SWAD-Europe deliverable 12.1.5: Semantic Portals - 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 portals demonstrator - requirements specification
URI:
☞ http://www.w3.org/2001/SW/Europe/reports/requirements-demo-2/

Authors:
☞Dave Reynolds, HP Laboratories, Bristol, UK
Paul Shabajee, Graduate School of Education and ILRT, Bristol, UK

Abstract:

This report describes the aims and requirements for the second of the two open demonstrators which make up workpackage 12.1. This, second, demonstrator is designed to illustrate decentralized creation and enrichment of cross-community information portals. We first describe the general problem class and the broad technical approach which the semantic web makes possible. Then we describe a specific application domain, a directory of wildlife, environmental and biodiversity organsations, which will form the basis of the demonstrator. Finally we examine the requirements raised by this demonstration application domain, the infrastructure components that will be required and sketch our approach to each of the issues raised by the requirements.

Status:

Initial 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 problem
3 ☞The approach
4 ☞Demonstration example - directory of wildlife, environmental and biodiversity organisations
5 ☞Requirements and target scenarios
6 ☞Summary and component requirements
A ☞References
B ☞Selective glossary
C ☞Demonstration - additional use cases and scenarios
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 define the specific requirements for the second of the two proposed demonstrators - *semantic portals*. We first describe the broad problem area we are addressing, community information portals, and the approach to this enabled by the semantic web. We then describe a specific instance of this broad problem area - the development of a national directory of wildlife, environmental and biodiversity organizations - and outline the match between that application and the semantic web approach. Finally we explore in more detail the requirements and issues raised by this demonstration application.

2 The problem context

Web *information portals* provide a point of access onto an integrated and structured body of information about some domain. They range from very broad domains (e.g. all web pages - [☞YAHOO] and [☞DMOZ]), to topic specific domains (e.g. cognitive psychology [☞COGNET], mathematics [☞MATHNET], fish species [☞FISHBASE]).

Community information portals are information portals which are also designed to support and facilitate a community of interest. They typically allow members of the community to contribute news and information to the pool, either by submitting information to the portal (via some editing or reviewing process) or by posting the information on some associated web bulletin board or other collaboration tool.

Whilst web information portals are very successful there are several ways in which the current state of the art can be improved. In particular, the emergence of the semantic web [☞SEMWEB] brings a new set of tools and techniques that could radically change the way portals are built, integrated and used. Below we identify some key dimensions along which existing portal practice could be assisted, or further assisted, by use of semantic web technologies. These dimensions are interconnected but we will build up the picture in stages before finally summarize the key points [☞figure 1].

Some of the dimensions we identify below are well-known trends and are the focus of many active development efforts; some are more novel, or at least less well represented by the existing portal projects that we are aware of.

1. Multi-dimensional search and browsing - This is an area of active and continued development of portal systems. Since the primary job of an information portal is to bring a large body of information items into one accessible structure then the ease and power of the navigation and search tools are crucial aspects of portal design.

Many portals use a hardwired navigation structure based on a single rich classification scheme coupled to hyperlinking of related items and free text search. [☞YAHOO] is a canonical example of this. This works well when the information being indexed is well described by a single classification structure or is is well suited to free text retrieval (e.g. when browsing and accidental discovery are more important than retrieval precision).

Portals handling more structured information typically have some more richly structured internal descriptive schema (often directly mirroring the database schema used for storage) and offer a rich search interface which can exploit this schema. This allows search to be tied to specific facets of the descriptive metadata and to exploit controlled vocabulary terms - leading to much more precise searches. An example of such a rich interface is fishbase [☞FISHBASE_SEARCH].

An emerging approach to rich information portals is to describe the properties, relationships and classifications of the various information items via some form of external declarative schema or ontology [☞ONTOLOGY]. This allows the portal maintainers to update the structure of the portal more easily and allows searches to exploit the structure of the ontology - for example by automatically narrowing or broadening search terms. Good examples of this approach include PlanetOnto and related projects and the Open University Knowledge Media Institute [☞PLANET-ONTO], the SEmantic portAL work from AIFB Karlsruhe [☞SEAL] and the web community portals from ICS FORTH [☞FORTH-PORTALS]. An interesting way of combining the very general web directory search interfaces with underlying domain ontologies is being developed by the TAP project [☞TAP]. For a good illustration of the sort of faceted navigation interface enabled by such technology see the online demonstrations at Siderean software [☞SIDEREAN].

A key benefit of adopting semantic web representations for such ontologies is the greater ability to share and reuse ontologies, link across ontologies and reuse tools for constructing, managing and applying these ontologies. The basic notion of using ontologies to separate portal content from structure and provide rich structural links is not new. What the semantic web offers is the standardization of the languages for expressing such ontologies (and the associated instance data) and the possibilities for sharing and reuse that such standardization opens up.

2. Evolution and extension of information structure - Information requirements change over time. If a portal is constructed out of linked free text pages then those pages have to be edited, often manually. However, if, as discussed above, the portal adopts a more structured approach with an explicit ontology then these changing requirements mean changing the information structure.

For example, in a portal including information on technical reports the community might decide that additional fields concerning the funding body and grant code which supported the report might be needed. This would in turn lead to a possible requirement for an ontology of types of funding bodies and an authority list for funding body names.

In principle, the ontology-driven approach to portal construction facilitates this. It should be easy to augment the report ontology with the additional properties and to link those properties to existing or newly constructed sub-ontologies for funding bodies and grant types. The data entry forms, validation procedures and search forms all need to be updated but if these are generated from the system ontology this may be automated or semi-automated.

This sort of evolutionary change requires us to change the instance data and any associated database schema, not just the descriptive ontology. This can be complex. We need to permit data to be added in the new format, without invalidating existing data and do so in such a way that both original and extended formats can be used interchangeably. This is greatly facilitated by use of semi-structured data representation such as that offered by RDF [☞RDF]. RDF enables incremental additions of properties and relations to the instance data by virtue of its property-centric, rather than record-centric, approach to representation.

This opens up the possibility of approaching portal information structure design bottom up. Rather than forcing a long top-down design cycle we can create an initial portal with seed ontology and core information structure and then extended it incrementally over time as the detailed representation needs change or become clearer.

3. Community extensions of structures and views - When an information portal serves a specific set of needs for a specific and stable community then modest evolutionary extensions to the information structure may be all that is required. Typically such changes would involve the addition of new properties for existing concepts or the incremental additional of new concepts easily related to the existing portal structure.

A more challenging, but exciting, set of problems arises when we consider different communities of interest sharing access to the same underlying information portal. In such situations one community may need to view the portal information using a completely different navigation structure, search facility and presentation from another community. If a portal is successful, and of broad relevance, we might expect to commonly find new communities of interest forming around the portal whose needs would typically not have been anticipated when the portal was designed.

To illustrate this consider the example of the ARKive project [☞ARKIVE]. This is a rich repository for multimedia information on UK and worldwide endangered species. Its primary organization and navigation structure is by species. For specific important interest communities (e.g. school children) specific navigation routes and views have been constructed. However, surveys of potential communities of interest [☞ARKIVE-COI] reveal many different possible user communities - further and higher education lecturers, researchers in particular species, researchers in environmental issues, hobbyist groups - each with different needs for structuring and navigation of the information. For example, university lecturers would want to be able to index by concepts such as particular behaviours, habitats, physiological adaptations and so forth; UK school teachers would want to be able to navigate by National Curriculum topics; researchers in narrow species areas would require much more comprehensive and precise indexing terms for their particular species groups that would be irrelevant to non-specialist users..

This raises two issues. First, there is the simple divergence of these specialist needs from the core, species-based, structuring of the initial portal. These require more that just the addition of extra links on a concept page or field in a search form but a radical extension to the information indexing, in turn raising interface and access issues. Second, there is the issue that the central organization managing the portal cannot be expected to be able to accommodate all of the potential user groups. Even just coordinating the collection and organization of the required information may be beyond the available resources of the central organization.

Both of these issues are potentially addressable by exploiting the decentralized approach to data and ontology sharing which the semantic web aspires to offer. One can imagine the central organization just providing the backbone structure of resources (in the ARKive case, multimedia objects depicting endangered species) and the communities of interest themselves supplying the additional classification and annotations to suit their needs. This external enrichment could be maintained externally. It does not necessitate allowing external communities of interest to directly edit the hosting portal. It "simply" requires the ability to generate community specific views by combining the information from the core site with the additional ontologies, annotations and extensions created by the community.

Many existing portals do allow direct submission of information by community members, however these are typically limited to annotation of items with simple ratings and comments or submission of well understood records within the main portal ontology - e.g. events or news items. The notion that specific sub-communities might be able to create complete new navigation structures and new types of entry goes much farther than existing systems. The nearest example to this is possibly the Open Directory project [☞DMOZ] in which a large group of editors collaborate in an open-source style to incrementally develop and populate the Open Directory index. Whilst this just involves a single classification scheme, and a single group of developers, the sheer scale of the number of editors involved, and the fact that they are drawn directly from the end user community, sets it apart from other portals maintained by small core teams.

