STERNA

Semantic Web-based Thematic European
Reference Network Application

Technology Watch Report

Guntram Geser | January 2009

WWW.STERNA-NET.EU

STERNA

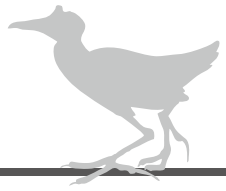
Semantic Web-based **T**hematic **E**uropean
Reference **N**etwork **A**pplication

Technology Watch Report

A Report on Semantic Approaches for Including
Digital Cultural and Bio-Heritage Resources
in the European Digital Library Initiative

Guntram Geser | January 2009

Table of content



Context, perspectives and recommendations

1 Context, perspectives and recommendations

1.1 Project context

This Technology Watch report has been produced by Salzburg Research in the period July to November 2008 as part of the STERNA project task 6.5: Cluster Activities, which forms part of work package 6: Network Extension and Deployment.
The STERNA (Semantic Web-based Thematic European Reference Network Application) project is co-funded under the *e*Content*plus* programme as a Best Practice Network in the target area of digital libraries (cultural and scientific/scholarly content).
STERNA has a formal duration of 30 months, running from June 2008 to November 2010, hence, the Technology Watch activity has been among the first tasks of the project.

The STERNA consortium

The STERNA consortium comprises 12 European natural history museums and other institutions that collect and hold content on biodiversity, wildlife and nature in general, the project coordinator Salzburg Research (Austria), and the technology provider and implementer Trezorix (Netherlands).
The natural history museums and other institutions are: Archipelagos (Greece), DOPPS BirdLife (Slovenia), Heritage Malta, Hungarian Natural History Museum, Icelandic Institute of Natural History, Natural History Museum of the Municipality of Amaroussion (Greece), Natural History Museum of Luxembourg, Naturalis, Natural History Museum of the Netherlands, Netherlands Institute of Sound and Vision, Royal Museum for Central Africa (Belgium), Teylers Museum (Netherlands) and Wildscreen/ ARKive (UK).

Pioneering semantic enhancement and integration

As a Best Practice Network project co-funded under the ***e***Content***plus*** programme, STERNA is pioneering the semantic enhancement and integration of digital resources from different partners' databases based on Semantic Web standards and techniques.
STERNA is understood as a showcase project of using such semantic enhancement methods and the capability they provide to link, search and access content from distributed and heterogeneous databases in novel ways.

Relation to the European Digital Library initiative

The STERNA consortium aims to contribute to the objectives and realisation of the European Digital Library (EDL). The overall objective of the EDL is to make Europe's cultural and scientific heritage accessible to all. The EDL will serve as a common multilingual access point to the digitised heritage resources that are held in databases of the participating organisations across Europe.
While over the last about ten years ever more cultural and scientific resources have been digitised and made accessible via the Internet, integrated semantic search of, and access to, resources across many heterogeneous databases is still difficult to achieve.
It is envisioned that the Semantic Web approach of using a machine-processable semantic information layer will leverage the access via the European Digital Library to the digitised content of both large and small cultural and scientific heritage institutions.
The goal to realise interoperability of cultural and scientific heritage resources based on a semantic layer is recognised by the Europeana project that has developed an EDL prototype portal.
The portal was launched on the 20th of November 2008, but it was not Europeana's intention to demonstrate semantic capability. Their technological roadmap for the future EDL, however, suggests to achieve semantic search and other capability by making use of Semantic Web standards such as RDF (Resource Description Framework) and SKOS (Simple Knowledge Organisation System).
This is the approach taken by STERNA and, as the report shows, also some other projects that focus on cultural heritage content.

1.2 Function and approach of the technology radar

Function of the Technology Watch

The function of the Technology Watch activity is to provide STERNA and other projects related to the European Digital Library with a "radar" that identifies initiatives in content/metadata enrichment and integration of heterogeneous digital collections based on Semantic Web languages and technologies.
As STERNA is an initiative of organisations from the fields of natural history and biodiversity, the radar was extended to ongoing developments in these fields. The aim was to create a wider picture of the digital environments natural science and history organisations and practitioners use to create, manage and share information resources.

Criteria for project indentification

For the STERNA technology radar, the identification and selection criteria for relevant projects were the following:

- projects that develop and/or use applications for semantic enhancement and integration of digital resources which are based on Semantic Web standards, and in particular, ...
- projects that use Simple Knowledge Organisation System (SKOS), which is a Semantic Web standard for thesauri, classification schemes and other knowledge organisation systems,
- work with cultural heritage content, which is the major focus of the Europeana project as well as the future European Digital Library (EDL),
- and/or interesting digital content and data environments in the fields of natural history and biodiversity, which is the focus of STERNA.

General focus of projects

In general, such projects aim to port to the Semantic Web legacy metadata as well as term lists, thesauri, classification schemes, etc. They implement advanced search and other capability that draw on the semantic layer of the created RDF metadata and "SKOSified" thesauri and other knowledge organisation schemes. Moreover, some projects use higher-level Semantic Web languages such as OWL (Web Ontology Language) to allow for some reasoning over the semantic layer.

Character of identified projects

Semantic Web languages and technologies have only in recent years found a wider adoption, however, quite a number of projects could be identified.
On the spectrum from pure and applied research projects to fully operational implementations under real world conditions, these projects are situated in the middle ground. Most often they are research projects that develop, implement and test novel applications using cultural heritage content to demonstrate their case. However, there are also projects of leading organisations, e.g. in the field of biodiversity, that promote using Semantic Web languages and technologies.

Description of identified projects

For the projects that have been identified, the report describes the project context, approach to semantic enhancement and integration of content, and available results – e.g. tools, services, experiences – that could be taken into account by STERNA and similar initiatives. Furthermore, projects, organisations and experts identified in the course of the Technology Watch activity may be of interest for extending the STERNA network or/and cluster activities, e.g. expert workshops in the context of the European Digital Library initiative.

Intended users of the report

While named Technology Watch, this report is not specifically intended for technological researchers or technical personnel of scientific and cultural heritage organisations. Rather, the audience for this report has been envisioned to include directors, project managers, curators of collections, librarians and other personnel who may or may not have some technical background.
This required to include chapters that explain some concepts in more detail, for example, how Semantic Web languages build on each other, details of SKOS, or how Life Science Identifiers (LSIDs) are implemented.
The descriptions of identified projects should provide enough detail to allow non-technical readers to understand the context and aims of interesting projects, and technical personnel to understand the technical approach taken and to consult the project website and other references for more specific technical information.

1.3 Two perspectives

The Technology Watch report comprises two parts with different perspectives in terms of technologies, content domain, and character of presentation, and recommendations.
The first part, **Knowledge organisation systems for leveraging access to cultural and scientific heritage:**

- presents mainly semantic technologies,
- describes (mainly) projects in the domain of cultural heritage, and
- provides several introductory chapters (e.g. Semantic Web, Knowledge Organisation Systems, SKOS standard).

The second part, **Natural history and biodiversity resources for the European Digital Library initiative:**

- presents a wider spectrum of technologies,
- focuses on natural history and biodiversity resources, and
- provides only one introductory chapter (on the Life Science Identifiers standard).

From the perspective of the European Digital Library it also may be argued that the first part represents a declared short to medium term goal (use of basic semantic languages and techniques), whereas the second part either has a short-term or a long-term horizon: Short-term, if the goal is to include more natural history and biodiversity content in the EDL; long-term, if the goal also is to incorporate advanced information services of these fields of knowledge.
In the sections below the two perspectives are introduced and results of the study summarised in recommendations for stakeholders in the EDL initiative.

1.3.1 Knowledge organisation systems for leveraging access to cultural and scientific heritage

The first part of the report (Part A, chapters 2–7), focuses on knowledge organisation systems for leveraging access to cultural and scientific heritage. The first five chapters of this part set the scene by describing:

Background and introductory chapters

- the European Digital Library initiative, in particular, the technological roadmap and the current approach to cross-domain content access (chapter 2);
- the basic setup of a semantic digital library, and the Semantic Web approach STERNA implements to allow for semantic enrichment and interoperability of information resources (chapter 3);
- the "layer cake" of Semantic Web languages, i.e. the different languages that build on each other to realise advanced resource discovery and access (chapter 4);
- the relevant Knowledge Organisations Systems (KOS) that may be ported to the Semantic Web, such as thesauri, classifications schemes and others (chapter 5);
- and, as last introductory element, the SKOS standard and the road it provides to semantic search and access across distributed and heterogeneous information resources (chapter 7).

State-of-the art projects

Chapter 7 then describes state-of-the-art projects that have transformed legacy metadata to RDF format and thesauri to SKOS.
Most of these projects are in the field of cultural heritage and concern art, archaeological, ethnographical and other museum collections.
Some of them also have implemented higher-level Semantic Web languages such as the Web Ontology Language or/and used the CIDOC-CRM, a core ontology that has been developed to facilitate the integration, mediation and interchange of heterogeneous cultural heritage information.

Selected tools and services

The final section of chapter 7 provides details on selected tools and services that have been developed and used in some of the projects described.

1.3.2 Recommendations on semantic approaches to leverage heritage content integration and access

Recommendation 1: Focus on metadata enhancement and provision in RDF format

In the development of the European Digital Library (EDL), most importantly issues of metadata quality and cross-domain interoperability need to be addressed. For example, there is a legacy of different metadata standards and other factors that make enabling cross-domain content search and access a particularly difficult task.

To allow for such search and access through the EDL, institutions that want to participate will often need to clean and enhance legacy metadata. Even a perfect technical, semantic and multilingual framework of the future EDL would face considerable limitations of interoperability if it operates on "dirty" heterogeneous data.

There are available powerful techniques that support a relatively easy creation of metadata in RDF format from different original encodings. However, while it is true that "a little semantics goes a long way" (Jim Hendler), first the ground for this must be prepared by the institutions. In particular this also includes the use of Uniform Resource Identifiers (URIs).

With rich metadata in RDF format, richly interlinked, users will be able to go in very different directions, i.e. explore, appreciate, and learn about European cultural and scientific heritage in many ways.

Recommendation 2: Capitalise on the rich legacy of thesauri, classification schemes and other knowledge organisation systems (KOS)

The Simple Knowledge Organisation System (SKOS) standard is intended to provide a light-weight conceptual modelling language and low-cost migration path for porting existing KOS to the Semantic Web. Hence, SKOS allows for re-using of, and capitalising on, the rich legacy of existing KOS in the Semantic Web environment.

However, experience from a number of projects shows that the task to represent legacy KOS in SKOS format without loss of important features is often difficult to accomplish.

The mapping between SKOS representations of different thesauri and other KOS can provide a semantic reference network that allows for enhanced search and other capability (e.g. faceted searching and browsing).

However such mappings generally require domain experts and may be time-intensive, and hence costly. Often detailed mapping work at the concept level is necessary for useful results, and automated assistance typically helps to accomplish only parts of the task.

There is a need of skills development in translating legacy KOS, and there is also a need for (semi-)automatic mapping techniques that are better tuned to semantically light-weight SKOS representations.

Recommendation 3: Establish a copyrights clearance mechanism for knowledge organisation systems

Thesauri, classification systems and other KOS are subject to copyright clearance. While SKOS representations of proprietary thesauri and classification systems such as Getty thesauri and Iconclass have been produced in the framework of research projects (and sometimes are available from project websites), copyrights may not be cleared sufficiently to allow re-use.

Many available KOS are intended as freely usable resources, however, it would be preferable to have a formal clarification and an appropriate licence for each KOS that is prepared for use in semantic information services (for example, of a future semantic European Digital Library).

Establishment of a central copyrights clearance mechanism should be considered, instead of a situation where many institutions and projects approach copyright holders to negotiate and receive a licence on an individual basis.

Recommendation 4: Establish a task force that collects and disseminates know-how and best practices in the creation of RDF metadata and SKOS representations of legacy KOS

There is need to enable many more cultural and scientific heritage institutions to create RDF metadata of content and SKOS representations of legacy thesauri and other KOS. For example, recently started large *e*Content*plus* projects such as Athena and EuropeanaLocal among other objectives intend to prepare participating institutions to provide such metadata and knowledge organisation schemes to the emerging semantic European Digital Library.
However, know-how and best practices are not easily available and may be unevenly distributed across Europe. Therefore a task force should be established that collects and disseminates know-how and best practices, for example, by providing guidelines and offering training workshops. Also a brokerage system for available expertise and services may be established.

Recommendation 5: Exploit domain and upper-level ontologies for cross-domain semantic integration

On top of the semantic layer provided by interlinked thesauri, classification schemes and other knowledge organisation systems, ontologies will be needed to allow for some higher-level integration, reasoning and other capability.
There are ontologies available, in particular, the CIDOC Conceptual Reference Model. However, to be able to make use of this complex ontology requires in-depth understanding of its event-centric modelling approach as well as how to extend or specialise the ontology, if required.
Again, this is a field where the need may be more in skills on the side of the subject experts (e.g. museum curators) than in ontology creation and management tools, however, also more user-friendly intelligent support is required.

1.3.3 Natural history and biodiversity resources for the European Digital Library initiative

The second part of the report (Part B, chapters 8–13), focuses on natural history and biodiversity resources for the European Digital Library initiative.

Differences to initial EDL content resources

STERNA contributes to the objectives and realisation of the European Digital Library, but it is important to recognise that with regard to the project partners' content, there are some important differences in comparison to the initial contributors to the EDL initiative.
Most STERNA project partners are natural science and history museums and the content they and other partners will make available is related to what may be called bio-heritage.
There is a high interest of people throughout the world in issues of ecology, biodiversity and species conservation, and a lot of progress has been made in the last 10 years in making related digital information resources available for research, education and other communities.
The website of Biodiversity Information Standards (TDWG) lists 600 bioinformatics projects of data providers, data aggregators, and facilitators woldwide, and there are tremendous volumes of information resources held in a multitude of databases. Increasingly these resources also are shared worldwide.

Natural history content, an enormous potential for the EDL

The EDL initiative so far has been mainly driven by cultural heritage institutions, in particular, the national libraries of Europe. But large and small museums and other organisations in the field of natural history and biodiversity own an enormous wealth of knowledge and content, some of which are also relevant to broad user groups interested in certain animals (e.g. birdwatchers), species conservation, and nature and wildlife in general.

Some illustrative examples

The following examples may illustrate the wealth of knowledge and content that is available in digital form or in the process of being digitised:

Natural history museum specimens

About 1.8 million species are currently known to science, while estimates of the total number of species on Earth range from five to fifteen million. Knowledge of species is largely based on the collections of the worldwide 6500 natural history museums that are estimated to hold between 1.3 and 3 billion specimens; for example, the Natural History Museum in London alone holds some 70,000,000 specimens. However, the building of digital collections of specimens is a huge task. It is estimated that worldwide below 5% of specimen collection records have been digitised so far.

Biodiversity literature

The taxonomic and other natural history literature has accumulated over the last 300 years, and researchers make use of publications from the past and present. As older publications are often only available in a few select libraries, there are international efforts such as the Biodiversity Heritage Library (BHL) project to make the literature accessible on the Web.

Taxonomic databases

There are taxonomic databases that record the names, synonymy, classification, geographic distribution and relationships of biological organisms. The existing global species databases presently account for some 60% of the total known species. These databases are integrated in initiatives such as the European BioCASE network and the Global Biodiversity Information Facility (GBIF).

Creation of entry points to a vast array of knowledge and content

The integrated databases provide the taxonomic backbone to initiatives such as the Encyclopedia of Life (EOL). The EOL project aims to create a webpage for each of the known species that will provide the entry point also for non-professional users to a vast array of knowledge (e.g. geographic distribution, evolutionary history, behavior, ecological relationships, etc.).

A living heritage

In short, natural history museums and related organisations such as specialised libraries and audiovisual archives provide an enormous potential for the European Digital Library. In the case of natural history, scientific heritage is often "living heritage", which means that many older information resources, in particular, the taxonomic knowledge developed by generations of researchers, are still important to our current understanding of biodiversity and species conservation.

How to integrate natural science and history resources in the EDL

The question arises how natural science and history resources can be integrated in the European Digital Library initiative most effectively, both extending its current thematic scope and interlinking parts of its cultural heritage material with relevant natural science and history content.

It is clear, that only a small fraction of the available natural history and biodiversity resources are of interest to non-scientific users and could be meaningfully interlinked with cultural heritage content. Yet, even this small fraction represents a high potential for adding value to the European Digital Library initiative.

In general this will be content such as explanatory texts, illustrative images (e.g. of museum type specimens), species distribution maps, audiovisual documentaries, 3D models of natural history objects and other such content that may appeal to wider user groups.

Though there may also be the question if a future European Digital Library in addition could promote knowledge acquisition and learning about ecology, biodiversity, biological evolution and other higher-level conceptual understanding – a question that also is evident with respect to the large volume of cultural heritage material that is intended to become accessible through the EDL (i.e. cultural concepts, diversity, change, etc.).

With regard to natural history resources, a future European Digital Library, for example, could provide the opportunity to link users into knowledge resources of natural history taxonomy, biodiversity, and themes such as species conservation.

Reinforcing the relevance of scientific heritage

The European Digital Library has been meant to provide access to cultural and scientific heritage, however, the scientific heritage part was somewhat lost in the effort to ramp up available digital content and develop a functional prototype of this library, named Europeana.
Nevertheless, projects funded under the *e*Content*plus* programme such as AquaRing and STERNA, that are expected also to make content available to the European Digital Library, indicate an interest in natural science and history resources. We expect that this interest will grow in the future, because of the high relevance of bio-heritage themes and the demonstrated added value of related resources.

Importance of recognising ongoing initiatives

Therefore, it will be important to keep ongoing initiatives for digitising, enhancing and integrating information resources in the fields of natural history and biodiversity on the technology radar.

1.3.4 Recommendations on the integration of natural history and biodiversity resources

Recommendation 1: Reinforce the importance of scientific heritage in the European Digital Library initiative

The European Digital Library (EDL) has been meant to provide access to cultural and scientific heritage, however, the scientific heritage part was somewhat lost in the effort to ramp up available digital content and develop a functional prototype of this library (Europeana).
The EDL Foundation statutes include that members are committed to provide access to Europe's cultural and scientific heritage through a cross-domain portal, to stimulate initiatives to bring together existing digital content, and to support digitisation of Europe's cultural and scientific heritage.
In the next phases of building the EDL and acquiring further resources, the importance of the scientific heritage of Europe should be reinforced.
With respect to scientific heritage resources held by museums, it may be interesting to note that the EDL Foundation Board of Participants currently (November 2008) has 16 members of whom only two are from the museum sector (European Museums Forum and ICOM Europe). While they may also represent the interests of natural history museums, the list of current content providers of the Europeana website only includes one museum from this domain, namely the Natural History Museum in London.
If the future European Digital Library is intended to extend the current focus on cultural heritage content to incorporate more scientific heritage held and curated by natural science and history museums, museums of this domain would need to be addressed specifically.

Recommendation 2: Recognise the potential of natural history and biodiversity resources for the European Digital Library initiative

Natural history content – an enormous potential for the EDL

There is a high interest of people throughout the world in issues of ecology, biodiversity and species conservation (bio-heritage), and a lot of progress has been made in the last 10 years in making related digital information resources available for research, education and other communities.
The EDL initiative so far has been mainly driven by cultural heritage institutions, in particular, the national libraries of Europe. But large and small museums and other organisations in the field of natural history hold an enormous wealth of knowledge and content. Some of this knowledge and content also is relevant to broad user groups that are interested in topics such as biodiversity, wildlife and species conservation.

A living heritage

In the case of natural history, scientific heritage is often "living heritage", which means that many older information resources and, in particular, the taxonomic knowledge developed by generations of researchers is still important to our current understanding of biodiversity and species conservation.

Recommendation 3: Clarify which knowledge and content resources from the fields of natural history and biodiversity are of particular interest

Natural science and history resources can considerably extend the EDL's current thematic scope. There is a broad spectrum and a huge volume of such resources available and increasingly shared throughout Europe and beyond.

Which resources to integrate in the EDL

Only a fraction of these resources will be of interest to non-scientific users and could be meaningfully interlinked with cultural heritage content. Therefore it is important to clarify which resources present the highest potential for adding value to the European Digital Library initiative.
In general this will be content such as explanatory texts, illustrative images (e.g. of museum type specimens), species distribution maps, audiovisual documentaries, 3D models of natural history objects and other such content that may appeal to wider user groups.

Criteria and priorities

However, it is not clear which content is of particular interest and should be given priority, considering criteria such as user interest, content type, relationships with cultural heritage material, technical integration, etc.

Recommendation 4: Clarify how to align technically with current developments in the fields of natural history and biodiversity

It is important for the EDL and participating organisations to recognise ongoing initiatives in digitising, enhancing and integrating information resources in the fields of natural history and biodiversity, in particular, novel approaches that are used by these initiatives.

Digitisation of content

In the field of natural history museums and libraries there has been much progress recently with techniques capable of extracting named entities from textual resources (e.g. specimen labels and taxonomic literature) and semi-automatic creation of metadata for such resources. This considerably reduces the cost of information extraction and metadata creation.
Techniques similar to Taxonomic Name Recognition may also be applied to certain textual cultural heritage resources. Such approaches are developed for example in the EU FP7-ICT IMPACT (Improving Access to Text) project for lexicon building from historical dictionaries and historical texts. The potential for know-how and technology transfer should be examined.

Provision of URIs

The provision of Uniform Resource Identifiers (URIs) by the participating organisations is a major challenge in the European Digital Library initiative.
In the fields of natural history and biodiversity, major organisations increasingly make use of Life Science Identifiers (LSIDs), which can be used for any information resource also from other domains. Experiences with this approach, for example, with regard to service provision and content integration should be examined.

Metadata standards

The current approach to metadata of the European Digital Library (Europeana) is to develop Metadata Application Profiles for the different domains that participate in the EDL initiative (i.e. libraries, archives, museums, audiovisual collections).
These profiles use Dublin Core as their basis, which is the preferred standard when, as Europeana does, using the Open Archive Initiative Protocol for Metadata Harvesting.
With regard to natural history museums, a specific Application Profile may be required which, for example, could draw on Darwin Core or ABCD (Access to Biodiversity Collections Data).

RDF metadata vocabularies

To integrate the European Digital Library in the emerging Semantic Web, the participating organisations would need to provide their metadata in RDF format.
In the fields of natural history and biodiversity, the recommendation of the Biodiversity Information Standards (Taxonomic Database Working Group – TDWG) to provide the Life Science Identifier metadata in RDF format will greatly support the integration of collections via the Semantic Web. To promote such integration, the TDWG provides LSID metadata vocabularies, which are also loosely connected by a core ontology.
The experiences with this setup should be examined in detail, considering how such natural history and biodiversity resources might be included in a future semantic European Digital Library.

Ontologies

The ontological layer of the Semantic Web plays a key role for knowledge representation, data integration and advanced search and other services spanning databases of distributed information providers. The realisation of such a layer with the capability to support some reasoning over RDF resources requires the implemention of domain and core ontologies.
With respect to natural history and biodiversity resources, the core ontology developed by the Taxonomic Database Working Group (Technical Architecture Subgroup) and/or simple classes from the LSID metadata vocabularies will allow for some ontological alignment.

Taxonomic backbone

The basic organisational units of biological knowledge in the fields of natural history and biodiversity are taxa (i.e. the scientific names designnating an organism or group of organisms). In the digital environment taxa are used to virtually tie together the available data about species and to provide search and other information services.
Content access websites in the fields of natural history and biodiversity typically make use of a taxonomic backbone. This applies to large-scale portals such as the Encylopedia of Life as well as small specialised websites. To provide access to a larger part of natural history and biodiversity resources, the European Digital Library may also need to make use of such a taxonomic backbone. Most likely this would be the Catalog of Life (CoL).

Recommendation 5: Consider to move on from content access to support active learning and knowledge creation

The European Digital Library initiative currently follows the library paradigm of mainly providing access to content. In the next phases, 5-10 years ahead, opportunities of promoting active learning and knowledge creation should be considered.

Importance of conceptual understanding – link users into knowledge resources

Such learning is about developing a higher-level conceptual understanding, in the case of natural history resources, for example, about ecology, biodiversity, biological evolution, etc. In the future, the EDL could at least link users from content pieces into knowledge resources, natural history taxonomy and systematics, for instance.
The issue of higher-level conceptual understanding is also evident with respect to the large volume of cultural heritage material that is intended to be become accessible through the EDL (e.g. cultural concepts, diversity, change, etc.).

Online collaboration tools

Active learning and knowledge creation may be promoted by offering Web-based spaces and tools that allow individuals and communities of users to effectively study, share, and work with, different types of content that is made available. Experiences with such environments in the field of natural history (e.g. Scratchpads) should be taken into account.

Part A:
Knowledge organisation systems for leveraging access to cultural and scientific heritage

Part A

Knowledge organisation systems for leveraging access to cultural and scientific heritage

Part A of the report (chapters 2–7), focuses on knowledge organisation systems for leveraging access to cultural and scientific heritage. The first chapters of this part set the scene by describing:

Background and introductory chapters

- the European Digital Library initiative, in particular, the technological roadmap and the current approach to cross-domain content access (chapter 2);
- the basic setup of a semantic digital library, and the Semantic Web approach STERNA implements to allow for semantic enrichment and interoperability of information resources (chapter 3);
- the "layer cake" of Semantic Web languages, i.e. the different languages that build on each other to realise advanced resource discovery and access (chapter 4);
- Knowledge Organisations Systems (KOS) that may be ported to the Semantic Web, such as thesauri, classifications schemes and others (chapter 5);
- and, as last introductory element, the SKOS standard and the road it provides to semantic search and access across distributed and heterogeneous information resources (chapter 6).

State-of-the art projects

Chapter 7 then describes state-of-the-art projects that have transformed legacy metadata to RDF format and thesauri and other KOS to SKOS format.
Most of these projects are in the field of cultural heritage and concern art, archaeological, ethnographical and other museum collections.
Some of them also have implemented higher-level Semantic Web languages such as the Web Ontology Language or/and used the CIDOC-CRM, a core ontology that has been developed to facilitate the integration, mediation and interchange of heterogeneous cultural heritage information.

Selected tools and services

The final section of chapter 7 provides details on selected tools and services that have been developed and used in some of the projects described.

2 The European Digital Library initiative

2.1 Focus point and driver of cultural and scientific heritage digitisation and unified access

EDL start in 2005

In April 2005, an initiative was started by the Heads of State and of Government of France, Germany, Italy, Hungary, Poland and Spain for building a virtual library of European dimension comprising cultural and scientific heritage content. The initiative in part was a reaction to Google's digital library project that had the announced aim of digitising and making accessible online 15 million books. Consequently, the initiative for a European Digital Library (EDL) was quickly followed by commitments of most of the national libraries of the EU Member States.

The initiative was welcomed by the European Commission that considered it as a flagship project under the i2010 European Information Society policy framework which was adopted on 1 June 2005. A press release in March 2006 informed about the steps already taken and planned for the EDL, the rollout of which described as follows:

"By the end of 2006, the European Digital Library should encompass full collaboration among the national libraries in the EU. In the years thereafter, this collaboration is to be expanded to archives and museums. Two million books, films, photographs, manuscripts, and other cultural works will be accessible through the European Digital Library by 2008. This figure will grow to at least six million by 2010, but is expected to be much higher as, by then, potentially every library, archive and museum in Europe will be able to link its digital content to the European Digital Library." (Europa.eu 2006)

Based on co-ordinated digitisation activity across Europe since 2001

In April 2005 the aim of making Europe's heritage accessible online was not a new topic, because at that time much digitisation work was already carried out by institutions of the EU Member States based on the Lund Principles and Lund Action Plan.

Issued in 2001, these documents established an agenda for actions to be carried out by Member States and the European Commission. These actions aimed at promoting a higher level of digitisation and online availability of cultural content and included mechanism for coordination and cooperation among the Member States, national inventories, centres of competence, and good practice guidelines.

In particular, the National Representatives Group (NRG) of European Ministries of Culture was established and from March 2002 onwards received operational support by the MINERVA (Ministerial Network for Valorising Activities in digitisation), MINERVAplus and MINERVAeC thematic network projects.

The EDL initiative as new focus point and driver

In a relatively short time much progress was achieved in making guidance material and reports on digitisation activities of the Member States available (see the MINERVA progress reports 2002-2007). However, around 2004/2005 the initiative was felt to have lost its momentum. One renowned expert, who has been involved in the initiative from the very start of the Lund Principles and Action Plan (2001), in September 2004 noted: "Progress towards widespread adoption and take-up of the principles (...) has, it is fair to report, been patchy", and warned that a number of key issues such as collaboration, metadata creation, and long term access to the digital assets needed sustained efforts. (cf. Ross 2004)

Also the update in November 2005 of the Lund Action Plan through the so called Dynamic Action Plan (DAP) for the EU co-ordination of digitisation of cultural and scientific content mentions that "many of the barriers identified within Lund continue to exist" and suggests a broad spectrum of actions in the areas of Users and content, Technologies for digitisation, Sustainability of content, Digital preservation, and Monitoring progress. (DAP 2005)

Hence, the European Digital Library (EDL) initiative came at the right time for the European Commission to give new impetus to the Member States efforts of making more digitised cultural and scientific heritage resources accessible online.

