data Grids](#)

Discussion

This class of applications primarily illustrates the common data-format aspects of the semantic web. Clearly there is an element of distribution but many near term practical applications will be intranet scale and applied to carefully selected clusters of databases - the network effect of webness is only partially present. The depth of semantics and ontology support can be quite significant here. The data sources will have typically been designed with a specific narrow set of queries in mind and integrating the different data schemas to support cross-source query may require nontrivial concept translation.

2. Data-dependent agents

This cluster of applications is one where some software entity (which we shall, informally, call an agent) is providing a service to a user that is only possible if a rich and inhomogeneous set of data can be integrated. In terms of technical work and challenges this is virtually identical to the data integration cluster but in this case it is the operations carried out by the agent that defines the end-user value - the data integration itself is but a means to that end.

Examples

- [Financial Assistant](#)
- [ITTalks](#)
- [Jema](#)
- [Shopping assistants](#)
- [Virtual Travel Agent](#)

Discussion

Whilst the semantic web issues here are very similar to those in the pure data integration category, the data sources tend to be less homogenous and more widespread so that this category has a greater "webness" score. For example, the shopping assistant has to not only aggregate price and offer information but also ratings and evaluations from prior customers and has to translate between the customer's specification of the desired article or service and the different descriptions used in the data sources. This is certainly an important class of applications for the semantic web - witness its prominent role in [\[SCIAM\]](#). It is also one of the easiest to communicate - the shopping assistant scenario is probably one of the most compelling we have looked at. However, from the point of view of a demonstration of value it suffers from a high cost (you have to build an effective software agent as well as succeed in the data integration challenge) and that the value delivered from the semantic web aspects of the work is indirect; it enables "cool" stuff but is neither itself "cool" nor visible.

In practical terms this category of application is likely to be slow to take off in the semantic web. The value to the data providers in making their data available is often not clear and the agents cannot deliver their value until a sufficiently comprehensive set of data sources is available. This can be partially overcome by using screen scraping techniques to artificially make data available, e.g. in the Isoco personal financial aggregator [\[GetSee\]](#). Once enough such applications are available and the economic model becomes clearer then network effect should make this a high value area for the semantic web in the long run.

3. Knowledge management

Knowledge management is a well defined field of research and technology [\[KnowledgeManagement\]](#) which comprises several different classes of application - from community formation, through collaboration support to enterprise knowledge preservation. It would be thus be possible to subdivide this category further into these component subfields.

The essence of the term *knowledge* as used in the knowledge management field is *applied information*. Simply storing or organizing information is not sufficient to turn it into knowledge, knowledge in this context is taken to be the ability to harness that information to solve actual problems. Unless people are able to apply the information to the task at hand the knowledge is useless.

The common factor in these examples is that the collective knowledge of some community is expressed in some information form such as a set of documents (case studies, past problem reports, notes on a bulletin board) and semantic web techniques can be used to classify and structure the document set to allow it to be matched against a problem. Ontologies provide the key tool for this classification.

Examples

- [Community formation](#)
- [Community portals](#)
- [Helpdesk support](#)
- [OntoShare - community of practice support](#)
- [PatMan](#)
- [PlanetOnto](#)
- [Sun GKE](#)
- [ePerson](#)

Discussion

This is an very important application area of substantial commercial importance which is certainly the target of many semantic web related projects. It primarily exploits the ontology management aspects of the

semantic web stack. The documents themselves will often not be particularly structured and there may be no machine processable information beyond the document classifications. It is also an area that is typically applied to a specific community, often within a single discipline and within a single organization - and as such may neither benefit from nor illustrate the web nature of the semantic web. However there will be applications of this class which transcend that generalization and so could illustrate data integration and the global "webness" of the semantic web to an adequate degree.

4. Semantic indexing and semantic portals

The web is already replete with examples of document integration, providing organized access to large collections of information in the form of web links. These include topic-specific portals, generic structured directories like [Yahoo!](#) or [DMOZ](#) and information retrieval directories like [Google](#). The semantic web offers the possibility that such portal services could be based on deeper categorizations of links exploiting rich ontologies. In particular, the categorizations, topic tags and other annotations associated with the indexed resources may be drawn from many locations and communities and integrated by the portal, rather than, as at present, being entirely synthesized by the portal. Further, if the indexing is based on some deeper underlying semantics, many different structured views could be synthesized to map onto the same resources via the same set of semantic tags. Curriculum Online [\[CurriculumOnline\]](#) is an example of this where educational resources are tagged according a 2,000 term topic ontology, which is then mapped onto the curriculum structure. This allows the tagging to remain valid despite changes in the curriculum. One interesting variation on this theme is where the semantic-driven lookup is applied in parallel with standard web operations like searches (e.g. [TAP](#)) or link following (e.g. [context aware links](#)). Here there is the added challenge of mapping a user's unstructured query onto a structured semantic space and using that additional semantic information to both disambiguate and enrich the user's query.