4. Sustainability by aggregating decentralized sources - This decentralized approach can be taken further.

Given the ability of the semantic web standards to provide interoperability between systems it is possible for the portal information itself to be an aggregation of a large number of small information sources instead of being a single central location to which people submit information.

This would address a key problem of several community portals - the fragility of the central organization. A typical pattern in development of specialist portals is that some organization obtains funding to put together a directory of information on some topic. It collects the information from a range of provider organizations and individuals and creates a web portal to access it. However, funding is generally finite and few topic areas are of such broad interest to be self funding. At best the portal remains in existence but is not updated and rapidly goes out of date. At worst the site cannot be maintained. Even in areas of very high visibility and public interest such as environmental science and biodiversity we see several such examples, see ☞section 4.

Suppose, instead, that the organizations supplying the source information make that information directly available over the semantic web and that the portal is constructed not as a central site but as an aggregator service which merely caches and provides access to the raw material. Central organization is still needed in the initial stages to provide the start-up impetus and ensure that appropriate ontologies and controlled vocabularies are adopted. However, once the system reaches a critical mass it can more easily be self sustaining. Anyone can run an aggregator services and ensure continued access to the data. A new supplier can add data to the pool without a central organization being a bottleneck. Communities with particular needs can take the aggregated information and combine it with community specific extensions to provide specialist services.

As well as technical benefits of local control and robustness this decentralized approach has soci-polical implications for the funding and maintenance of information portals, giving more options for how initial seeding can take place.

This approach is very much in the spirit of the RSS syndication language [☞RSS], which in part was originally developed as a means to aggregate news items within personal information portals. The difference here is that whereas RSS is aimed at providing a text summary onto a separately hosted textual item, like a news article, we are describing a situation where the items aggregated are complete and structured information blocks.

From the point of view of the information providers an aggregation-based portal using open standards could be an attractive time-saving alternative to centralized systems. For example, anyone involved in the management of a hobbyist or sports club will be familiar with handling frequent requests for information on club activities for inclusion in web and print based directories - directories that will rapidly disappear, go out of date or re-ask for the same information time-after-time. If the information could be published once in machine readable form, using an appropriate ontology, then each of these external directories could reuse the same published data with the provider confident that their information will be up to date without the overhead of repeated resubmission.

5. Cross-portal integration - The same approach, used to aggregate multiple separate information items into a single integrated portal, can be used to allow multiple portals to be accessed as an integrated whole. If several portals in related areas make their aggregate data accessible over a web service interface, using shared or compatible ontologies, then it becomes possible for users or software agents to perform searches across portals which use identical or compatible ontologies. We replace a set of disjoint, self-contained, topic-specific portals with a web of information repositories coupled to search and aggregation services with open-ended uses.

A good example of the need for this in the biodiversity area is the National Biodiversity Network [☞NBN]. This is a consortium of government agencies and voluntary bodies that is creating a common pool of biodiversity information by combining data from many sources. Even when the source data was already online in the form of organization-specific portals those online portals could not themselves be used as the source of the data for integration. Instead separate agreements on data formats and access protocols were needed before data could be imported into the central pool.

Traditional design approach	Semantic Portal
Search by free text and stable classification hierarchy.	Multidimensional search by means of rich domain ontology.
Information organized by structured records, encourages top-down design and centralized maintenance.	Information semi-structured and extensible, allows for bottom-up evolution and decentralized updates.

Community can add information and annotations within the defined portal structure.	Communities can add new classification and organizational schemas and extend the information structure.
Portal content is stored and managed centrally.	Portal content is stored and managed by a decentralized web of supplying organizations and individuals. Multiple aggregations and views of the same data is possible.
Providers supply data to each portal separately through portal-specific forms. Each copy has to be maintained separately.	Providers publish data in reusable form that can be incorporated in multiple portals but updates remain under their control.
Portal aimed purely at human access. Separate mechanisms are needed when content is to be shared with a partner organization.	Information structure is directly machine accessible to facilitate cross-portal integration.

Figure 1 - contrast semantic portals proposal with typical current approaches

3 The approach

3.1 Overview of aims - To sum up the above discussion:

> *The aim of this demonstrator is to enable decentralized provision of community information portals by providing tools to enable aggregation of data provided directly by community members.*

The key attributes of the solution we are looking for are:

Explicit rich structure representation - separate from content

The use of explicit ontologies enables the separation of content from structure and simplifies the implementation of the integration and extensibility requirements (see below). Using a rich ontology formalism, rather than a simple schema or classification hierarchy, enables us to support multidimensional search, richer description of content structure and interrelationships, and "intelligent" search features such as term normalization/generalization/narrowing.

Openly extensible

It should be possible to incrementally add fields and relations to existing concepts or instance data. It should be possible for user groups to develop alternative views or complete new classification and annotation structures involving new ontologies, thus facilitating a bottom up approach to portal development.

Decentralized and sustainable

There should be no reliance on a fixed infrastructure. All of the data formats, ontology representation and access protocols should be based on open standards. It should be possible for each information provider to host their own data. It should be possible for anyone in the community to run any part of the infrastructure, including the aggregation and viewing servers. Any initial infrastructure tools should be open source to reduce the costs of taking over running of the infrastructure. Any foundation ontologies should be openly accessible and freely reusable.

Machine accessible

The aggregated data and ontologies should be accessible over similar protocols to enable cross-portal queries and navigation.

Human accessible

At a minimum the user interfaces for search and browsing of the information in the portal should not be compromised compared to traditional portals. Ideally the user interface descriptions should be declarative and separate from the content and ontological structure so that different communities can develop and exchange different user interface views onto the same pooled data.

Low barrier to entry

The cost to an individual or organization to contribute information should be as low as possible. It should be possible to do so without installing specialist tools and without requiring deep knowledge of semantic web standards and technologies. This does not prelude providing installable tools; it just means there should be at least one, easy to use, zero-install route.

Adequate controls

The design should support a level of validation (accuracy of the data), moderation (elimination of inappropriate content) and privacy (restricting onward aggregation of sensitive data such as email addresses) appropriate to the needs of the community being supported. It should not rule out the option of implementations with strong security measures nor should it require commercial grade security for communities with more modest needs and resources.

3.2 Architecture layers - The critical parts of the semantic portals architecture are the conventions, protocols, data formats and ontologies used to represent the different information layers. These need to be supported in the demonstration by working (and open source) components for creating, editing, aggregating and viewing the data. However, in keeping with the decentralization requirement, any of these components should be replaceable by similar standards-compliant versions that exploit the same conventions. Hence we first sketch the set of representation layers that need to be supported before outlining the infrastructure components.

We conceive of the semantic portal representations as layered with each layer being more domain specific than the layer below. This layered breakdown is designed to promote interoperability so that portals in similar domains can be integrated or cross searched using the common underlying ontologies. The following figure illustrates the overall layered requirements.

Specific portals	● **directory of wildlife and environmental organizations** ● biodiversity publications from such organizations ● hobbyist wildlife photography ● …
Domain specific	● thesaurus of types of environmental organizations ● thesaurus of types of environment activities ● thesaurus of environmental services ● species ontology ● …
Cross-domain	● Person ontology for individuals ● Organization ontology ● Location ontology ● Time ontology ● Event ontology ● Publication ontology ● …
Portal infrastructure	● ontology hinting conventions for view, edit, search templating ● conventions on provenance tracking ● ontology for policy descriptions (validation, moderation, privacy) ● protocols and conventions to support access policies ● representation for thesauri (TIF) and controlled terms
Semantic web	● semi-structured data representation (RDF/XML) ● generic ontology representation (RDFS, OWL/(f)lite) ● network access conventions (HTTP, NetAPI)

Figure 2 Summary of architecture layers

We will briefly discuss each of these layers, in order from general to most specific.

Semantic web - The proposal is that the entire edifice should be based upon the W3C Semantic Web standards.

This means that all information for the portals should, at heart, be represented in RDF [☞RDF]. This is well suited to the requirements of *open extensibility, machine accessibility* and *decentralization*. But leaves requirements like *adequate controls, human accessibility* and *ease of entry* needing to be addressed by additional layers.

Secondly, the data formats and domain-specific representations should exploit a common semantic web ontology language. The OWL candidate recommendation [☞OWL] provides for a set of different levels. We propose to use the RDF Schema language [☞RDFS] augmented with the constructs from OWL/lite profile of the OWL language - the OWL/lite subset of OWL/full. This is different from strict OWL/lite in that we don't require enforcement of the OWL/DL syntactic constraints such as the separation of instances and classes.

Thirdly, we rely on the standard web infrastructure to permit reading of source RDF documents hosted by standard web servers. Large datasources, such as the aggregator itself, should ideally permit selective access to subsets of the RDF information. We propose to adopt the RDF Net API [☞NetAPI], at least its HTTP profile, for this purpose.