EC adjustment of instruments

The European Commission contributed at the European level by adjusting the required instruments, which included:

- the definition of the EDL as the flagship project of the "i2010: Digital Libraries" initiative (EC 2005);
- the Commission's Communication on "Digitisation and online accessibility of cultural material and digital preservation" of 24 August 2006 (EC 2006a; see also the Commission's impact assessment document EC 2006b);
- funding from 2005 onwards of digital content and metadata enrichment projects under the *e*Content*plus* programme;
- funding of related research and technological development projects under the 6th Framework Programme (in the relevant last call for proposals) and the 7th Framework Programme, Challenge 4: Digital Libraries and Content;
- furthermore a High Level Expert Group on Digital Libraries was established that advises the European Commission on organisational, legal and technical issues such as matters of IPR (e.g. orphan and out-of-print works) and access to results of publicly funded research.

Many projects clustered around the EDL initiative

The intended integrative effect of the "i2010: Digital Libraries" initiative is excellently presented in a brochure of the European Commission's DG Information Society and Media (EC 2006c), which describes the fields of policy actions and 25 selected projects under the *e*Content*plus*, eTEN, 5th and 6th Framework Programmes (FP5 and FP6).
Among the projects are (full project titles and URLs in section 17.2):

- the ones in support of the European national libraries' effort to create a common infrastructure for making available their digitised collections: EDL (eCp), TEL (FP5) and TEL-ME-MORE (FP6);
- more general digital library and repository infrastructure projects: BELIEF (FP6), DELOS (FP5), DRIVER (FP6) and DILIGENT (FP6);
- projects with a focus on tangible cultural heritage, such as monuments and archaeogical sites: BRICKS and TNT - The Neanderthal Tools (both FP6);
- some that deal with specific content like audio and audio-visual content and related material such as Braille music sheets: CONTRAPUNCTUS, EASAIER, MEMORIES and PRESTOSPACE (all FP6);
- projects with a focus on multi-lingual access: MICHAEL (eTEN) and MultiMATCH (FP6), and
- projects that aim at ensuring the long-term preservation of digital assets: CASPAR, DPE and PLANETS (all FP6).

There have been many more projects which can be seen to relate to the EDL initiative, in particular, most of the 25 projects that have been funded under the 2005, 2006 and 2007 calls of the *e*Content*plus* programme in the areas of digital libraries and cultural and scientific/scholarly content (see: *e*Content*plus* Programme: Projects).
Moreover, there are several relevant research and technological development projects funded under the FP7-IST programme's first and third call addressing Challenge 4: Digital Libraries and Content.

2.2 EDL technological roadmap for interoperability

Europeana

The European Digital Library (EDL) initiative aims to build a common multi-lingual access point to Europe's distributed cultural and scientific heritage, including digital content from all types of heritage institutions (archives, libraries, museums and audio-visual collections).
A prototypic showcase website of the EDL has been developed by the Europeana project and was formally launched on the 20th of November 2008, http://www.europeana.eu. The Europeana version 1.0, which would be developed in a new project, is expected to see its first release early 2010.

Europeana (originally named EDLnet) is a project funded under the *e*Content*plus* programme for a period of two years (07/2007-06/2009). The project is run by a core team the National Library of the Netherlands, the Koninklijke Bibliotheek. It builds on the project management and technical expertise developed by The European Library (TEL), which is the common portal of the Conference of European National Librarians. Overseeing the Europeana project is the EDL Foundation, which includes key European cultural heritage associations.

Technological roadmap for EDL interoperability

The Europeana project has been entrusted to find consensual technical solutions to interoperability issues of the emerging European Digital Library (EDL). Such solutions need to be found as the EDL should be able to handle data from the different cultural and scientific heritage domains such as archives, libraries, museums and audio-visual collections. It is fully considered that common solutions can not be imposed from above and progress can only be made by consent.

From January to June 2007, before the official start of Europeana, a working group on digital library interoperability comprising technological researchers, cultural heritage experts and representatives of the European Commission identified areas for short term actions (2008) in the context of the European Digital Library initiative as well as key elements for a long-term strategy (2010 and beyond).

The following summary of suggested actions toward EDL interoperability is based on presentations from September and December 2007 by one of the lead technological researchers in the Europeana project (Gradmann 2007a and 2007b):

User requirements

(1) Existing use cases should be used as input for a systematic and generalised process of identifying EDL user requirements. Examples given are use cases in operation with The European Library (TEL) and the Bibliothèque nationale de France (which indicate a focus on user requirements as perceived from a library online services point of view).

Object models

(2) Object models – granularity and structure: With respect to models of digital information objects in the short-term only complete objects are considered, e.g. "books" (librarian), "records" (archival) and "artefacts" (museum). In the longer term the level of granularity should be refined to allow for dealing with intra-object reference structures. For complex, multimedia objects description and packaging standards such as METS, MPEG 21 (DIDL) or XFDU may be used.

Persistent identifiers

(3) Persistent identifiers are seen as a key element of interoperability. A technical solution was envisioned to make it "technically impossible to create new resources in EDL without applying standard identifiers".

As the EDL mediates access to content held by the participating institutions such identifiers will need to be implemented by the content providers. This is a critical issue, because, today many potential content providers do not have persistent identifiers in place. For example, in an explorative survey among the 26 regional content co-ordinators of the EuropeanaLocal project in June 2008 it was found that less than a quarter use persistent identifiers. (Davies 2008)

Whatever identifier framework will be suggested by the EDL, it must be applied systematically and the resolving mechanisms need to be transparent. Application of the CENL (Conference of European National Librarians) European Resolution Infrastructure was suggested for resolving purposes and for identifier referral.

Metadata standards

(4) Domain-specific Dublin Core Application Profiles should be developed that take into account the needs of the different heritage domains and support object-level search and retrieval across digital collections. Each application profile should include provision for rights metadata as well as some technical metadata.

For the provision of collection level descriptive metadata, a harmonisation of existing description formats (e.g. MICHAEL, TEL, Archival Grid, etc.) was suggested. Furthermore, development of a metadata registry for the EDL was considered important.

A higher level interoperability application profile was understood to be not appropriate for the purposes of the EDL. Instead semantic interoperability techniques should be used to implement semantic mappings between metadata schemas and support cross-searching of descriptive metadata (see also below point 8).

Service registry

(5) Implementation of a service description framework was considered as an important element of the EDL, allowing for systematic service integration. For the development of such a framework, the JISC IESR (Information Environment Service Registry) was considered as a possible starting point.

Licensing

(6) Licensing policies: For all freely available content and metadata a suitable licence should be used that clearly specifies the respective rights and use conditions.

Authentication

(7) Authentication data exchange: SAML (Security Assertion Markup Language) and Shibboleth-enabled methods are suggested as the standard solution for trust based exchange of authentication data within the EDL network and towards the outside. A "What Federation Are You From" (WFAYF) service should thus be implemented as part of EDL.

Semantic interoperability

(8) Basic semantic interoperability: A data layer ready for semantic query methods should be created through making existing metadata and the controlled terminology used therein machine understandable. The suggested method of choice for the conversion of controlled vocabularies is SKOS, but also use of OWL was thought to be appropriate in some near-term application scenarios. In the longer term, advanced semantic interopera-bility, based on a layer of ontologies, rules and reasoning mechanisms, and mapping to object modelling standards should be aimed at.

Semantic functions as USP

(9) Awareness building regarding semantic interoperability: Short term viability and the value added of providing basic semantic interoperability for searching and browsing should be demonstrated. Semantic interoperability functions are considered as a unique selling point of the emerging EDL.

Interoperation with generic WWW services

(10) Interoperation of EDL and WWW services: The EDL architecture should allow for maximum exposure of services and content via general-purpose WWW services (e.g. Google, Yahoo, and others), making sure that EDL provenance is clearly identifiable.
Details about the suggested practical implementation of some of the points above are to be found in two Europeana project deliverables: "Initial Semantic and Technical Interoperability Requirements" (EDLnet, December 2007) and "Europeana Outline Functional Specification. For development of an operational European Digital Library" (EDLnet, August 2008).

2.3 Current approach to cross-domain content access

Focus on enhancement of legacy metadata

The European Digital Library (EDL) will become a multi-lingual common access point to the digitised content that is held in the distributed repositories of libraries, archives, museums and audiovisual collections across Europe.
In this context, the importance of content/metadata enrichment is emphasised. For example, the recent calls of the *e*Content*plus* programmes, which now works to a large part in support of the EDL, specifically invited proposers to suggest projects that focus on such enrichment.

Leveraging interoperability

Indeed, in the development of the EDL in the first place issues of metadata quality and cross-domain interoperability need to be addressed. For example, there is a legacy of different metadata standards and other factors that make cross-domain content search a particularly difficult task.
To allow for such search, exploration and access through the EDL, institutions that want to participate will often need to enhance legacy metadata. And from the perspective

of the future EDL, even a perfect technical, semantic and multilingual framework would face considerable limitations of interoperability if it operates on "dirty" heterogeneous data. (cf. Gradmann 2008, who calls this "the nasty bit" of several challenges of the future EDL)
General best practices for the generation and sharing of metadata include to use an established metadata standard (or create an application profile based on existing metadata schemes) and to employ controlled vocabularies and authority files for data values. However, there are many challenges with regard to actually providing metadata that can be effectively shared within large-scale projects involving many partners with different content and metadata schemes. (cf. Shreeves et al. 2006)

Current Europeana metadata specification

The current contributors to the Europeana showcase website use the "Europeana Semantic Elements" (v2.0) specification, which is based on Dublin Core, but has two additional refinements for the DC Relation element ("isShownBy" and "isShownAt") and an additional element "UserTag" for public tags created by registered users. (Europeana 2008)

Dublin Core

The reason for building on Dublin Core (http://dublincore.org) is that this metadata standard has been specifically developed to support cross-domain provision and searching of metadata, and that it is already widely used for this purpose. With 15 elements the Dublin Core Metadata Element Set is a rather lightweight, but extendable, standard for describing and sharing information resources.

Dublin Core metadata in RDF

The technological roadmap of the EDL considers semantic interoperability as a future unique selling point of the library. Therefore it may be important to note that the Dublin Core Metadata Initiative (DCMI) already in 2002 had provided guideline recommendations for encoding simple and qualified Dublin Core metadata in the Semantic Web standard Resource Description Framework (RDF). In January 2008 these have been replaced by the recommendation "Expressing Dublin Core metadata in the Resource Description Framework (RDF)". (DCMI 2008)

OAI-PMH

Dublin Core also is the basic metadata standard to be used with the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). Dublin Core (unqualified) was chosen by the Open Archives Initiative as mandatory minimal requirement, but it is also possible to use much richer metadata schemes.
The OAI-PMH specifies a method for digital repositories ("data providers") to expose metadata about their objects for harvesting by aggregators ("service providers"), which then provide search and other services based on the aggregated collections of metadata.
The OAI-PMH method is widely used and what constitutes best practice is well documented (see: http://www.openarchives.org). The method also particularly has been a success with many cultural and scientific heritage organisations and networks (cf. Foulonneau 2003 and 2004), and now with the EDL initiative.

Domain-specific Dublin Core Application Profiles

With regard to the emerging EDL, the decision has been taken to use domain-specific Dublin Core Application Profiles which, however, had not been issued at the time of completion of this report.
In general, a metadata application profile is a combination of data elements from different metadata schemas, often customised for use by a network of data providers in a particular domain. (Heery and Patel 2000; Dekkers 2001) An application profile also can be understood as a considerable extension of a widely used metadata schema such as Dublin Core, adding domain or context-specific elements, wherever possible from other established schemas.
There are already a number of large projects that make use of a Dublin Core based application profile, for example, the MICHAEL (Multilingual Inventory of Cultural Heritage in Europe) portal or the CulturaItalia portal (Masci, Buonazia and Merlitti 2007)

In the field of natural history and biodiversity, extensions of the Darwin Core standard (see chapter 14) are often used to create a customised application profile. For example the Avian Knowledge Network (AKN, http://www.avianknowledge.net) uses a Darwin Core extension called Bird Monitoring Data Exchange, and their nodes have contributed so far over 50 million observation records. An other example of a network of data providers that uses an extension of Darwin Core is the Ocean Biogeographic Information System (IOBIS, http://iobis.org), which provides access to 16 million records of 102,000 species from 441 databases.

Dublin Core and CIDOC-CRM

Finally, with regard to the EDL's goal of realising cross-domain interoperability at a higher semantic level, some use of the CIDOC Conceptual Reference Model should be considered. The CIDOC-CRM is a core ontology that formally describes concepts and relations used in the documentation of cultural heritage. In September 2006 it became an official ISO standard (see section 7.7.1).
A mapping of the Dublin Core element set to the CIDOC-CRM is available, as is a "crosswalk" between the CIDOC-CRM and the Dublin Core Collection Application Profile. (Lourdi, Irene et al. 2007; Lourdi and Papatheodorou 2008; see also the official CIDOC-CRM website, http://cidoc.ics.forth.gr/crm_mappings.html)

3 Semantic content / metadata enrichment and interoperability

3.1 Towards semantic digital libraries

The technological roadmap for making content accessible through the future European Digital Library includes that semantic interoperability techniques should be used to implement semantic mappings between, and searching across, the metadata of the different cultural and scientific heritage domains. Indeed, semantic interoperability is understood to be one of the unique selling points of the emerging European Digital Library.

Semantic content/ metadata enrichment

In this report we will mainly focus on semantic content/metadata enrichment in the context of distributed, interoperable cultural heritage and natural science and history collections.
Semantic content/metadata enrichment is understood to make the intended meaning of, and the relationships between, information resources explicit and machine-processable, to allow machines and humans to better identify, access and (re-)use the resources.
The main focus of content/metadata enrichment for the Semantic Web is to create a network of machine-processable information resources whose syntax and semantics are understood by machines in order to provide services such as search & retrieval, information integration and recommendation.

Semantic digital libraries

Semantic Web standards and tools allow for implementing semantic approaches and functionality of digital libraries. Semantic digital libraries extend first generation digital libraries by describing the resources they hold (or only provide access to), and relationships between them, in a formal, machine understandable way. For this formalisation the Semantic Web standard Resource Description Framework (RDF) is used.
The resources will also comprise taxonomies, classifications schemes, thesauri and other Knowledge Organisations Systems (KOS), which are used to organise information and provide terms, keywords, etc. for element fields of metadata schemes. KOS will be formalised with the Semantic Web standard Simple Knowledge Organisation System (SKOS) or, even, the more expressive Web Ontology Language (OWL).
Moreover, there may be ontologies that provide the conceptual framework of domains of knowledge for which the semantic digital library provides information resources. Such ontologies will typically be formalised with OWL.
Based on this setup a semantic digital library is capable of providing a semantic layer across various heterogeneous sources, connecting different digital repositories, and supporting novel search paradigms such as faceted or concepts-based searching and browsing.

3.2 The STERNA approach to semantic content/metadata enrichment and interoperability

STERNA is pioneering the integration of semantically enriched digital resources from the domains of natural history, biodiversity and related fields with a view to make the resources accessible via the European Digital Library (EDL).

A federated approach based on RDF/SKOS

While traditional approaches to provide one-stop-access to distributed digital collections have focused on applying encompassing metadata schemes, STERNA takes a different approach.
STERNA uses the basic Semantic Web language Resource Description Framework (RDF) and the Simple Knowledge Organisation System (SKOS) to create a semantic layer that allows for searching and accessing content held in the heterogeneous databases of

the local autonomy of institutions and leaves their organisational and data processing environments intact.
However, it requires to convert legacy metadata to RDF format and thesauri, classification schemes and other Knowledge Organisation Systems (KOS) to SKOS/RDF format, and to implement search and other facilities that draw on the semantic layer of the combined RDF data.

Use of SKOS to represent controlled vocabularies

A key component of the STERNA approach is to make use of SKOS. SKOS provides a standard, low-cost migration path for porting existing thesauri and other controlled vocabularies to the Semantic Web. Such vocabularies are used to create metadata for information objects (e.g. documents collected in a database, Web pages, etc.). More specifically, they provide appropriate terms, keywords, etc. for certain metadata element fields, such as the "subject" element of Dublin Core, for instance. In turn, the vocabularies can also be used to form queries for search and retrieval of information resources.
Generally, controlled vocabularies such as a thesauri, classification schemes and other KOS can be understood as a network of linked concepts, and publishing these conceptual links in SKOS format makes them part of the Semantic Web. Their role then is to provide a semantic layer for faceted search, where the facets are concepts of the thesauri, classification systems, etc. used by the institutions for describing and organising their content.

Creation of RDF metadata

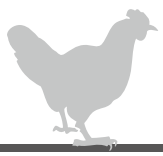
The purpose of implementing semantic search and other functionality is to discover and access related content that is held in distributed heterogeneous databases of different cultural and scientific heritage organisations. However, to allow for such discovery and access, the organisations must provide the metadata of their collections in RDF format.
Today only few organisations already have their metadata also available in RDF format. Hence, a number of activities must be carried out to evaluate and enrich the legacy metadata, also taking into account available thesauri, classification schemes, etc.
This includes to evaluate existing metadata with regard to their data models (e.g. entities, metadata fields, etc.), and their relations with thesauri, classification schemes, etc. that are in use at the organisations or/and in their domain of knowledge.

Focus on interesting common use cases of diverse collections

It must be emphasised, that the evaluation needs to be made for each of the partners content databases that are considered to be included in the project work, and driven by the goal to realise relevant use cases of related content.
Hence, in order to support such use cases, this may require to enrich legacy metadata and reference schemes, e.g by adding element fields and terms or other data not so far covered, before they are transformed to RDF and SKOS format. For the transformation, mechanisms such as database connectors, conversion rules (converters) as well as manual editing procedures will be used.
In short, STERNA is a "workshop" that examines and showcases approaches to realise interesting common use cases of distributed diverse collections that are enabled by semantic linking, searching and accessing content. Together with other such workshops, STERNA aims to provide the European Digital Library initiatives with feasible approaches of, and lessons learned in, building semantic interoperability among distributed and heterogeneous cultural and scientific heritage collections.

Technical architecture

In addition to the brief explanation above, the STERNA technical architecture is detailed in section 7.2, followed by descriptions of other completed and ongoing projects that have developed similar or complementary approaches to semantic interoperability, using RDF, SKOS or/and the Web Ontology Language (OWL).



Overview of next chapters

Readers with little technical background may benefit from firstly consulting the following chapters which include:

- an introductory overview of the so called "layer cake" of Semantic Web languages (chapter 4),
- a brief presentation of Knowledge Organisation Systems (KOS) that may be converted to SKOS format (chapter 5), and
- a detailed presentation and discussion of the SKOS standard (chapter 6).

The latter chapter covers the creation and publication of SKOS representations of existing KOS, mapping of SKOS representations, and opportunities to combine SKOS with OWL-based ontologies.

4 The "layer cake" of Semantic Web languages

The Semantic Web vision

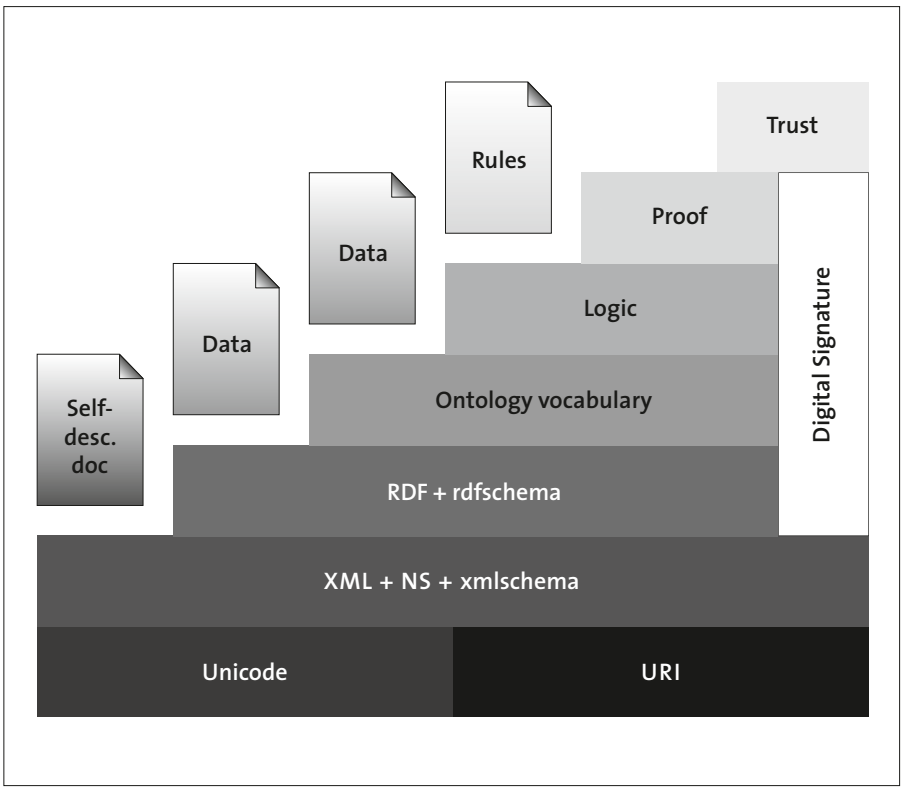
The Semantic Web is a vision of the Internet as a "distributed machine" that allows computer programmes to understand semantic relations between Web resources in order to seek and process relevant information and perform transactions for humans. Contrasted with the established, human-readable Web (e.g. Web pages), the Semantic Web is envisaged as a web of data that is expressed with certain languages in a machine processable form. Key to the understanding of the Semantic Web, therefore, is how these languages work, how information is expressed in order that computers can automatically process Web resources and assist in making the Web more useful for humans. (cf. Berners-Lee 1998a and 1998b; Berners-Lee, Hendler and Lassila 2001).
The aim of this chapter is to provide an overview of the Semantic Web concept by describing its so called "layer cake" of languages and other important elements. The explanations are not intended to give in-depth definitions of these elements. Such definitions are provided in the relevant W3C specifications that are all available from http://www.w3c.org.

Introductory material

There also is a wealth of introductory materials on Semantic Web languages available. Particularly useful are the primers of the W3Schools, http://www.w3schools.com, and for more advanced purposes the Semantic Web primer by Antoniou and Van Harmelen (2004). Guntram Geser (2003) provides a primer for the Semantic Web of cultural heritage content, which is based on the example of the Finnish Museum on the Semantic Web project.

The Semantic Web "layer cake"

The architecture of the Semantic Web is usually represented as a "layer cake" or hierarchy of languages, each language both exploiting the features and extending the capabilities of the layers below.



Source: Tim Berners-Lee 2000,
http://www.w3.org/2000/Talks/1206-xml2k-tbl/slide10-0.html

More recent figures present the "layer cake", in particular, the middle layers, somewhat differently (cf. Berners-Lee 2003 and 2005), however, the figure above is the most useful for our purpose of giving a basic overview of the Semantic Web languages.

URIs An URI (Uniform Resource Identifier) is a compact string of characters for identifying a resource on the Internet. URIs can be given to anything (physical or abstract), and anything that has a URI can be said to be "on the Web". An URI can be further classified as a locator, a name, or a combination of both. The familiar URL (Uniform Resource Locator) tells a computer where to find a resource, whereas an URN (Uniform Resource Name) is the name of a resource that is required to remain globally unique and persistent. An example of a standardised URN scheme are Life Sciences Identifiers (LSIDs) which we will address later in this report (chapter 12).

Unicode Unicode is a standard allowing computers to consistently represent and manipulate text expressed in most of the world's writing systems. The standard Unicode character encoding for the Web is UTF-8.

XML XML (eXtensible Markup Language) is a markup language for describing structured data (documents) and transporting it on the Internet between a sender and a receiver.
XML shares the syntax and bracketed tags of the well-known HyperText Markup Language (HTML), but XML serves a different goal. While HTML is used to define the layout of pages on the WWW, XML is used to define the content of documents.
XML has been created to allow anyone to design the structure of their own documents. Elements of an XML document are defined with start and end tags such as <book> and </book>, that can contain other (child) elements (e.g <author> and </author> or text content (e.g. an author's name). Furthermore, elements may have attributes that provide additional information about elements (e.g. <book category="fiction">).

XML Namespaces Web applications need to be able to recognise the XML elements and attributes which they are designed to process. Namespaces provide a method for qualifying element and attribute names used in XML documents. XML has been designed to allow for combining markup vocabulary while avoiding clashes if different vocabularies contain the same element or attribute names which, however, are intended for different applications. To keep markup vocabularies distinct, element and attribute names used in XML documents are associated with namespaces that are uniquely identified by URI references.

XML Schema An XML Schema is a means of specifying how an XML document should be structured – which elements are permitted where, which elements are optional or required, and what the elements and their attributes can contain. This specification of the building blocks of an XML document includes, but is not limited to, which elements are child elements, as well as their order and number, and the data types for elements and attributes. One of the greatest strengths of XML Schema is that it allows for data typing. The most common data types are xs:string, xs:decimal, xs:integer, xs:boolean, xs:date and xs:time.
But, as XML has no formal semantics, it is impossible for a computer application to understand how information represented in one XML document relates to information represented in an other, which means that the application cannot meaningfully merge the information content of two XML documents. To allow for such merging is one of the important roles of RDF.

RDF In order to make resources semantically interoperable on the Web, they must provide machine-understandable statements about themselves. In the Semantic Web architecture, the Resource Description Framework (RDF) provides a data model for such statements.
Its base element is the "triple", which takes the form of subjectNode–propertyArc–objectNode. Such a triple is a directed graph between resources, where the subject and

property are Uniform Resource Identifiers (URIs), and the object is either an URI or a literal (such as a string value).
Triples become connected whenever the object of one is the subject of another, but literal values cannot be the subject of new triples and so are always at the edge of the RDF graph.

RDF Schema (RDFS)

Above we have described the data model provided by RDF for expressing statements about Web resources. The semantics of such statements clearly depends on the named properties of the RDF triples.
RDF Schema now provides a mechanism that can be used to declare properties, to define the classes of resources they may be used with, to restrict possible combinations, and to detect violations of those restrictions. Basically, RDF schema defines properties in terms of the classes of resources to which they apply, and resources are defined as instances of one or more classes. In addition, classes can be organised in a hierarchical fashion.
RDFS complements and extends RDF by providing a declarative, machine-processable language that can be used to formally describe (domain-specific) metadata schemes or simple ontologies, supporting a potential merging on a more general level.
RDF and RDFS can be used for this as they provide a neutral, general-purpose knowledge representation method, i.e. they do not make assumptions about content or incorporate semantics from any particular domain.

Ontology vocabulary

The ontology vocabulary layer is reserved for more recent Semantic Web languages that have been developed to overcome the limitations RDF Schema shows when it comes to expressing and reasoning over complex ontological relationships.
The most important is the Web Ontology Language (OWL), which has different "dialects", OWL Lite, OWL DL and OWL Full. OWL Full goes beyond RDFS by providing more advanced constructs to describe the semantics of RDF statements. OWL DL is based on description logic and so brings more reasoning power. OWL Lite was intended for users primarily needing a classification hierarchy and simple constraints (e.g. thesauri and other KOS), however, it did not find a wider use in practice.

Logic and proof

On top of the ontology layer, a logic framework (e.g First Order Predicate Logic) should provide axioms and rules to support the checking of the consistency, soundness (i.e. possible inferences) and logical validity of complex, interrelated statements. The logic framework allows a Semantic Web application, often called "agent", to use an inference engine to derive conclusions and, based on the results, provide an answer to a search task or suggest the further course of action.
The proof layer should allow to check an agent's reasoning mechanism and justify as valid the answer given by the automated agent. The integrity of the proof should be tracable down through the ontological layer to individual RDF statements, which the agent uses for reasoning and task completion.
Among the current logic frameworks and languages for rules and axioms are the Semantic Web Rule Language (SWRL), the Knowledge Interchange Format (KIF), and OWL DLP (Description Logic Programs).

Trust

In the trust layer, mechanism need to be in place to ensure that the results delivered by a Semantic Web application based on inferences can be trusted. If the application can draw on logic and proof mechanisms, trustworthiness of a Web agent's answers and suggestions may be ensured by them.
However, those layers of the Semantic Web are currently not fully established, and trust of course is not only important on top of the "layer cake". Therefore other trust mechanisms are suggested that do not build on formal proofs, but draw on the data providers (e.g. certificates of trusted data respositories) or user groups (e.g. systems for rating sources).

Digital signature / encryption

The figure of the Semantic Web "layer cake" includes digital signatures as a vertical component that runs from RDF statements up to the proof layer. In more recent figures,

also encryption is included as such a vertical component. These components make clear that also the Semantic Web needs mechanisms that ensure security and authentication. These mechanisms support the "web of trust" among machines and between humans and the distributed machinery of the Semantic Web.

You don't need the full "layer cake"

Finally, it is important to note that projects that want to make use of the Semantic Web do not need to establish the full "layer cake" before any useful applications can be realised. In fact, one can build useful Semantic Web applications by using URIs, XML/S and RDF/S.

What will be rather limited, though, is the reasoning capability of such applications. Because there will be a lack of semantic depth and logical support to enable a reasoner to infer new relationships or new information from the underlying web of data.

Indeed, the expressivity of RDF and RDF Schema has considerable limitations: RDF is (roughly) limited to binary ground predicates, and RDF Schema is (again roughly) limited to a subclass hierarchy and a property hierarchy, with domain and range definitions of these properties.

Why to go beyond existing database schemas

Furthermore, there may be the question why anyhow to implement RDF, RDFS and OWL ontologies on top of existing robust database schemas.

If we consider current generation natural history and biodiversity databases, most work in this area concentrates on using relational databases to store data, and XML schema for exchanging data (e.g., Darwin Core or ABCD).