Examples

- [Community Arkive](#)
- [Community portals](#)
- [Context aware links](#)
- [Curriculum Online](#)
- [Distributed topic portals](#)
- [HP Portal](#)
- [ITTalks](#)
- [Museum portals](#)
- [MusicBrainz](#)
- [PlanetOnto](#)
- [Score](#)
- [Semantic tagging](#)
- [TAP semantic search](#)

Discussion

This group of semantic web applications is particularly strong at showing the connections between the semantic web and the human-oriented world wide web.

If we are dealing with a single portal with a single modest underlying ontology then the research challenges are small and there already several examples of such systems. However, despite the connections to the existing web, this class of applications may not fully illustrate the webness of the semantic web because often the ontologies and the original data remain hidden behind the scenes. There are exceptions to this where these structures are deliberately exported and exposed (e.g. TAP) or where data and ontologies from multiple sources or multiple communities needs to be combined (e.g. community Arkive, distributed topic portals). The latter cases, where data spanning multiple ontologies needs to be combined, rapidly pushes such applications over into the higher research content (and thus higher risk) zone.

5. Personal information management

This category is concerned with applying semantic web techniques to help individual users manage their own information. Several of these have a strong community aspects of sharing both information and categorization schemes, but they are all person-centric - they are managed by the user primarily for their own benefit. In contrast the examples in knowledge management category are typically managed by some specific organization such as an employer on behalf of the organization's collective good.

Examples

- [Bibliography workbench](#)
- [Community bookmarking](#)
- [Event tracking](#)
- [Genealogy assistant](#)
- [Haystack](#)
- [Ideas workbench](#)
- [Jema](#)
- [Mozilla](#)
- [ePerson](#)

Discussion

The applications in this category are primarily exploiting the semi-structured data representation layer of the semantic web. By translating the many different data objects an individual has to manage (events, appointments, phone lists, mail lists, mail headers, filing categories, action lists) to a common RDF format then it is easier to build highly reusable and extensible tools and to link data across multiple formats. As in the last category, an issue with these applications from the point of view of our demonstrator goals is that the transformed data may remain hidden inside the application and the benefits of the approach may not be visible to anyone other than application developers (see Mozilla for example). These applications do involve some format translation and hence some schema mapping element but are typically not exploiting the semantics aspect of the semantic web in a particularly deep way.

However, for applications where the sharing of information is a key feature and where some rich categorization or ontological structure is involved then this category can be an excellent source of demonstrators. It has the strong attraction that such applications can be immediately useful to a single individual or a small group and can then grow in value as the network of users grows - there is not the barrier of making substantial external data sets available or artificially stimulating a substantial "ignition" community that arises in some of the other categories.

6. Metadata for annotating and enriching

The foundation layer of the semantic web, RDF, was originally designed as primarily a format for metadata. So many semantic web applications are aimed at some aspects of metadata management that we found it useful to distinguish between several subclasses of metadata applications. In this category we see applications where the metadata is intended to be directly visible to the end user; it is there to annotate or enrich the data itself. Typically this is used as a means for the viewer of a resource to also see the comments, opinions and ratings from a larger community of users.

Examples

- [🔗 Annotea](#)
- [🔗 Assumption tracker](#)
- [🔗 Bibliography workbench](#)
- [🔗 Community Arkive](#)
- [🔗 Community bookmarking](#)
- [🔗 Distributed topic portals](#)
- [🔗 EARL](#)
- [🔗 Gene Ontology](#)
- [🔗 MIT/HP SIMILE project](#)

Discussion

This category is an excellent demonstration of the webness and semi-structured data aspects of the semantic web. The common data format allows many different annotations and metadata to be attached to the same underlying object without barriers or restrictions ("*anyone can say anything about anything*"). Those annotations can then be aggregated and organized for access, opening up new channels of communication between users.

Typically the depth of the semantics is low. The annotations and enrichment are often generated by humans for humans and are not machine processable in any important way (except perhaps for numerical rating schemes). One exception to this is where the annotations include some classification of the annotated resources - combining different classification schemes is a core semantic web issue. However, such applications are typically in the overlap between this category and the next one - metadata for discovery & selection - and are discussed there.

7. Metadata for description, discovery and selection

In this category the metadata is primarily used to help a user locate a resource, product or service to meet their needs. The division between this and the *metadata for enrichment* category is blurred but essentially the idea is that in this category the metadata is primarily of value during some search task and adds less value during any subsequent phases. This category is also closely related to those of *semantic indexing* and *knowledge management*.

Examples

- [🔗 B2B trading market-places](#)
- [🔗 DCMI registry](#)
- [🔗 Edutella](#)
- [🔗 HP Portal](#)
- [🔗 MIT/HP SIMILE project](#)
- [🔗 MUSE](#)
- [🔗 Recommendation Networks](#)
- [🔗 Scholnet](#)
- [🔗 SeLeNe](#)
- [🔗 Semantic tagging](#)
- [🔗 Sun GKE](#)
- [🔗 Web service description and discovery](#)

Discussion

It is possible to treat any of the text annotations created by applications in the earlier - metadata for enrichment - category as additional descriptive terms to search on. However, the essence of this group of applications is that some more structured representation of the descriptive properties is used - classification into a defined taxonomy, property annotations using a controlled vocabulary, numeric and symbolic descriptive properties. This enables search tools to offer discovery, comparison and selection functions which are semantic based and are thus more selective and less ambiguous than those based simply on text retrieval techniques.