Portal infrastructure - The portal infrastructure layer augments the semantic web standards with a set of conventions and representations needed for any semantic portal of the type discussed here. The components needed here divide into three groups.

Firstly, a set of conventions is needed to enable the combination of RDF instance data and OWL ontologies to be used to generate human accessible views, edit screens and search tools. These need to address issues such as how the data should be grouped into viewable pages and how the data should be ordered and laid out on the pages.

Secondly, the requirement of *adequate controls* needs to be supported at this level by explicit representation of the access, moderation and aggregation policies for a given community and access protocols to support adequate enforcement of these policies. For this application area openness and free contribution is prized above security so that strong enforcement is not required but some controls and policies are needed.

Thirdly, whilst the generic OWL ontology language can represent the information models for the portal and the relevant domain conceptualizations there is also a need to express classification schemes in the form of thesauri. Whilst this can sometimes be do using OWL that approach is overly complex and rigid for representation of less formal (and ill-structured) classification structures and controlled terms. We can use the Thesaurus Interchange Format developed as part of SWAD-E workpage 8 [☞SWADE-THESAURUS] to represent such thesauri but may also need a format for representing flat, but extensible, authority files.

Cross-domain ontologies - Many of the domains that might exploit the semantic portal approach are likely to share a need to represent common concepts such as people, organizations and events. For any such concept there are many ontologies currently available but for very few of them is there a dominant ontology. Worse, in many cases the ontologies lack the modularity that makes it easy to combine or even map them.

At a minimum we need a set of recommendations for the particular ontologies to build upon. Preferably, for each core concept, we would also be able to generate mappings from the chosen ontology to the core properties of the most common alternative ontologies. The ideal case would be to be able to make the ontologies modular enough to enable different combinations to be viable. Without careful design for modularization ontologies become entangled (for example it would be difficult to use the vCard representation for Person in combination with the DRC Orlando representation of people Locators). For more discussion on some of the issues that arise in the case of matching ontologies for personal profiles see appendix 4 of [☞EPERSON]. Some decoupling is possible through a mixture of careful design and indirection through some abstract shared vocabulary such as Wordnet [☞WORDNET].

Domain-specific representations - For our chosen demonstration example - ☞directory of wildlife and environmental organizations - there are several domain specific ontologies and thesauri which are more general than the specific portal to be constructed. These should be factored out as reusable components so that future portals in the same domain can exploit them to achieve interoperability.

3.3 Infrastructure components - These protocols, conventions and ontologies do not themselves implement a semantic portal. They need to be embodied in working software components. A simplified deployment picture of the proposed components is shown below.

Figure 3 - outline of infrastructure components

The core components are the aggregator (which takes data from a broad variety of sources and makes the complete filtered, integrated assembly of information available) and the viewer (which takes the data, ontologies and templates and uses them to generate a portal web site). In addition, to meet the *ease of entry* requirement, we need a web-based edit tool (also driven by the ontologies and associated hints) to avoid the need for local edit tools and an optional data hosting component to make it possible for individuals to contribute annotations and other data without having to run their own web site.

All four of these key components could be implemented as a single portal toolkit or as a decentralized collection of independent tools. We will return to the more detailed requirements for these components ☞below.

3.4 Design issues

- The problem and approach outlined so far raises a number of requirements that will need to be addressed in any implementation design. We enumerate and briefly discuss the issues here in the general setting. Later, once we have outlined the specific chosen demo topic, we will discuss in more depth the issues which are critical to address for the demonstration.

a. Access control - In a centralized portal the question of who is able edit any given data record is a crucial part of the design. In the decentralized approach we take the view that the data should be hosted by the provider organizations and they therefore have complete control over the editing of their information, no special infrastructure is needed to support this. For organizations or individuals unable to host their own data, who use a shared hosting server, then some traditional access-controlled account will be needed on the hosting server.

However, the fact that other community members cannot change the supplied data does not at all preclude them from adding annotations, links and classifications to the supplied data. The property-centric RDF approach means that these additions are simply additional data assertions than can be hosted elsewhere. The display template mechanisms (see below) should make it easy for users to see the provenance of data and thus not confuse third party annotations with authoritative statements.

b. Privacy - There is a second, harder, access control issue - that of privacy. There may be information that an organization is willing to make available in human readable form for particular purposes but unwilling to allow that data to be aggregated - email addresses and phone numbers of staff contact members are good examples of this. Whilst there are some technologies that can limit onward aggregation of some types of disclosed information [☞VORA] they don't apply to this sort of situation.

We see three solutions to this.

First, sensitive data can simply not be included in the RDF files. The machine accessible data can reference back to the supplier's normal web site and be just as accessible and open to aggregation, or not, as at present.

Second, it is possible to disclose private information via a one-way hash function. This means the information is not humanly readable but it can be searched for and you can identify when, for example, the same email address is being used in two separate locations. This approach is used in foaf [☞FOAF]to enable use of email addresses as unique identifiers without exposing email addresses to spammers.

Thirdly, it is possible to disclose information such that it can be human readable while being hard to machine process. For example, a phone number can be rendered as a bitmap with sufficient noise and distortion to defeat simple OCR approaches but not so much as to render it humanly illegible. This approach should only be regarded as an partial-barrier to machine processing. It also raises accessibility issues since such information would not be accessible to screen readers.

c. Moderation - Related to access control is the issue of moderation. The open decentralized approach is at risk of exploitation by people outside of the community using the aggregation network to transmit inappropriate content (e.g. spam, pornography). Thus the aggregators should be controlled though a policy which limits the sources which will be aggregated - so that sources which are contaminated by inappropriate content can be removed from the aggregation list. Some measure of centralized control will be needed to maintain this policy file and any community setting up a portal will need to determine the appropriate authority structure for this remaining aspect of central control.

d. Validation - The final security related issue is that of validation. When the information providers are the appropriate arbiters of the accuracy of their data then the simple decentralized design will work well.

There will be cases where this is not the case and third-party validation is required before the supplied information is incorporated into the main portal view. This is still possible in a decentralized design. The validating organization can publish its validation information itself - again they retain control over the validation data they publish into the system. The portal aggregator and viewer can then employ a policy of only showing the validated data or also allowing the raw supplied data to be accessible depending on requirements. So long as the provenance of all aggregated data is recorded then a range of validation policies can be employed - including more radical schemes such as peer-to-peer reputation based validation [☞P2P REPUTATION].

There are challenging design issues here as to how the validation organization links its validation to the source data being validated and how well that link is protected against attack after it has left the validator's site. Whilst there are cryptographic approaches to signing such validations they do require an infrastructure such as a PKI or an accepted peer-to-peer alternative such as PGP [☞PGP].

e. Template and UI issues - One disadvantage of the decentralized aggregation of semi-structured data is that, whilst good for machine processing it is actually quite hard to create a good user interface onto it. Things like ordering, graphical layout and relative prominence of different elements are important to a user interface but hard do with an open ended data format like RDF. It is perfectly possible to design a templating system where the templates define the human readable page structures whose content is populated from the RDF data sources. The hard challenge is to do this while still allowing open-ended extension of the data and structure.

We hope that a generic template solution to give the overall page structure coupled to the ability to add user interface hints to the entries in the ontology will be sufficient. The ontology hints should be able to influence ordering and embed/link decisions in the interface so that extensions to the ontology can be made incrementally without breaking the effectiveness of the basic user interface.

The user interface design will also need to cope with the decentralized nature of the portal. In particular, it should clearly signal when information is being drawn from different sources and make it easy to examine the provenance of that information. In particular, the difference between an organization description page provided by that organization and an commentary or annotation added externally by a third party should be clear.

f. Ease of creation and edit - Closely coupled to the issues of the user interface for data consumption is the requirement that adding new data items should be as simple as possible.

When creating a brand new entry on some concept it should be easy to locate an appropriate structured data entry form. Then extending the initial entry with additional links and annotations should be possible incrementally. In particular, it should be easy to search the set of community ontologies to find the appropriate property and class to use to express a given concept. When filling in the value for that property it should be easy to use controlled vocabularies (with autocompletion) when available but also possible to submit extensions to that vocabulary as part of the input process.

This support for reuse of existing ontologies and controlled terms will be quite crucial. The danger of allowing end user extension of the ontologies is that they become tangled and unmanageable. We tackle this in several ways. First, we encourage reuse of existing ontologies by making them modularly composable and providing good tools to support discovery of terms to reuse (possibly employing a linguistic reference ontology such as WordNet to facilitate that indexing). Second, we can exploit the term mapping capabilities of OWL (with extensions for property composition) to permit duplications to be rectified after the fact.

g. Base data model - We have referred to the semantic web standards of RDF and OWL as providing the base data representation. We need to adopt additional conventions on top of these to meet the requirements outlined above. First, the discussion on moderation, validation and privacy shows that a standard convention for provenance tracking and representation will be required. Second, we cannot rely on any unique URIs to identify concepts, especially concepts such as people, instead we anticipate concepts being identified by patterns with a variety of inverse-functional-properties being used for the identification. For example, *the personal with mailbox hash xxxx* or *the organization whose name within this controlled vocabulary is yyyy.*

h. Uptake issues - Finally we note that the broad approach of information sharing, extensive aggregation and onwards machine accessibility will not be acceptable in all domains. In many commercial applications the information suppliers will actively resist making their information directly comparable to competitor's information. The decentralized approach will work best initially in noncompetitive areas where the benefits of information sharing and universal access are either intrinsically accepted or mandated (e.g. by governments).