As Rod Page notes: "Both these technologies have a role to play. Relational databases support data integrity and a sophisticated query language (SQL), however they have limitations – database schema can rapidly become large, complex, and domain specific. Furthermore, the emphasis in designing such schema is on internal data integrity, rather than relationships with external data sources. This is a major limitation in an environment where most data is stored elsewhere. XML schema are good at describing messages, but poor at communicating meaning. Like relational database schema, XML schema can rapidly become large and unwieldy." (Page 2006, 14)

The Resource Description Framework (RDF) offers a different, though, complementary approach, in that RDF triple stores may be created that contain the semantic relationships among information resources encoded in RDF. However, it must be noted that currently RDF triple stores may not scale well enough to replace many existing relational databases and the powerful query language SQL.

With regard to ontologies that are used on top of such data stores, the key point is that ontologies are designed to evolve over time and to facilitate integration of data, while database schemas are not. Database schemas are typically considered an internal design decision for a given application and rarely, if ever, are reused when implementing other databases and applications. In comparison, an ontology is an external resource that may rather easily be reused, extended and integrated with other ontologies.

5 Knowledge Organisation Systems (KOS)

The Technology Watch activity of STERNA focused on relevant projects that develop and/or use applications which make use of the Semantic Web standard Simple Knowledge Organisation System (SKOS). SKOS was developed by a working group of the World Wide Web Consortium (W3C) to allow for "webifying" a range of Knowledge Organisation Systems (KOS).
In this chapter we give a brief overview of KOS, focussing on the ones for which the SKOS standard is intended. Examples of KOS are included and it is noted if there already are SKOS versions available for them.

5.1 Overview of relevant KOS

A Knowledge Organisation Systems (KOS) is a means to organise scientific or professional resources. A major use of such systems is to describe the content of resources which is expressed as appropriate keywords, key phrases or classification codes. For example, with Dublin Core metadata, such keywords or codes are used to fill the element field "subject".
KOS eliminate ambiguity, allow for controlling synonyms in use, and also often make clear some relationships that may exist between resources. KOS vary in function, structure and complexity, but, in general, they are used to support resource discovery and access.
In the overview below, we describe different KOS with regard to the type of systems they represent.

Different types of KOS

In the physical and digital environment of libraries, archives and museums, many different KOS are used. KOS provide a more or less formalised controlled vocabulary of concepts and terms, and relationships between them, that is used to describe, classify and organise objects.
The different types of KOS can be seen to represent a continuum of systems between low levels of term control and lacking relationships between terms (and terms and concepts) at one end and systems with higher level conceptualisation, formal definition of terms and relationships and, even, inference rules to support reasoning applications at the other end.
An often quoted overview by Gail Hodge (Hodge 2000, 4-7) distinguishes KOS according to growing degree of language control and growing strength of semantic structure as follows:

- Term lists: Authority Files, Glossaries, Gazetteers, Dictionaries. Such KOS emphasise terms often with definitions.
- Classifications and categories: Subject Headings Systems, Classification Schemes (also called Taxonomies), Categorization Schemes. Such KOS emphasise the creation of subject sets.
- Relationship schemes: Thesauri, Semantic Networks and Ontologies. Such KOS emphasise the connections between concepts.

This grouping would need to be discussed further, however, it has proved to be a useful starting point for a more systematic taxonomy of KOS taking into account their different purposes and characteristics. (cf. DELOS 2005; Tudhope 2006).
It is in fact very important to consider the particular purposes of different KOS as these determine what level of formalisation is needed. Generally a higher level of formalisation implies higher development cost, which need to be invested to allow a KOS to provide a more rigid term control and formalised relationships between concepts and terms.

SKOS scope of KOS

The Simple Knowledge Organisation System (SKOS) standard has been specifically developed to represent thesauri, but it may also used for subject headings, classification and categorisation schemes. Hence, its scope does not comprise term lists such as auhority files, glossaries, dictionaries, gazetteers, and also not formalised conceptual reference models or ontologies.

Relevant KOS will typically provide a controlled vocabulary, may provide synonym links, and may organise their conceptual units into hierarchies and/or networks of association. Ontologies are sometimes viewed as a type of KOS, however, they are fundamentally different because of their formal semantics.
Below we briefly describe the KOS that may be represented in SKOS:

Subject heading systems

A subject heading system provides a set of controlled terms to represent the subjects of items in a library or other collection. Such a system can be extensive and cover a broad range of subjects, but it has a rather limited hierarchical structure. However, subject headings can be combined to describe to some details the subjects of collection items.
One example is the Library of Congress Subject Heading (LCSH) system, the world's largest and most widely used general subject terminology list. The LCSH already has been converted to SKOS. (Summers et al. 2008)

Taxonomies, classification and categorisation systems

In the library and information science communities the terms taxonomy, classification or categorisation system are often used interchangeably, although there may be subtle differences from example to example (a detailed examination of the systematic properties of, and differences between, classification and categorisation is to be found in Jacob 2004).
Generally, these KOS allow for separating entities according to topical or other levels. The hierarchy of these levels also often is represented by a numeric or alphabetic notation system, but may lack an explicit definition of the hierarchy such as is provided by a thesaurus (i.e. "broader term" and "narrower term").
A well-known example of a classification system in the field of nature protection and biodiversity is the EUNIS Habitat Classification of the European Environment Agency (http://eunis.eea.europa.eu/habitats.jsp; Davis, Moss and Hill 2004), which currently is not available in SKOS format.
In the field of cultural heritage Iconclass is an example of a hierarchical, subject specific classification system (http://www.iconclass.nl). Iconclass supports the documentation of images, in particular art historical images, by providing a systematic collection of 28,000 ready-made definitions of objects, persons, events, situations and abstract ideas that can be the subject of an image. The definitions consist of an alphanumeric classification code and its textual correlate.
Iconclass today is maintained by the Rijksbureau voor Kunsthistorische Documentatie (RKD) in the Netherlands. It is not publicly available in SKOS, though, an experimental web service has been developed that serves a full SKOS record. (Drenth 2008; http://iconclass.org) A modelling of Iconclass in SKOS was also done in the FinnONTO project: http://www.seco.tkk.fi/ontologies/iconclass/.

Thesauri

Thesauri are controlled vocabularies that are based on concepts and show relationships among terms. Relationships commonly expressed in a thesaurus include hierarchy, equivalence (synonymy), and association or relatedness. There are ISO (ISO 5964-1985, ISO 2788-1986) and NISO (1998) standards for the development of thesauri, however, their definition of a thesaurus is fairly narrow and often at variance with schemes that are traditionally called thesauri. Most thesauri were developed for a specific scientific or professional domain of knowledge and many of them are rather large, comprising more than 50,000 terms.
Examples of thesauri that are available in SKOS format include the General Multilingual Environmental Thesaurus (GEMET), the CSA/NBII Biocomplexity Thesaurus and the CAIN Invasive Species Management Thesaurus (see chapter 15).
An often quoted example of a major thesaurus in the field of cultural heritage is the Art & Architecture Thesaurus (AAT), one of the Getty Research Institute's vocabulary databases, that provides a structured vocabulary of 34,000 concepts and 131,000 terms (http://www.getty.edu/research/tools/vocabulary/aat/).
The Getty Research Institute currently does not offer the AAT, or its Thesaurus of Geographic Names, for licensing in SKOS format. The Dutch version of the AAT (http://www.aat-ned.nl) was converted to SKOS format in the E-Culture project (Omelayenko 2008; for background information on the electronic version of the Dutch AAT see Drenth 2008).

5.2 Formal ontologies

Ontologies (or Conceptual Reference Models) are not covered by the W3C SKOS specification, however, in this study they are also of interest because their Web based representations may be combined with SKOS applications (e.g. thesauri services) or metadata standards in RDF/S format.

Ontologies

The most frequently quoted definition of an ontology is from Tom Gruber who describes an ontology as "an explicit specification of a conceptualization", and conceptualization here means "an abstract, simplified view of the world that we wish to represent for some purpose". (Gruber 1995) For this representation, a language is needed that allows for declaring what types of relevant things exist, and what types of relationships they have with each other.

Degree of formality

Regarding the degree of formality, the language used and, hence, the ontology created, may range from informal to rigorously formal exemplars (cf. Unschold and Jasper 1999):

- a (highly/semi) informal ontology is expressed loosely in natural language or in a restricted and structured form of natural language,
- a semi-formal ontology is expressed in an artificial, formally defined language; and
- a (rigorously) formal ontology has meticulously defined terms with formal semantics, theorems and proofs of soundness and completeness (for example, axiomatised logic theories that include rules to ensure the well-formedness and logical validity of statements).

Degree of specialisation

Furthermore, an important aspect for distinguishing ontologies is their degree of specialisation (cf. Guarino 1998):

- top-level ontologies: describe the basic concepts and relationships invoked when information about any domain is expressed; the concepts on this level are very general like space, time, matter, object, event, action, etc., which are independent of a particular domain or problem (i.e. they are generally applicable across a wide range of domains and tasks);
- domain ontologies and task ontologies: describe, respectively, the vocabulary related to a generic domain (e.g. biology) or a generic task or activity (e.g. analysing), by specialising the terms introduced in the top-level ontology;
- application ontologies: describe concepts depending both on a particular domain and task, which are often specialisations of both the related ontologies.

Examples of such ontologies are:

- Top-level ontology: DOLCE (Descriptive Ontology for Linguistic and Cognitive Engineering), developed by the Laboratory for Applied Ontology (Trento/Italy) as part of a Foundational Ontologies Library of the WonderWeb project. DOLCE provides a domain-independent framework to build ontologies on the basis of highly-reusable patterns. Website: http://www.loa-cnr.it/DOLCE.html
- (core) domain ontology: The CIDOC Conceptual Reference Model formally describes concepts and relations that are used in the documentation of cultural heritage; the CIDOC-CRM is an official ISO standard (ISO 21127:2006 - A reference ontology for the interchange of cultural heritage information). CIDOC-CRM is aligned to DOLCE. Website: http://cidoc.ics.forth.gr
- Application ontology: A combination of the domain-specific aspects of the CIDOC-CRM and the MPEG-7 model into a single ontology for describing and managing multimedia in museums has been developed by Jane Hunter (2002). A simple version of this ontology is used by Museo24, a semantic virtual museum of the Jämsä region in central Finland. (Szász et al. 2006; see section 7.7.4)

A selection of ontologies that have been developed in the fields of natural history and biodiversity is included in chapter 13.

Ontologies for the Semantic Web

Ontologies are part of the W3C standards stack for the Semantic Web and there are languages, in particular, the Web Ontology Language (OWL) and a variety of tools for creating and working with machine-processable ontologies. In terms of formal expressiveness, ontologies are the "high road" to semantic content/metadata enrichment.
Most available ontologies are domain ontologies, that allow for expressing, constraining and analysing the intended meaning of the shared vocabulary of concepts and relations in specific domains of knowledge. Such vocabularies are used to exchange data among systems, publish reusable knowledge bases, provide semantic search & retrieval services, and offer services to facilitate interoperability across multiple, heterogeneous systems and databases. (cf. Gruber 2007)

Lexical-semantic networks

Finally, it may be important to distinguish Web applications and services that draw on the formal, semantic layer provided by ontologies from so called semantic networks. The considerable progress in recent years in processing natural language expressions has allowed for identifying words that are used synonymously, organise them into sets of synonyms, which represent different concepts, and capture different semantic relations between such sets or concepts. The web of such conceptual-semantic relations lacks the apparatus of formal ontologies (e.g. conceptual hierarchies, axioms, rules, etc.), however, can greatly enhance data mining and search & retrieval applications. It is also possible to map high-level concepts of a semantic network to classes of a formal ontology.
The most noted example of a freely available semantic network is WordNet, a large lexical database of English developed by researchers at Princeton University (http://wordnet.princeton.edu). This network is used in a variety of search engines. In the WordNet database nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. The resulting network of meaningfully related words and concepts can be navigated with a browser. There also are a number of other wordnet projects. For example, EuroWordNet has produced wordnets for several European languages and linked them together, however, these are not freely available.

5.3 Folksonomies

Folksonomies as emergent semantics

Web platforms for storing and sharing content (e.g. Flickr for image sharing) or bookmarks (e.g. del.icio.us) and widely used "social software" tools such as Weblogs have brought about an explosion in user generated content categories, keywording and other annotations.
In contrast to a formalised classification of resources that uses a controlled vocabulary, in these Web environments so called "folksonomies" emerge through an unconstrained process in which many people use their own freely chosen categories or keywords.
Although most tagging systems do not implement vocabulary control there is almost always a cognitive or social feedback that influences tagging behaviour towards consensus. This process also is known as "emergent semantics" or "wisdom of the crowd".

"Trees" versus/and "leaves

There has been much discussion about the value of folksonomies and, inevitably, many contributions contrast them with formal classification systems such as taxonomies or ontologies. In the comparison often the formal systems are criticised as "top-down", "exclusive", and "overrated" ways of organising Web resources. (cf. Kroski 2005; Shirky 2005; a neutral contribution is Mathes 2004)
A more appropriate comparison may be "trees" (taxonomies) versus "leaves" (keywords) and to admit, "This is not an either-or. The old way – trees – make sense in controlled environments where ambiguity is dangerous and where thoroughness counts. Trees make less sense in the uncontrolled, connected world that cherishes ambiguity." (Weinberger 2005)
The analogy also suggests that the two approaches may be combined which actually has become an important topic of technological research. (cf. Mika 2005; Quintarelli, Resmini and Rosati 2007; Specia and Motta 2007)

There are many interesting aspects as well as shortcomings in folksonomies, but the following points may be of particular interest:

Reduction of cognitive effort

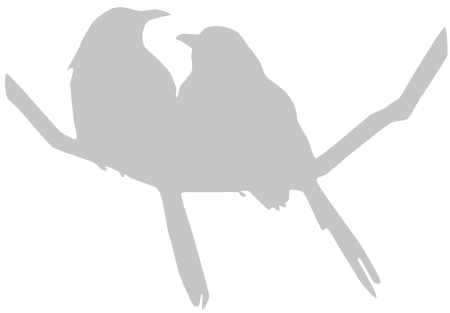
Tagging resources with freely chosen keywords requires little cognitive effort and allows for some personal benefit (Sinha 2005), while the task of turning this "metadata" into a useful resource is off-loaded to the computing system of the platform that is used to share content, bookmarks or other information resources.

Exploitation of user created tags

The "leaves" that are raked together by the computing system for the most part are simple tags in a flat namespace, but can be exploited through mechanisms such as clustering keywords (e.g. "tag clouds") and presenting resources that have been tagged with the same keyword/s. This can allow for identifying some interesting resources, although there is "no semantics inside". In general, users will not be interested in all resources that are available on a topic but the most popular or the latest additions.

Ethno-classification

One of the most important strengths of a folksonomy is that it based on the vocabulary of the content users, which is particularly useful if they form a community of interest. A folksonomy that emerges in such a community may be a starting point for creating a professionally designed controlled vocabulary. Peter Merholz notes: "A smart landscape designer will let wanderers create paths through use, and then pave the emerging walkways, ensuring optimal utility. Ethnoclassification systems can similarly 'emerge.' Once you have a preliminary system in place, you can use the most common tags to develop a controlled vocabulary that truly speaks the users' language." (Merholz 2004)
Indeed, collaborative tagging could be a catalyst for improvement and innovation in creating and using knowledge organisation systems.



6 The SKOS road to semantic interoperability

The chapter above gives an overview of different types of Knowledge Organisation Systems (KOS) and notes the ones for which the W3C Simple Knowledge Organization System (SKOS) standard is intended.
These do not include simple term lists (e.g. glossaries or gazetteers), folksonomies such as result from simple keyword tagging, lexical-semantic networks (e.g. WordNet), and formal ontologies.
This chapter now presents the SKOS road to semantic content/metadata enrichment, which is about how to exploit available KOS such as thesauri and classification systems in Semantic Web enhanced information services.
SKOS has been designed to provide a light-weight conceptual modeling language and low-cost migration path for porting existing KOS to the Semantic Web. Hence, SKOS allows for re-using of, and capitalising on, the rich legacy of existing KOS in the Semantic Web environment. SKOS is not intended to replace existing KOS, however, it also supports developing and sharing of new KOS.
It should be noted that SKOS is not the only format for encoding and exchanging KOS data over the Web. Other notable formats are the MARC21 formats for authority data and classification data and the Zthes specification for thesauri. However, it is generally understood that SKOS provides more flexibility with regard to KOS representation, extension and access.

6.1 Aims and current status of SKOS

Capitalising on the rich legacy of domain-specific KOS

The focus of Simple Knowledge Organisation System (SKOS) is on controlled vocabulary that is used to describe information resources. In domains of knowledge such as cultural and scientific heritage, there is a rich legacy of domain-specific KOS and collections indexed by using this controlled vocabulary.
In oder to capitalise on existing KOS in the emerging Semantic Web environment, there is a need to make them machine-processable and to integrate them in indexing and search systems. "SKOSifying" controlled vocabulary allows to develop semantically enhanced indexing, search & retrieval, browsing, recommendation and other services.

Designed for semi-formal hierarchies of concepts

SKOS has been specifically designed for expressing in RDF the semantics of controlled vocabularies that have a semi-formal hierarchy of concepts, such as used in thesauri or classification systems.
The goal of the SKOS design was to provide a formalisation powerful enough to support semantically enhanced search and other functionality, but simple enough to be undemanding in terms of the cost and expertise required to create the formalisation. Ideally, for thesauri following international standards (see below) or typical classification systems "SKOSification" should require little or no remodelling of the original SKOS structure.

Focus on thesauri and selected other KOS

SKOS initially has been primarily intended for thesauri, however, its scope was extended to also include other semi-formal KOS such as taxonomies, classification and categorisation systems, and subject heading systems.

Thesaurus standards

SKOS at present is most often applied to thesauri broadly conforming to the ISO (ISO 5964-1985, ISO 2788-1986) and NISO (NISO Z39.19:1993) standards for the development of thesauri. It should be noted that these standards were developed in "pre-internet" times with little consideration of Web-based interoperability and current generation search & retrieval applications.
Recently the British Standards Institution's committee IDT/2/2/1 has developed the "Structured Vocabularies for Information Retrieval" (BS 8723) standard. The proposal to adopt BS 8723 was submitted to the committees of all the national standards bodies participating in ISO 2788 and ISO 5964. The proposal was accepted in August 2007 and some countries have agreed to participate in the ISO standardisation,

process (project ISO NP 25964). A revision of the U.S. standard for controlled vocabularies NISO Z39.19:1993, was initiated by NISO in 2002 and is an ongoing process. (Dextre Clarke 2007)

W3C status of SKOS

The initial development of SKOS was done as part of the EU-funded Semantic Web Advanced Development for Europe (SWAD-E) project (FP5-IST, 05/2002-10/2004). The results of the SWAD-Europe Thesaurus Activity were taken up by the W3C Semantic Web Best Practices and Deployment Working Group to prepare SKOS for formal W3C status.
It must noted that SKOS still has only Working Draft status, however, it has advanced considerably on the review process of the W3C recommendation track. End of August 2008, the Semantic Web Deployment Working Group has published the Last Call Working Draft of SKOS Simple Knowledge Organization System Reference.

Where does SKOS sit in the Semantic Web "layer cake"

Chapter 4 above briefly describes the "layer cake" of Semantic Web languages and other important elements. SKOS is located in the RDF/ RDFS layer, building on the syntactic XML layer, but not aspiring to model complex domains of knowledge, which is the remit of the Web Ontology Language (OWL).

A "bridging technology"

However, SKOS also is understood to provide on the one hand a gateway into the onotological layer of the Semantic Web (i.e. OWL domain or top-level ontologies) and, on the other hand, a reference point for less formal keywording and categorisation practices such as "social tagging".
As summarised in the W3C SKOS Primer: "SKOS can also be seen as a bridging technology, providing the missing link between the rigorous logical formalism of ontology languages such as OWL and the chaotic, informal and weakly-structured world of Web-based collaboration tools, as exemplified by social tagging applications." (W3C / Isaac and Summers 2008)

SKOS and folksonomies

Folksonomies that emerge from social tagging behaviour are discussed in section 5.3. In the following we do not elaborate further on potential bridges where folksonomies and formal approaches of providing controlled vocabulary might move closer together.
However, some interesting research questions with respect to SKOS may be: How could "SKOSified" controlled vocabularies be integrated in the computing backbone of social tagging platforms in a way that leverages their capability to capture and expose semantic relationships between tags? Are there feasible, ideally (semi-)automatic approaches for "SKOSifing" folksonomies?
The overall approach should on the one hand not impose language control on taggers and, on the other hand, dynamically leverage semantic structure and depth. Ideally, the approach would create a feedback loop with user groups who want to benefit from added semantic intelligence of the tagging platform.

SKOS and OWL

With regard to possible ways of combining SKOS and OWL, some notes are provided in section 6.3.3.

Who uses SKOS?

SKOS is increasingly used in many fields of knowledge from astronomical entities (e.g. the International Virtual Observatory Alliance – IVOA, 2008) to biodiversity on Earth (e.g. the CSA/NBII Biocomplexity Thesaurus).
In this report only uses of SKOS in the fields of natural history and biodiversity and cultural heritage are covered (a number of suggested use cases in other disciplines are provided in W3C / Isaac et al. 2007).
In chapter 15, the CSA/NBII Biocomplexity Thesaurus, the General Multilingual Environmental Thesaurus (GEMET) and the CAIN Invasive Species Management Thesaurus are described. The CSA/NBII Biocomplexity Thesaurus is freely available for application developers as SOAP based web service while for the other thesauri SKOS files are freely available for download.
Examples from the field of cultural heritage are mentioned in the sections 5.1, 7.3, 7.4 and 7.7.2 (e.g. Getty thesauri, Iconclass and English Heritage thesauri). Here it is important

to note, that SKOS representations of these thesauri and classification systems have been produced in the framework of research projects (and sometimes are available from project websites), however, copyrights may not be cleared sufficiently to allow re-use.

6.2 Brief description of SKOS

SKOS basics

Simple Knowledge Organisation System (SKOS) provides a standard way to represent KOS such as thesauri and other controlled vocabulary in a machine-processable form by making use of Resource Description Framework and Schema (RDF/S). (The following description of SKOS is based on Isaac 2008; Miles 2005; Miles et al. 2005; W3C / Isaac and Summers 2008)

The SKOS Core Vocabulary is a set of RDF properties and RDFS classes, that can be used to express the structure and content of a KOS. Encoding this information in RDF/XML allows a KOS to be published, KOS information passed between applications, used for purposes such as resource discovery and retrieval, and linked or merged with other RDF data on the Semantic Web enabling wider re-use and better interoperability.

The model underlying the design of SKOS assumes that the basic purpose of a controlled structured vocabulary is to establish a set of distinct meanings or concepts, and to provide a way of referring to those concepts that is unambiguous at least within the scope of the vocabulary.

The W3C SKOS Primer summarises how this is enabled: "In basic SKOS, conceptual resources (concepts) can be identified with URIs, labelled with lexical strings in one or more natural languages, documented with various types of note, semantically related to each other in informal hierarchies and association networks and aggregated into concept schemes." (W3C / Isaac and Summers 2008)

SKOS concept classes

SKOS provides only two concept classes: The skos:ConceptScheme class is used for representing a set of concepts, and skos:Concept is used to declare individual concepts, which are linked to the concept scheme using the skos:inScheme property. One important feature of SKOS is that it is possible for the same concept to be linked to several concept schemes.

Labelling Properties

SKOS provides properties to attach labels to concepts. The basic type of label is a lexical label, i.e. a string of Unicode characters. Each lexical label may also be associated with a particular natural language (e.g, German or French), which allows for multilingual labelling of concepts.

Each lexical label is either preferred, alternative or hidden. There can only be one preferred label per language. Alternative labels may be used for synonyms but also abbreviations and acronyms. Hidden lexical labels are usually not rendered when generating a visual representation for users, rather, they are used by search applications for dealing with often mis-spelled or mis-typed words.

Documentation Properties

SKOS also provides properties for documentation purposes, which are primarily intended for human-readable documentation.

A skos:note property for general documentation purposes is further specialised into the following properties for more specific types of documentation:

skos: scopeNote for some, possibly partial, information about the intended meaning of a concept (especially to inform indexing practice);

skos: definition for a more complete explanation of the intended meaning of a concept;

skos: example for an example of the use of a concept; and skos:historyNote for significant changes of a concept.

In addition to these types of documentation, which are intended for users of a concept scheme, skos:editorialNote and skos:changeNote are meant for purposes of administration and maintenance of the thesauri or classification system.

As SKOS allows for extension, also additional types of documentation may be defined.

However, also other, non-SKOS properties – for example, from the Dublin Core Element Set (e.g. dc:creator) – may be used.

Semantic relations

Most interestingly with respect to the Semantic Web is that SKOS allows to define semantic relations between concepts. Semantic relations play a crucial role for defining concepts: The meaning of a concept is defined not just by the natural-language words in its labels, but also by its links to other concepts in the vocabulary.
The basic SKOS standard offers built in support for three types of relationships: broader, narrower and related (however, it may be extended by defining custom relationship types):

- skos:broader: is used to assert that one concept is broader in meaning, i.e. more general, than another, where the scope of one (e.g. "mammals") falls completely within the scope of the other (e.g "animals");
- skos:narrower: is used to assert the inverse, that one concept is narrower in meaning, i.e. more specific, than another;
- skos:related: is used to assert an associative, non-hierarchical relationship between two concepts, for example: "birds" and "ornithology". The property skos:related is a symmetric property which is not transitive (see below).

It is important to note that the SKOS model does not state that the properties skos:broader and skos:narrower are transitive; which would mean, for example, if concept A has a broader meaning than concept B which itself has a broader meaning than concept C, it would follow that concept A also has a broader meaning than concept C.
Yet this does not imply that these properties are intransitive; some SKOS concept schemes may state conceptual hierarchies that are transitive. To declare and exploit such hierarchies (e.g. for inferencing), specific (super-) properties, skos:broaderTransitive and skos:narrowerTransitive can be used.
Finally, to allow an efficient access to the entry points of broader/narrower concept hierarchies, SKOS provides a skos:hasTopConcept property. This property allows for linking a concept scheme to the most general concepts it contains (e.g. concepts such as mammals, fish, etc. of a classification system for animals).

Advanced features

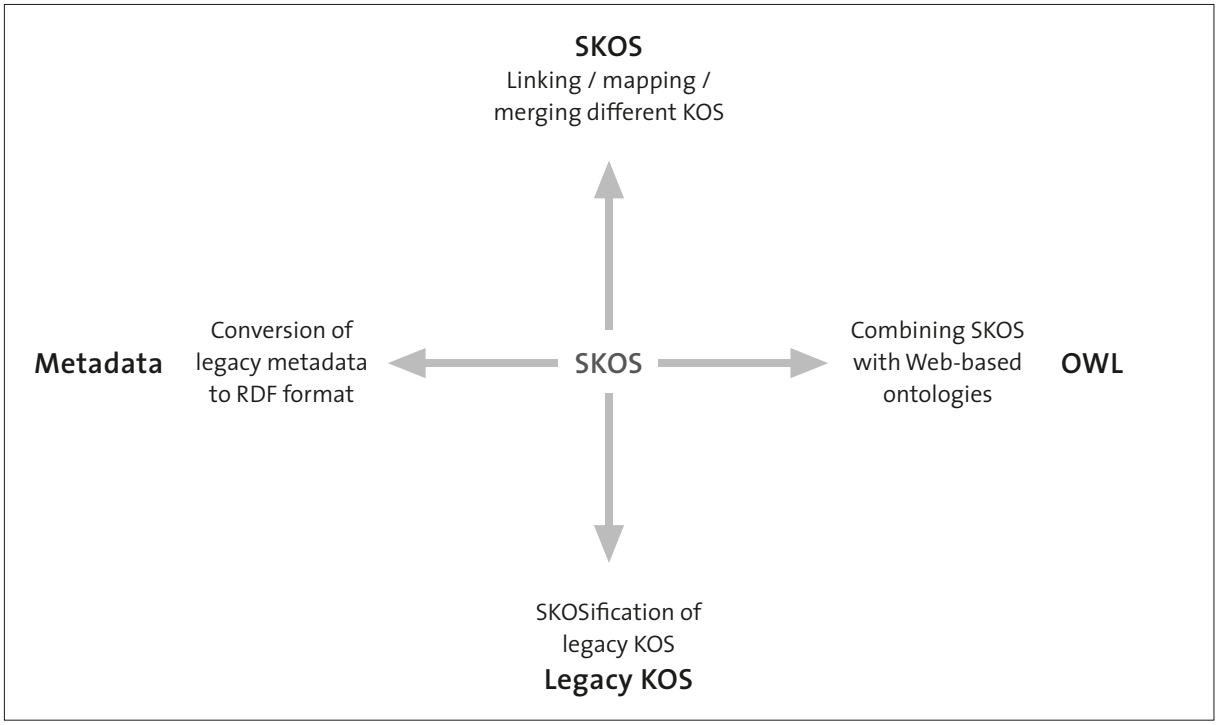
Furthermore, advanced SKOS provides some support for representing meaningful groupings of concepts such as labelled or ordered collections.
However, more important may be that SKOS also allows to map concepts across concept schemes. These additional semantic relations for mapping and merging different SKOS are addressed below in section 6.3.2.

Support of subject indexing dropped in 2008

It may alos be important to note that the current W3C SKOS documents (e.g. SKOS Reference, SKOS Primer) do not contain the four subject indexing properties that formed part of previous material and many presentations and publications. These properties were skos:subject, skos:primarySubject, and their inverses: skos:isSubjectOf and skos:isPrimarySubjectOf.
Although one of the main applications of SKOS would be subject indexing, and skos:subject is already deployed in some applications, e.g. DBpedia (http://wiki.dbpedia.org/Datasets?v=1ec1#h18-7), these properties were dropped in May 2008.
The rationale of the decision was that "1) it's the role of SKOS to publish vocabularies and not to indicate how they should be used for indexing purposes, 2) there appear to be enough support from existing metadata vocabularies to handle links between resources and SKOS concepts". (Miles [SKOS issues review] 2008)

6.3 The SKOS "cross road"

Following the brief introduction to SKOS above, we now address what may be called the SKOS "cross road" and is shown in the figure below.