This is a core semantic web application area and nicely combines the web-of-metadata features of the last category with clear illustration of the value of the more explicit semantics. Furthermore the user is directly accessing the metadata in formulating searches and viewing results and so the semantic web features are less hidden than in, say, the *data-dependent agents* category.

The drawbacks to this category are firstly that it can be hard to bootstrap - a sufficient fraction of the universe of resources being searched need to be semantically annotated before the discovery and selection tools become useful. Secondly, it is quite well represented already by existing and current projects indeed there is already an annotation application within the SWAD-E workplan.

8. Metadata for media and content

An important use for metadata is as a tool for content or media management. Here the value of the metadata is mostly seen during the creation and archiving of the content - it may not be very visible or valuable to the end users of that content.

Examples

- [☞ Adobe XMP](#)
- [☞ Arkive internal](#)
- [☞ MIT/HP SIMILE project](#)

Discussion

Here the metadata is primarily used for management of the annotated objects. It is similar to the personal information management category in that it emphasizes the common metadata format aspects but can lack rich semantics or webness - the metadata helps the content producers and the archive maintainers but is not visible to the end user as much as the earlier two metadata categories.

This "RDF inside" category of applications is still extremely important in the development of the semantic web - the use of RDF for such technical and management metadata eases its use for richer external metadata. However, it is not a direct illustration of the whole semantic web vision.

9. Knowledge formation

In application classes such as *semantic indexing* and *knowledge management* we saw that resources were classified and indexed as a means to an end such as improved search or better problem solving. However, in the knowledge formation category this classification and relationship knowledge is of primary value in its own right. Those involved in these communities are consciously creating new organizational structures which have value beyond the resources themselves.

Examples

- [☞ Assumption tracker](#)
- [☞ Bibliography workbench](#)
- [☞ ClaiMaker/Scholonto](#)
- [☞ Community Arkive](#)
- [☞ Community formation](#)
- [☞ DMOZ - Directory Mozilla - open directory](#)
- [☞ Ideas workbench](#)
- [☞ SWAP - semantic web and peer-to-peer](#)

Discussion

This is an intriguing category. It is putting the representation of structured knowledge at the forefront and highlights the semantic side of the semantic web and the differences between that and the current web quite well. It is appealing to think that by enabling communities to cross-link and structure collective information this way, new insights will be gained that would not have been apparent from simple text indexing. Even without such speculative benefits this class of applications is building the web-based semantic structures that other categories of applications can then exploit - for example the TAP Knowledge base is foundation for the TAP semantic search application.

A possible drawback to this category is those applications aimed narrowly at knowledge formation rather than its application may appear a little niche and may thus be less convincing concerning the broader value of the semantic web. However, those broader applications that fit in the overlap between this category and some of the others are strong potential candidates for our demonstrators.

10. Catalogue and Thesaurus management

Structured and controlled vocabularies of terms play an important role in many applications ranging from digital libraries (Thesauri) through to B2B market places (catalogues). The management of these structures

can be challenging, as the continued evolution of world being described causes the creation of new terms and the restructuring of existing branches. Further, the ontology representation techniques used in the semantic web offer the possibility of richer representation of such categorization schemes than the simple hierarchical keyword trees often used. This application category is focused on just this issue of management of vocabularies as a distinct requirement from their creation (e.g. *knowledge formation*) or application (e.g. *semantic indexing*).

Examples

- [☞ Catalogue Management](#)
- [☞ Catalogue integration](#)
- [☞ DMOZ - Directory Mozilla - open directory](#)
- [☞ Thesaurus management](#)

Discussion

This is certainly an important and challenging application area. There is significant commercial interest in tasks such as product catalogue management and integration, and there are many interesting research challenges in the semi-automated mapping of such catalogues ☞ [Fensel2002]. As a demonstration of the overall semantic web vision, however, such applications are not ideal in that they focus almost exclusively on the ontology representation issues to the exclusion of the webness and common data representation aspects. This is also an area already being explored with the SWAD-E program (work package 8) so for us to build another demonstrator focusing entirely on this issue seems unnecessary. However, this is such a core problem that many of the potential demonstrators will have some aspect of vocabulary management.

11. Syndication

In this final category lie applications where the semantic web representations are used as a common format for broadcasting metadata around some network of users. In this case we are not necessarily indexing, classifying or annotating - merely disseminating the information.