4 Demonstration example - national directory of wildlife, environmental and biodiversity organisations

Background and Context - As part of background research to support the specification of this demonstrator it was decided to investigate the nature of information flow and use within a specific domain, that of biodiversity/wildlife [☞SWARA]. One finding of the survey was that at present there is no single comprehensive directory of organisations focused on wildlife, environmental and biodiversity in the UK. Such a directory would have many potential uses and be valuable to a number of different groups of users and across a number of sectors. This section details a proposed system that would provide a comprehensive, up-to-date and easily maintainable and (possibly most importantly) robust and sustainable system that enabled the creation, collation and publication of this information in Web-based and other formats based on the Semantic Community Portal approach outlined above.

In the UK, the Environment Council [☞EnvCouncil] produced paper-based directories for each UK country 'Who's Who in the Environment', [☞WWE] and an electronic database version covering the whole of the UK. This directory provided detailed information including; contact details, a text description and other data such as membership and volunteering opportunities, publications and services. The paper based version was last published in 1995 (detailing some 600 organisations) and the database version in 1998 (detailing some 1050 organizations). Other paper-based directories do exist the most recent is 'the greendirectory' [☞TGD] available in both paper based and a restricted Web base version. It contains names and contact details for a variety of environmentally related organisations, however this does not cover the many 'wildlife' or other small specialist organisations. A partial list of other paper based directories can be found at [☞SLV], the majority of these date 1995 or before. The most recent publication that we can identify is the "World Directory of Environmental Organizations" published in 2001 by EarthScan [☞EarthScan], which covers over 2000 organisations worldwide, however this does not seem to cover small specialist organisations (e.g. those with particular species or local foci).

There are a number of large Web-based directories e.g. World Directory of Environmental Organizations [☞WDEO], Envirolink [☞EnviroLink], Environmental Organization WebDirectory! [☞EOWD], Directory of Organizations and Institutes Active in Environmental Monitoring [☞DOIAEMER]. However, none of these provides comprehensive data related to the UK. A brief survey of these and the paper based resources discussed above identified limitations and issues detailed in the next section.

Existing Approaches and Issues - The existing directories seem to be largely created using traditional means e.g. questionnaire surveys, and data collation by researchers working on the projects, using a wide range of information resources including, organisational Web sites, reference books and existing contacts databases. Both these approaches and associated directories have a number of significant limitations.

- **Limited Scope**: In general directories are created for specific purposes and thus with specific foci, e.g. "Who's Who in the Environment" [☞WWE] was largely focused on non-comercial organisations and international directories tend to have a national foci e.g. *Envirolink* and '*Environmental Organization WebDirectory!*' both have a disproportionate number of USA based organisations.
- **Incompleteness**: This is a major issue, even relatively comprehensive directories (within their scope) such as "Who's Who in the Environment" [☞WWE], miss many important organisations e.g. the 1995 edition missed Butterfly Conservation (☞http://www.butterfly-conservation.org/) a very significant organization, and the 'World Directory of Environmental Organizations' [☞WDEO] is one of the most comprehensive Web-based directories and yet seems to have omitted 'English Nature' (a very major UK based organisation) from its directory. Similarily 'The Green Directory' [☞TGD] has many ommisions in the wildlife and biodiversity domain e.g. Butterfly Conservation. Such omissions have a number of likely causes, e.g. if a questionnaire survey is used to collect the data, organisations must be contacted and respond to be included in the directory over a limited timescale - i.e. while funding is available to conduct the collation of data and the publication is in production.
- **Accuracy**: This is tightly inter-related with currency (see below). Validation of *all* information is problematic even where the organisations themselves are providing the data - there may still be errors. It is likely that different types of data will have different requirements for 'accuracy' depending on the specific application e.g. an accurate URL (Web address) for an organisation is probably very important, because that is how most users can begin the process of verifying the other data, while small mistakes in the description of an organisations are likely to be less important. Obviously if data is out of date it is inaccurate.
- **Provenance of Information**: Provenance of data is intimately related to accuracy - the source of data is a primary means for users to assess the likely validity or reliability of data e.g. is this from a source that they believe to be reputable? In general existing resources do not tend to provide data about the sources of the individual pieces of data, or indeed the data in general. Thus making it difficult for the user to assess the likely validity or reliability of the data or verify that the data is accurate.
- **Currency**: The paper directories of their nature rapidly become out of date, however Web based directories suffer from similar issues as the data must be very actively maintained by the

organisation that collated and holds the data. The negative consequences of information being out of date are many, ranging from inconvenience for users when they try to contact an organisation using incorrect information to propagation of errors as incorrect information is re-used by other information providers.

- **Lack of contextual information**: The majority of Web based directories tend to provide only links to organisational Web sites, with little or no contextual information. This is also the case with some paper based directories e.g. The Green Directory [☞TGD] provides only contact details in the majority of cases. This is probably related to the issue of *currency* above, as any contextual data is likely to become out of date and be significantly harder to maintain than simple URL or contact details. With respect to Web sites, where there is supporting data (e.g. contact details, addresses) these are often incomplete or out of date.
- **Short lived organisations or projects and those that have ended/ceased to exist are rarely if ever listed**: For example funded research projects are not generally listed. This is probably because by the time they become known they may have ended or coming to an end. The non-inclusion of organisations and projects that have ceased to exist is probably because the directories aim to provide current information, however it is often useful to know that an organisation or project has existed and what they did e.g. validating references, historical research, etc.
- **Relationships between organisations tend not to be detailed**: This means that it is difficult to understand how organisations dealing with similar 'topics' or locations are related and to gain an understanding of the structure or workings of organisations with a given domain. Even where these are available e.g. via the organisational Web sites themselves, it is still often difficult to gain an overview of these.
- **The User Interfaces (UI):** These are often unintuitive or use inappropriate classification schemes for many users. They [Web based directories] tend to be based on hierarchical classification systems and are often unintuitive and difficult to use, for those who are not specialists in an area. They are also often problematic to use where the system was devised with a specific use(s) in mind e.g. in the '*World Directory of Environmental Organizations*' directory, finding organisations that focus on particular species or group of species is problematic as this is not one of the pieces of data indexed in their internal classification scheme. Broadly speaking they tend to lack flexibility being based on focused and relatively narrow classification schemes. This is especially the case if they don't have any contextual information e.g. text description, since they can't use free text searching except on the names of the organisations.
- **Sustainability**: The sustainability of such directories is a very significant and problematic issue. In the vast majority of cases such directories are funded and maintained by individual organisations or projects. There is a clear and demonstrable risk that if those organisation or projects either have a change of priorities, funding problems or simply come to an end, then the directories themselves will be lost or go out of date quickly with little opportunity or incentive for others to take them over. For example this seems to have been the case with the 'Who's Who in the Environment' publications in the UK.
- **Copyright and Reusability**: In general the data is collected for the needs of the collating organisation, this is the copyright of the collating organisation so cannot be reused by others, without express permission. Even where re-use is allowed the data is not generally in a format that can easily be reused
- **Sensitive Data**: In general Web based directories do not deal with data that might be thought of as 'sensitive [because of potential misuse], e.g. e-mail addresses, used by spammers who harvest this data and use it to send spam to the e-mail addresses. In some cases where specialist directories might be developed by particular communities there are other types of data that they wish to share [between them] but do not want to be publicly accessible e.g. details of location of rare or endangered species, budgetary issues etc. While harvesting of data such as e-mail addresses can happen already via the organisations Web sites, clearly data provided by any specialist directory is more focused and centralised, thus potentially more easily harvested and valuable to spammers.

This list is not exhaustive but does demonstrates that current models for the creation, access and maintenance of such directories, have significant limitations, many of which map very well to potential solutions offered by the generic architecture and approach in sections 2 and 3 above. While it is clear that these approaches cannot offer complete solutions, to what are in many cases fundamentally complex and difficult issues (e.g. keeping data *absolutely* up-to-date), they do in many cases provide significant advantages e.g. once an organisation updates its own profile that update can be almost immediately propagated to all systems that make use of that data. The limiting factor is the organisation itself updating the information. Which, using these approaches, they only have to do once, as opposed to systematically notifying all those who hold the out of date information.