Looking into four directions

Standing on this cross road we can look in four directions:

- legacy controlled vocabulary that should be converted to, and published in, SKOS/RDF format;
- possible mappings with other controlled vocabulary in SKOS format that extend and enrich the semantic reference network;
- combining SKOS with ontologies in OWL to further formalise semantic relations of the reference network, thereby providing for enhanced capability in cross-domain search, browsing, reasoning, etc.;
- legacy metadata that was created based on a standard or "homegrown" metadata scheme (and terms from some controlled vocabularies) that needs to be converted to RDF format.

In the sections below we address the first three of these directions on a general level, while in the next chapter they are presented with selected cases, including the approaches and tools that are used in current best practice. Some of these cases also comprise to convert legacy metadata to RDF format, because, the final goal of course is to discover and retrieve via the metadata relevant content items.

6.3.1 SKOS creation and publication

In order to benefit from the rich legacy of cultural and scientific heritage KOS (e.g thesauri, classification systems, etc.) in the Semantic Web environment, it is necessary to convert these KOS from other formats to SKOS. Hence, in this section we address the creation and publication of a SKOS representation of an existing KOS that is available in some digital format (e.g. XML, relational database, CSV file, spreadsheet). The focus is on the conversion of thesauri, which is the field where so far most conversion projects have been carried out.

SKOS Core Guidelines for Migration of thesauri

The SWAD-Europe "SKOS Core Guidelines for Migration" (2004) provide a useful guide for generating SKOS/RDF based serialisations of existing thesauri, for both standard (i.e. ISO 2788:1986 compliant) and non-standard thesauri, and from a number of existing formats. The Guidelines comprise three case studies of non-standard thesauri, one of which is the General Multilingual Environmental Thesaurus (GEMET).

Complex classification schemes will often need specialisations of SKOS

In general, thesauri conforming to the ISO or NISO standards should map in a fairly straight forward manner to SKOS. Also simple taxonomies may be encompassed within SKOS with relatively little specialisation, if any. However, complex classification schemes may require considerable specialisations and extensions if their full content is to be captured.

Issues in the conversion of thesauri

Below we briefly describe issues in the conversion process that have been identified in a number of thesaurus conversion projects. (Byrne 2008c; Omelayenko 2008; Tudhope, Binding and May 2008, Van Assem et al. 2006)

Thesaurus analysis

In order to allow an as effective as possible conversion, the first step should always be to analyse the existing thesaurus if it adheres to a thesaurus standard or has any non-standard features. Some departures from the standards may entail some loss of original features, others may be accommodated by specialisation of the core SKOS elements.
Particular problems may pose older term-based thesauri (i.e. based on the ISO 2788-1986 standard) or thesauri employing specific, non-standard relationships or properties. In some cases there will also be the need to consult with the thesaurus providers or experts on the aims of such non-standard features.

Different conversion routes

Based on the analysis of the thesaurus elements and decision on a strategy for non-standard features, different thesaurus formats will require the use of different conversion routes:

KOS available in XML

If the KOS is already available in an XML representation conforming to a published XML Schema, this can greatly facilitate the conversion. If the thesaurus conforms closely with the thesaurus standards and is reasonably compatible with the SKOS data model, it may even be possible to use XSL Transformations (XSLT) to achieve the conversion.
XSLT is the XML transformation language that allows to create rules for translating one XML document to another (http://www.w3.org/TR/xslt). RDF, the target format of the legacy KOS, uses a specific kind of XML and, hence, RDF documents can be rather easily constructed at the XML (syntactical) level with XSLT.
Though, there have been experiences that an XSL transform approach worked well for smaller thesauri but not for large ones. In such cases a SWI-Prolog program was used to convert the XML data to SKOS/RDF. (cf. Tudhope, Binding and May 2008)
In some cases it may also be useful to import the XML distribution of a KOS into a database and create a custom SKOS output generator.

KOS available in a relational database

If the KOS is available in a relational database format, one can generate a RDF/XML report, or can use RDB-RDF mapping (e.g. D2RQ). Making use of a relational database schema may also be necessary, or the easiest approach, if the KOS is distributed as a spreadsheet or CSV (Comma Separated Value) file. For example, CSV files may be imported into a MS Access database and a small custom C# application written to export the data from this database into SKOS/RDF format.

KOS available in a spreadsheet

If the KOS is available in a spreadsheet one can chose output to CSV, XML or other intermediate form, and proceed from there.

Specific issues in the conversion process

Some specific conversion issues of note are:
Character encodings may be problematic as, for example, encountered in an attempt to use an XSL transform to convert between MARC-XML and SKOS RDF/XML; the solution was to create an XSL 2.0 transfrom using the Saxon XSLT 2.0 processor (Vizine-Goetz, Houghton and Childress 2006).
Concept or term identifiers may also pose problems as many controlled vocabularies either do not have identifiers – the preferred term acts as the identifier – or the internal identifiers are not Web actionable URLs.
There generally is the need to create unique identifiers (URIs) for the SKOS respresentation as part of the conversion process. Actually, unique identifiers may need to be

invented if the thesaurus has no notion of identifiers (general suggestions for URI creation are to be found in W3C/ Sauermann and Cyganiak 2008).
The importance of validation must be emphasised: The W3C provides a SKOS validation service that consists of a series of SKOS compatibility and thesaurus integrity tests (http://www.w3.org/2004/02/skos/validation)

Publication on the Internet

The simplest way to make a SKOS representation of a controlled vocabulary available on the Web is to publish the entire vocabulary as a single RDF/XML document on an HTTP server. The vocabulary can then be retrieved by Web clients via issuing an HTTP GET request.

Use of a SPARQL service

However, this may not be a practical solution if the vocabulary is large and clients only need small parts of it. In this case a solution is to make the vocabulary available via a SPARQL service. SPARQL Query is an RDF query language that allows data from one or more RDF graphs to be queried and selected.

W3C Best Practice Recipes for Publishing RDF Vocabularies

The W3C's "Best Practice Recipes for Publishing RDF Vocabularies" (W3C / Burrueta and Phipps 2008) describes the basic steps needed to publish an RDF vocabulary on a Web server. All of the recipes give example configurations for the Apache HTTP server, however, also other such servers as well as dedicated RDF servers such as Joseki or Sesame may be used. The document also contains a set of requirements that should be met to allow the data to be used with Semantic Web applications.
The recipes differ according to the types of content one wants to provide (only machine processable RDF or also single or multiple HTML documents) and the URIs of the concepts and properties of the vocabulary. With respect to the URIs the question is if a hash namespace or a slash namespace is used.
SKOS uses a hash namespace: This means that the URIs for the concepts and properties are constructed by appending first the hash character (#) and then a "local name" to the vocabulary URI. The "local name" is a string of characters that uniquely identifies that concept or property within the scope of the vocabulary; this also is known as a "fragment identifier" (example: http://www.w3.org/2004/02/skos/core#Concept).

6.3.2 SKOS – SKOS mapping

A major goal of the Semantic Web approach is to allow for uniform search & retrieval across distributed heterogeneous content databases. Often the metadata of these databases has been created using different controlled vocabularies (e.g. thesauri, classification systems or other KOS).

Increasing interest in mappings between controlled vocabulary

Through semantic mappings between concepts of different controlled vocabularies in SKOS format, queries on available content across the metadata can be enabled, if the metadata itself is available in RDF format.
Therefore, there is an increasing interest in such mappings and better integrated library and other terminology services. (cf. the OCLC terminology services project, 2008; Vizine-Goetz, Houghton and Childress 2006; Tudhope, Koch and Heery 2006, Si 2007)

SKOS mappings as a key element for enhanced service provision

Library, archive and museum information is rapidly evolving into XML services environments, like for example the library sector standard MARC-21 has done. (cf. McCallum 2005) It is expected that building on this evolution, next generation services will use RDF and exploit semantic relations for enhanced service provision. The capability of SKOS to support semantic mappings between controlled vocabularies makes it a key element in such services.

SKOS properties for mapping between controlled vocabularies

The current SKOS Reference defines five properties that can be used to state mapping (alignment) links between SKOS concepts in different concept schemes (W3C / Miles and Bechhofer 2008):
The properties skos:broadMatch and skos:narrowMatch are used to state a hierarchical mapping link between two concepts.

The property skos:relatedMatch is used to state an associative mapping link between two concepts.
The properties skos:closeMatch and skos:exactMatch are used to assert that two concepts have a similar meaning:
skos:closeMatch is used to link two concepts that are sufficiently similar that they can be used interchangeably in some information retrieval applications. However, skos:closeMatch is not declared as a transitive property, which prevents such similarity statements to propagate beyond the two concept schemes.
skos:exactMatch is used to link two concepts that are considered to have equivalent meaning and, hence, can be used interchangeably in retrieval applications. skos:exactMatch is a a sub-property of skos:close Match, but is declared as transitive. This means that, if a concept A is an exact match for another concept B, which is itself an exact match for concept C, it does follow from SKOS semantics that A also is an exact match for C.

Mapping may be costly

Mappings between SKOS representations of different thesauri, classifycation schemes and other KOS can provide a semantic reference network that allows for enhanced search and other capability (e.g. faceted searching and browsing).
However such mappings generally require domain experts and may be time-intensive, hence, costly. Often detailed mapping work at the concept level is necessary for useful results, and automated assistance typically helps to accomplish only parts of the task.
Below we summarise some results from experimental SKOS mappings in the Ontology Alignment Evaluation Initiative (OAEI) 2007 campaign. A more detailed case study is provided in section 7.4.

Automated thesauri alignment excercises in the OAEI 2007 Campaign

In the OAEI 2007 Campaign some thesauri had to be matched using relations from the SKOS mapping vocabulary:
The campaign comprised alignments between SKOS versions of the UN Food and Agriculture Organization's AGROVOC thesaurus and the US National Agricultural Library's Agricultural Thesaurus (NALT), and the European Environment Agency's GEMET thesaurus and AGROVOC and NALT, respectively. Furthermore, two library thesauri for books (GTT and Brinkman) in SKOS format had to be matched. (Euzenat et al. 2007; on the library case see the detailed analysis in Isaac et al. 2008)
OAEI campaigns aim at comparing ontology matching systems on precisely defined test sets in order to reliably assess their capability of finding correspondences between entities (i.e. thesaurus concepts) that suggest possible alignments.
In the thesauri mappings, the following tools where employed in one or more excercises: Falcon-AO 0.7 (South East University) and DSSim (Knowledge Media Institute) participated in all exercises; RiMOM (Tsinghua University), Scarlet (Knowledge Media Institute) and X-SOM (Politecnico di Milano) in the AGROVOC-NALT alignment, and SILAS (Roelant Ossewaarde) in the library thesauri alignment.
The exercises show that the algorithms used in these systems are good in finding correspondences between thesauri that suggest using a skos:exactMatch, but suggestions for skos:broadMatch and skos:narrowMatch were only provided by Scarlet (Knowledge Media Institute) in the AGROVOC-NALT alignment, and skos:relatedMatch only provided by SILAS in the library thesauri alignment.

6.3.3 SKOS – OWL ontologies

SKOS allows for porting thesauri and other KOS to the Semantic Web in a way that is often suitable enough to implement some enhanced search capability (e.g. faceted searching and browsing). However, SKOS also provides a gateway into the semantically more expressive world of ontologies that are built with the Web Ontology Language (OWL).

SKOS allows for only little formalisation of

From the explanation in section 6.2 it should be clear, that SKOS allows for only little formalisation of semantic relations, because, it has been specifically developed for the

semantic relations

rather shallow concept schemes of thesauri, classification schemes and other KOS.
SKOS provides three properties for declaring semantic relations between concepts:

- The inverse properties skos:broader and skos:narrower are used for asserting that of two concepts one is broader or narrower in meaning than the other (e.g. "animals" and "mammals").
- With the (symmetrical) property skos:related it is asserted that two concepts are related somehow, i.e. without defining the semantic relation in any way (e.g. "birds" and "ornithology").

Hence, these properties allow for only generic extension or restriction in searching and browsing applications and to suggest "related resources".
The question now is, what could be expected from combining a thesauri or classification system that is represented in SKOS with a formal ontology in OWL. In this context it is important to emphasise the different purposes of SKOS and OWL (cf. Vatant 2008):

Different purposes of SKOS and OWL ontologies

The focus of SKOS is on the relation between content and controlled vocabulary. Hence, SKOS represents a librarian view of the world, where the main purpose of SKOS concepts is to classify, index, search and retrieve content, based on a limited but extensible set of attributes and relationships.
OWL supports a knowledge representation or ontological view of the world. The main purpose of OWL according to this view is to model domains of knowledge with ontological hierarchies, subclass and subproperty relationships, domain and range restrictions, and instances of classes of entities.
However, using SKOS to represent controlled vocabularies does not necessarily mean that an information system may not benefit from OWL-based expressivity for some part of its knowledge base. Similarly, if a system is OWL-driven it does not necessarily mean that it may not benefit from incorporating vocabularies in SKOS format.

Different options to combine SKOS and OWL

Currently different options of how SKOS and OWL may be used together are explored and discussed by the experts, i.e. there are at present no standard solutions of how to best combine SKOS and OWL in practical applications.
A working document of the Semantic Web Deployment Working Group (W3C SWDWG 2007) distinguishes possible design patterns for working with SKOS and OWL:

- 1. Going from less to more formal as well as from more to less formal (i.e. SKOS to OWL or OWL to SKOS),
- 2. Formal / semi-formal hybrids (part OWL, part SKOS), and
- 3. Adding labels and documentation (notes) to a formal ontology.

SKOS to OWL, OWL to SKOS

1. The document suggests that going from less to more formal and vice versa may be implemented by "overlay" or "transformation", but concludes that overlays should better be avoided.

Overlay: In an overlay of SKOS with OWL, in addition to a skos:broader / skos:narrower hierarchy an OWL/RDFS class/sub-class hierarchy of the same vocabulary concepts is created. This leads to a situation where an instance of skos:Concept also is an instance of owl:Class, which may result in unpleasant consequences if the two sets of RDF triples are merged in the same RDF graph.
Transformation: In the case of a transformation, the concepts of a thesauri are defined as OWL classes and again an OWL/RDFS class/sub-class hierarchy is created. Here the representations are completely separate worlds, though, the definition of some bridges may be useful to express existing correspondences.
In the FinnONTO project, some light-weight thesaurus-to-ontology transformations have been implemented to define more accurately the meaning of semantic relations. Actually, the semantics of "broader term" (BT) relations in thesauri are ambiguous: in ontological terms it may mean a subclass-of relation, part-of relation, or instance-of relation. In the FinnONTO project, some BT relations of thesauri were transformed into subclass-of and part-of relations, instance-of relations were not used. (Hyvönen, Eero et al. 2008)

Hybrids

2. Formal / semi-formal hybrids: In such cases, SKOS and OWL are used side-by-side to model different parts of a conceptualisation. Here unpleasant consequences can be avoided as the SKOS and OWL representations are effectively kept separate in the RDF graph. An example of such an application is the Semantic Web Environmental Directory (see section 7.5).

Annotation of an ontology

3. Adding labels and documentation to a formal ontology: This is considered to not pose any problem, because, it does not involve the use of skos:Concept, only the labelling and documentation (notes) properties are used. (However, see Jupp et al. 2008 for some details that need to be taken into account.)
Some general discussion of how to combine SKOS and OWL also is provided by researchers involved in the STAR project. (Tudhope, Binding and May 2008; see section 7.7.2)

7 State-of-the-art projects

7.1 Introduction

In the Technology Watch activity we have identified a number of projects that have developed and implemented approaches similar to STERNA.
These projects have ported to the Semantic Web legacy metadata as well as KOS and implemented advanced search and other capability that draw on the semantic layer of the created RDF metadata and "SKOSified" thesauri and other knowledge organisation schemes. Furthermore, some projects have used higher-level Semantic Web languages such as OWL (Web Ontology Language) to allow for some reasoning over the semantic layer.
The sections below describe interesting approaches, tools and services that have been developed by these projects.

Character of identified projects

On the spectrum from pure and applied research projects to fully operational implementations under real world conditions, the identified projects are situated in the middle ground. Most often they are research projects that have developed, implemented and tested novel applications using cultural and scientific heritage content to demonstrate their case.
Interestingly, the larger part of identified projects that make use of SKOS are situated in the field of cultural heritage and concern art, archaeological, ethnographical and other museum collections. One reason for this may be that in the field of natural history and biodiversity the key entry point to collections such as observation records, specimen, etc. is the taxonomic classification of organisms. This classification generally is not represented in SKOS (skos:broader / skos:narrower), but with ontological class-hierarchies in OWL or OBO (Open Biomedical Ontologies).
However, SKOS has been used to represent a number of thesauri that could be employed in projects aiming to provide semantic search of, and access to, natural history and biodiversity databases (some examples of such thesauri are described in chapter 15).

Focus on cultural heritage projects

In the sections below we mainly focus on the identified projects from the cultural heritage domain that develop semantic access to heterogeneous collections. In addition, three projects related to natural science and history are covered, the Semantic Web Environmental Directory, AquaRing, and STERNA.

Limited coverage of RDF metadata creation

Most of the projects included to convert legacy metadata to RDF format and controlled vocabularies to SKOS/RDF, starting from whatever formats they were encoded. Furthermore, in order to create the layer for semantic search and browsing, some mapping or alignment between the SKOSified thesauri or classification schemes needed to be achieved.
The sections 6.3.1 to 6.3.3 above describe how legacy KOS in different formats are converted to SKOS and published, and how controlled vocabularies in SKOS format may be mapped. Furthermore, some issues in integrating SKOS with OWL are addressed.
Furthermore the intention was to describe how the leading projects have converted different legacy metadata to RDF format. However, when consulting the available literature it became clear, that an appropriate description of these approaches would require a level of detail only experts may digest and appreciate. (For example, see Byrne 2008a-c, who details methods to convert relational databases to RDF).
Therefore, in the project descriptions below, we address the conversion of legacy metadata only very briefly, and invite experts interested in any details to consult the references that are provided.

Some important points concerning RDF metadata

However, for our purposes the following points may be important to note:
Initiatives that aim to port different cultural heritage repositories with similar content (e.g. historic photographs) to the Semantic Web, typically will map their legacy metadata to a solid target schema, i.e. create RDF/XML metadata based on a common XML schema.

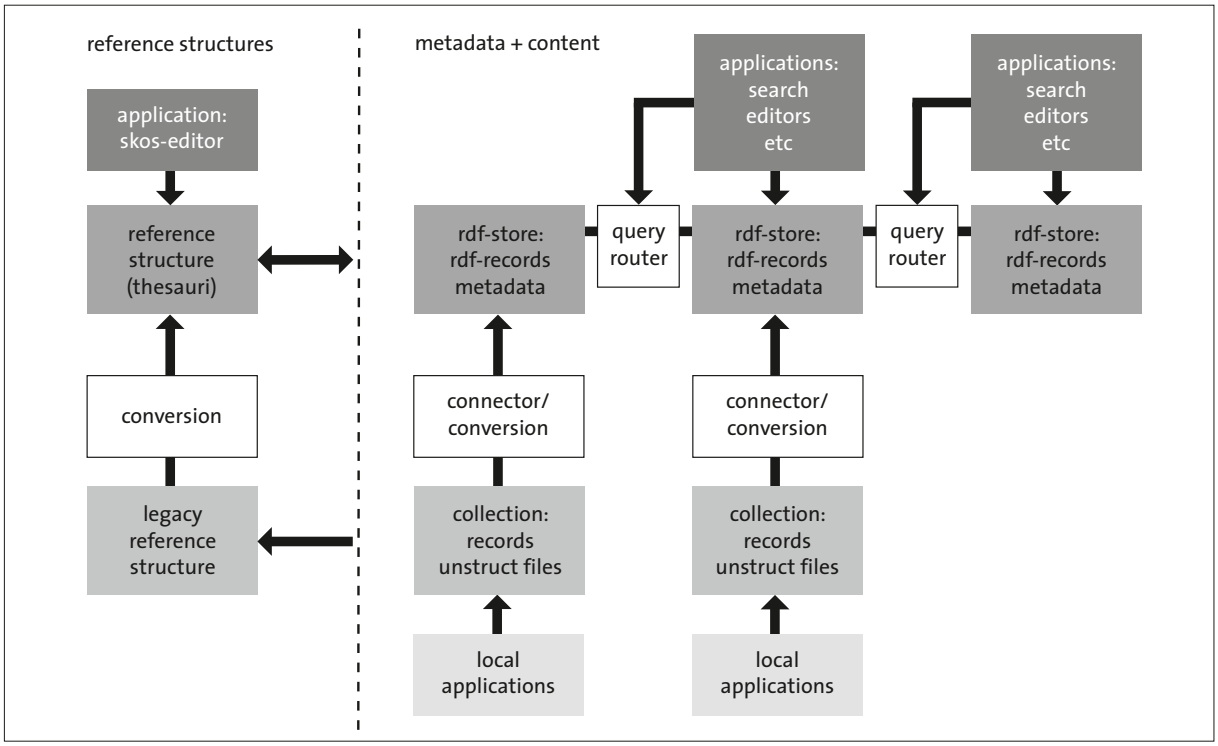
For example, in the MultimediaN E-Culture project (see section 7.3), which focused on image collections, the VRA Core standard of the Visual Resources Association (VRA, www.vraweb.org) was used as the target metadata schema. VRA Core provides a set of 17 elements for describing visual cultural works (e.g. art works, artifacts, architecture) and images of those works; there also is a mapping available from VRA Core (3.0) to Dublin Core.
In the European Digital Library (EDL) initiative, partners here may use as target schema one of the (forthcoming) domain-specific Dublin Core metadata profiles, for example, the profile for museum content.

Conversion of legacy metadata – an important area for know-how transfer

The quality of the common metadata pool of the future EDL is a point of major concern, independent of the question if the metadata is made available in basic XML or RDF/XML format.
With regard to the creation of RDF metadata from legacy databases, the practical state-of-the-art is to use converters, i.e. to create some custom code that specifies rules of how the legay metadata should be transformed in order to further process it to RDF/XML.
In the MultimediaN E-Culture project, which converted several datasets from different institutions, it was found that nearly every dataset required some dataset-specific code to be written and integrated. However, by identifying and separating conversion rules that may be reused, the overall effort can be reduced considerably. Nevertheless, it is estimated that a skillful professional who uses a state-of-the-art conversion support system (in this case, AnnoCultor) would need around four weeks to convert a major museum database, creating for this purpose a dedicated converter of 50-100 conversion rules plus some custom code. (Omelayenko 2008)
Hence, this is an important area where a systematic exchange of know-how, methods and tools could greatly help cultural heritage institutions to port legacy metadata to the Semantic Web as cost-effectively as possible.

7.2 The STERNA architecture for semantic interoperability (SKOS)

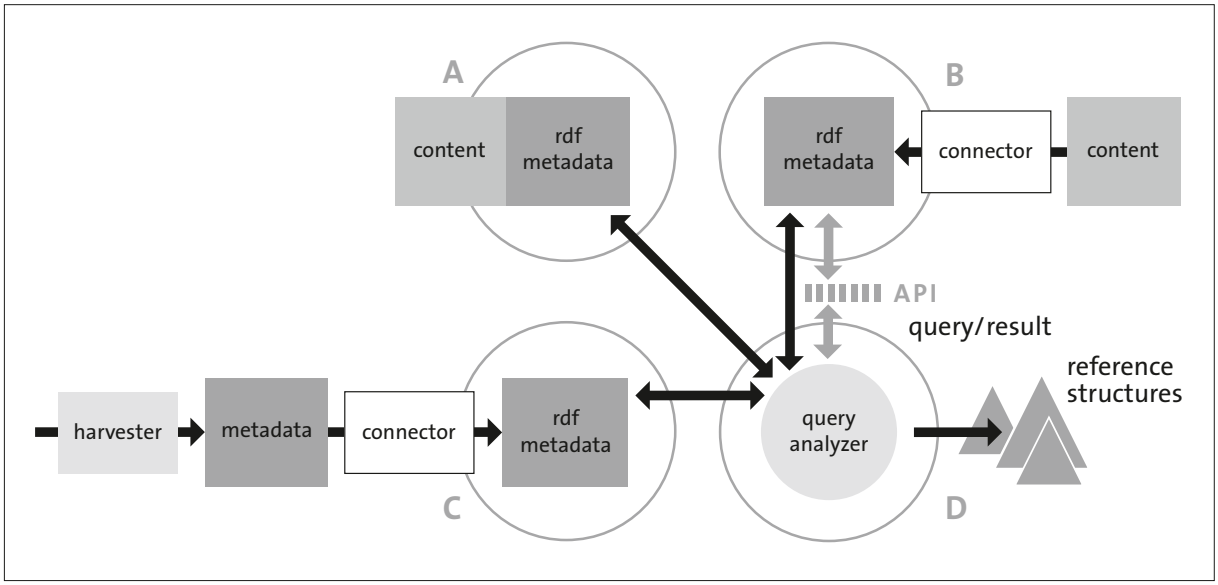
The STERNA approach builds on and extends some of the methods that have been developed in the Dutch Reference Network Architecture project (01/2005-12/2007), which was funded by the Dutch Ministry of Economic Affairs under the auspices of the Ministry of Education, Culture and Science.
The overall aim of the RNA project was to develop practical methods, tools and techniques for building dynamic knowledge systems, based on sets of reference structures (like thesauri, taxonomies, etc.) and content metadata.
The project involved several heritage organisations, research institutes and companies, which worked on a number of different application cases. These cases are described in detail on the RNA website, and there is a publication available that points out the practical approach of the project, with reference to individual cases and lessons learned. (Wester and Nederbragt 2007)
The paragraphs below mainly describe the technical architecture of the STERNA project. A description of the work required to evaluate and select the most appropriate collection databases and reference structures (thesauri, classification schemes, etc.) of the STERNA partners is not included. Actually, this work is carried out at present, driven by end-user scenarios that should allow to combine the most interesting related content.
The selected collection databases and reference structures will be trans-formed to RDF and SKOS and integrated as schematically represented in the figure below:



Source: Hans Nederbragt, © Trezorix (2008): Introduction to the STERNA architecture (available from the STERNA website)

Conversion and aggregation procedures

Legacy reference structures are converted to SKOS format and conceptually related structures mapped, using a SKOS editor. The combined reference structures will be aggregated and held at one of the nodes of the STERNA federated network, where a central query analyzer is implemented.
Collection database records and other files are transformed to RDF format with converters (i.e. conversion rules and some custom code) and aggregated in local RDF triple stores.



Source: Hans Nederbragt, © Trezorix (2008): Introduction to the STERNA architecture (available from the STERNA website; note: in the figure presented above, the part "reference structures" has been added).

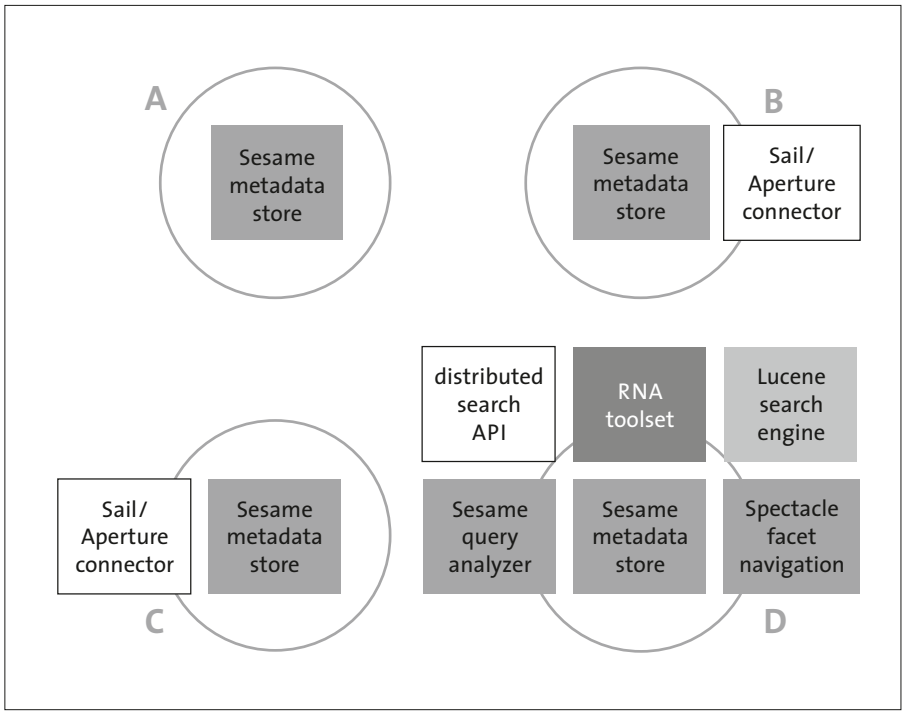
Federated search

The figure above illustrates the setup from the perspective of the network node where the query analyzer sits. In order to respond to received search queries, the query analyzer uses the combined reference structures and draws on the RDF triple stores at the partners' sites.

Incorporation of other data providers

The illustration also comprises a partner (C) that harvests metadata from other data providers, and uses a connector (or conversion routine) to translate the metadata to RDF format. Such other data providers, for example, may include institutions that make metadata available for the Europeana website. If some of their collections would fit particularly well to be combined with STERNA natural history content, connectors could be implemented to produce RDF metadata and incorporate it in STERNA's federated search environment.