Examples

- [☞ Event tracking](#)
- [☞ Rich Site Summary/RDF Site summary](#)
- [☞ Syndication](#)

Discussion

This is a challenging application area to analyze. On the one hand the whole world of *blogging* ☞ [Appendix C] and ☞ RSS is an immensely successful and interesting growth area for the world wide web and is an excellent illustration of how a simple common metadata format can support aggregation and filtering of content streams in useful and effective ways. On the other hand, for the bulk of current applications any centrally agreed representation would have worked and indeed there is still violent disagreement over whether the pure XML or the RDF based approaches to RSS are most appropriate. It terms of developer evangelism the use of RDF here has been of mixed success - the apparent additional complexity of RDF has not yet been fully offset by real exploitation of the additional power it brings. Curiously some of the webness aspects of the semantic web do not come across that clearly from this application. A single common global schema is useful, but the power of the semantic web in supporting the combination of different sorts of data is only beginning to be explored. Despite these reservations this area does offer an excellent infrastructure and design approach for lightweight publishing and dissemination of structured metadata. Building upon this but extending it towards applications which involve richer and more varied structured data - *semantic blogging* - is a prime candidate for a demonstrator.

Given this discussion of semantic web application categories as related to our selection criteria (outlined in ☞ Section 4) we can see that the most relevant application categories are those of semantic indexing, knowledge formation and to some extent personal information management, knowledge management and syndication.

From this analysis we created a short list of 10 applications:

- bibliography workbench
- community arkive
- ideas workbench
- digital library metadata applied to DSpace
- Gene database ontologies
- virtual community support
- distributed product naming
- personal information and content management based on the ePerson prototype
- semantic portals
- distributed topic portals

These were then filtered and combined to arrive at the final two proposals.

6 Demonstrator 1 : semantic blogging and bibliographies

Our first chosen demonstrator takes the semantic blogging ideas touched on in the last category above and applies them to a specific application domain: bibliography management.

The semantic blogging core of this demonstrator will develop a generic framework that could be applied to many different tasks where a user community is incrementally publishing structured and semantically rich (categorized and cross-linked) information. It could thus be extended to encompass other proposals on our short list - such as the ideas workbench or the distributed topic portals. This generality is a source of risk; unless a specific domain is chosen there is not enough application feedback to enable the team to focus on just core values and key technical challenges.

The bibliography management domain has the attraction of being very specific with much available data, both personal data and network accessible resources such as [CiteSeer](#). It helps to focus the semantic blogging area down nicely. Whilst bibliography management is an important task in the research community, it could be seen as a niche application in the wider community. However, the same tools and approaches will be as applicable to dissemination and management of other content such as business documents or news items. By starting with a specific, but widespread, task of personal interest to the developers we aim to keep the work focused and relevant. Generalizing the results to related areas will be straightforward.

Semantic blogging

Web-logging, typically abbreviated to "blogging", is a very successful paradigm for lightweight publishing which has grown sharply in popularity over the last two years. The notion of semantic blogging builds upon this success and clear network value of blogging by adding additional semantic structure to items shared over the blog channels. In this way we add significant value allowing navigation and search along semantic rather than simply chronological or serendipitous connections. We provide extra background on the blogging phenomenon and its extension to semantic blogging in [Appendix C](#).

Blogging, as it stands, already offers many compelling values. 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.

Semantic blogging exploits this same personal publishing, syndication, aggregation and subscription model but applies it to structured items with richer metadata data. The metadata would include classification of the items into one or more topic ontologies, semantic links between items ("supports", "refutes", "extends" etc.) as well as less formal annotations and ratings. There are several ways this more structured data could extend the power of blogging:

- **Discovery.** At present it is not easy to discover either a channel of interest (e.g. "I would like to find blog channels about the semantic web") or a collection of specific items of interest (e.g. "Are there any more blog entries describing this application idea?").
- **Cross-linking.** Current blogs support a single link between the channel record and the blogged item. By extending this mechanism to support linking between items (using a property hierarchy) we can create a network of topic interconnections that supports more flexible navigation. These links can themselves form part of the disseminated content - for example to represent the structure or scholarly discourse (c.f. [ClaiMaker/Scholonto](#)).
- **Flexible aggregation and selection.** The current blog subscription mechanisms are in some ways both too fine (being bounded by the individual blogger's channel of posts) and too coarse (e.g. I might like Ian's technology channel but am only interested in the semantic web bits). The richer categorization and structure of semantic blog channels would make it easier for users to create virtual blog channels which aggregate across multiple bloggers but select from that aggregate according to other criteria such as topic (or community rating).
- **Integration with other sources and applications.** The structured nature of semantic blog channels makes it possible to develop automated blog robots that can process and enhance the blogged items. For example, in the bibliography domain transducers would enable import and export via existing bibliography schemas like BibTex [BibTeX](#) and automatic linking to large repositories such as CiteSeer.

Bibliography management

Management of citation databases is a recurring problem in scientific and research domains. Many tools exist which support good integration between a personal bibliography and word processors, e.g. Endnote [EndNote](#) and ProCite [ProCite](#). However, the ability to index and annotate the citations in these tools is often limited with a lack of support for structured or controlled indexing vocabularies.