Need and Application Areas - As part of the survey of Biodiversity and Wildlife information [☞SWARA] a wide range of applications for an environmentally focused organizational directory were identified - depending on the exact nature of the information held. Below are listed some of these:

- **General public** seeking information about wildlife/biodiversity or environmental topics, locations, species etc.
- **Educationalists and Students** teaching and learning in these and related areas who wish to find teaching and learning materials, resources and information.
- **Academic Researches** seeking specialist organizations, contacts or partners as part of their research activities.
- **Media or Other Non-specialist Researchers** seeking specialist resources or advice on particular areas or species etc.
- **The Organizations Themselves** [including voluntary sector, research institutions, government departments and other statutory bodies] wishing to find or make contact with related organizations or find specialist information.
- **Businesses** seeking environmental advice and services e.g. as part of environmental impact assessment processes.
- **Other 'directory' providers** who currently provide (or wish to provide) directory listsings including environmentally related organisations and projects and or supplementary information would be able to draw on the distributed data as a component of the data that they provide. In many cases such directories form part of a larger publication or Web site e.g. The Sustainable Careers Handbook (☞ http://www.cat.org.uk/catpubs/book.tmpl?sku=SCH&cart=327658133122482) and The Stakeholder's Guide to Sustainable Waste Management (☞http://www.wasteguide.org.uk/) both of which contain substantial contacts sections to compliment their content. These could utilise the environmental directory data (adding their own complementary data where necessary), thus reducing maintenance costs.

The list above is based on the kind of use that might be made of an environmental directory such as Who's Who in the Environment [☞WWE], which contains generic data e.g. names and alternative names, basic contact details, short text descriptions and legal status (type of organization). However with the addition of supplementary specialized data the potential uses of the data become much more broad.

5 Requirements

High Level Requirements - Given specific issues outlined above both generic and specific to the Environmental domain an initial list high level requirements for the demonstrator, is given below. *This list is not intended to be definitive but can act as the basis for a more focused set of requirements, that would be developed as part of stakeholder and user requirements analysis and a user study.*

- **Data should be up to date and easily maintainable -** or where up-to-date data is not available indication of how recent the data is.
- **Data should be validated to at least a basic level -** the information should be validated (or its validation status indicated) by the organisation itself and an external authority should confirm that the organisation exists and the core information is correct. This is necessary to ensure that at a minimum that 1) at least users can begin to validate the information themselves via the organisations Web site and 2) that inappropriate organisations do not attempt to register as part of the directory.
- **There should be a low [ideally very low] barrier of entry to organisations wishing to be in the directory -** including very low financial, time or personnel cost and low levels of technical expertise, on the part of organisations in the directory.
- **The system as a whole should be robust and sustainable** with respect to changes in priorities, policies or disappearance of any individual service providers, projects or funding streams.
- **The system should provide a simple to use but extensible classification schema** based on open metadata schemas and vocabularies where possible.
- **It should be possible to create a range of effective user interfaces (UI)** to provide easy and intuitive access to data via different views e.g. browsing and searching via organisation type, locations, areas of interest, species, etc.
- **There should be visualisation tools** [as part of UI] that allow users to obtain graphical representation of data e.g. geographic, topic relationships and organisational relationships.
- **The system should enable other services and data to interoperate easily -** the system should be set up such that the core data (see below) is maintained by the individual organisations within the directory, or their nominated proxies. The data should be open (i.e. it is always possible to down-load the original RDF file) and aggregation service(s) should be based on open standards and ideally be an Open Source project. This is so that even if the primary/initial service provider

ceases to provide the service a new provider can set up quickly or competing/ complementary services can run simultaneously.

Once again these requirements map well onto the characteristics of the generic approach and architecture detailed in sections 2 and 3 above.

Requirements Specific to the Environmental Directory - As discussed in ☞section 3.2 (Architecture layers) above, the generic architecture for the demonstrator has been designed to be customizable to meet the specialized needs specific to a community or group of communities. This section details an initial list of requirements for the environmental directory case and relevant communities. However it is important to understand that *these will almost certainly change and be added to, once deeper engagement with the relevant communities begins.*

Core Data & Ontologies, Thesauri and Vocabularies

It is envisaged that the directory will contain core data that will provide the basic level of directory data, *some* elements of which should be mandatory. At this stage we believe that this might include: (see also ☞section 3.2 above)

Organisational Ontology. It is envisaged that there will be a cross-domain organisational ontology that can act as the basis for interoperability between communities and domains. e.g. current name, primary address, primary telephone and e-mail addresses, textual description, etc. However, within this generic structure specialisation may be required. Specialisations might include:

- **Secondary Organizational Names**: it is known from existing directories (e.g. Who's Who in the Environment' [☞WWE] that many environmental organization have a*lternative names* and *previous names* and *acronyms* by which the organization is very commonly known, and should be available as part of the data set. It may be that these requirements can be made part of the cross-domain ontology.
- **Organizational Type:** It is likely that it is possible to create a cross domain ontology of 'organisational type' e.g. Government Department, Non-Departmental Public (Statutory) Body, Commercial Sector, Voluntary Sector, Educational Establishment, Research Institute, etc. In many cases a particular type of organisation has associated data e.g. Voluntary sector organisations may be charities and have a Charity Commission Number in the UK. In addition in the case of the environmental directory it is desirable to describe *'projects'* and *'initiatives'* as well as organization and it may be that there are specialization of the generic organisational types that are specific to the environmental domain, e.g. Grant Giving Trust/Body. It is envisaged that organisations may have multiple entries..
- **Organisational Events**: There are a range of significant events in the life cycle of organisations and projects, e.g. date of formation, date ended or ceases to exist, date of name change and date of merger with other organisations. Such events are particularly common at the level of projects.
- **Members of Staff -** [Number of]: This piece of data was provided in the Who's Who in the Environment [☞WWE]. It is an example of contextual data about an organization, that in the case of an organizational directory provides contextual information about the nature (e.g. size), of the organization. Other more specialist data might include annual financial statistics or areas of land owned.
- **Volunteer Activity**: Volunteering activity is a major part environmental activity in the UK especially in the voluntary sector, where many organizations may be run entirely by volunteers or volunteers with a small number of paid workers. Information that is likely to be required includes; does the organisation take volunteers, if so how many and for what types of activity - an ontology of types of volunteer activity would be required.
- **Membership**: Membership organizations play a very large part in the environmental sector. Indeed the National Trust and the RSPB are two of the largest membership organizations in the UK. Information about membership might include; does the organization take members, if so are there any restrictions, what types of membership are available, what are the benefits and what are the costs.
- **Local Branches/Secondary Locations**: Many environmental organisations have 'local branches' that is, local offices or contact points or groups. These often provide particular services to members, customers or others. Data may include the location and contact details of the branches and their role(s) or facilities available.
- **Relationships between Organisations**: In many cases there are highly significant relationships between organization within the environmental sector. This is especially the case with the relationships between projects, initiatives and organization, where organizations are partners or funders or initiators, etc. Other examples include trade associations, or federations of smaller organisations (e.g. Federation of City Farms - ☞http://www.farmgarden.org.uk/). It would be very useful for many types of query to be able to have access to information about such

relationships. It may be that such relationships are generic enough to be included in a cross-domain ontology.

Ontology of Type of *environmental activities*: While 'environmental activities' could be seen as part of the organisational ontology, it might be more usefully thought of as independent from organisations, since such categories can be reused in contexts other than organisations. The environmental sector is very broad with regard to the type of activity(s) that are undertaken by organisations/projects e.g. volunteer coordination, regulation, practical conservation, etc. Such activity is a key means by which organisations are differentiated. Examples include: regulatory, advisory, consultancy, conservation (sub-divided into habitat and species), educational (subdivided by phase or type of education), information provider, campaigning, learned society, volunteer coordination etc. In general, for specialist use, the broad categories can be subdivided into specialist categories.

Geographic Range: In the case of many types of organisation the concept of location is restricted to the physical contact address of the organisation. However in the environmental domain (and many others) the concept of location is more complex. In the case of the environmental directory the concept of 'geographic range' (or similar) is necessary. Many organisations work only within particular regions - countries, counties, and other administrative borders, but also across other types of region e.g. a National Park or Local Nature Reserve that might span a number of administrate borders or all National Parks. Such information is important for many types of user query related to location. Therefore representation of geographic range may be an area where a balance must be found between expressiveness and practicality both technical and for organisations to submit their data.

Topics: All organizations have particular foci, i.e. what their activity is *about*. In the environmental sector these might include; particular environmental issues e.g. pollution or energy production, or particular types of species or habitats e.g. birds or wetlands. These foci are likely to be critical for the majority of users of an organizational directory especially where they do not already know of relevant organizations and their inquiry is focused on a particular interest or problem or requirement. This categorization of organizations is likely to be a significant issue for this project - there appears to be no existing categorization that has the breadth and yet specialization required for such a directory, e.g. GEMET [☞GEMET] an Environmental Thesaurus designed for the categorization of environmentally related documents, has breadth (and is possibly too wide) but does not provide significant depth where it likely to be relevant to a directory e.g. lower levels of groupings of animal and plant species. It seems likely that a topic schema may best be created from a set of external vocabularies where they exist with the facility to extend when necessary. Of course where specialist versions of the directory/portal are developed it is very likely that the developers will require the ability to create their own or extend existing schema in order to meet their specific needs.