Technologies used in the STERNA architecture



Source: Hans Nederbragt, © Trezorix (2008): Introduction to the STERNA architecture (available from the STERNA website)

The figure above provides an overview of some key technologies that are used at the different partner sites. The following list provides some additional information on these technologies and links where more details may be found:

Details and links

- Sesame RDF framework for storage and querying of metadata: see http://www.openrdf.org;
- Sesame query analyzer for intelligent distribution of queries;
- Spectacle facet navigation for building Semantic Web search interfaces: see http://www.aduna-software.com;
- Lucene and Solr for high-performance, full-featured text search and Semantic Web search: see http://lucene.apache.org and http://lucene.apache.org/solr;
- Sail and Aperture connectors for extracting and querying full-text content and metadata from various information systems (e.g. file systems, websites, etc.): see http://www.aduna-software.com and http://aperture.sourceforge.net;
- Semantic networking toolset for editing and maintainance of metadata and reference structures: see www.rnaproject.org;
- Distributed search API for incorporating smart search functionalities in websites.

References Wester and Nederbragt 2007; Nederbragt 2008; Tresorix 2008; some further interesting publications are available from http://www.rnaproject.org/whitepapers.aspx

Websites RNA project, http://www.rnaproject.org
STERNA, http://www.sterna-net.eu

7.3 MultimediaN E-Culture project (SKOS)

Project brief / context The E-Culture project developed a search portal and engine that served as a joint prototype Semantic Web application for subsets of digital collections and thesauri from a number of heritage institutions. The demonstration project focused on semantic interoperability, information access and context-specific visualisation. The E-Culture search portal demonstrator has won the Semantic Web Challenge award at the ISWC Conference 2006.
E-Culture was part of the MultimediaN group of projects that were funded through the BSIK (knowledge society) program of the Dutch Government. The project was led by Guus Schreiber (Vrije Universiteit) and Jacco van Ossenbruggen (Centrum Wiskunde & Informatica) involving a number of senior researchers and PhD students. Cooperation partners from the heritage sector were Digitaal Erfgoed Nederland (Digital Heritage Netherlands) and Instituut Collectie Nederland (Netherlands Institute for Cultural Heritage).
Below we provide an overview of the converted datasets and KOS, the basic conversion process and key tools that have been used in the conversion process as well as for the semantic search portal.
The key tools, AnnoCultor and ClioPatria are described in more detail in chapter 7.8. These tools also have been identified as candidates to be re-used in the development of the Europeana v1 prototype that is expected to be launched in early 2010. (Cousins and Siebinga 2008)

Collection databases In the E-Culture project several datasets from different Dutch art and ethnographic collections have been ported to the Semantic Web. These datasets comprise (incl. number of objects): Artchive.com (>3,000), Rijksmuseum.nl (>16,000), Volkenkunde.nl (>10,000), Tropenmuseum.nl (>78,000) and Bibliopolis.nl (>1,600).

KOS Thesauri and other controlled vocabularies used in the E-Culture project comprise (incl. the number of concepts): Getty AAT (>31.000), Getty ULAN (>130.000), Getty TGN (>890.000) and SVCN (Dutch ethnology, >11.000). Furthermore, the Bibliopolis collection (1,645 images related to book-printing) of the National Library of the Netherlands uses a "home-grown" bilingual thesaurus (English and Dutch) containing 1,033 terms for indexing images.
The core Getty vocabularies (AAT, TGN and ULAN) have been converted from the Getty XML files into RDF using a procedure called gettyconvert. The RDF data is available for download as a zip-file that also includes the RDF schemas used for the Getty vocabularies.

Key conversion support tool AnnoCultor, a generic Java-based framework for converting collection metadata and controlled vocabularies into RDF/SKOS, has been developed and is available from SourceForge.
Details about the conversion rules and methods employed to align terms from legacy metadata to standard vocabularies have been published, including interesting statistics that show the success rate of the conversion and the costs implied. (Omelayenko 2008)

Semantic search portal tools The RDF metadata of the converted datasets and the SKOSified thesauri and other controlled vocabularies form the RDF graph underlying the E-Culture semantic search portal demonstrator. The specific software developed for the portal comprises:
ClioPatria: a semantic search Web server (available under the GPL license) developed using the SWI-Prolog SeRQL engine; libraries that are widely reusable such as the Semantic

Web library also have been developed as part of SWI-Prolog (which offers a comprehensive Free Software Prolog developer environment).
/facet: a generic browser that allows users to explore the databases along any facet such as artist, genre, period or otherwise.
MyArt: an application for personalising the semantic search; users that search with any of the available options (/facet, basic/advanced search or local view) can collect topics of interest to store and further personalise their search.
Other interactive features: The demonstrator portal also features interactive timelines of art works and the lifespan of the artists, that can be used for semantic navigation and search.

References Omelayenko 2008; Schreiber et al. 2006; Tordai et al. 2007; van Ossenbruggen et al. 2007; Wielemaker et al. 2007

Website Project website: http://e-culture.multimedian.nl
Search portal: http://e-culture.multimedian.nl/demo/search

7.4 STITCH – Semantic Interoperability to access Cultural Heritage (SKOS – SKOS mapping)

Project brief / context The STITCH project examined to what extent current Semantic Web techniques can solve issues presented by the heterogeneity of cultural heritage collection databases and controlled vocabularies. To this purpose, STITCH developed methods for aligning and browsing reference structures such as SKOSified thesauri and classification systems.
STITCH has been funded under the CATCH (Continuous Access To Cultural Heritage) programme that is managed by the Netherlands Organisation for Scientific Research (NWO).
The project has been included in our sample of state-of-the-art projects following a suggestion by Hans Nederbragt from Trezorix (STERNA's technology partner) that STICH methods might be adapted for the domain of natural history.

Collection databases STITCH specifically worked on two collections,
- the Aria Masterpieces collection of the Rijksmuseum, Amsterdam,
- the Medieval Illuminated Manuscript collection of the National Library of the Netherlands.

A further use case (not addressed below) explored semantic correspondences between the vocabularies of the Dutch collection of illuminated manuscripts and the French National Library's Mandragore collection, which contains a broader spectrum of illuminated manuscripts.

KOS In the use case addressed here, the following two KOS were employed:
- the Rijksmuseum's Aria thesaurus, and
- the Iconclass classification system.

Approach The STITCH researchers created SKOS representations of Iconclass and the Aria thesaurus, aligned these representations using two state-of-the-art mapping tools, Falcon and S-Match, and implemented a faceted Web browsing environment to visualise and examine the results.
In the development of the SKOS representations, Aria proved almost fully compatible with the SKOS schema, while Iconclass could be converted only partly: subject hierarchies worked well, however, Iconclass idiomatic elements such as keys could not be represented.
In the Web browsing environment different "views" where realised. With these "views" users can browse the vocabularies and retrieve documents from both collections in parallel: Aria single view; Iconclass single view; combined view (where the results correspond

to the conjunction of subjects selected in both subject hierarchies); and merged view (which is based on the fusion of Iconclass and Aria correspondences identified in the mapping of the two SKOS representations).

Considerable limitations in automatic mapping of specific Cultural Heritage KOS

The evaluation of the automatic alignment of the SKOSified Aria subject vocabulary and Iconclass classification system revealed considerable shortcomings of state-of-the-art mapping tools. Falcon only showed 16% and S-Match 46% correct mappings for a selected subset of Iconclass (1500 concepts) and the complete Aria thesaurus (500 concepts).
The reasons for this are that the two vocabularies do not use simple terms but glosses for describing concepts and, generally, current mapping tools expect to be fed with rigidly formalised ontologies rather than loosly-defined conceptual structures.
In order to do justice to Falcon, the following should be noted: S-Match has been purpose-built for thesaurus-like structures, and one researcher of the S-Match developer group was involved in the case study work.

Tools used

Ontology/vocabulary mapping tools:
- Falcon: is an ontology matching system that has been developed by researchers at the South East University's Institute of Web Science (http://iws.seu.edu.cn/projects/matching/).
- S-Match: is a mapper for tree-like vocabulary structures that has been developed by researchers at the University of Trento's Department of Information and Communication (details are given in Giunchiglia, Shvaiko and Yatskevich 2005).

STITCH faceted Web browsing environment: This implementation uses SWI-Prolog and the Sesame RDF repository for storage and querying: http://www.openrdf.org

Lessons learned

Important lessons learned in the STITCH project include:
- there is a need for best-practices to overcome the loss of semantics when translating legacy KOS into available Semantic Web standards such as SKOS;
- ontology alignment techniques need to be better tuned to SKOS representations (e.g. current techniques do not exploit labelling information);
- current generation tools such as mappers and inference engines do not necessarily scale for handling the volume of data present in rich cultural heritage KOS.

References

Van Gendt, M. et al. 2006; Isaac 2007a

Website

http://www.cs.vu.nl/STITCH/
Demonstrator: http://www.cs.vu.nl/STITCH/KB_Rijks_demo.html

7.5 Semantic Web Environmental Directory (SKOS + OWL hybrid)

A prototypic showcase

SWED is a (prototypic) Semantic Web directory of mostly UK based environmental, natural history and community organisations and projects. This application was developed as part of the Semantic Web Advanced Development – Europe (SWAD-E) project with the goal to showcase the use of emerging Semantic Web standards.
SWED uses a combination of RDFS/OWL ontologies and SKOSified thesauri or taxonomies to organise the information into different topic hierarchies.

Use of RDFS and OWL ontologies and SKOS

Of two OWL ontologies used in the SWED one defines the properties of organisations and projects and the other the types of relationships between them. One simple RDFS based ontology describes the facet Operational Area, mainly defining which areas are contained within other areas (e.g that the area of Essex is contained in the area of East of England, East of England in the area of England, and England in the area of the United Kingdom).
SKOS is used for three thesauri or taxonomies: Type of Activity, Type of Organisation and Type of Project.

Faceted search

The "Browse Directory" page of the SWED portal provides a set of facets which can be selected as filters to aid searching. These facets are Topic of interest, Organisation type, Project type, Activity, Operational area and Name. The facet Topic of interest, for example, classifies entries according to environmental topics that organisations or projects are interested in.

One may chose to start a search with one of the listed concepts such as "species" (which produces 30 results), then limit the search results with a concept of the facet Activity, such as "Education and Training" (which reduces the number of relevant entries to 25), and then a narrower term such as "Education for Sustainable Development" to finally arrive at two relevant organisations for which information is available in the database.

The SWED project was completed in October 2004, however, the portal is kept running as a useful demonstration prototype.

Website

http://www.swed.co.uk

7.6 AquaRing (KOS in OWL)

Project brief / context

AquaRing (full title: Accessible and Qualified Use of Available Digital Resources about Aquatic World in National Gatherings) is an *e*Content*plus* project that runs from September 2006 to February 2009.

The consortium includes the Aquarium of Genoa, University of Genoa, Lithuanian Sea Museum, Nausicaa, Royal Belgian Institute of Natural Sciences, Rotterdam Zoo, ECSITE (the European Network for Science Centres and Museums) and the World Ocean Network. Furthermore the technology companies Fundación Robotiker and Softeco Sismat are project partners.

AquaRing develops a semantic information portal (demonstrator) for research and education in marine and aquatic sciences. The Semantic Web tools and expertise are provided by Robotiker-Tecnalia. The tools allow for for semantic annotation, search and navigation.

Initially about 20,000 content items from the aquariums, natural history museums and science centres should be made accessible. The content is annotated using thesauri and other classification systems in OWL. It is also expected that AquaRing makes content available to Europeana.

KOS in OWL

The AquaRing semantic portal will draw on the following KOS:

- the UN Food and Agriculture Organization (FAO) classification schemes for Biological Entities, Fishing Areas, Land Areas and Vessels Types;
- the FAO's ASFA (Aquatic Sciences and Fisheries Abstracts) thesaurus;
- the EUNIS Habitat Classification of the European Environment Agency (only using the habitat types classification);
- the EDUcational ontology, created by AquaRing partners using concepts from different resources such as the Learning Object Metadata (LOM) standard, Learning Resource Exchange (LRE) metadata, Bloom's Educational Taxonomy and others.

The first four FAO concept schemes were received in OWL from the NeOn - Lifecycle Support for Networked Ontologies project (FP6-IST, 03/2006-02/2010), that together with the FAO works on a use case on "Ontology-driven stock over-fishing alert system" (http://www.fao.org/aims/neon.jsp).

OWL formalisations of the other concept schemes were created by Robotiker-Tecnalia using as input for the FAO ASFA thesaurus an XML file, for the EUNIS Habitat Classification an Excel file, and for the EDUcational ontology a modelling made by AquaRing partners.

It is not expected that the combined OWL representations of the concept schemes will cover all concepts that are of interest to AquaRing. Therefore it was decided to allow for extending the coverage, or detailing of concepts, by (controlled) free tags. During annotation of resources, editors can add free tags (i.e. terms, keywords, etc.) to the concepts of the classification schemes as formalised with OWL.

For example, FAO Fishing Areas "_22011_12 (EC Atlantic) may be detailed with "Rio Deva;Rio Asón". The free tag is placed (hierarchically) as an instance of the corresponding root concept (OWL class) that contains "Rio Deva;Rio Asón" as value.
Moreover, AquaRing implements an "ontology learning" technique that allows for integration the ontologies via (supervised) automated semantic relationship creation. The technique takes content annotations as information input and exploits the relations that are implicitly established when ontology instances are used to annotate contents (some details are provided in González 2008a).

References González 2008a+b; González, M., Bianchi, S. and Vercelli, G. 2008

Website http://www.aquaringweb.eu
At the time of finalising the STERNA Technology Watch report, there was no publicly accessible AquaRing demonstrator website available.

7.7 CIDOC-CRM based applications

7.7.1 Purpose and scope of, and issues with, CIDOC-CRM

CIDOC-CRM basics The CIDOC (International Committee for Documentation) Conceptual Reference Model formally describes concepts and relations that are used in the documentation of cultural heritage. It was developed with the intention to facilitate the integration, mediation and interchange of heterogeneous cultural heritage information. (CIDOC-CRM, http://cidoc.ics.forth.gr)
More specifically, the CIDOC-CRM follows object-oriented design principles and provides a conceptual framework of 80 classes and 132 properties for describing common high-level semantics that allow for information integration at the schema level.
Although the CIDOC-CRM was initially engineered from data structures in the cultural heritage sector, most of the classes and properties are surpriseingly generic. Actually, they are characteristic for the logic of retrospective documentation as it occurs generically in most scientific, cultural and other domains. (Doerr, Ore and Stead 2007)
The driving principle of the CIDOC-CRM is the explicit, formal modelling of events, which allows to connect facts into coherent descriptions of historic events. In the domain of natural history, such events may be expedition events, field observation events, object collecting events, object curation events, etc. (examples of how such events are represented in CIDOC-CRM are provided in Lampe 2006; Lampe and Krause 2008).

ISO standard development Work on the CIDOC CRM began in 1996 under the auspices of the CIDOC Documentation Standards Working Group, and in 2000 was delegated to the CIDOC-CRM Special Interest Group. This group collaborated with the ISO working group ISO/TC46/SC4/WG9 to bring the ontology to the form and status of an international standard. In September 2006, the ontology became an official OSI standard (ISO 21127:2006 - A reference ontology for the interchange of cultural heritage information).

Difficulties in the application of the CIDOC-CRM: Some examples The ability of the CIDOC-CRM to support information integration has been demonstrated in a number of demonstration projects in different domains including e-science, cultural heritage, archaeology, biodiversity, and others.
However, it is a rather abstract, high level conceptual model, which has shown difficult to apply for researchers and practitioners that have not been involved in its development and related demonstration projects. Moreover, the model may need to be specialised when warranted for particular information integration purposes.

SCULPTEUR For example in the research and technological development project SCULPTEUR (FP5; 05/2002-04/2005), museum databases were mapped to the CIDOC-CRM (with some extensions) to implement cross-collections, concepts-based search & retrieval. As in

other cases the researchers reported, that "mapping is complex and time consuming. The CRM has a steep learning curve, and performing the mapping requires a good understanding of both ontological modelling as well as the source metadata system. Eventually the assistance of a CRM expert was required to complete and validate the mappings." (Sinclair et al. 2005)

English Heritage CRM specialisation and extension

The practical difficulties in developing the required understanding and representations (e.g. spreadsheets, UML diagrams, etc.) to allow subject-experts from a particular domain apply the CIDOC-CRM is also well documented for the work carried out at the English Heritage' Centre for Archaeology. (Cripps et al. 2004; see section 7.7.2 below for current information on this work)

BRICKS

Researchers of the BRICKS project (FP6, 01/2004-06/2007) identified two major issues with the CIDOC-CRM that might impede the goal of enabling interoperability across heterogeneous databases: "The first issue is the abstractness of the concepts (e.g. Time Appellation, Man-Made Object) defined by the global ontology, which makes them ambiguous to any human user. Even expert users have produced ambiguous mappings and have required several iterations to produce consistent mapping definitions. If several experts specify mappings independently from each other, it is very likely that they will produce incompatible mappings and fail the goal of enabling interoperability. (...) The second issue is the lack of technical specifications in global ontologies such as the CIDOC CRM. Without any detailed instructions of how to implement the mappings, represent instances, and process data during run-time, it is likely that each institution applies its own interpretation on a standardised global ontology. This again causes heterogeneities in scenarios that initially have aimed at providing interoperability." (Nußbaumer and Haslhofer 2007a)

Option to implement CIDOC-CRM in simple ways

It is hoped for, that based on the available documentation, use cases and know-how transfer, an increasing number of projects large and small will be able to implement the CIDOC-CRM in some way or other.
For example, there is the option to use CIDOC-CRM in very simple ways, while at the same time allowing for future interoperability with more complex implementations. One such example is the Museo24, a semantic virtual museum that is described in more detail in section 7.7.4.

Some projects may aim to high

Some projects may aim to high and underestimate the required sustained efforts to build a semantic portal based on CIDOC-CRM and a multitude of metadata standards. One example may be the Cantabria Cultural Heritage initiative, that aims to build such a portal for the Cantabria region in Northern Spain. A major problem with such projects is that some companies (such as in this case iSOCO) have acquired a lot of expertise in ontology development, Semantic Web languages and technologies, but often there is a mismatch with the existing capability, resources and skills of the heritage organisations that are involved "on the ground" (see section 7.7.3).

7.7.2 STAR – Semantic Technologies for Archaeological Resources (SKOS and CIDOC-CRM in RDFS)

Project brief / context

STAR – Semantic Technologies for Archaeological Resources (01/2007-12/2010) is a research and technical development project led by the University of Glamorgan's Hypermedia Research Unit. The project work is carried out in collaboration with English Heritage and the Royal School of Library and Information Science (Denmark), and funded by the UK Arts & Humanities Research Council (AHRC).
Some parts of the work build on results of the Hypermedia Research Unit's research tasks in the Knowledge Extraction and Semantic Interoperability cluster of the DELOS Network of Excellence (EU FP6-IST).

Precursor projects

STAR also particularly draws on tools developed in the UK-based precursor project FACET, a collaborative research project of the Hypermedia Research Unit with the Science Museum, MDA and J. Paul Getty Trust, funded by the Engineering and Physical Sciences Research Council (EPSRC).
The focus of this project was on automatic expansion of thesaurus-based, faceted search queries, integrating measures of semantic closeness/distance into the matching function. The main thesaurus used in this project was the Getty Art and Architecture Thesaurus. (Detailed descriptions of FACET are given in Binding and Tudhope 2004, and Tudhope et al. 2006; FACET website and web demonstrator, http://www.comp.glam.ac.uk/~FACET.)

General aim of STAR

The general aim of the STAR project is to investigate the potential of semantic terminology tools for improving access to digital archaeology resources, including disparate datasets and associated grey literature (the overview below in particular draws on Tudhope, Binding and May 2008).
An immediate goal is to use a domain-specific extension of the CIDOC-CRM as an overarching common schema to which different archaeological datasets may be mapped, where the datasets are indexed by domain thesauri and other controlled vocabularies.
The CIDOC-CRM specialisation and extension, called CRM-EH, has been created by English Heritage's Centre for Archaeology to reflect the processes and events involved in archaeological excavation and analysis. (Cripps, P. et al. 2004) In the STAR project, the most elaborated part of the CRM-EH, which focuses on environmental archaeology (May 2006), has been produced as a modular RDF extension referencing the published (v4.2) RDFS implementation of the CIDOC-CRM.

SKOS web services

STAR has developed a pilot set of SKOS web services (written in C# and running on Microsoft .NET framework) that builds on a subset of the SWAD Europe SKOS API, with extensions for concept expansion. The pilot set of the web services provides facilities for term look up in vocabularies, browsing and semantic concept expansion. Queries may be expanded by synonyms or by semantically related concepts.
A more detailed technical description of the services is to be found in section 7.8.2.

Domain KOS used

The current services operate on a MySQL Triplestore database backend comprising six separate thesauri that have been converted to SKOS format:
These include four English Heritage thesauri – Archaeological Sciences Thesaurus; Evidence Thesaurus; Building Materials Thesaurus; Monument Type Thesaurus – provided by the English Heritage National Monuments Record Centre in CSV format files.
Furthermore, the MDA Archaeological Objects Thesaurus and the Alexandria Digital Library's Feature Type Thesaurus (comprising terms used to categorise geographic places/features) are included.
Moreover, 27 glossaries were created from archaeological recording manuals in SKOS format (using MultiTesXSL transformation).

Collection databases

The content used in the STAR project includes datasets and "grey literature" from the Roman (and some Iron Age) field work reports of the English Heritage Raunds Project and other UK excavations.
The datasets comprise the Raunds Roman Analytical Database (RRAD), the Raunds Prehistoric Database (RPRE), and the York Archaeological Trust's Integrated Archaeological Database (IADB).
The STAR work for including "grey literature" involved information extraction based on Named Entity Recognition rules, supported by thesauri and flat gazetteer lists. Extracted and annotated terms were connected to thesaurus concepts and ontology classes of the CRM-EH. (Binding, Tudhope and Vlachidis 2008 provide some more details on this work)

Data extraction, mapping and aggregation

Data extraction from the three archaeological databases focused on selected key data concerning contexts (i.e. artefacts such as a wall or pit) and their associated finds. The approach was to extract modular parts of the larger data models of the databases via

SQL queries, and to store the data retrieved in a series of RDF files.
The utility used in the data extraction and mapping process consists of a form allowing the user to build up the SQL query incorporating selectable URIs that represent specific RDF entity and property types (including CIDOC-CRM, CRM-EH, SKOS, Dublin Core and others). The output is a RDF format file, with query parameters saved in XML format for subsequent reuse.
For the RDF metadata creation, also an URI format needed to be defined. The solution was a simple dot delimited notation which (although verbose) allowed the use of existing ID values of the database records without introducing ambiguities. In addition, date/time and spatial location formats were defined.
The RDF files of the mappings, the CIDOC-CRM (with alternative language labels), the CRM-EH extension, and the SKOSified English Heritage thesauri were aggregated and combined into a single SQLITE database, using the SemWeb/RDF library for.NET.
The database of aggregated data was 193MB overall and consisted of 268,947 RDF entities, 168,886 RDF literals and 796,227 RDF statements (triples). The SemWeb library supports SPARQL querying against the database, however, the SQLITE database itself also supports direct SQL queries.

Work in progress

The STAR research on how to most appropriately connect the thesauri expressed in SKOS to the database items and to the CIDOC-CRM / CRM-EH is work in progress. Currently, the linking of SKOS concepts and information items is modeled by a project specific is represented by relationship (which is the most flexible option).
With regard to the integration of the English heritage thesauri with the CIDOC-CRM / CRM-EH, it was found that they may not fit neatly under these ontologies. Therefore, it was suggested that the appropriate connection may be a loose SKOS mapping (broader) relationship between groups of concepts rather than complete thesaurus hierarchies. The current approach seems to be a mapping of the data items (instances) to the ontology where the data items are indexed with thesaurus concepts.

Tools used in the STAR project

XML, RDF, SKOS and OWL and related tools used in the STAR project include:

- Altova XMLSpy, http://www.altova.com/products/xmlspy/xml_editor.html
- Drive RDF parser: C# RDF Parser, provides API to parse RDF/XML into an in-memory RDF graph for manipulation; fully compatible with the .NET platform.
- Altova SemanticWorks: Visual RDF and OWL editor, http://www.altova.com/products/semanticworks/semantic_web_rdf_owl_editor.html
- Protégé: Open source ontology editor and knowledge-base framework, http://protege.stanford.edu
- Semantic Web/RDF Library for C#/.NET, http://razor.occams.info/code/semweb
- W3C RDF validation service, http://www.w3.org/RDF/Validator
- W3C SKOS validation service, http://www.w3.org/2004/02/skos/validation

References

The above overview draws on Tudhope, Binding and May 2008, and information on the STAR project website; in addition, the following publications were useful: Binding, Tudhope and Vlachidis 2008 (a very detailed presentation of the STAR project); AHRC ICT Methods Network 2008 (in particular discusses the relevance of the STAR project results for archaeological data integration).

Website

STAR project, http://hypermedia.research.glam.ac.uk/kos/star
STAR Semantic Services, http://hypermedia.research.glam.ac.uk/kos/terminology_services

7.7.3 Cantabria cultural heritage ontology (CIDOC-CRM in RDFS and FRBRoo)

Project brief / context

The Cantabria Cultural Heritage initiative aims to build a semantic portal for cultural heritage of the Cantabria region in Northern Spain, incorporating sources from excavations of prehistoric sites to industrial partrimony. The project is funded and led by the

Marcelino Botín Foundation and carried out together with the University of Cantabria, the semantic technology provider iSOCO, and domain experts from heritage organisations. iSOCO has been a partner in several major EU funded research projects dealing with Semantic Web technologies and ontologies.

Implementation

A semantic portal and search engine under development should draw on the CIDOC-CRM (v4.2 version in RDFS) and FRBRoo (the object-oriented version of the Functional Requirements for Bibliographic Records), as well as a multitude of metadata standards and protocols in use with the Cantabria cultural heritage sources, e.g. Dublin Core, MARC21, Encoded Archival Description (EAD), EAG (Encoded Archival Guide) and others.

References

Hernández 2007; Hernández et al. 2007; Hernández et al. 2008

Websites

Whereas the "Cantabria Cultural Heritage Ontology" was presented on several occasions (see above), no publicly accessible demonstrator website has been launched so far. However, the initial project will run until 2009, so the next year may see the launch of a leading edge cultural heritage semantic portal.
Marcelino Botín Foundation, www.fundacionmbotin.org
iSOCO, http://www.isoco.com/innovacion_web_semantica_d.htm
Proyecto Ontología del Patrimonio de Cantabria, 2006/2009, http://80.34.0.78:8080/c/portal/expire_session

7.7.4 Museo24 – semantic virtual museum (a little CIDOC-CRM in OWL)

Project brief / context

Museo24 is a virtual museum that aims to present the socio-cultural memory of the Jämsä region in central Finland.
The project has been initiated by regional heritage associations and supported by the towns of Jämsa and Jämsänkoski, University of Jyväskylä and UPM (the international forest products group). Some funding for the project also was provided by the European Regional Development Fund (ERDF).
The technical implementation was carried out 2004-2005 mainly by the Finnish-Czech Web applications development company ARTIO.

Implementation

The virtual museum features historic buildings, work places (forestry and aircraft industry), local news and stories, historic maps, photographs, etc. A rather simple ontology was implemented in OWL using the CIDOC-CRM "is about" property with four subproperties for "who", "what", "where" and "when". Moreover, building on the approach suggested by Jane Hunter (2002), this was extended with a MPEG-7 class hierarchy to support semantic annotation of multimedia objects.
Specific applications that have been developed in the project include a simple semantic annotation tool, an AJAX based LiveSearch tool and a semantic timeline. However, the current Museo24 website shows a rather classic presentational approach of themes with texts and images, that in the backbone are organised in hierarchical (semantic) folders. The only more appealing application is the interactive timeline for use with Adobe Flash Player.

References

Heikka, Juhani et al. 2006, Szász, Barnabás et al. 2006

Websites

http://www.museo24.fi
http://www.artio.net/en/projects/museum24

7.8 Selected tools and services

The following sections provide details on selected tools and services that have been developed and used in some of the projects described above. The intention is clearly not to

provide representative examples of all types of tools that need to be used in such projects.
The sections 7.2 and 7.7.2 above include lists of tools that are employed in the STERNA and STAR projects, however, also these lists are not intended to be complete.
With regard to certain types of tools, overviews on the following may be of particular interest:

- semantic annotation (Maynard et al. 2007 present a benchmarking of many such tools);
- ontology creation and management (Denny 2004 provides a detailed survey), and
- matching of SKOS or ontologies (http://www.ontologymatching.org provides a current overview of tools with links).

7.8.1 AnnoCultor – a library of metadata/vocabulary conversion operations

Brief description

AnnoCultor has been developed in the MultimediaN E-Culture project (see section 7.3) to assist conversion of legacy datasets and vocabularies represented in XML or databases to RDF. AnnoCultor has been identified as an interesting candidate for inclusion in a portfolio of Europeana/EDL tools. (cf. Cousins and Siebinga 2008)
AnnoCultor is an open source library that provides a set of conversion operations that

- convert XML trees to (linked) RDF objects,
- filter the objects to be converted,
- treat part-of structures,
- rename object properties,
- affix property values,
- interpret values with regular expressions,
- look up property values in external vocabularies (with disambiguation),
- allow faceted property conversion, and
- development of own conversion rules.

AnnoCultor has been developed as a generic Java-based platform (based on Java 5) that is available from SourceForge under a GNU General Public License (GPL).
The platform provides programming infrastructure and basic conversion rules that are open to incorporate custom rules. It also is open to other systems, such as GATE, for instance.