A researcher's database of monographs they have read, together with their annotations and categorization, is a valuable resource not only to that researcher but potentially to others in the same field. It may help others discover references they were not aware of themselves and the commentary and evaluation associated with the records can be an invaluable summary and guide. This collaborative discovery and evaluation is currently limited due to the inaccessibility of personal bibliographies and the weaknesses of current bibliography standards when it comes to representing rich community annotations.

The semantic web approach to representation of bibliography entries, their annotations and their classifications could have several benefits outlined below. This could be approached in several ways - as a centralized citation

repository, as a personal information management tool or as a community sharing tool. It is the latter option that interests us, both because of its intrinsic value and because it more clearly indicates the semantic web values than either of the other two schemes. Thus we propose to take the semantic blogging approach sketched above and apply it initially to the management and dissemination of citations and associated commentary. We see several benefits in this approach:

- **Community categorization and 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.
- **Inter-document relationships.** Traditional bibliography schemes represent just properties of citations but not relationships between them. As well as the direct citation relationship ("this paper cites this set of prior papers") represented by the Science Citation Index [\[ISI\]](#) and CiteSeer [\[CiteSeer\]](#) there is value in representing the implicit links between the referenced works and the semantics of those links such as "provides evidence for", "refutes", "supports". Some groups have already started to explore this area both in general and more recently in a semantic web context ([ClaiMaker/Scholonto](#)).
- **Annotation.** By indexing bibliography entries by URI (or indirectly by property patterns) we enable rich annotation of the entries from different sources to be integrated. For example, comments and ratings from a peer community can be created decentralized and aggregated. Further more their origin or provenance of such annotations can be explicitly captured and traced.
- **Subscription.** By using the semantic blogging approach with its inherent time-based channel structure we make it possible to subscribe to categories such as "what's going on in field x". Very often the personal reading patterns and comments of a group of colleagues with similar interests is an excellent guide to the important papers in a field and enables one to keep up to date with background topics without being overwhelmed. Informal email networks are useful for this, but they tend to be closed and hard to discover. Centrally maintained topic portals are easier to find, but tend to be out of date and may not offer the best tradeoff between completeness and selectivity. The semantic blogging infrastructure, with its ability to discover and selectively aggregate feeds, offers an excellent alternative, enabling subscribers to plug into the recommendations and highlights from peer readers.
- **Citation of components and other media.** Finally the use of web infrastructure to designate and refer to the content items means that the citation infrastructure can be extended to include references to other web addressable objects such as media objects and to sub-components of objects (though use of XPATH [\[XPATH\]](#) and similar mechanisms).

Clearly not all of these features will be achievable within the bounds of this project. However, a functional and interesting core demonstrator is manageable and future extension of this core to deliver some of the other features listed above could be the subject of further open source community development. Defining the precise boundaries of the initial core demonstrator will be the subject of the next package of work and will be reported as part of deliverable 12.1.2 - *requirements analysis*.

7 Demonstrator 2 : semantic community portals

Given the inherent extensibility of our first choice of demonstrator it is tempting to make the second demonstrator a variant on the same theme. We felt, however, that this would lack balance and that it was better to choose a reasonably distinct second application to explore a different cut at the semantic web development and research issues.

For our second application we have chosen the broad area of semantic community portals.

Again we need to select a specific application domain to ground this application and turn it into a feasible demonstrator. There is a difficulty in doing so for this application in that we really need an external user community that is in a position to provide requirements, feedback on early prototypes and most importantly the metadata content itself. Our initial choice is to develop an external community portal for a subset of the Arkive [\[ARKive\]](#) media repository. However, at this stage it is not certain that the appropriate community links can be put in place. We may need to switch our focus to another similar application in a different domain - such as a related environmental biology topic like birds or 'mini-beasts' as studied in the UK National Curriculum for schools, or a more generic repository such as the DSpace digital library [\[DSpace\]](#). Our proposal is to explore the practical issues of establishing a suitable set of community links in parallel with the development of demonstrator 1 so that these issues will have been resolved before the scheduled start of demonstrator 2.

The notion of semantic portals was introduced earlier in [Section 5](#). The idea is that a collection of resources is indexed using a rich domain ontology (as opposed to, say, a flat keyword list). A portal provides search and navigation of the underlying resources by exploiting the structure of this domain ontology. There may be an indirect mapping between the navigation view provided by the access portal and the domain semantics - the portal may be reorganized to suit different user needs while the domain indexes remain stable and reusable. This indirection is exploited, for example, in the Curriculum Online project [\[Curriculum Online\]](#) in which the a 2,000 term ontology of education concepts is used in the annotation of educational resources whereas the access portal navigates these annotated resources according the current UK national curriculum requirements. The mapping from user search or navigation terms to the domain ontology may itself be an inferred step - as in the TAP semantic search demonstrator where free text search terms are matched to property and class labels in the domain ontology to support semantic augmentation of a conventional keyword search.