Products: Using the term very generically this might include publications, equipment, software etc. it is possible that comprehensive high level product ontologies/vocabularies exist, however it is likely that any existing system will require refining to describe products related specialised sub-communities. It may be best to use and/or develop a range of ontologies related to different types of product e.g. publications and environmental monitoring equipment.

Services: Environmental organisations and projects provide many types of service e.g. consultancy, advice, grant giving, educational, providing speakers and presentations, image or media libraries. As with products (see previous paragraph) it may that a single existing ontology/vocabulary exists however it it likely that highly focused portals might require the ability to develop their own specific categories.

Relationships between organisations and projects: As discussed in the previous section relationships between organisations, projects and initiatives are often important in many types of query, some examples of such relationships might include:

- Part_of / has_part
- Initiative_of / has_initiative
- Member_of / has_member
- Partner_in / has_partner
- Affiliated_to / has_affiliate

Validation and moderation

It seems likely that for the initial demonstration a simple level of validation will be sufficient. While the accuracy of the data is important, since it is being provided by the organisations themselves they can reasonably be expected to be responsible for its accuracy. However it is necessary that a validation process should ensure that 1) the organization is a bone fide 'environmental organisation' (e.g. it is not an Internet porn site) and 2) that the data is provided by the organization that it purports to be about.

In the case of 1) this requires that there is some explicit criteria for being included in the directory and have some mechanism for ensuring that new entries meet these criteria. 2) is more complex but a basic level could simply be that the RDF profile file is located in a directory of the organisation's own

Web site.

In practice both 1) and 2) might be done in a number of ways ranging from individually inspecting the submissions as the come in and before they are entered into the directory though to automatic entry into the directory with an 'unvalidated' flag, used to signal that the data is unvalidated in the user interface. Once they are validated the flag can be set to 'validated'. We would expect that such simple validation would take about approx. 5mins per entry.

It is also possible to signal sources of data using the user interface e.g. via colour - one colour for data provided by the primary organisation and another for other secondary or added data from other sources.

Sensitive Data

The only piece of data that is likely to be sensitive in core data is probably the e-mail addresses which could be used as a target for e-mail spam as discussed above. Any one of the generic solutions to this discussed in ☞section 3.4b above should be sufficient to significantly reduce the risk of this.

Core Scenarios for Demonstrator? - The requirements outlined in this section provide the basis for the specification of exactly what the Semantic Community Portal Demonstrator will do. It is helpful to provide illustrative scenarios to help focus those requirements. This section details scenarios which will form the basis of the functionality of the demonstrator. This is in four parts:
0) the creation an uploading of the RDF profile file by an environmental organisation
1) the basic directory that provides access to the core data about the environmental organisations and
2) annotation of the existing data using the existing properties e.g. filling in 'missing data' and
3) extensions to this basic data that demonstrate how specialist user communities can customise and augment the data and interfaces to meet their needs.

0) Tools to allow organisations, projects and initiatives create their profile

Summary: Representative of an organisation finds out about the directory. The organisation decides to be involved and a representative creates RDF record and has it uploaded it to the organisation's Web server. The data is then validated and made available

1. Organisational representative (Jo) receives an e-mail notifying her of the directory and how to add the organisation to it.
2. Jo looks up the directory Web site, browsing and exploring to decide if they want to be involved and if so what is involved on their part - this includes going to the 'add organisation' section and reading how it works and looking at, and experimenting with (and possibly printing off) the data entry form.
3. Jo talks to other members of the organisation, including their Web master, who checks out the technical issues on the directory Web site. They agree that the organisation wishes to have their information in the directory.
4. She returns to the Web site and follows the 'add organisation' link. This takes her back to the data entry form. She re-reads the 'how to' section and completes the form and presses the 'send' button. She is notified that she has not completed all the mandatory fields and that the e-mail address is not valid. She corrects these errors and presses 'send' again. This time the request goes off.
5. 2-10mins later she receives an RDF file by email. She forwards this to the Web master, who uploads the file to the root directory of the organisation's Web site.
6. The directory harvester program downloads the file and adds it to the directory's RDF database, setting the 'data_validated' property on the record to 'no'. This would happen nighly by default but step 4 could also notify the harvester to check for this new entry more frequently, e.g. hourly.
7. At the end of the overnight 'harvesting' process. The validation module is initiated and sends an e-mail to the person responsible for validation, noting that 4 records have been added to the database.
8. The validator checks that the organisations a) are genuine environmental organisations (using pre-defined criteria), b) that the URL is that of the domain name of organisation or an appropriate proxy organisation and c) that the data does not contain any inappropriate content.
9. This is the case for our record and the validator signs the records off for being made accessible.
10. The 'data_validated' property on the record is set to 'basic' and other administrative metadata about the record is updated (e.g. date and time of validation and ID of validator) to indicate that a basic level of validation has taken place.

11. The system makes the data available to users.

0b) Organisation edits their profile

Summary: The organisation in Scenario 0) above recieves an e-mail by a user notifiying them that their directory listing contains an error - part of their address is misspelled. They edit the RDF profile file and up-loaded the new file to over write the old incorrect version.

The organisation in Scenario 0) recieves an e-mail by a user notifiying them that the address in their directory entry is spelled incorrectly.

1. An organisational representative (Jo) confirms that this is the case by visiting the directory Web site.
2. She then visits the 'update or edit a record' page on the directory Web site - this provides a Web from into which she pastes the address of the RDF profile of the organisation (held her organisations server). The form then returns a data entry form for the directory filled in with the data from the RDF profile.
3. Jo corrects the error and presses the 'send' button as in step 4. of scenario 0) above.
4. 2-10mins later she receives an RDF file by email. She forwards this to the Web master, who uploads the file to the root directory of the organisation's Web site.
5. When the directory harvester program next downloads the file it replaces the original (incorrect file) with new one. As the organisation and file location have already been validated it is not necessary to re-validate the data, unless the location of the RDF file or the URL of the organisation are changed.

1) Basic directory Web site showing aggregate of self-published data:

Summary: A user comes to the Directory aiming to find out if there is a membership organisation that might support their recent interest in butterflies, they search the site and use the filtering facilities to locate potential organisations.

A user is conducting background research into the organisations focused on a particular topic or area of work (e.g. conservation of a specific type of animal or plant species, waste management, campaigning about a particular environmental issue…). This may be for a very large number of underlying reasons e.g. personal interest, part of a formal course of study, seeking advice or expertise, careers planning, seeking project partners...

Such activity is likely to be motivated by a desire to identify relevant organisations, assess which (if any) are appropriate to help answer their particular need and then follow up this 'explorative' searching with more focused interactions with the directory data and external Web sites and other information sources.

Context: The user may well have been directed to the Environmental Directory via a link from another related site or a search engine such as Google. They are thus likely to have some idea of at least some of the organisations that they are seeking information about and are likely to use this knowledge as part of their evaluation criteria of their initial interactions e.g. do their initial searches come up with the organisations that they are already aware of?

Examples of specific information sought might include:

Gaining an Overview:

- A sense of just how many organisations are active or have an interest in, their particular area of interest
- What types of organisations they are
- Which are the most active in their 'topic' area
- If there are many results they may experiment with various filtering activities (e.g. filter by organisation type to see which/how many are commercial companies or voluntary or by geographic range or location) depending on their particular query. [This may be the end of their interaction, as they may have answered their query]

Focusing Down

- Once this overview is obtained, it is likely that the user will want to identify a small(ish) number of organisations that are most relevant to their need and find more detailed information about them.
  They may wish to 'collect' this data together for later use (e.g. by printing off their details) or compare organisations.
- It is likely that once a small number of organisations have been identified the user will follow links to the Web pages of the organisations themselves - returning to the directory in between if

their need is not met.

- They may wish at any point to print off or import the results that they have - these might be basic search results or full records.
- They may find it helpful to 'collect' records from various searches and print or import those at the end or an intermediate stage of their query.

Example of a session:

Specific Context: the user is a member of the public, they have developed an interest in butterflies and wish to find out what organisations exist that might be able to support that interest - they have in mind something like the equivalent of RSPB (in the case of birds).

- The user enters 'butterflies' in the main search box
- They are returned results [algorithm based on text search of main properties, e.g. name, long description, topics, …] prioritised in order of relevance [tbd. - initially possibly simply on number of occurrences] - the results show the name, brief description and other data e.g. legal status, geographic range, top-three topics.
- They browse the results, reading the names and other data.
- They click on the most likely organisation e.g. 'Butterfly Conservation' - they browse/read down the page and notice that there is a 'membership' property.
- They click on the 'yes' of the membership field in the record and are taken to a new set of search results filtering the current results by 'membership=yes' - this returns a small list <5 organisations. They browse this list and view the full record of each one.
- They follow the URL links to Web sites of two of the organisations. Returning from the first but not the second.
- User makes contact with the organisation of their choice via e-mail address gained via the organisation's Web site.