Examples of use

AnnoCulture has been used in the E-Culture to convert the metadata of several Dutch art and ethnographic collections to RDF, using VRA Core as target common metadata scheme. It has also been used to (fully or partially) convert large Getty vocabularies (AAT, TGN and ULAN). These vocabularies have been converted from the Getty XML files into RDF; there is a zip-file available that also includes the RDF schemas used for the Getty vocabularies.

Websites

http://sourceforge.net/projects/annocultor
http://borys.name/tools.html#annoCultor

7.8.2 STAR semantic terminology services

Brief description

The STAR semantic terminology services (SKOS_WS) have been developed in the STAR - Semantic Technologies for Archaeological Resources project (see section 7.7.2), also building on earlier and related efforts in the FACET and DELOS projects. For example, an earlier version of the current services was integrated with the DELOS prototype Digital Library Management System.
SKOS_WS provides SOAP-based web services for vocabularies represented in SKOS Core vocabulary. The services are written in C#, running on Microsoft .NET framework (version v2.0.50727), and are based on a subset of the SWAD Europe SKOS API, with extensions for concept expansion.

The services currently consist of 7 function calls, which may be integrated into a textual or metadata based search system. The services provide term look up in vocabularies known to the system, along with browsing and semantic concept expansion.
In combination with a search system, the services allow queries to be expanded (automatically or interactively) by synonyms or by expansion over the SKOS semantic relationships. Expansion is based on a measure of "semantic closeness".
A detailed description of the available API and function calls is provided on the STAR website, where also a client demonstrator can be downloaded. This client is specifically configured to operate with a subset of English Heritage thesauri, but compatible with any thesaurus (or other KOS) expressed in SKOS.

Website http://hypermedia.research.glam.ac.uk/kos/terminology_services/

7.8.3 ONKI-SKOS web server

Brief description The ONKI SKOS web server is intended to provide "out of the box" support for publishing and utilising SKOS vocabularies and lightweight concept ontologies in RDFS/OWL format. Web applications that make use of the server functionalities do not need to implement application specific user interfaces for end users.
Using ONKI SKOS, a SKOS vocabulary can be published and used in applications cost-efficiently with little extra work as AJAX mash-up and Web Service support are provided. ONKI-SKOS allows to browse, search and visualise any vocabulary conforming to the SKOS specification and also RDFS/OWL ontologies. It also supports simple reasoning, e.g. transitive closure over class and part-of hierarchies.
ONKI SKOS has been piloted using various KOS and ontologies, e.g., Medical Subject Headings, Iconclass and the General Finnish Upper Ontology.

Examples of use At present, ONKI-SKOS is mainly used with applications developed in the Finnish National Semantic Web Ontology and Ontology Service Infrastructure projects (FinnONTO). For more details see the references below.

References Tuominen et al. 2008; Hyvönen, Eero et al. 2008

Website http://www.seco.tkk.fi/services/onkiskos/

7.8.4 ClioPatria – semantic search web server

Brief description On the MultimediaN E-Culture website ClioPatria is described as follows:
"ClioPatria is a SWI-Prolog based platform for Semantic Web Applications. It joins the SWI-Prolog RDF and HTTP infrastructure with a SeRQL/SPARQL query engine, interfacing to the Yahoo! User Interface Library (YUI) and libraries that support semantic search.
The platform combines a high performance in-core RDF store with flexible reasoning in Prolog, query optimization. Prolog's interactive usage and capabilities of recompiling modified source code while the system remains alive greatly speed up development.
Key figures: Up to about 25 million RDF triples on 32-bit hardware, only limited by memory on 64-bit hardware. Exploits multi-CPU and multi-core hardware to answer requests over HTTP concurrently. Runs on Windows, MacOS X, Linux and most Unix flavours, supporting both 32-bit and 64-bit operating systems. 64-bit systems are recommended for servers with lots of data or many users."
Currently the software is only made available through a Source Code Management system. The repositories are stored in GIT (a software version control environment) and can also be accessed through anonymous CVS (Concurrent Versions System) addresses to view history, files and download snapshots.
An early release of the software was made in October 2007 under the GPL-2 license in order to promote and simplify cooperation.

Examples of use ClioPatria has been developed in efforts related to the MultimediaN E-Culture project (see section 7.3), where it was first used to implement thesaurus-based searching across heteregenuous cultural heritage collections.
Other projects reported to have used ClioPatria are:

- CATCH CHIP: to power the search engine underlying their Rijksmuseum art recommender and personalised museum tour guide;
- DBtune: to create semantic mashups of music-related information;
- K-Space Network of Excellence: to access semantically annotated news-related articles and photographs.

ClioPatria also has been identified as an interesting candidate for inclusion in a portfolio of Europeana/EDL tools. (cf. Cousins and Siebinga 2008)

Website http://e-culture.multimedian.nl/software/ClioPatria.shtml

7.8.5 /facet browser

Brief description /facet has been developed within the Dutch MultimediaN E-Culture project (see section 7.3) and also received support under the K-Space Network of Excellence contract.
/facet is a generic browser for heterogeneous semantic web repositories, that works on any RDFS dataset without any additional configuration.
Some unique features of /facet are described as follows:
"Select and navigate facets of resources of any type. Facets are associated to each type. The type facet is used to navigate the hierarchy, typically organized by rdfs:subclassOf, and a selection in this facet automatically selects which other facets are also active.
Make selections based on properties of other, semantically related, types. For example, select a set of artworks based on the properties (facets) of their creators.
Semantic autocompletion in three flavours:
1) search on all instances, helping to select the right type,
2) search within a single facet, helping to move in complex facet hierarchies,
3) search across all active facets, showing the user the different uses of a keyword in different facets.
/facet allows the inclusion of facet-specific display options. We have developed a timeline plug-in to visualize time-related facets. Geographical information can be displayed on yahoo maps."

References Hildebrand, van Ossenbruggen, Hardman 2006

Website http://slashfacet.semanticweb.org

Part B:
Natural history and biodiversity resources for the European Digital Library initiative

Part B Natural history and biodiversity resources for the European Digital Library initiative

Part B (chapters 8–13) presents digital environments natural science and history organisations and practitioners use to create, manage and share information resources. In particular, the focus is on novel technological approaches, tools and information services that may be of interest to the European Digital Library initiative.

Issues and progress in the digitisation of natural history and biodiversity resources

Chapter 8 provides an overview of issues and progress in the digitisation of natural history and biodiversity resources:
It is noted that in the digitisation of natural science and history objects (e.g. physical specimen) mass digitisation methods such as used by libraries for printed material cannot be applied. Therefore, only a slow growth in digital representations of such objects (e.g. images or 3D models) can be expected.
On the other hand, there has been considerable progress in the digitisation of specimen labels and taxonomic literature, in particular, with respect to information extraction and metadata creation. A specific focus is on the extraction of taxa (i.e. the scientific names designating an organism or groups of organisms) for which Taxonomic Name Recognition and other sophisticated techniques are used.

Taxonomic databases and services

Chapter 9 focuses on taxonomic databases and services:
Taxonomic databases play an important role as they record the scientific names, synonymy, classification, geographic distribution and relationships of organisms. Such databases also are understood to help overcome the so called "taxonomic impediment", the lack of taxonomic information and practical capacity particularly in the developing countries.
Selected highlights are the Catalogue of Life project of the Species 2000 programme which aims to compile and make openly accessible a single unified and validated index of all the world's known species. Furthermore the role of "taxonomic intelligence" services such as uBio in leveraging access to a variety of information resources is emphasised.

Online collaboration tools

Chapter 10 presents selected online collaboration tools for taxonomic and other biological studies:
Such tools allow individuals and communities of practice to create, manage, and share study results. One major application field is work on taxonomies of groups of organisms with the aim of revising and consolidating them. A leading project in this area is Creating a Taxonomic e-Science (CATE).
Other important collaboration environments are Scratchpads, which have been developed as part of the EDIT cybertaxonomy platform, and LifeDesks, which provide a collaborative component to the Encyclopedia of Life project. Furthermore, the state-of-the-art Web repository of scientific images Morphbank is presented.

The Encyclopedia of Life – lessons learned in large-scale content aggregation and access

Chapter 11 presents the Encyclopedia of Life (EOL) as an example of a large-scale programme of content aggregation and access:
The EOL aims to create within 10 years a webpage for each of the estimated 1.8 million known species on Earth. These webpages are intended to provide the entry point to a vast array of knowledge and high-quality content for a wide audience that includes scientists, natural resources managers, conservationists, teachers and students around the world.
The EOL may provide some lessons for other large-scale initiatives such as the European Digital Library. The EDL uses different technologies but may face similar problems with respect to the expected richness of content.

Life Science Identifiers

Chapter 12 addresses Life Science Identifiers (LSIDs) which are increasingly used in the fields of natural history and biodiversity to provide globally unique resource identifiers. The Taxonomic Database Working Group (TDWG), the international biodiversity data standards setting group, has adopted LSIDs as recommended standard for such identifiers and suggests to provide related metadata in RDF format. This will greatly help pave the way of natural history and biodiversity resources to the Semantic Web. In particular, the TDWG also has developed LSID metadata vocabularies that formally describe the metadata that should be provided for particular classes of information objects.
There is a growing number of LSID implementations, for example, the Biodiversity Collections Index, Catalogue of Life, Global Biodiversity Information Facility, Index Fungorum, International Plant Names Index, Morphbank, ZooBank and many others have implemented LSID based information services.

Semantic Web ontologies

Chapter 13 focuses on Semantic Web ontologies for natural history and biodiversity domains.
The ontological layer of the Semantic Web plays a key role for knowledge representation, data integration and advanced search and other services spanning databases of distributed information providers. The realisation of such a layer requires the development and implemention of domain and upper-level ontologies.
In the fields of natural history and biodiversity there are already some projects that have used the Web Ontology Language (OWL) to develop such ontologies and implement prototypic applications. Most notably, the TDWG Architecture Group is developing a biodiversity core ontology, which is intended to semantically integrate the TDWG LSID metatdata vocabularies mentioned above.
Furthermore, a selection of other ontologies is presented that have been developed by research projects. The selection illustrates the wide range of ontologies that have been created as well as some prototypic applications.

8 Digitisation and enrichment of natural history resources

8.1 General aspects, requirements and funding of digitisation of natural history resources

Lack of overview

In comparison to the cultural heritage sector, where the MINERVA projects since 2002 have worked on promoting and co-ordinating digitisation initiatives of EU Member States (see section 2.1), there is no comprehensive overview available of already completed and ongoing digitisation of natural history and biodiversity resources in Europe. Though, a number of digitisation projects and approaches are described in the ENBI and GBIF manuals for the digitisation of natural history collections (ENBI / Häuser et al. 2005; GBIF 2008b; see section 8.4)

Below we attempt to provide an overview of issues and progress in the digitisation of such resources. Firstly, we explain why in comparison to the library sector museums in general show a slow progress in the digitisation of resources. Secondly, we address the digitisation of different natural history resources, such as observation records, specimen, taxonomic literature and databases, and note the sections in this report where some projects are described in more detail.

Slow progress in the digitisation of museum artefacts

In recent years there has been considerable progress with respect to mass digitisation of cultural heritage holdings on the national level, particularly regarding collections of libraries and archives of visual media. In comparison, archaeological, historic and other museums that hold physical artefacts have seen a rather slow progress in digitisation. Consequently, it is understood that the European Digital Library initiative will for some time to come mainly build on digitised resources such as books, manuscripts, historic photographs and some other visual media.

Uniqueness of museum artefacts

One reason for this is that a large part of museum artefacts are unique for being individual pieces and the context from which they come, e.g. objects found in archaeological excavations. This is a major difference to the library sector, where copies of books, series, journal issues, etc. are usually kept in several public, university or research libraries. This also the case with incunabla and other rare printings, though, not with unique historical manuscripts. Hence, there is the possibility and, indeed, need to coordinate digitisation activities in order to prevent the digitisation of the same resources (e.g. series or journals) in different places while others may be neglected.

Specific digitisation requirements

A second major reason for the slow growth in the digitisation of museum artefacts is that handling the objects in the digitisation process is much more complicated than with printed material, and there is a need for specialised digitisation equipment and skills. In particular this concerns digitisation of museum artefacts in 3D formats (Arnold and Geser 2008). Also in the field of natural history we would expect a rather slow growth in 3D digitisation of specimens and other physical objects (some examples are included in the following section).

Unclear impact of funding of natural history digitisation initiatives

In the last 20 years or so there has been made available on the national and European levels a considerable amount of funding for the digitisation of cultural and scientific heritage resources. With regard to cultural heritage resources, the MINERVA reports provide a good impression of the impact of this funding. However, in an extensive online search we could not find an overview of the use and impact of funding made available for natural history digitisation initiatives.

National and European funding sources

To create such an overview would indeed be difficult: generally, most of the funding will have come from national funding agencies, requiring to collect and analyse reports of these agencies from across Europe. On the European level, we found that projects of natural history organisations were not funded under the eContent programme (2001-2005).

Under the current *e*Content*plus* programme (2005-2008), which focuses on the enrichment of, and access to, existing digital content, there are two examples of funded projects, AquaRing (see section 7.6) and STERNA. Some relevant digitisation and enrichment work may also have been carried out in the context of projects that have been funded under the EU Framework Programmes for research and technological development.

GBIF DIGIT programme

A more widely known example of global funding is the Global Biodiversity Information Facility (GBIF) Seed Money programme. This programme was started in 2003 and comprises two components:

- GBIF DIGIT (Digitisation of Natural History Collections) supports digitisation of information associated with specimens in natural history collections as well creation of species level observational databases;
- GBIF ECAT (Electronic Catalogue of Names of Known Organisms) supports efforts in increasing access to authoritative taxonomic checklists, nomenclatural data, and other useful names list compilations, for example, for regional, invasive or endangered species.

Volume of funding

In the global context, the resources that GBIF has available to fund digitisation activities are very limited. However, it has always been recognized that the vast majority of funds would have to come from national and/or regional funding sources. Accordingly, GBIF has funded only a smaller part of the project costs (e.g. 20% in 2004, 30% in 2005-06) Until March 2007, GBIF has provided nearly USD 4 million in seed-money awards. However, the latest request for proposals for 2007/2008 only had a total amount of funding of € 350,000 available for both DIGIT and ECAT (with a max. size of awards of € 50,000). (GBIF 2007)

8.2 Issues and progress in the digitisation of natural history resources

Wide range of natural history knowledge and content

The field of natural history comprises a wide range of knowledge and content resources including nomenclature data, taxonomies and phylo-genies, specimen collections, field observations, ecological data sets (e.g. species distribution maps), diagnostic keys and character data, molecular sequence data, databases of scientific literature, images and audiovisual content.

Natural history observation records

Natural history observation records document the observation and collecting of an organism from the field, which is then preserved in a curated collection. The record of the observation event serves as primary reference point of the collected object, which may be prepared in various ways (e.g. skin, skeleton, microscope slides, etc.).

There is already a huge amount of such records available in digital form. For example, the United Kingdom currently serves nearly 15,000,000 data records through the Global Biodiversity Information Facility (GBIF) network of which 14,761,000 are records from observational initiatives.
In comparison, the number of served records of prepared specimen is only 174,000. To put this figure in perspective, the Natural History Museum in London alone holds some 70,000,000 specimens. (cf. GBIF 2008)

Museum specimen collections

Knowledge of species is largely based on the collections of the worldwide about 6500 natural history museums that are estimated to hold between 1.3 and 3 billion specimens. Hence, the building of digital collections of specimens is a huge task.
It is estimated that worldwide below 5% of specimen collection records (i.e. not the specimen themselves) have been digitised so far. (Some information on the volume in 2002 of digital specimen records in several countries is given in Chavan and Krishnan 2003).
The Biodiversity Collections Index (BCI) project aims to build a central index to specimen reference collections (see section 12.3).

Primary type specimen

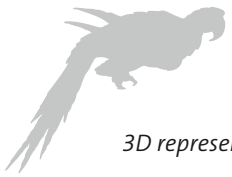
The "unique pieces" principle mentioned in section 8.1 above also applies to the primary type specimens of natural history museums. A type specimen serves as the scientific name-bearing representative of an animal or plant species, providing the objective standard of reference for the identification and naming of the species.
Collections of such specimens are a prime target for digitisation efforts. There already exist a number of so called "e-type" collections, for example, the Linnean types of the Swedish Museum of Natural History and the type specimen image collection of the Herbarium Berolinense. (Speers 2005)
However, even digitising type specimen collections is a huge task, including not only the rather difficult process of digitising specimen labels and records, but also photographing or 3D imaging of the specimens.

Limited availability of digital representations of specimen

This explains why a large part of available digital information are observation records while digital representations of specimen are comparatively rare. The following are but two examples of museums with bird collections that can illustrate the progress in digitising natural history collections:

Images of type specimen

The Smithsonian National Museum of Natural History, with nearly 125 million specimens, around 2003 had 9 million specimens cataloged online (up from over 3.7 million in 2001), however only relatively view images of specimens are available.
At present, the birds collection of the Smithonian seems to be rather well documented online. It is the third largest bird collection in the world with over 625,000 specimens, comprising representatives of about 80% of the approximately 9600 known species in the world's avifauna. It also holds nearly 4000 primary type specimens (i.e. specimen that serve as the scientific name-bearing representative of a species). Approximately 65% of the bird collection is completely or partially listed in the Smithonian specimen database that is accessable on the Web, however, only a small number of specimen records have images associated with them. (Smithonian 2008)
The STERNA project partner Royal Museum for Central Africa (RMCA) has a unique African bird specimen collection of about 145,000 specimens (as flat skins, specimens in alcohol, mounted specimens and skeletons), including 987 type specimens of 226 nominal species. For the latter 500 high-quality digital images are available that have been captured and documented with financial support of the Federal Belgian Science Policy Office. Of the whole specimen collection records 78% are available in digital form, 73% are geo-referenced. (Mergen et al. 2008)



3D representations

Due to lack in funding, specialised equipment and expertise, there is a rather slow growth in 3D digitisation of natural history specimens and other objects. Though, there are some high-profile showcase projects such as the Digital Morphology database (http://digimorph.org) of the University of Texas at Austin that has been funded under the US NSF Digital Libraries Initiative.
With regard to the focus of STERNA, 3D imaging of bird type specimen for example has been carried out by the ETI BioInformatics centre for the Zoological Museum of the University of Amsterdam. (Veldhuijzen van Zanten et al. 2005; http://ip30.eti.uva.nl/zma3d/)
In general, there is a need to broaden the expertise base in acquisition and management of 3D objects. One example of a training initiative that includes such tasks is the Marie Curie research training network European Virtual Anthropology Network (EVAN; 2006-2009; http://evan.at).

Taxonomic literature

The cited half-life of publications in the field of taxonomy is longer than in any other scientific discipline. Indeed, taxonomists regularly need to use old and new publications. This can be a costly, time-consuming process because older publications are often only available as hard copies in a few libraries.
It is estimated that there are over 5.4 million volumes on biodiversity dating back to 1469 comprising some 800,000 monographs and 40,000 journal titles. Fifty percent were published before 1923 and are in the public domain in the United States. (Gwinn

and Rinaldo 2008) A major international effort in digitising taxonomic literature is the Biodiversity Heritage Library project, which is described in section 8.3.2 below.
Of core interest to taxonomists are publications that contain taxonomic treatments, i.e. systematic species descriptions. The volume of such treatments is estimated at 100+ million pages of scientific literature. The Plazi.org project focuses on the extraction of taxonomic treatments from digital and born-digital literature (see section 8.3.4).

Taxonomic databases

The existing and increasingly integrated global species databases presently account for some 60% of the estimated total known species. There is a growing volume of digital information in taxonomic databases such as, for example, Index Fungorum, International Plant Names Index (IPNI) and Integrated Taxonomic Information System (ITIS) (see section 12.3). Such databases record the scientific names, synonymy, classification, geographic distribution and relationships of biological organisms.
Taxonomic databases allow for leveraging access to authoritative taxonomic checklists, nomenclatural data, and useful names list compilations, for examples, for regional, invasive or endangered species.
Taxonomy databases also increasingly provide a systematic backbone for websites that are meant for use by non-professional users such as the Encyclopedia of Life (see chapter 11), Animal Diversity Web (see section 13.2.6) and many others. The Encyclopedia of Life, for example, wants to create a webpage for every known species on Earth that provides the entry point to a vast array of knowledge (e.g. geographic distribution, evolutionary history, behavior, ecological relationships, etc.)

8.3 Digitisation of specimen labels and taxonomic literature

Much progress in (semi-) automatic techniques

In the field of natural history museums and libraries there has been much progress recently with (semi-)automatic techniques that are able to extract information from digitised textual resources (e.g. specimen labels and taxonomic literature) and to create metadata for such resources. This considerably reduces the cost of information extraction and metadata creation.

Potential for know-how and technology transfer

Similar techniques also are developed for textual cultural heritage resources. Innovative research and development work, for example, is carried out in the IMPACT (Improving Access to Text) project, a 4-year FP7-ICT project (01/2008-11/2011) coordinated by the National Library of the Netherlands (http://www.impact-project.eu).
In this project advanced OCR, named entities recognition and lexicon building technologies are developed to allow for enrichment of digitised cultural heritage resources. IMPACT makes use of identification, extraction and classification of named entities (incl. variants), lexicon building from historical dictionaries and historical texts, and deployment of lexicon content in enrichment (e.g. for dealing with historical spelling variation).
The potential for know-how and technology transfer between the application fields of cultural heritage and natural history should be examined.

8.3.1 HERBIS, digitisation of specimen labels

Digitisation of specimen labels

Digitisation of specimen collection information is rather difficult because for a large part this requires digitisation of legacy specimen labels and migration of intricate legacy metadata (e.g. catalogue entries of the scientific names of species, the location and date of specimen collection, habitat information, etc.).
A specimen label can be digitised and stored in an image format, however, more useful would be to automatically capture and process the textual information of the label, thereby reducing much human labor.
A key role in the digitisation of specimen labels therefore plays Optical Character Recognition (OCR) technology. However, automated extraction of information from specimen

labels is difficult because of the high variability of museum label formats and a high degree of OCR errors with such labels.

The HERBIS "one button" specimen imaging and data capture system

In the HERBIS (Erudite Recorded Botanical Information Synthesizer) project researchers from the University of Illinois at Urbana-Champaign in collaboration with the Peabody Museum of Natural History at Yale University have developed a "one-button" specimen imaging and data capture system for herbarium specimen. Specific challenges in the development of the system have been: 1) rapid image capture, 3) image to text conversion of label data, 4) text markup into data elements to simplify database loads, 4) georeferencing, and 5) web services development.

With the HERBIS system, clicking the shutter on a digital camera initiates a sequence of processes that culminates with the population of label data and a specimen image into a collection database. Museums anywhere in the world can create digital images of specimen labels on their site and transfer them to the Yale Peabody Museum OCR processing unit where the label is detected and converted to a string sequence.

This text packet then is passed to the HERBIS Learning System at the University of Illinois at Urbana-Champaign through a web services connection. The text is converted to an XML document with appropriate element information and returned to the museum that sent the digitised specimen label.

The system uses Hidden Markov and Naïve Bayes models, data cleaning procedures, field element identifiers, and special learning mechanisms to automatically extract Darwin Core and other metadata from the specimen labels.

References Heidorn and Wei 2008a and 2008b

Website http://www.herbis.org

8.3.2 Biodiversity Heritage Library

Making accessible the rich legacy of taxonomic literature

The Biodiversity Heritage Library (BHL) is a partnership of ten UK and USA based natural history museum libraries, botanical libraries, and research institutions that joined forces in 2005 to digitise and make accessible on the Web the taxonomic literature held in their respective collections. All digitised publications will be openly accessible to the public, unless they are copyrighted. Headquartered at the Smithsonian Institution Libraries, the BHL also is one of the cornerstones of the Encyclopedia of Life initiative (see chapter 11).

In October 2008, the BHL portal already provided access to 8.4 million images of digitised pages (more than 20,000 volumes) and the text of the literature, captured with Optical Character Recognition (OCR) technology. The publications are scanned by facilities of the Internet Archive and partner institutions.

The BHL portal ingests low-resolution JPEG files and available bibliographic data encoded in MARCXML which is used to provide search and browse capability. High resolution JPEG2000 files are retrieved on the fly from the Internet Archive when requested by a user and decoded at the portal for viewing via a web-browser. The underlying architecture of the BHL is a .Net application environment which, however, is planned to be moved to the open source Fedora Commons architecture.

uBio "taxonomic intelligence" services

One aspect of the Biodiversity Heritage Library (BHL) project that distinguishes it from other mass digitisation projects is the use of "taxonomic intelligence" to identify scientific name strings in the digitised content and to provide names-based interfaces into the taxonomic literature.

The BHL system employs the "taxonomic intelligence" (species name finding) services that have been developed in the uBio (Universal Biological Indexer and Organizer) project (see section 9.4). The OCR text of the digitised literature is sent to uBio to identify and extract likely scientific names (text strings that match the characteristics of Latin binomials) which are displayed in real time with the page image.

To identify the names of species, the uBio TaxonFinder (a named entity recognition application) compares the OCR text with uBio's NameBank, which is a database of over 11 million name strings of recorded biological names and identifiers that link those names together.
As of 20 November 2007 more than 6.8 million potentially relevant name strings were identified throughout the BHL corpus, with more than 3.8 million matched to a corresponding NameBank identifier. Iterative processing of BHL texts both increases the number of name strings in NameBank and increases the accuracy of name string recognition.
The BHL applies Globally Unique Identifiers (GUIDs) for linking to other taxonomic information services; this linking can be done at the bibliographic record, volume, and page levels.

BHL end-user services

BHL's goal is to allow a user to search its collection of biodiversity literature using any form of an organism's name, i.e. scientific, common or vernacular names. This also will allow non-professional users searching the Encyclopedia of Life (see chapter 11) to draw in literature related to the species they are interested in. Furthermore, the Scratchpads project uses BHL content as part of their "Panels" feature (see section 10.2).

References

The overview above is based on Gwinn and Rinaldo 2008.
Several illustrative presentations of the BHL system are to be found at http://biodiversitylibrary.blogspot.com/search/label/Presentations

Website

http://www.biodiversitylibrary.org

8.3.3 INOTAXA – Integrated Open Taxonomic Access

Development of an online workspace

The INOTAXA project aims to create a Web-based workspace in which taxonomic descriptions, identification keys, specimen data, images and other resources can be accessed simultaneously according to user-defined needs. To realise such a workspace, INOTAXA builds on a distributed data model that makes use of a set of interoperable XML schemas, which allow for linking data of different types and from different sources.

Testbed: Biologia Centrali-Americana

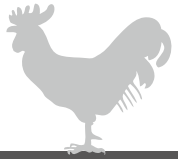
As a testbed the project focused on Mesoamerican biodiversity, drawing on a major literature resource, the out of print Biologia Centrali-Americana (BCA). The BCA consists of 63 volumes that describe a total of 50,263 species of which 19,263 were described in the BCA for the first time. The volumes also contain 1677 plates (of which more than 900 are coloured) depicting 18,587 subjects.

Development of an XML schema for taxonomic literature: taXMLit

In the first phase of the project, the BCA has been digitised and made accessible on the Internet. The project team also developed an XML schema for taxonomic literature, taXMLit, which is taken into account in the development of an TDWG standard (see section 8.3.5).

INOTAXA prototype

The INOTAXA system is in prototype form and has been tested by a panel from different taxonomic and other backgrounds. The system provides the following functionality for search and other purposes:
"INOTAXA allows extraction of parsed data on names, authors, places of publication, places of use; specimens cited, nomenclatural types, relationships (taxonomic, nomenclatural, phylogenetic and ecological) with other taxa, etc. Taxon names may be restricted to valid (accepted) names only, synonyms or, of course, all names may be returned. Specimen data, extracted from the literature according to user-set conditions, can be viewed and downloaded. In addition to fine-detail content treatments and keys can be retrieved, again according to more or less complex criteria and restricttions, according to user needs." (Lyal and Weitzman 2008b)



References Lyal and Weitzman 2008a and 2008b; Weitzman and Lyal 2004 and 2005

Websites Electronic BCA, http://www.sil.si.edu/digitalcollections/bca/
INOTAXA prototype, http://www.inotaxa.org

8.3.4 Plazi.org

Open access to species descriptions

Plazi is a Swiss-based non-profit organisation that advocates, promotes and supports the development of persistent and openly accessible taxonomic literature. The work programme of Plazi comprises to create and maintain a digital taxonomic repository that allows for archiving of taxonomic treatments (species descriptions), to enhance submitted treatments by creating TaxonX XML versions, and to participate in the development of new models for publishing species descriptions that maximise interoperability with other e-infrastructure components (e.g. taxonomic name servers, biodiversity databases, etc.)
Species descriptions are highly structured and rich in data, essentially a quality controlled summary of what is known at any specific time about a particular species. In best cases, this information includes a detailed morphological description, drawings and images, a summary on behavior and ecology and a detailed list of all the specimens studied. In more recent publications, links to DNA sequences, multimedia documentation and other forms of information are provided.

Technical approach

Plazi promotes open access to taxonomic literature by extracting and making available species descriptions that are not subject to copyright. XML documents of the descriptions are generated with the GoldenGate editor according to a taxonomic literature specific XML schema, called TaxonX (details on GoldenGate and TaxonX are provided in the section below). Plazi also enhances the descriptions with Life Science Identifiers (LSIDs) for taxonomic names, bibliographic citations, and if available for specimens, Consortium for the Barcode of Life (CBOL) sequences. All the descriptions are linked to the original publications and a proper citation is provided.