We used the qualifier *community* in the description of this demonstrator for several reasons. Firstly, we are

particularly concerned with applications where some external community is cooperating to develop the semantic indexing - both developing the ontology itself and the categorization of the resources. Secondly, we are looking at applications where in fact several communities with different interests in the same underlying resource set need different but overlapping categorizations. This combination enables us to emphasize the web connectedness of the ontologies and indexed resources and gives us an opportunity to explore the ontology development, reuse and mapping issues raised by the semantic web.

Our preferred starting point for this application is the Arkive media repository. This is a long term archive of rich multimedia about worldwide endangered species and a large cross-section of non-endangered UK species. Prior work [Shabajee2002] has indicated that many different user groups have interest in structured access to such data. These include:

- School teachers, support staff, pupils and third party resource developers - who need to relate the content to the UK national curriculum;
- Higher Education (HE) and Further Education (FE) tutors and students - for example, to illustrate lectures with appropriate multimedia content depicting relevant animal behaviours;
- Researchers - whose interests may be specific to only some species or phyla and may span behavioural, habitat and environmental issues; and
- "Hobby" groups - for example bird watchers who are both interested in accessing such multimedia resources and who have much to offer in terms of current information on sightings and distribution.

This is a domain where there is a rich taxonomy of species information (though some scholarly disagreements remain) but a lack of agreed ontologies to cover other aspects such as behaviour or habitat. Further, different user communities have different depths of interest in particular areas. There are also many external portals and repositories relating to biological concepts and descriptions to which the Arkive repository could be usefully cross-linked.

All of this is substantially beyond the scope of the Arkive project itself. The proposal for this demonstrator is explore the approach of creating an external index and annotation store, through which a subset of the different user communities can create classifications, cross-links and annotations which reference the same underlying repository. This multi-community enrichment of a shared repository is a common usage pattern, which appears in other areas such as academic digital libraries [DSpace] or museum and heritage portals [Museum portals].

The challenge of this demonstrator is to balance the desire to begin exploring some of the technical issues involved in the cross-community categorization and ontology development, with the limitations of project resources and timescales. One way to reduce the project scope in order to improve feasibility would be to look at only a subset of the repository by species (for example focus just on birds or 'mini-beasts'), by media (for example, concentrate on still photographs and side step the complex problems of annotating time-based media), and by user community. It is the task of the initial requirements study phase to pick the precise focusing subset and to build the links with potential annotators, user groups and external ontology and data sources to make this project feasible.

As noted above, we plan to manage the risk associated with this by (a) beginning this requirements study phase earlier than scheduled and do some work in parallel with development of demonstrator 1, and (b) by retaining the option to switch to an alternative application domain of the same category should the community arkive specification prove intractable.

A References

[SEMWEB]

W3C Semantic Web activity
<http://www.w3.org/2001/sw/>

[SCIAM]

The Semantic Web, Scientific American, May 2001, Tim Berners-Lee, James Hendler and Ora Lassila

[RDF]

Resource Description Framework (RDF) Model and Syntax Specification, O. Lassies 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](http://www.w3.org/TR/REC-rdf-syntax) is available at <http://www.w3.org/TR/REC-rdf-syntax>.

[RDFS]

RDF Vocabulary Description Language 1.0: RDF Schema, D. Brickley, E.V. Guha, Editors, World Wide Web Consortium W3C Working Draft, work in progress, 19 March 2002. This version of the RDF Primer is <http://www.w3.org/TR/2002/WD-rdf-schema-20020430/>. The [latest version of the RDF Primer](http://www.w3.org/TR/rdf-schema/) is at <http://www.w3.org/TR/rdf-schema/>.

[OWL]

Web Ontology working group - <http://www.w3.org/2001/SW/WebOnt/>.

[DAML]

DAML + OIL ontology language - <http://www.daml.org/>

[SEMWEB LAYERS]

Semantic web architecture roadmap
<http://www.w3.org/2000/Talks/1206-xml2k-tbl/slide10-0.html>

[Mozilla]

Mozilla web browser.
<http://www.mozilla.org/>

[NETWORK EFFECT]

Definition of the term *network effect*.

http://www.marketingterms.com/dictionary/network_effect/

[KnowledgeManagement]

Knowledge Management Tutorial: An Editorial Overview, Antony Satyadas, Umesh Harigopal, Nathalie Cassaigne, IEEE Trans Systems, Man and Cybernetics - part C, 31, #4, November 2001.

[Fensel2002]

Semantic web application areas. Dieter Fensel, Christoph Bussler, Ying Ding, Vera Kartseva, Michel Klein, Maksym Korotkiy, Borys Omelayenko, and Ronny Siebes. In Proceedings of the 7th International Workshop on Applications of Natural Language to Information Systems (NLDB 2002), Stockholm, Sweden, June~27-28, 2002.

[CiteSeer]

CiteSeer, Scientific Literature Data Library.