2) Annotation with data about relationships between organisations, projects and initiatives to give second view and augmented search

Summary: The organisation that has created the directory aims to add to and make more complete the data available via the Directory.

The organisation that has created the directory aims to add to it and make it more complete. They commission a researcher to add to what are likely to be incomplete 'relationship' details - these are not mandatory within the basic profile - the aim is to provide this data for the most popular [most accessed] 50 organisations within the directory. This data is used to create a customised browsing facility that allows users to navigate though the data via relationship data.

- A researcher is commissioned by the organisation that has created the directory to augment their existing core data by adding missing data about the simple relationships between the organisations that are already represented within the core set of properties. (e.g. suggested relationships include: Part_of / has_part, Initiative_of / has_initiative, Member_of / has_member, Partner_in / has_partner, Affiliated_to / has_affiliate).
- The researcher conducts research into these relationships using the Web sites of the organisations themselves.
- The directory developers then develop a new interface that is focused on the relationships between organisations - this may be visual or text based. The interface is designed such that the source of the data in any particular case is signaled e.g. by the organisations themselves or by the directory organisation.
- The relationship property is now added as default to the faceted search interface of the directory to allow users to search using this property. The interface signals that this is not a mandatory piece of data and many not be complete for all entries.
- The researcher contacts the organisations themselves and asks them to confirm that 1) the relationships are correct and 2) that they are happy for the data to be made available to the public - this is not done prior to the research as it will not be known which organisations will be affected.

3a) Extension: adding additional data (e.g. information about publications from organisations)

Summary: A 3rd party organisation decides (for their own needs and that of their academic discipline) to compile a list of 'gray literature' on nature conservation activities in the UK. They use the environmental directory infrastructure and distributed data to bootstrap the process, working in partnership with the publishing organizations.

A 3rd party organization (e.g. a specialist University Research Library) wishes to create a bibliographic database of publications about nature conservation activities in the UK. However much of these publications are 'gray literature' i.e. "That which is produced on all levels of government, academics, business and industry in print and electronic formats, but which is not controlled by commercial publishers." (source ☞http://www.nyam.org/library/greylit/whatis.shtml). Such literature is often very hard to locate and access.

The Library decide to contact the relevant organisations and ask if they can publish their bibliographic data so that it can be collated - the organisations agree. The organisations already publish their profiles for the environmental directory as RDF. The Library therefore decide to use the environmental directory architecture as the basis of this activity.

Then the Web developer from the Library visits the environmental directory Web site and reads the documentation. They download the development kit and install the software on their development server. They begin by creating a standard 'directory' out of the box' using the default settings.

They then customise the RDF profile creation tool to produce the bibliographic data using (based on an external schema) as RDF. This is then used in collaboration with the organisations to produce the bibliographic data in an RDF format.

The data is published as with the basic profile data by the organisations - this is harvested by the Library and the data collated and made available in a searchable format via the Libraries Web site. It is also converted from RDF into the format used by the Library for their own Internal bibliographic data management.

3b) Extension: indexing of additional data (e.g. publications by topic classification)

Summary: The original environmental directory organisation decides that since the bibliographic data now exists in RDF that they will extend the data in the directory to include access to this data and index the new bibliographic data using the internal topic ontology.

The original environmental directory organisation decides that since the bibliographic data now exists in RDF that they will extend the data in the directory to include access to this data. They add a specialist search (based on free text in the title, keywords and description) for bibliographic items to the site and link to publications from the data records of the relevant organisations.

The environmental directory organisation decides that it would be very helpful (to their users) to have the bibliographic data indexed using the directory's topic classification scheme. They create a customised version of the organisational profile data creation system to allow indexing of the bibliographic data.

One of their voluntary staff indexes the items under the topic schema. The data is integrated with the existing data and the bibliographic search interface is extended to allow searching and browsing of the data by topic.

3c) Extension: indexing of additional data by 3rd parties (e.g. indexing publications by National Curriculum (☞http://metadata.ngfl.gov.uk/) topics and commentary from teachers or relevant organisations)

Summary: An organisation working to support the school curriculum, indexes the new bibliographic data using the National Curriculum indexing scheme, to improve access to the resources for schools.

An organisation specialising in Education for Sustainable Development notes that the bibliographic database is potentially very useful for members and other teachers since it provides details of normally 'invisible' resources such as posters, leaflets and teaching packs produced by conservation and sustainable development organisations - they decide that it would be more useful to teachers it the items were indexed these using the National Curriculum indexing scheme (☞http://metadata.ngfl.gov.uk/). Their Web developer goes though a similar process to that of the developer in 3b) above, but creating a data entry form interface to facilitate the indexing of the bibliographic items using the National Curriculum metadata data terms.

The developer also uses the development tools to customise an interface template to add the National Curriculum data to the search and navigation interfaces of the bibliographic data interfaces from the original environmental directory.

Given the scale of the information involved an complementary scenario would be for the initiating organization to simply configure the data entry and search tools and then to encourage groups of school teachers to index resources they find useful *as they go*. This incremental, peer-based, approach may result in imperfect data but in the long run may be more maintainable than the centralized approach.

6 Summary and component requirements

Finally we summarize these discussions by enumerating the components that will be needed for the proposed demonstration and the key features required for each.

The aim of the demonstrator is to provide the system design, starting ontologies and open source component implementations sufficient to support the basic portal described by scenarios 0-2 above, and to provide a foundation for future proof-of-concept demonstration of the extension scenarios 3(a)-(c).

This will require the following components:

Type	**Component**	**Comments**
System design	Representation and dataflows for validation and moderation requirements.	Iterative approach. Start with simple black list and self-validation, then add third party validation workflow.
	UI templating scheme	Set of ontology annotation properties used to enable template-driven UI for viewing and editing.
Core ontologies	● organization ● organization relations ● organizational event ● geographic reach ● time of event	Should be structured for modular reuse. Ideally should include partial mappings to related ontologies for the same topic.
Domain ontologies	● environmental activity ● environmental topic ● environmental resource	Minimal sets of description used as controlled vocabularies for describing the activites, topic areas and resources make available by the organization. Probably expressed as thesauri rather than full ontologies.
Software components	Editor	● template driven creating of new records ● find records to extend/link using identifying patterns ● support for controlled vocabularies (flat, structured) ● support for extending controlled vocabularies ● support for discovering correct term to use from across known ontologies
	Aggregator	● Aggregate from web pages, semantic blogs, other aggregators. ● Records provenance. ● Ability to block sources based on moderation policy. ● Support aggregation of ontologies and (extensible) thesauri of controlled terms ● Optional "live push" to enable fast corrections as well as overnight harvesting ● Logging of access patterns.
	Viewer	● Create browsable page views onto data based on the hinted-ontologies and layout templates ● Multidimentional search capability based on the ontologies ● Linking to editor component for annotation of entries ● Future extensions to support graphical visualization of whole dataset - beyond scope of demonstrator.

	Optional hosting	● user accounts for hosting data files posted by individuals ● base on Joseki

A References

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

[YAHOO]
☞http://www.yahoo.com/

[DMOZ]
The Open Directory project
☞http://www.dmoz.com/

[COGNET]
MIT Press COGNET
☞http://cognet.mit.edu/

[MATHNET]
The Math-Net initiative
☞http://www.math-net.de/

[FISHBASE]
FishBase - a global information system on fishes
☞http://www.fishbase.org/

[FISHBASE-SEARCH]
FishBase - search interface
☞http://www.fishbase.org/search.html

[PLANET-ONTO]
KMi Planet technologies
☞http://kmi.open.ac.uk/projects/kmi-planet/

[SEAL]
Semantic portal - The SEAL approach, A. Maedche and S. Staab and N. Stojanovic and R. Studer and Y. Sure, In *Creating the Semantic Web*. D. Fensel, J. Hendler, H. Lieberman, W. Wahlster (eds.) MIT Press, MA, Cambridge, 2001.
☞http://citeseer.nj.nec.com/maedche01semantic.html

[FORTH-PORTALS]
Querying CommunityWeb Portals, G. Karvounarakis, V. Christophides, D. Plexousakis, and S. Alexaki (2000)
☞http://citeseer.nj.nec.com/karvounarakis00querying.html

[SIDEREAN]
Siderean software demonstrators
☞http://www.siderean.com/demos.jsp

[TAP]
TAP: Building the semantic web
☞http://tap.stanford.edu/

[OWL]
Web-Ontology working group
☞http://www.w3.org/2001/SW/WebOnt/

[RDF]

The Resource Description Framework
☞http://www.w3.org/RDF/
☞http://www.w3.org/2001/SW/RDFCore/

[RDFS]

RDF Vocabulary Description Language 1.0: RDF Schema
☞http://www.w3.org/TR/rdf-schema/