Ant species as application case

Plazi's application case are ant species of which it already holds some 4000 descriptions of more than 3,000 taxa, with a goal of archiving all forthcoming new descriptions. The ant species names are added to the Hymenoptera Name Server/antbase.org where all the known names are stored.
A longer-term goal is to archive all the descriptions of the known ant species listed in the server, enhanced with globally unique species numbers. On the 2nd of April 2008 antbase reported that there were 44,614 names associated with ants in the Hymenoptera Name Server, and 12,359 considered by the experts covering accepted species.
The Plazi.org website has been officially released on the 20th of February 2008 in London at the "IPR and the web: challenges for taxonomy" meeting of the European Distributed Institute of Taxonomy (EDIT).
The project has been partially funded through a binational US National Science Foundation – German Deutsche Forschungsgemeinschaft digital library grant, and more recently by the Global Biodiversity Information Facility (GBIF).

References Agosti 2008; Plazi.org 2008; Antbase 2008

Website http://plazi.org

8.3.5 XML schemas and editors for taxonomic literature

A highly standardised resource

Taxonomic and biosystematics literature typically has a highly standardised structure, in particular, the sections comprising the taxonomic treatment of species (including character X species data matrices, images and distribution records), tools for identification

(keys), and phylogenies. Therefore, taxonomic literature offers a unique chance for data extraction, database creation, and online access.

XML based approach

In the digitisation of legacy taxonomic literature for online access XML plays an important role. The digitised documents are marked up with XML for two purposes: Firstly, to preserve the original document structure as well as publication-related information like publisher, title, issue, etc.; and secondly, to facilitate deployment of standard query languages like XPath to search and retrieve information from literature databases.
In general, three basic information needs in biosystematics should be supported: taxonomic names, collection locations (i.e., where specimens of a particular taxon were collected), and concepts of morphological features.

XML schema development

A variety of XML schemas have been suggested for encoding the information, notably ABCD (Access to Biological Collections Data), SDD (Structure of Descriptive Data), TaxonX and taXMLit. The current work of the TDWG Taxonomic Literature Interest Group on a common standard focuses on TaxonX and taXMLit (cf. http://wiki.tdwg.org/Literature).
TaxonX provides a light-weight approach as it focuses on the core components of taxonomic treatment information, whereas taXMLit covers entire works, providing very detailed markup for document as well as data structures. Hence, the work of TDWG's Taxonomic Literature Interest Group aims at developing an optimal solution starting from these two suggested schemas. (cf. Catapano and Weitzman 2007)
A common limitation of both TaxonX and taXMLit is that they do not support well queries over morphological features, because, they lack in markup for identifying individual concepts within morphological descriptions. Therefore, some detail-level extensions of the XML schemas would be required. (cf. Sautter, Böhm and Agosti 2007a)

GoldenGATE editor for semi-automated XML markup

A leading-edge tool that has been specifically designed to support the digitisation of legacy taxonomic literature is the GoldenGATE editor, developed by researchers at the University of Karlsruhe's Department of Computer Science.
GoldenGATE supports all the steps from OCR output to full machine readability: OCR cleanup, semi-automated markup (both structural and semantic), including the detection of treatment boundaries and the markup of the internal structure of treatments. It also allows the application of automated external markup tools, like TaxonGrab, FAT or FindIT (see section 8.3.6 below) for the markup of scientific names.
GoldenGATE currently supports XML encoding according to the TaxonX format, but not taXMLit, though it may become an important tool for a future TDWG standardised XML schema.
The GoldenGATE editor has proven to simplify and accelerate the XML markup creation process significantly. These advantages result from both the semi-automated, token-wise XML editing and the integration of existing Natural Language Processing (NLP) tools for automated detail-level markup. It has been shown that marking up an OCR document using GoldenGATE is three to four times faster than with an off-the-shelf XML editor like XMLSpy. Using domain-specific NLP-based plug-ins, the speed of markup creation even can be higher. (cf. Sautter, Agosti and Böhm 2007b)
An example of a large project that uses the GoldenGATE editor is Plazi.org (see section 8.3.4 above).

GoldenGATE editor website

The current version of GoldenGATE, including all the resources needed to convert OCRed biosystematics documents into XML content marked up in the TaxonX XML schema, is available from http://Idaho.ipd.uni-karlsruhe.de/GoldenGATE/.

MARTT markup rules learning system

An other interesting system is MARTT (MARkuper for Taxonomic Treatments) which aims to enhance the automated conversion of taxonomic publications to XML format. MARTT makes use of machine-learning mechanisms that allow the system to learn markup rules from training examples and apply the rules to tag new descriptions.

The system has a knowledge induction component, which takes a tagged collection to induce semantic association rules from it. Furthermore, the system allows for storing and managing association rules learned over time. In addition, MARTT provides a number of utilities for reducing the effort for training example preparation, creation of a comprehensive schema, and predicting system performance on a new collection of descriptions.
The system has been tested with several plant and alga taxonomic publications including Flora of China and Flora of North America. (cf. Cui 2008)

8.3.6 Taxonomic Name Recognition tools

Named Entity Recognition (NER) is a subtask of information extraction that seeks to locate atomic elements in natural language text and classify them into predefined categories. A wide range of computational techniques, linguistic grammar-based, statistical and other approaches have been applied for this task. (cf. the survey of Nadeau and Sekine 2007)
Scientific names are a special case in NER and the term Taxonomic Name Recognition (TNR) has been coined to cover methods and algorithms for identifying and extracting names from taxonomic publications. (Koning, Sarkar and Moritz 2005)
Extraction of organism names from digitised (OCR) or "born-digital" texts is essential to allow for enhanced content management, linking content related to particular taxons, search & retrieval and other services. Because organism nomenclature and taxonomic publications conform to prescribed rules, TNR and related applications are particularly useful for extracting names and leveraging indices of taxonomic names.
Below we briefly describe three such applications, TaxonGrab, FAT (Find All Taxon Names) and FindIT / TaxonFinder.

TaxonGrab

TaxonGrab has been developed in a NSF-funded project at the American Museum of Natural History by an informatics group led by Drew Koning.
TaxonGrab draws on the rules conventionally used for taxonomic nomenclature and uses a combination of contextual rules and a language lexicon to implement a set of computational techniques for extracting taxonomic names. (Koning, Sarkar and Moritz 2005)
Basically, TaxonGrab uses a list-based exclusion approach in combination with contextual rules. List-based exclusion means that a lexicon of common words such as WordNet serves as a list of known negatives, i.e. words that should in principle not be part of taxonomic nomenclature. On top of this, rules for what counts as regular taxonomic expressions are used to identify relevant phrases.
TaxonGrab has been tested with a corpus of 5,000 pages from "The Birds of the Belgian Congo", volume 1, by James Paul Chapin (published in four parts in the Bulletin of the American Museum of Natural History, 1932-1954). Extraction of taxonomic names from this corpus previously had been conducted manually by a team of experts who identified over 8,000 taxonomic names.
One problem with list-based exclusion such as used with TaxonGrab is that taxonomic expressions that include common language words are exluded. Besides OCR errors and manuscript typos, this was the main reason for the majority of errors in the evaluation of TaxonGrab. However, it performed at greater than 96% precision and 94% recall from the documents examined.
Precision is defined as the ratio of correct taxonomic names (TP) to the sum of correct and false taxonomic names (TP+FP): TP/(TP+FP). Recall is defined as the ratio of the sum of correct taxonomic names (TP) to the sum of correct and missed taxonomic names (TP+FN): TP/(TP+FN).
With regard to the speed of extraction, the manual extraction was reported to have taken 80 hours, while the automated method took approximately 330 seconds.
The TaxonGrab project website is to be found at: http://research.amnh.org/informatics/taxlit/apps/.

FAT – Find All Taxon Names

FAT (Find All Taxon Names) has been developed by researchers from the University of Karlsruhe's Department of Computer Science under a research grant by Deutsche Forschungsgemeinschaft.

FAT combines several computational linguistics and learning techniques to automatically extract taxonomic names from legacy documents. In particular, the techniques make use of structural rules, dynamic lexica with fuzzy lookups, and word-level language recognition. They are applied sequentially so that each technique can use the results from the preceding ones. FAT has been tested with legacy documents from different sources and times to evaluate its performance. The experimental results showed greater than 99% precision and recall.

We do not attempt to summarise here the complex sequential application of the different techniques and the heterogenous corpus of taxonomic literature from which names have been extracted. These are described in detail in an available research report. (Sautter, Böhm and Agosti 2006) This report also provides a useful overview and assessment of the limits of techniques that have been used in various approaches to extract taxonomic and other scientific names.

FindIT / TaxonFinder

TaxonFinder has been developed by the Universal Biological Indexer and Organizer (uBio) research team and is described on the uBio website as "an attempt to merge the TaxonGrab and FindIT algorithms". Yet, a lengthy survey of resources available online did not produce any details about this merging.

The original FindIT application, which is still in use at uBio, uses name recognition methods for parsing free text and identifying scientific name and author combinations. These methods are enhanced with the capability to recognise author citations, taxonomic rank and nomenclatural annotation that may occur within a scientific name string.

In a first step, the application discriminates possible name and author combinations from non-name/author text sequences. Pattern-matching expressions and a lexicon of English words are used to identify likely scientific names and author combinations. Based on an analysis of the millions of scientific names strings recorded in the uBio NameBank, some 3,000 English words have been flagged as co-occurring within taxonomic nomenclatural.

The result of the first step is an array of text strings that represent potential scientific name and author combinations. In the second step, the results are parsed, evaluated and given a confidence score, using several taxonomy specific and other lexica (biological genera, species and infra-species names, suprageneric names, genus and species suffixes, strings that are both text words and scientific names, etc.)

The scoring of the results is based on the presence of known names within the scientific name and author combination or if an unknown name falls within the probability range of known latin name suffixes.

Sources: FindIT documentation: http://www.ubio.org/tools/recognizeHelp.php; TaxonFinder, http://www.ubio.org/index.php?pagename=soap_methods/taxonFinder; taxonfinder2, http://code.google.com/p/taxonfinder2/

Evaluation of FAT and TaxonFinder

A comparison of the performance of FAT and TaxonFinder has recently been conducted in the framework of an evaluation of the Biodiversity Heritage Library. (Wei, Freeland and Heidorn 2008; on the BHL see section 8.3.2 above)

In the evaluation, 392 OCRed taxonomic literature pages were randomly selected from the BHL database. A group of biologists manually identified taxonomic name strings in these pages, producing 3,003 valid names (2,610 unique names). For this sample, the OCR error rate for name strings was 35.16%.

Against this sample, the performance of FAT and TaxonFinder was evaluated, applying two measures: Precision and Recall. Precision was defined as the proportion of algorithm identified strings that are valid names (i.e. the capability of the algorithm to identify the valid names as well as exclude the non-valid name at the same time). Recall was defined as the proportion of valid names in the sample that are recognized by the algorithms (i.e. the capability of finding all valid names from the collection).

In this setup, TaxonFinder found 1,540 names; 674 of them were correct names; FAT found 1,603 names; 517 of them were valid names. The precision for TaxonFinder and FAT was 43.77% (=674/1540) and 32.25% (=517/1603) respectively. The recall for TaxonFinder was 25.82% (=674/2610) and for FAT 17.21% (=517/3003). In short, TaxonFinder overall performed considerably better than FAT.
A further interesting finding was that for TaxonFinder the uBio NameBank omission rate was 5.4%, which means that 5.4% of the correct names found by TaxonFinder were not already in NameBank. According to the evaluators, "[T]his demonstrates that names missing from the NameBank authority file are not the major source of information loss in converting the data from images to a structured searchable database. Our results indicate that improving the performance of TNR [Taxonomic Name Recognition] algorithms is the main challenge for producing an index to taxonomic names within digital library projects like BHL." (Wei, Freeland and Heidorn 2008)

8.4 Natural history collection digitisation manuals

The following are useful manuals for the digitisation of natural history collections:

ENBI manual of best practice in type specimen digitisation

The European Network for Biodiversity Information (ENBI) project provides a manual of best practice in the digital imaging of biological type specimens (300+ pages). The manual covers some general topics (e.g., image metadata standards and practices, requirements of "taxonomic-grade" images, colour management, etc.), presents digitisation approaches for different groups of organisms, and provides information on equipment and standards used for selected taxa and projects. (ENBI / Häuser et al. 2005).

GBIF training manual on natural history collection data

Recently, a GBIF working group compiled a training manual for digitisation of natural history collections data (500+ pages), which covers topics such as possible uses of digitised collection data, initiation of a digitisation project, data quality, data cleaning, georeferencing and others. (GBIF 2008b)

9 Taxonomic databases and services

9.1 Reducing the "taxonomic impediment" through easier access to taxonomic databases

Growing importance of taxonomic knowledge

Taxonomy is an accumulation of information and expertise about plants and animals. It includes the names of organisms which are governed by Codes of Nomenclature, methods of identifying organisms, and hypotheses of their evolutionary relationships.
The recent decades have seen a growing importance of taxonomic knowledge to address issues in ecology, agriculture, biodiversity and species conservation. Indeed, ever more research, professional and policy communities are looking towards taxonomy as a key scientific knowledge base and, in turn, taxonomists are challenged to contribute to major questions of bio-ecological change. (cf. CETAF 2004)
In Europe, the Consortium of European Taxonomy Facilities (CETAF) has initiated major projects that aim to leverage taxonomic capability in technological and organisational terms, SYNTHESYS and EDIT, the European Distributed Institute of Taxonomy (see section 16.1).

Taxonomic impediment

The Convention on Biological Diversity (1992) recognised the crucial role of taxonomy in promoting sustainable development, however, it also became clear that there is a "taxonomic impediment", the lack of taxonomic information, skills, personnel and capacity particularly in the developing countries, impeding the implementation of policies and practices of sustainable management and conservation of biodiversity.
The Global Taxonomic Initiative (GTI) was installed as a high-level mechanism under the Convention to remove or at least reduce the impediment, however, on the ground some countries (for example, Mexiko) seem to have addressed the challenge more thoroughly than others. Indeed, augmenting taxonomic capacity is no simple endeavour and demands sustained investment to educate additional cohorts of taxonomists, transfer of knowhow and provision of easy to use tools.
There are an estimated number of 6,000 professional taxonomists worldwide, which is a small number compared to the challenge and, even worse, they are distributed very unevenly. Most taxonomists are located in the industrialised countries, while there are often only a few in the biodiversity-rich but economically poorer countries (the website of the German National GTI Focal Point provides some illustrative figures). One acknowledged effort to ease this situation is BioNET International's work on establishing and operating partnerships for taxonomy in developing countries.

Providing easier access to taxonomic information

It is understood that online taxonomic databases and information services can help to reduce the "taxonomic impediment" by at least allowing for easier access to the stock of digital information that already has been accumulated. (Kim and Byrne 2006; Remsen and Lane 2008). Though there are also voices emphasising that the massive investment in such databases and services and the focus on "big" IT-based biodiversity projects of large museums and universities is consuming a too large part of available funds that is lacking where the taxonomic groundwork is done. (Flowers 2008)

9.2 Taxa as the basis of integrated information services

Taxa as basic organisational units of biological knowledge

Taxa are the basic organisational units of biological knowledge in the fields of natural history and biodiversity. A taxon is a scientific name designating an organism or group of organisms, which is assigned a taxonomic rank and can be placed at a particular level in a systematic hierarchy reflecting evolutionary relationships.
In the digital environment taxa are used to virtually tie together the available data about species and to provide ever more sophisticated information services. Taxonomically-informed services are expected also to increaseingly make accessible species-related information that is embedded in the growing array of medical, agricultural, ecological and other scientific information.

Taxonomic inconsistencies and inflation

While the taxonomic name of an organism is a key link between different databases, such names have serious limitations as identifiers, because one organism can have many scientific names over time and the same name can have been used to refer to different taxa. Moreover, there may be a taxonomic inflation in some regions of taxonomy, i.e. an accumulation of scientific names due to processes other than new discoveries of species. Such processes are "splitting", i.e. separating a species into two or more species, or elevation of taxa, creating inflation at the higher level. (Mallet 2004; a list of 314 out of 545 potential "splits" of bird species identified in August 2002 [often subspecies proposed to receive full species status] is provided by de By 2002)
Thus it can be difficult to retrieve information about an organism even if a scientific name is known. This is one major reason for the implementation of unique identifiers such as Life Science Identifiers (LSIDs), which are described in chapter 12.

Taxonomic and other databases

There are a large number of taxonomic databases available worldwide that record the scientific names, synonymy, classification, geographic distribution and relationships of biological organisms. Such databases allow for leveraging access to authoritative taxonomic checklists, nomenclatural data, and other useful names list compilations such as regional species lists, invasive or endangered species lists.
We do not intent to provide a detailed overview of available databases or description of the technical setup of such databases. With regard to the number of existing authoritative databases,

- the Global Biodiversity Information Facility (GBIF) lists 27 global taxonomic databases, http://www.gbif.org/links/taxo,
- the Catalogue of Life project currently draws on 52 databases (some of which overlapping with the ones listed by GBIF), and
- the Digitaltaxonomy.infobio.net lists 387 biodiversity databases and database access providers worldwide (some representing several databases while others are facilitators or sponsors).

9.3 The Catalogue of Life project

Creation of a validated index of all the world's known species

The Catalogue of Life (CoL) is the flagship project of the Species 2000 programme which it promotes together with the Integrated Taxonomic Information system (ITIS). ITIS North America is a partnership of U.S., Canadian and Mexican agencies and other organisations that have joint to provide authoritative taxonomic information on plants, animals, fungi, and microbes.
CoL's goal is to compile and make openly accessible a single unified and validated index of all the world's known species. The declared target for coverage of the estimated 1.8 million species is 2011.
To this end, CoL brings together an array of global species databases covering each of the major groups of organisms. The participating data-bases are widely distributed throughout the world and currently number 52. The 2008 Annual Checklist aggregates information on 1,105,589 species.
CoL is used by the Global Biodiversity Information Facility (GBIF) and Encyclopedia of Life (EOL) as the taxonomic backbone to their web portals.

Websites

Species 2000, http://www.sp2000.org
Catalogue of Life, http://www.catalogueoflife.org/info_about_col.php

9.4 Universal Biological Indexer and Organizer (uBio)

In two sections above we have described how the uBio TaxonFinder application is used in the Biodiversity Heritage Library project to identify and extract scientific names from OCRed taxonomic literature (section 8.3.2), and how this application compares to other Taxonomic Name Recognition tools (section 8.3.6). Here we document uBio in more

detail, as it provides an example of leading-edge taxonomic information integration and service provision.

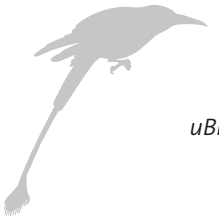
Research-driven taxonomic service development

uBio is a research-driven taxonomic services project that has developed a series of innovative solutions demonstrating the value that taxonomy provides as an organisational framework for all online information related to biological species.
uBio is located at the MBLWHOI Library (Massachusetts, USA) that serves the library needs for the Marine Biological Laboratory (MBL) and the Woods Hole Oceanographic Institution (WHOI). uBio's research and development work has been supported by the Andrew W. Mellon Foundation.

Core components of the uBio system

The uBio system has the following core components:

- the Taxonomic Name Server (TNS) – acts as name thesaurus that ingests and catalogs biological names and classifications; more specifically, TNS maps alternative names of organisms (whether they are taxonomically correct, colloquial names or mis-spellings) against each other, and places them within flexible multiple hierarchical structures;
- the NameBank – is a repository that stores and serves recorded biological names, at present over 11 million names; the NameBankIDs are served via Life Science Identifiers (LSIDs);
- the ClassificationBank – stores multiple classifications and taxonomic concepts, i.e. it allows different experts' views on the classification and circumscription of the same taxon to coexist in one system.



uBio RSS

uBio has developed a number of innovative algorithms and applications, of which we have already presented uBio TaxonFinder. An other interesting application that draws on the different components of the uBio system is uBio RSS (http://www.ubio.org/rss). uBio RSS daily filters hundreds of RSS feeds of scientific journals and other scientific sources for new content that references scientific taxa against the Annual Checklist of the Catalogue of Life Partnership (see section 9.3) and other taxonomic sources.
Users of uBioRSS can create personalised profiles and, for example, choose to display content that refers to organisms from a particular taxonomic group, or from a regional or thematic list, such as the IUCN Red List of Threatened Species. Users can also receive updates on new content that matches their criteria by subscribing to a custom uBioRSS feed. Websites with a biological focus may also receive and present dynamic updates on literature referencing organisms in their domain.
References for uBio RSS: Leary et al. 2007; Remsen and Lane 2008.

Website

http://www.ubio.org

9.5 Taxonomic Search Engine

Brief description

The Taxonomic Search Engine (TSE) is a widely acknowledged pioneering search engine to query multiple taxonomic databases using web services.
The TSE has been developed by Roderic D.M. Page from the University of Glasgow's Division of Environmental and Evolutionary Biology, Institute of Biomedical and Life Sciences.
TSE queries multiple taxonomic databases (ITIS, Index Fungorum, IPNI, NCBI, uBIO and others), summarises the results in a consistent format, and supports further drill-down queries to retrieve a specific record. The TSE can also optionally suggest alternative spellings the user can try.
TSE also can act as Life Science Identifier (LSID) authority for source taxonomic databases, creating and serving globally unique identifiers for each name.

References

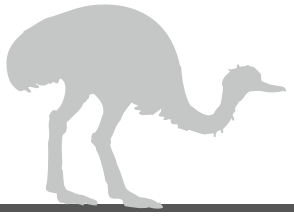
Page 2005 (gives a detailed technical description of the TSE);
Page 2006 (provides background on the development of TSE and discusses the use of LSIDs and RDF with taxonomic names).

Website http://darwin.zoology.gla.ac.uk/~rpage/portal/

9.6 NHM Nature Navigator

Brief description We include the Nature Navigator of the Natural History Museum, London, as an example of how a taxonomic backbone has been implemented in an application for broader, non-professional user groups such as teachers and students.
The development of the Nature Navigator has been funded by the UK New Opportunities Fund (Digitise Programme), to provide a single access point to information on more than 8,000 of the best-known species that occur in Britain. The Navigator uses the ITIS (Integrated Taxonomic Information System) as its taxonomic backbone, but only includes species that also have a common name.
The intention of the Navigator is to guide users through the mass of names of organisms, showing the preferred scientific and common names, related organisms and where they fit into the classification of the natural world.
The application allows browsing access to the taxonomy, expanding and collapsing branches. The taxonomy is integrated with display of fact sheets and provides links that carry out searches on external websites such as the UK National Biodiversity Network (that provides distribution information), Google Images and others.

Websites http://www.nhm.ac.uk/nature-online/biodiversity/nature-navigator/
http://www.itis.usda.gov



10 Online collaboration tools for taxonomic and other biological studies

Alongside the development of regional and global databases of taxonomic information and enhanced taxonomic information services, tools have been created that allow for conducting online taxonomic and other biological studies.
These tools are understood to be a means for tackling both taxonomic impediment and taxonomic inconsistencies. They provide taxonomists, of whom many work outside well-equiped organisations, with a state-of-the-art workbench, and allow to collaboratively work on taxonomies of groups of organisms with the aim of revising and consolidating them (creating so called consensus taxonomies).
However, existing and emerging collaborative tools need not necessarily be used only for taxonomic work, indeed, they increasingly provide a flexible environment in which different interest groups can pool their efforts and share information sources and expertise.

10.1 Creating a Taxonomic e-Science (CATE)

Project goals

Creating a Taxonomic e-Science (CATE) is a project funded under the UK Natural Environment Research Council's e-science initiative. The project partnership comprises the University of Oxford, the Natural History Museum and the Royal Botanic Gardens, Kew.
CATE tests the feasibility of Web-based consensus taxonomy using two model groups, one from the plant (Araceae) and the other from the animal kingdom (Sphingidae).
CATE explores practically the idea of "unitary" taxonomy and promotes Web-based revisions as a source of authoritative information about groups of organisms (for background on the why and how of unitary taxonomies see their website).

Technical implementation

The project develops all required layers of a Web-based system that allows to collaboratively conduct revisionary taxonomic work. This comprises the underlying data model, persistence layer, service layer, Web controller, JavaScript widgets, view component, Web interfaces, specific tools, and so forth. The system development and software releases are carefully documented on the CATE website.
There also are two CATE demonstrator websites of online taxonomic revision, http://www.cate-araceae.org and http://www.cate-sphingidae.org, which resolve Life Science Identifiers (LSIDs) for the taxonomic concepts presented. This has been implemented with funding by the Global Biodiversity Information Facility (GBIF).

References

Godfray et al. 2007, discuss the option of moving revisionary taxonomic work completely to the Web, and present CATE as a prototype model.

Website

http://www.cate-project.org

10.2 Scratchpads

Scratchpads is a project managed by the Natural History Museum (NHM) in London with funding through the European Distributed Institute of Taxonomy (EDIT), the Global Biodiversity Information Facility (GBIF), and from core funding within the NHM.
Scratchpads are integrated workbenches and open access spaces on the Web that allow research communities to create, share and manage biodiversity information. But they also provide the freedom for individuals to work in different ways, at their own pace, without necessitating consensus.

Technical platform, modules and key features

Scratchpads rely on the open source Drupal Content Management System which in part is a social networking application that enables communities to manage, share and publish taxonomic information online. The Scratchpad project team works on making Drupal's underlying infrastructure better suited to the needs of biodiversity communities.

The team develops modules that support specific taxonomic data types (e.g. specimens, literature, etc.), templates for import and export of data (e.g. taxonomic classifications), and by making web services of other data providers readily accessible (e.g. Biodiversity Heritage Library, NCBI Genbank, etc).
Key features of Scratchpads include tools to manage Classifications, Phylogenies, Specimen records, Bibliographies, Documents, Image galleries, Maps, and Custom data.

Data hosting and organisation

The Scratchpads sites are hosted at the Natural History Museum. Data added to a Scratchpad are automatically classified and grouped around a taxonomy that is supplied by the users. This is optionally supplemented with information from Web accessible databases to automatically construct content rich web pages about any documented taxon. Currently these sources include Genbank, Global Biodiversity Information Facility (GBIF), Biodiversity Heritage Library, Google Scholar, Yahoo! Images and Flickr.

Scratchpads users and licensing of content

Scratchpad users include academic societies, journals, scientists, students and amateurs. Indeed, Scratchpads are offered free to anybody who com-pletes an online registration form, an academic affiliation or professional qualification is not required.
In less than two years the Scratchpad project has enabled the self-assembly of more than 70 research communities with over 700 registered users. Collectively these scientists have built more than 130,000 pages of content. (Rycroft et al. 2008)
Scratchpads assign ownership to the users generating the content, but enforce a licensing framework through which others can reuse this output. More specifically, the content must be made available under a Creative Commons "Attribution-NonCommercial-ShareAlike" license.

References

The Scratchpads website provides very detailed information about the project, including many presentations and some publications, http://scratchpads.eu/about
Scratchpads have been developed under Workpackage 6 of the EDIT – European Distributed Institute of Taxonomy (EDIT) project:
http://editwebrevisions.info

Website

http://scratchpads.eu

10.3 Encyclopedia of Life – LifeDesks

The Encyclopedia of Life project (see chapter 11) includes the development of tools for participation which are called LifeDesks. The technical development work is done by EOL's Biodiversity Informatics Group, based on the open source Drupal Content Management System.

Features

LifeDesks should allow groups interested in particular species to compile and further develop structured information for eventual aggregation on EOL species pages.
Initial implementation of the LifeDesk environment focuses on tools for the expert user. Data may be entered, linked and curated through a set of graphically-rich tools that interface to a relational database. Initial functionality will include the creation of "stub" species pages given a list of names, inclusion of text and images, and literature tools.
The approach has been influenced by the Scratchpads project and it is intended to develop compatible modules to interface with their existing study groups.

Project status

The LifeDesk environment is currently under development, the beta testing phase for a few LifeDesks will be launched mid-December 2008.

References

Schopf et al. 2008; EOL Taxonomy Sprint: Goals and Progress, http://groups.drupal.org/node/14749

Website

http://lifedesk.eol.org

10.4 Morphbank – Sharing of scientific images

Project background

Morphbank is a growing Web repository of scientific images that receives its main funding from the Biological Databases and Informatics program of the National Science Foundation (USA).
The Morphbank project is currently housed at the School of Computational Science at Florida State University and includes a team of 15 biologists, computer and information scientists who are working on developing the system using open-source software.
Images in Morphbank are deposited and often shared by scientists for a wide variety of research, including specimen-based studies in comparative anatomy, morphological phylogenetics, taxonomy and related fields.

Software

The software used in the current Morphbank system includes PHP, ImageMagick, MySQL, Apache, Java, and JavaScript.

Features

Morphbank provides templates to describe uploaded images of specimens in detail (taxon name, specimen part, sex, stage, imaging technique preparation, etc.), and annotate them with comments. For example, taxonomic descriptions of new species or other nomenclatural acts can be documented by images and image comment tools. There is a taxonomic tree for browsing the database and different strategies for searching specimen images are offered.

Fair Use principle

Morphbank is designated as a Fair Use website. The images in Morphbank that are not password protected can be used for private, education, research or other non-commercial purposes for free, provided that the source and the copyright holder are cited. Currently, Morphbank provides access to more than 63,000 public images of about 216,000 in total.

Website

http://www.morphbank.net

11 Strategies in content aggregation and access: The Encyclopedia of Life example

The Encyclopedia of Life (EOL) is an example of a large-scale program of content aggregation and access. This example may provide some lessons for other initiatives such as the European Digital Library, which uses different technologies, but may face similar problems with respect to the expected richness in content.