[Shabajee2002]

Adding value to large multimedia collections through annotation technologies and tools: Serving communities of interest, Shabajee, P., Miller, L. and Dingley, A. 2002, In Museums and the Web 2002: Selected Papers from an International Conference (Eds, Bearman, D. and Trant, J.) Archives & Museums Informatics, Boston, USA. p101-111. Available:

<http://www.archimuse.com/mw2002/papers/shabajee/shabajee.html>

[BibTeX]

LaTeX: A Document Preparation System by Leslie Lamport, 1986, Addison-Wesley.

BibTeXing ([btxdoc.tex](#)), by Oren Patashnik, February 1988, (BibTeX distribution).

[EndNote]

EndNote Bibliography software tool.

<http://www.endnote.com/>

[ProCite]

ProCite Bibliography software tool.

<http://www.procite.com/>

[ISI]

ISI Web of Knowledge

<http://isi2.isiknowledge.com/>

[XPath]

XML Path Language Version 1.0

<http://www.w3.org/TR/xpath>

[ARKive]

The ARKive project.

<http://www.arkive.org.uk/>

[DSpace]

The MIT *DSpace* digital repository.

<http://web.mit.edu/dspace/>

[Winer]

The History of Weblogs, Dave Winer

<http://newhome.weblogs.com/historyOfWeblogs>

[Blood]

Weblogs: a history and perspective, Rebecca Blood

http://www.rebeccablood.net/essays/weblog_history.html

[Radio]

Radio Userland

<http://radio.userland.com/>

[MoveableType]

MoveableType

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

[DC]

Dublin Core Metadata Initiative

<http://dublincore.org/>

[RSS-DC]

RDF Site Summary 1.0 Modules: Dublin Core

<http://web.resource.org/rss/1.0/modules/dc/>

[RSS-Syndication]

RDF Site Summary 1.0 Modules: Syndication <http://web.resource.org/rss/1.0/modules/syndication/>

B Application survey

The appendix contains a summary of an RDF database of notes and example applications and suggestions which was developed during the course of this work. The survey is in no way comprehensive - there are many applications that we have missed or not had time to create a record for. Even for those that have been captured our short descriptions may well not do justice to the depth of research and development activities involved. Think of these as an example flags in a map of semantic web applications, not as in depth reviews.

The appendix content has been moved to a separate document ([http-applications-survey.html](#)) in order to keep the size of the primary document manageable.

C Blogging and semantic blogging

In this appendix we provide more back ground on the application of Semantic Web technology to enhance the lightweight publishing paradigm known as "web-logging", typically abbreviated to "blogging". To ground the discussion, we concentrate on the use of semantically-enhanced blogging in the context of the development and sharing of bibliographic data by academics and researchers. We start with a review of the operation of blogging today, and try to identify some of the reasons why it has become popular. Then we identify some key enhancements that could improve standard blogging though the application of semantic web techniques, particularly the context of bibliography sharing.

Blogging today: a review - The roots of blogging go back to 1997 [\[Winer\]](#) [\[Blood\]](#). Its popularity, however, has risen sharply over the past two years. This can be mostly attributed to the emergence of better tools for bloggers (for example [\[Radio\]](#), [\[Movabletype\]](#)), and the *network effect*, in which the success of a community-based activity causes more participants to join in, further enhancing that success.

These success factors suggest the two key values that explain blogging. Firstly, a key driver for many is to address the high cost of maintaining up-to-date web sites. In the typical scenario for large web sites today, one or more dedicated individuals must be responsible for developing and maintaining the web site, and must possess a wide range of specialist skills from web application architecture and database, through to graphical design. Absent these skills and dedicated resources, web sites quickly become out-of-date, bug-ridden, or in thrall to "under construction" clip-art. If individuals with interesting stories to tell could be freed from having to worry about web design, and simply focus on content, such issues would - to a greater or lesser extent - resolve themselves. The first key value for blogging, then, is to provide a *very low effort publishing medium*, in which the individual author can provide content via a simple web form, and a back-end application would generate a polished, indexed view of each of the author's writings. The standard organisation of these contributions is as a series of diary or journal entries, indexed by calendar date and time. This approach has been very successful, and today a very large number of people generate such blogged journals, ranging in quality from professional-standard journalism to highly personal, subjective reflections.

The creation of the blogging tools has been facilitated by some key standards for the format of blog entries. Firstly, *RSS* (variously "rich site summary", "really simple syndication", "RDF site summary" or other variants) provides a basic set of structural metaphors. RSS structures blogs as series of *items*, where such series is termed a *channel*. An item, minimally, has a title, link, and body. A channel has a title and an ordered sequence of items. The link, if present, is taken to be a pointer to another piece of content that this item is commenting upon. Thus blog entries may be created that comment on an item of news (referring, perhaps, to the news report on Yahoo!), or, commonly is a comment on a blog entry by some other person. In particular, RSS defines an XML format for summarizing a user's blog entries. A second key standard is the *blogger API*, which is discussed further below.