[ARKIVE]

ARKive - images of life on Earth
☞http://www.arkive.org/

[ARKIVE-COI]

Adding value to large multimedia collections through annotation technologies and tools: Serving communities of interest., 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, Shabajee, P., Miller, L. and Dingley, A. (2002) Available:
☞http://www.archimuse.com/mw2002/papers/shabajee/shabajee.html

[RSS]

RDF Rich Site Summary - history and resources
☞http://www.oasis-open.org/cover/rss.html

[NetAPI]

RDF Net API - W3C Note, Member submission, Graham Moore, Andy Seaborne
☞http://www.w3.org/Submission/rdf-netapi/

[SWADE-THESAURUS]

SWAD-E Thesaurus activity
☞http://www.w3c.rl.ac.uk/SWAD/thesaurus.html

[ePERSON]

The ePerson Snippet Manager: a Semantic Web Application, Dave Banks, Steve Cayzer, Ian Dickinson, Dave Reynolds. HPLabs Technical report: HPL-2002-328 .
☞http://www.hpl.hp.com/techreports/2002/HPL-2002-328.html

[WORDNET]

WORDNET - a lexical database for the English language.
☞http://www.cogsci.princeton.edu/~wn/

[VORA]

Towards a theory of variable privacy, Poorvi Vora, in review.
☞http://www.seas.gwu.edu/~poorvi/plv_variable_privacy.pdf

[FOAF]

The *friend of a friend* project.
☞http://www.foaf-project.org/

[P2P REPUTATION]

Reputation, Richard Lethin, in "Peer-to-peer: harnessing the power of disruptive technologies", Ed. Andy Oram, O'Reilly, 2001.

[PGP]

The International PGP project
☞http://www.pgpi.org/

[WWE]

Who's Who in the Environment, Environment Council, 1995

[EnvCouncil]

Environment Council
☞http://www.the-environment-council.org.uk

[SLV]

Environmental Directories, State Library of Victoria
☞http://www.slv.vic.gov.au/slv/refresources/environment/direct.htm

[TGD]
The Green Directory, Green Directory Ltd, 2001
(Also avaliable at: ☞http://www.thegreendirectory.co.uk/home.cfm)

[EarthScan]
World Directory of Environmental Organizations, EarthScan, 2001
☞http://www.earthscan.co.uk/asp/bookdetails.asp?key=3426

[WDEO]
World Directory of Environmental Organizations, California Institute of Public Affairs
☞http://www.interenvironment.org/wd/

[EnviroLink]
Environmental Resources, EnviroLink Network
☞http://www.envirolink.org/categories.html?catid=5

[EOWD]
Environmental Organization WebDirectory!
☞http://www.webdirectory.com/ & ☞http://www.webdirectory.com/Wildlife/

[DOIAEMER]
Directory of Organizations and Institutes Active in Environmental Monitoring, Environmental Research Information System IHGE & UFIS
☞http://www.webdirectory.com/ & ☞http://www.webdirectory.com/Wildlife/

[GEMET]
GEMET
☞http://eionet.eu.int/GEMET

[SWARA]
Summary SWARA biodiversity information survey, P. Shabajee, 2003
Publication pending

[NBN]
National Biodiversity Network
☞http://www.nbn.org.uk/

B Selective Glossary

[ONTOLOGY]

We use the term *ontology* to refer to an explicit conceptual model of the domain in question. In this sense an ontology is an engineering artifact. It provides a description of the concepts (e.g. Person, Organization, AnimalSpecies) which are relevant to the domain, the properties associated with those concepts (e.g. name, activity) and the formal relationships between those concepts and properties (e.g. subClassOf, inverseOf).

This is the common usage of the term in computer science and the semantic web, it is related to, but narrower than, the broader philosophical concept of the metaphysics of *being* and *existence*.

Ontologies are related to, but not the same as, schemas. The term *schema* in databases or XML terms is typically used to describe a specific syntactic representation. This differs from the notion of an ontology in two ways. Firstly, a schema would typically not formally capture the logical relationships between concepts and properties such as sub-class relationships. Secondly, a given property or relation described in an ontology may be represented in several different ways, they are not tied to a specific syntactic representation.

[THESAURUS]

A thesaurus comprises a vocabulary of words and the relationships between them. Thesauri can, but don't always, separate the words themselves (indicator terms) from the concepts being represented. Thus they can relate closely to the notion of an ontology - though typically an ontology will include more structural information and will be more formal. In particular, it is common in thesauri to using the general relationships "broader term (BT)" and "narrower term (NT)" without formal restrictions on their usage. This can lead to them being used for any or all of *part-of*, *isa, subClassOf* and *intersecting-class* relationships.

C Demonstration – additional use cases and scenarios

This appendix contains aditional use cases for the environmental directory which are outside the scope of the demonstrator project but may usefully influence some of the detailed design decisions.

End User Focused Use Cases - Use case 2a: University student undertaking a project about recycling and waste management in the UK.

This situation would probably be in relation to the student needing to answer specific questions related to for example economics, policy development, planning, public perception, environmental impact, etc. The student may come with specific goals in mind or as part of their initial research to 'scope' the subject area. It is likely that this would be used in conjunction with other data sources.
If we take the first case, specific goals might include:

- Identification of organisations and projects that are involved in recycling and waste management in the UK
- The nature, aims, roles and activities of those organisations - in particular finding those that are most relevant to their specific focus
- Materials and data published by those organisations esp. grey literature

Given that part of the goal of conducting project work is to help students develop their research skills, it could be assumed that such students would use any standard and advanced search and browse facilities. Given the diversity of this group, it is unlikely that there are any particular common but specific types of functional requirement as distinct from other types of researcher activity.

Use case 2b: Member of general public seeking information about environmentally related volunteering opportunities in their area:

In this case the user wishes to find out about organisations and projects in their area that take volunteers related to a specific 'topics' e.g. conservation, animal welfare, environmental education, transport etc., and might have particular types of volunteering activity in mind e.g. office support, practical conservation, etc.
While it is possible for the user simply to use an advanced (faceted) search with 'volunteer accepted' = yes and their 'topics' of interest and browse the search results. It might be more effective to provide a specialist volunteering' interface. This might provide specialist browsing and search interfaces e.g. only make available organisations that have explicitly indicated that they accept volunteers and allow browsing via 'topics' and/or 'type of activity', providing search results with mini-records showing 'topics', address(es) and/or 'type of activity'.

Use case 2c: Conservation Organisation Project Officer researching background information related to issues within a new local conservation project:

In this case the user is seeking background information about a specific local or organisational conservation project, within which there are some specific aspects that are outside of their own (or their organisations) experience. The user wants to find contacts that will be able to help them answer their questions or offer expertise that they require (e.g. dealing with a species, piece of legislation, type of practical work…).

It is likely that the user could find potential organisations via the advanced search or a faceted browse, which of these is more appropriate will depend on their specific requirements. However in general the user is likely to use a 'simple search' as the starting point, unless the interface prompts them into other approaches. Thus feedback from searches that highlight the potential for advanced searches or browsing might be helpful.

Use case 2d: Media Researcher conducting background research on a species for a TV or Radio programme:

In this case the user is seeking specialist up-to-date information about a species. They are likely to be seeking 'consultant' contacts and information; they may also be seeking media to use as part of their programme.

In general the user is likely to come to the site after having conducted more basic research about the species using reference books and the Web. They are probably seeking contacts to obtain 'expert' information about the species, access to video or still images, and possibly potential experts to be part of the programme itself. This category of user is likely to be very experienced in using advanced search facilities - again the key issue is to signal that the terms that they wish to search on e.g. 'type of service - consultancy' and topic, are available to them.

3rd Party Use Cases - Use case 3a: PhD student conducting historical research into conservation in the UK:

In this case a 3rd party research (a PhD student) is conducting research into the history of conservation in the UK. As part of their background research they attempt to collate a timeline of the development of organisations involved in conservation since 1900. They come across the environmental directory and realise that the infrastructure provides a basis for capturing the information that they require.

The contact the directory organization and discuss the potential of using the directory infrastructure to collect and make available their research data. They begin by using the data entry/capture tool developed as part of the original directory project to capture their data.

They publish this data (RDF) files, which is harvested and becomes available to the users of the environmental directory. However they themselves wish to develop a new interface to locate and relate the data from a historical perspective. The work with a fellow student and technician in their department to download the development kit and create their own directory for internal use. They use the tools to customise data-based views allowing display of time ordered data and develop a new visualisation tool to help visualise the growth and changing relationships between organisations over time.

Once they have competed their research they publish their results in research reports and papers and make their internal version of the directory available externally. They are contacted by another research project and asked if the new organisation can duplicate their data and interface for a related project. This is done and the parent of the research project, with the agreement of the original PhD student, decide to run the historically focused interfaces as part of their own Web site, as it would be helpful to their users.

D Changes

25-10-2003
Initial draft.

3-11-2003
First release.