The second success factor in blogging is the sharing of and reuse of streams of such lightweight publications. Since the RSS file summarises the current set of blog entries, it is a simple matter to examine it, detect which entries are new or changed, and highlight them to the user. Monitoring the changes in an RSS XML file in this way is termed *subscribing* to that user's blog. An *aggregator* is a desktop tool that provides a user with a view of the new and changed items in all of the RSS channels to which he or she is subscribed. By subscribing to a set of channels that closely matches their interest, the user gains a highly selective flow of information most likely to be of both relevance and interest. As such channels get connected between members of the blogging community, a rich network of quality information flows is created.

There remains the problem of how users discover blog channels they wish to subscribe to. There are four main mechanisms:

- The blogging tool that the user employs notifies one or more central directories of the changes to the RSS XML file. While in principle each such directory could have its own means of submission, the blogger API has become the *de facto* standard that such directories adhere to. The directories publish various indexes of the blogs they know about, including the most recently updated (over a window of, say, the preceding 60 minutes), the most active blogs, or the blogs most subscribed-to by other users. In addition, most directories are searchable by keyword (from the blog title or description).
- On the published HTML containing the blog, many bloggers will list the influential or entertaining blogs to which they subscribe. A common pattern when reading one blogger's writings is to explore the writing of one or more of the blogs they subscribe to. The phenomenon of listing such subscriptions (and hence helping to propagate interesting writing) is termed *blog-rolling*.
- A blog entry frequently refers to an entry by another blogger. By following the link, the reader is taken to the source blog, to which they can then subscribe.
- Typically, individual bloggers will reference their blog from their personal home pages, email signatures, etc.

Given the metadata contained in the blog's RSS file, and the blog-roll subscriptions, a range of tools for exploring the meta-network of connections between bloggers have been developed, and continue to evolve.

Upgrading blogging to semantic blogging - In this section, we outline some of the key changes that must be added to standard blogging in order to achieve our vision of semantic blogging.

1. Currently, items have a very limited structure (title, link and body). Semantically meaningful items will have much more structure, and will be governed by one or more ontologies to give meaning to the structure. The RSS 1.0 specification provides for the use of *modules* to extend the content of the item, though currently this

is limited to Dublin Core metadata [\[DC\]](#) [\[RSS-DC\]](#) and syndication metadata [\[RSS-Syndication\]](#).

Fortunately, RSS 1.0 (but not all RSS variants) are standard RDF, so in principle it should be possible to add arbitrary additional structure to items.

2. Current blogs use channels to identify sub-streams of information from one contributor. For example, an author may have a blog which contains categories of information on Java programming, semantic web technology, RDF tools, etc. Each of these categories is available as a separate RSS channel. However, there is no formal semantic relationship between these channels, and the channel structure is relatively static and coarse-grained. A more flexible approach would be to label each item with one or more ontological categories, and use these to define implicitly the channels available from that source.
3. At present, links between items are limited to the single 'link' field in the item. There is no particular semantics associated with this link. For semantic blogging, there is likely to be a rich and extensible set of links between items. For example, in the bibliography domain, there will be such linkages as 'cites', 'isCitedBy', 'extends', 'replacesVersion', etc.
4. The aggregator tool, or the HTML presentation, will need to be extended to present these additional semantic capabilities. In addition, the capture tool must allow the user to enter such information as is available. Neither should, as far as possible, sacrifice the lightweight simplicity that provides one of the key values of the blogging approach.
5. The HTML presentation of the semantic blog will also need to be extended to metaphors other than the reverse-chronological order of the current calendar-based presentations.
6. Given the additional semantic information present in the network, there are probably better ways of discovering other blogs, or channels, or even individual items. For example, allowing distributed queries based on semantic terms, or persistent filters that notified the user of new matching occurrences, would allow much more effective discovery.
7. The network infrastructure, which for blogging today is centred around the human reader, could be extended to allow web-services or autonomous agents to contribute additional source data or semantic mark-up. Examples include the automatic querying of CiteSeer for details on a publication, or using a shared reference ontology to translate between the various source formats used by different reference formatting tools.

Issues - There are some hard research problems that will need to be addressed to some degree during the project.

- To commonly refer to a semantic item, it must have a consistent URI. This is not available for many publications. This naming problem occurs in other guises in Semantic Web projects, and is not unique to this project. It is also not the key research issue, though it will require some investigation. It may be that limiting the source data in the initial versions of the applications to those publications that have a well-known URL on CiteSeer or Amazon may be sufficient. Alternatively the *naming by property pattern* approach (as used in TAP or as implemented in the query-by-example system of [ePerson](#)) might be appropriate.
- Users will want to develop their own ontologies to markup the content of their publications. However, to have maximum value to the community, such taxonomies must be widely shared and reused. The problem of merging and maintaining such distributed ontologies is again not unique to this project. It will require some solution, else ontological balkanization will result. It is less clear that an effective work-around exists.

D Changes

29-10-2002

Revised initial draft to fix about 100 typos, tweak sections 6 and 7, and add appendix C.

5-11-2002

Added Appendix B, generated from RDF-based application survey.