Goal and funding of EOL

The Encyclopedia of Life (EOL) is an ambitious program to organise and make accessible online available information about all known species on Earth. The initial idea for this program came from the prestigious sociobiologist E. O. Wilson, Harvard research professor and two-time Pulitzer Prize winner. The idea became a working program in 2007 based on a funding commitment of $50 million by the MacArthur Foundation, the Sloan Foundation and six founding partners. The latter group comprises the Field Museum of Natural History, the Harvard University, the Marine Biological Laboratory, the Missouri Botanical Garden, the Smithsonian Institute and the Biodiversity Heritage Library (a group of American and UK natural history organisations; see section 9.3).

One website per species approach

The EOL aims to create within 10 years a webpage for each of the estimated 1.8 million known species on Earth that provides the entry point to a vast array of knowledge and high-quality data. This knowledge and data about species should, for example, comprise taxonomy, geographic distribution, collections, genetics, evolutionary history, morphology, behavior, ecological relationships, and importance for human well being. As its taxonomic backbone the EOL uses the Catalogue of Life (see section 9.3).
The EOL is intended to become a primary resource for a wide audience that includes scientists, natural resources managers, conservationists, teachers, and students around the world. The EOL programme also includes a participatory component called LifeDesks (see section 10.3).

Issues in the current development of EOL

The February 2008 launch of EOL included content from FishBase, AmphibiaWeb, Tree of Life, and Solanaceae Source in addition to 24 exemplar pages and more than a million stub pages for names in the Catalog of Life. The launch of EOL generated a tremendous interest which, however, dropped off markedly thereafter.
In mid April 2008, there was a first review of the work of EOL's Biodiversity Informatics Group in which the two most debated areas were how to acquire more content and the "vetting" of content (i.e. only to use information from trusted providers that is scientifically authenticated or verified by experts). Other topics discussed were the site design, globally unique identifiers, and organisational matters. (Page 2008a)

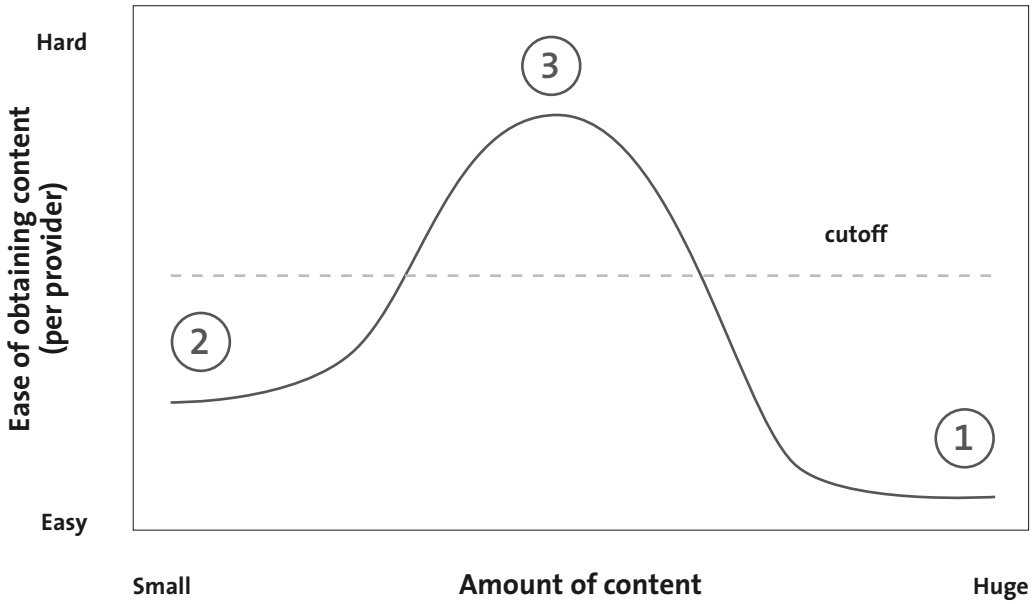
Relative lack of content in comparison to other websites

It was clear, that there was a relative lack of content on most of EOL's species pages, in particular, compared to other websites such as DiscoverLife, ZipcodeZoo or iSpecies. For example, iSpecies, is a simple mashup site which assembles automatically information from sources such as GenBank, GBIF, Google Scholar, Yahoo Images, and Wikipedia.
There is a concern that EOL risks being marginalised. The challenge for EOL is how to cover all estimated 1.8 milllion known species in the 10 year timeframe, which means that it would need to add around 500 content-rich species pages per day. It seems doubtful that EOL can achieve this with its current strategy to limit its content to "vetted" information from trusted providers and, even, trying to identify and adjust erroneous information. Indeed, by drawing on species distribution maps from GBIF, which are known to often contain errors, EOL implicitely acknowledges that it cannot produce interesting websites without running the risk to present erroneous information.

Strategies for content acquisition

The review report of EOL's Informatics Advisory Group (IAG) makes it clear that EOL "needs more content, fast, and needs to tackle the issue of vetting in a way that will scale". Rod Page, the chair of the IAG, provided a figure that plots the cost of obtaining certain types of content against the amount of content obtained (see figure below).

Costs to consider include developer time to import data, time spent negotiating intellectual property agreements, etc.



Source: Page 2008a, http://blog.eol.org/category/biodiversity-informatics/

Types of content and priorities

Content of type (1) are large, freely available and relatively easy to import data sources, (2) are small sources that require specific tools to make their content available, and (3) are data sources of well-established data providers that can require considerable effort to incorporate due to both IPR issues and idiosyncratic data structures. An arbitrary cutoff represents the level above which the effort required to obtain content outweighs the value that content would bring to EOL.

The report recommends going after content in category 1 first, which would include PubMed, GenBank, Wikipedia, ITIS, Flickr, and GBIF. Flickr and Wikipedia of course are not scientifically curated, however, it was noted that for example on Flickr there are some groups who build photo libraries of organisms that are tagged with scientific name and geographic location.
The example was "Field Guide: Birds of the World", http://www.flickr.com/groups/birdguide/, which as of 31 October 2008 hat 7487 group members who contributed 69,534 images. The Field Guide demands that the uploaded photographs must be tagged with the correct scientific name according to Clements 5th edition which is accessible via Avibase.
In content category 2, the next to go for, tools are needed to allow small providers to manage their own content, and contribute to EOL at the same time. The "LifeDesks" EOL is developing at present correspond to this strategy (see section 10.3)
Finally, for content sources in category 3, though representing large and valuable sources, in the short term the effort involved in incorporating it may outweigh the value it brings. It was also noted, that tools developed for small content contributors may facilitate acquiring some of the content from category 3 sources.

Additions to the EOL

A recent EOL status report describes already made and planned additions:
"In early September, content came online from Animal Diversity Web, AntWeb, ARKives, and others to reach about 38,000 taxa with text and another 15,000 with no text but images from several sources. At least 40,000 (perhaps up to 150,000) additional text pages will be released in December. Original literature from the Biodiversity Heritage Library is linked to several hundred thousand species pages.
To accelerate connections with existing resources, a registration process now allows prospective data partners to establish their own affiliation with EOL. Providers map

their schemas to the EOL Transfer Schema, which uses TDWG standards such as the Species Profile Model. (...)
A variety of tools and features that enable EOL participation are coming online. LifeDesk:Expert is a Drupal-based content management environment, modeled after and compatible with EDIT Scratchpads, which scientists can use to assemble and manage information useful to their own communities and to EOL. Later LifeDesk versions will support educational and citizen science audiences. In December 2008, we will launch commenting and tagging features. Flickr (http://flickr.com/) has been chosen as one way for users to contribute images to EOL.
Once the public begins contributing content, a robust curatorial network is necessary to decide which contributions are suitable for authoritative pages." (Parr 2008)

Website http://www.eol.org

12 Life Science Identifiers (LSIDs) in natural history and biodiversity

In the presentation of the "layer cake" of the Semantic Web we have addressed the importance of URIs which are used to uniquely identify Web resources. In the fields of natural history and biodiversity a standardised approach to globally unique identifiers are Life Science Identifiers (LSIDs).
This approach is expected to be increasingly used as the TDWG has adopted LSIDs as recommended standard for assigning globally unique identifiers.
It is also understood that LSID will form an important basis of building a Semantic Web for the life and natural sciences. (cf. Good and Wilkinson 2006)

12.1 Life Science Identifiers (LSIDs) basics

LSID specification

The Life Sciences Identifiers (LSIDs) specification provides a standardised way of naming and locating data sources based on a Uniform Resource Name (URN) scheme and retrieving metadata in a standard format. LSIDs as such are persistent, location-independent identifiers for uniquely naming biological data sources.
The LSIDs specification (OMG 2004) has been developed by the Interoperable Informatics Infrastructure Consortium (I3C) and OMG Life Sciences Research. The aim was to help overcome severe shortcomings of the many naming schemes in use in the life sciences and related domains, making integration between the multiple, distributed data stores very difficult.
More specifically, the LSID specification provides a solution for implementing a standardised naming schema, a service assigning globally unique identifiers complying with this schema, and a resolving service that specifies how to retrieve the entities identified by such naming schema from repositories, using web services.

Standardized naming schema

An LSID is represented as a Uniform Resource Name (URN) that consists of three scoping mechanisms: an authority, a namespace, and an identifier. It can also optionally contain a version, specified by a revision identifier.
These parts are combined to create an LSID string with the following form:
urn:lsid:<Authority>:<Namespace>:<ObjectID>[:<Version>]

- urn:lsid: is a mandatory prefix in which "urn" indicates that the LSID is a Uniform Resource name (URN), and "lsid" indicates that the identifier is resolved using the LSID protocol;
- Authority: is an unique string, usually an Internet domain name owned by the LSID data provider;
- Namespace: is an alphanumeric sequence that constrains the scope (e.g. to a particular database),
- ObjectID: is an alphanumeric sequence identifying the object;
- Version: is an optional alphanumeric sequence describing the version of the object.

Example: urn:lsid:ipni.org:names:302735-2 (the IPNI record for the taxon name Achillea millefolium), which can be accessed through http://lsid.tdwg.org/urn:lsid:ipni.org:names:302735-2.
Some LSID best practices are summarised in Smith and Szekely 2005.

No central LSID authority

There is no central authority for registering or resolving identifiers as for example with Digital Object Identifiers (DOIs). This means that there is no mechanism to prevent that different authorities create different LSIDs for a common resource such as a taxon name.

LSID metadata in RDF

A key benefit of using LSIDs is the clear separation of data and metadata, of which the data should never change whereas the metadata may be updated or changed. The data

behind an LSID can be any resource, such as a taxonomic concept or name, pecimen record, image, 3D model, audio recording, etc.
The LSID specification does not specify that the metadata for the resource should be in a particular format, however, the LSID metadata is, by convention, provided in RDF format. Furthermore, it is suggested to use existing metadata schemes rather than to create a new set of RDF properties.
The use of RDF allows for describing relationships between different LSID data resources (e.g. between taxon names and images) of the LSID authority or objects held in databases of other organisations. This greatly facilitates the linking and integration of information from multiple sources, i.e. a semantic layer is created that can be exploited by Semantic Web tools.

12.2 LSID service process and software

LSID resolver and client

A LSID resolver service is required that is capable of interpreting the LSID encoding to resolve and return the correct data. A LSID resolver is a software that implements the LSID resolution protocol and allows client applications to locate and access the data uniquely named by the LSID URN.
A LSID client accesses the data or metadata of a LSID in four steps (cf. Page 2005):

- Firstly the client needs to find the location of the service that can resolve a particular LSID. For this step it queries the Internet DNS service records to find the hostname and TCP/IP service port for the LSID authority.
- Secondly, with the returned location of the LSID authority server the client can then query the authority for available services and retrieve the authority WSDL (Web Service Definition Language) file that defines the LSID resolution service, including location and bindings. The LSID standard defines bindings for SOAP, HTTP GET and FTP, of which the HTTP GET binding is the mostly widely used.
- Thirdly, given the authority WSDL, the LSID client uses its preferred protocol to retrieve a second WSDL file that specifies how the metadata or data corresponding to the LSID can be retrieved.
- Finally, the client sends a getData or getMetadata call to the LSID data retrieval service, which uses the namespace and object identifier parts of the LSID to locate or build the corresponding data or metadata from local resources such as a database.

LSID authority setup guidelines

Details of the technical setup of LSID authority servers are given in TDWG / Pereira et al. 2008.

LSID resolver testing software

A software for testing LSID resolver services is the LSID Tester developed by Rod Page from the University of Glasgow's Institute of Biomedical and Life Sciences, Division of Environmental and Evolutionary Biology.
LSID Tester is a web application that, given a LSID, performs seven tests, reporting the results at each step. (Page 2008b) If all tests are successful the metadata associated with the LSID is displayed, and can be viewed in a range of formats. The application also displays a link to the W3C RDF Validation Service so that the user can validate the RDF metadata.
The LSID Tester performs seven tests:

1. Is the LSID correctly formed?
2. Is the resolution service discoverable?
3. Can it retrieve the authority WSDL?
4. Does the authority WSDL define a HTTP GET binding for the service WSDL?
5. Can it retrieve the service WSDL?
6. Does the service WSDL define a HTTP GET binding for the metadata?
7. Can it retrieve the metadata for the LSID?

The source code of LSID Tester is available under a GNU General Public License version 2 from http://code.google.com/p/lsid-php/, and a working version is online at http://linnaeus.zoology.gla.ac.uk/~rpage/lsid/tester/.

Overview of available server and client software

To support the implementation of LSIDs, TDWG researchers have carried out a number of activities, that included a gap analysis of LSID software, documention of useful existing software, and identification of additional components that need to be developed to deploy a production quality LSID solution for biodiversity informatics.
Useful software that have been identified are:

- Lean PHP Resolver (simple PHP server-side LSID framework);
- Perl LSID API (server-side and client-side LSID implementation)
- J2EE LSID API (server-side and client-side LSID implementation)
- MS .NET LSID API (server-side and client-side LSID implementation
- LSID Server Conformance Test Tool (simple check of protocol conformance for any LSID);
- LaunchPad for Internet Explorer (plug-in allowing Internet Explorer to handle LSIDs natively);
- LaunchPad for Mozilla Firefox (plug-in allowing Firefox to handle LSIDs natively).

Source: http://wiki.gbif.org/guidwiki/wikka.php?wakka=LsidSoftwareInventory

TDWG LSID Web resolver

Notably TDWG also offers a LSID Web Resolver that is available at http://lsid.tdwg.org

12.3 TDWG recommendation of LSIDs and some recent implementations

LSIDs are a TDWG recommended standard

The Taxonomic Database Working Group (TDWG), the international biodiversity data standards setting group, adopted LSIDs as its recommended standard for assigning globally unique identifiers to data records and suggests to provide the LSID metadata in RDF.
TDWG also has defined the deployment of Life Science Identifiers as one of the priorities of the TDWG community of organisations and developers. Ongoing and new projects should address the need for tagging their data with LSIDs and consider the use or development of appropriate metadata vocabularies (the LSID metadata vocabularies developed by TDWG are described in section 12.4 below).

A growing number of implementations

Until recently there were only a few implementations of LSIDs, such as the public first LSID resolution service of the Northern Temperate Lakes - Long Term Ecological Research Network, http://lsid.limnology.wisc.edu.
Today, a growing number of institutions and projects in the field of natural history and biodiversity are implementing LSIDs. The availability of LSID resolvers also became a push by a TDWG Prototyping Working Group that in 2006 has supported the development of a number of LSID resolvers. Taxon names LSID resolvers were given the highest priority and there are now such resolvers available for IPNI, Index Fungorum and others. Below we describe some recent examples, comprising implementations initiated by the TDWG and others.

IPNI

The International Plant Names Index (IPNI) is a database of the names and associated basic bibliographical details of seed plants, ferns and fern allies. IPNI's focus is purely nomenclatural, i.e. no opinions are given on what are currently accepted names or synonyms.
The data records in IPNI come from three sources: the Index Kewensis, the Gray Card Index and the Australian Plant Names Index. The data are freely available and are gradually being standardised and checked.
As well as offering a website for individual users to search and download selected records, IPNI since 2006 also acts as an LSID server, allowing the automatic resolution of IPNI LSIDs into RDF format metadata which can be used by other services such as the Global Biodiversity Information Facility (GBIF) and incorporated into other systems.
Website: http://www.ipni.org/lsids.html

Index Fungorum

Index Fungorum is a major database of fungal names at species level and below, indicating if the name has formal status or not (all names are linked to pages giving the correct name, with lists of synonyms).

This international effort is co-ordinated and supported by the following custodians: CABI Bioscience, CBS and Landcare Research.
In 2005, the Index Fungorum partnership implemented LSIDs for the records in the IF database, and in 2006 developed a prototype LSID resolver system building on its established web services. Some development work also has been invested in allowing for the provision of RDF metadata according to the Taxonomic Concept Schema (TCS).
Website: http://www.indexfungorum.org/Names/IndexFungorumLSIDs.htm
LSID Resolver for Index Fungorum Taxon Names, http://wiki.gbif.org/guidwiki/wikka.php?wakka=LSIDResolverForTaxonNamesIF

Catalogue of Life

The Catalogue of Life (CoL) project in 2008 has implemented LSIDs as recommended by TDWG. In the past, the CoL changed identifiers with every new version of their Annual Checklist, thus forcing database owners who make use of CoL names and identifiers to adapt their databases if they wished to maintain their external linking to an authoritative source.
CoL now has a unique LSID for every recognised taxon in their Annual Checklist, which provides a persistent and location independent means to access taxon metadata. LSIDs appear on CoL Species Details pages and in the CoL tree. The LSIDs can be resolved to obtain information expressed as TCS (Taxonomic Concept Schema) metadata in RDF format, using an LSID resolution service. The RDF documents are drawing reference to concepts from the TDWG Taxon Concept LSID vocabulary. (Orme, Jones and White 2008, provide a detailed description of the CoL LSID deployment)
Website: http://www.catalogueoflife.org/lsid/

Biodiversity Collections Index

The Biodiversity Collections Index (BCI) aims to become a central index to specimen reference collections worldwide. The initial data for populating the index came from three sources: Index Herbariorum, Insect and Spider Collections of the World (ISCW) and Biorepositories.org (a Barcode of Life Initiative). BCI provides LSIDs for the indexed collections, which can be used for the CollectionCode field in Darwin Core and ABCD specimen records.
The BCI beta versions of the index website and its web services have been launched in July 2008. BCI provides a LSID authority service (and associated HTTP proxy service) that handles the resolution of LSIDs into RDF metadata in accordance with the LSID specification and the TDWG LSID Applicability Statement.
For additional information on the implementation and collaborative use of the BCI see: BCI guidelines on using LSIDs: http://www.biodiversitycollectionsindex.org/static/citing.html
Website: http://www.biodiversitycollectionsindex.org

CATE

The CATE (Creating A Taxonomic e-Science) software since its September 2007 release includes an LSID Resolution Service, developed as part of a contract with the Global Biodiversity Information Facility (GBIF). The two CATE demonstrator websites of online taxonomic revision for the Araceae (http://www.cate-araceae.org) and the Sphingidae (http://www.cate-sphingidae.org) resolve LSIDs for the taxonomic concepts presented. More information on the CATE project is to be found in section 10.1.
Website: http://www.cate-project.org (details of the LSID implementation are given in CATE 2007)

Morphster

Morphster is a project under the Assembling the Tree of Life (AToL) grand challenge initiative of the US National Science Foundation, which aims at describing up to 10 million extant species and computing and analyzing a unified phylogenetic tree.
The Morphster project developed a prototypic service-oriented architecture enabling and supporting morphologically based phylogenetic studies. In this context, a primary issue was seen to be the complete and consistent distributed representation of ontologies (both taxonomic and morphol-ogical), for which the use of a LSID system was explored as one important implementation mechanism.
In particular, the implementation of the LSID system focused on a solution for mapping LSIDs to information held in legacy databases. The use case was the University of

Texas UTCT Data Archive and the information comprised Darwin Core metadata about specimen and both metadata and images from high-resolution X-ray computed tomographic scans of those specimens.
The goal was to implement a system that allows for integrating the resources with the LSID protocol as an add-on layer on top of relational databases. This was realised using a trigger-based approach to facilitate LSID assignments. A SQL-like domain-specific language is used to define an export schema (from an existing database schema) marking the data that needs to be assigned LSIDs (i.e. an equivalent of SQL view definitions is used). The export schema is compiled into appropriate runtime tables and triggers. These triggers assign LSIDs to the existing data when run initially as a batch process and on the fly to new data additions or updates.
A detailed description of the implementation is provided in Miranker, Bafna and Humphries 2006.

Many more LSID implementations

The examples above are but a few selected implementations of LSIDs. Other implementations are to found in taxonomy services such as uBio (see section 9.4), individual databases such as Morphbank (see section 10.4) and research projects such as SEEK (see section 13.2.3).

12.4 TDWG LSID metadata vocabularies

Purposes of TDWG's LSID vocabularies

To exploit the potential of LSIDs in the Semantic Web environment, the TDWG is developing a set of LSID vocabularies that allow to formally describe the metadata returned for particular classes of objects within the TDWG domain. This is part of a larger TDWG ontology effort that aims at describing how these classes of data are related (see section 13.1 below).
The TDWG LSID vocabularies enable the typing of metadata records associated with LSIDs and provide the RDF semantics of the metadata needed to describe the information objects that are exchanged. In the Semantic Web environment, this should also allow applications to combine data of different kinds from multiple sources, e.g. not just consume specimen or observation data from one database, but combining it with geographic, phylogenetic, molecular and other data.

Available LSID metadata vocabularies

At present four TDWG LSID metadata vocabularies are available, three of which are based on the Taxonomic Concept Schema (TCS) that is a TDWG recommended standard since 2005. This signals TDWG's priority to further standardise the exchange of taxonomic information via their strong promotion to implement LSIDs.

The already available TDWG LSID metadata vocabularies are:

- TaxonName: Based on TCS (already used by Index Fungorum, IPNI and ZooBank);
- TaxonRank: Derived from TCS; a vocabulary supportive to TaxonName that provides an enumeration of taxonomic ranks;
- TaxonConcept: Based on TCS; currently used as an embedded object by the TaxonOccurrence vocabulary (already used by the LSID service of the Catalogue of Life);
- TaxonOccurrence: Based on Darwin Core, provides the minimum required to exchange observation and specimen data (already used in Global Biodiversity Information Facility's web services).

Vocabularies under development concern Person, Team, PublicationCitation, Institution, and Collection.
For more detailed information see: TDWG LSID vocabularies, http://wiki.tdwg.org/twiki/bin/view/TAG/LsidVocs

LSID vocabularies are (small) OWL ontologies

Technically each LSID Vocabulary is a (small) OWL ontology containing one or more classes and a number of properties whose domain is in one of those classes.

The metadata returned when the LSID is resolved is an instance of one of these OWL classes containing some or all of the class properties and some general properties that can be used with any of the LSID Vocabulary classes.

TDWG Species Profile Model (SPM) development

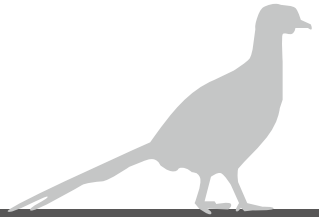
An interesting new development is the TDWG Species Profile Model (SPM) that is intended to complement metadata models which already are available for specimens and observations (i.e. Darwin Core and ABCD). SPM specifies data concepts and structure intended to support the retrieval and integration of data that describe species, e.g., facts about biology, ecology, evolution, behaviour, etc.
The SPM is developed in RDF and OWL and intended as one of the TDWG LSID metadata vocabularies that are loosely linked into and by the TDWG core ontology (see section 13.1 below).
An information object modelled in SPM "provides various named types of information about a taxon, or more precisely, about a Taxon Concept expressed in the TDWG Ontology controlled vocabulary. The associated information (SPM "InfoItems") comprise a collection of strongly typed attributes drawn currently from one of 37 classes of information about the taxonomic, ecological, and economic properties of the taxon. These include traditional morphological descriptions, information critical to the management of invasive or endangered species, and attributes important for field biology, for ecological science and for molecular studies." (Morris 2008)

SPM demonstrator cases

A demonstrator use case for SPM may be Plazi.org. Plazi conducts an experimental project funded by the Encyclopedia of Life (EOL) that explores how SPM could be used to serve content managed by Plazi (scientific species descriptions) to the EOL for inclusion in their species webpages. Similarly, the Cornell Lab of Ornithology explores how it could use the SPM to provide EOL access to its Birds of North America (BNA) multimedia collection. (Gerbracht and Kelling 2008)

Website

Species Profile Model (SPM), http://wiki.tdwg.org/SPM/ and http://rs.tdwg.org/ontology/voc/SpeciesProfileModel



13 Semantic Web ontologies for natural history and biodiversity domains

The ontological layer of the Semantic Web "layer cake" plays a key role for knowledge representation, data integration and advanced search and other services spanning heterogeneous databases of distributed information providers.
In chapter 7, we have addressed projects mainly from the cultural heritage domain that have developed such a layer based on the W3C SKOS standard. We have also pointed out limitations of SKOS for more complex demands than relatively simple semantic search functionality. Such demands require using domain and upper-level ontologies, which can be developed in Web Ontology Language (OWL).
In the fields of natural history and biodiversity, there are a number of efforts to develop and implement ontologies and other conceptual resources in OWL with the goal to leverage data integration and access.
In the sections below we first describe efforts by the TDWG Technical Architecture Group to develop a Biodiversity Informatics Core Ontology. This is intended to be an ontology above the TDWG LSID metatdata vocabularies which are presented in section 12.4. Because such an ontology would allow to semantically integrate, at a very generic level, a large part, if not all, biodiversity informatics, we describe this development to some detail.
Second, we briefly present a selection of other ontology development projects. The objective here is to illustrate the wide range of ontology developments, including examples of prototypic applications.

13.1 TDWG Biodiversity Informatics Core Ontology development

13.1.1 Towards a stack of biodiversity ontologies

Intended stack of TDWG ontologies

In 2006, the TDWG Technical Architecture Group started working on a stack of TDWG ontologies which was envisaged to comprise a Base Ontology, a Core Ontology and a Domain Ontology. (TDWG TAG 2006a and 2006b):
The Base Ontology would comprise classes that are not concepts generally discussed in the biodiversity research community, but provide base classes from which the Core Ontology classes would be derived.
The (Biodiversity Informatics) Core Ontology would comprise classes that correspond to the most common concepts used within the TDWG com-munity; it would include a basic class hierarchy and define some of the properties and relationships which are of greatest importance to the domain of biodiversity.
The Domain Ontology would be developed from the classes in the Core Ontology, and it was anticipated that it would comprise sub-ontologies that have a correspondence to a single class in the core ontology to encourage reusability of the main ontology classes (e.g. to prevent a Specimen ontology defining Place or a TaxonConcept ontology defining Descriptions).
Moreover, it was anticipated that Application Ontologies would map their more specific classes or data structures to classes and their properties in the Domain Ontology.

Development of the Core Ontology

The approach for the development of the Core Ontology was to derive the most important classes from four of the existing TDWG XML schemas, ABCD (Access to Biodiversity Collections Data), Darwin Core, SDD (Structured Descriptive Data) and TCS (Taxonomic Concept Schema).
In a TDWG Core Ontology Meeting, held 16-18 May 2006, suggested high-level classes and properties were presented and discussed. (TDWG TAG 2006b) Then the Core Ontology was drafted, represented in UML and realised in OWL Lite (the ontology, several UML representations, including a Base Ontology, and several explanatory documents

are available from the TDWG TAG Ontology Wiki). In October 2006, the results were presented at the TDWG 2006 Annual Meeting. (Kennedy et al. 2006)

Stagnation and reconsideration

However, since then only little further progress has been made. Rather, the original concept was dumbed down as is evident from the descriptions of the ontological layer development in the TDWG Technical Roadmaps for 2007 and 2008. (TDWG TAG 2007 and 2008)

The 2007 roadmap document, issued on 27 August 2007, explains that the ontology development could not be progressed at a sufficient level of detail allowing to provide a common ontological layer for the rolled out TDWG LSID vocabulary programme (described in section 12.4). Hence, the decision was taken to only loosely link the classes of the LSID metadata vocabularies into the higher classes of the core ontology.

The 2008 roadmap document, issued on 15 October 2008, states that the TDWG ontology "is more of a functional thing" that would have been better named a "dictionary", rather than giving the impression of "an expansive formalisation of the biodiversity informatics domain".

However, it is emphasised that there is the need to have a shared understanding of the kind of things that are behind LSIDs and at least some of the properties that are used to describe these things. This would be the function of the TDWG ontology, understood as "a rather trivial list of the things that we, as a community, can agree on the meaning of".

The 2008 roadmap stresses the tremendous benefits of having such a list of concepts, but, that even to keep it up to date, to manage the consensus building process around new concepts, and to educate the community on how to use them is an expensive enterprise. In fact, "[N]obody has been resourced to do this work in 2008 and therefore it hasn't happened as it should."

13.1.2 TDWG suggested technical architecture

The TDWG basic architecture

To put the development of the TDWG ontology in perspective, two important aspects of the general TDWG technical architecture as describe in the 2007 and 2008 roadmap documents should be noted.

Firstly, the architecture is meant to meet two needs: It should allow generic interoperability between data providers of the TDWG community as well as restricted validation of data for some networks. Therefore, a three pronged approach is proposed:

"1. An ontology is used to express the shared semantics of the data but not to define the validity of that data. Concepts within the ontology are represented as URIs (Universal Resource Identifiers).
2. Exchange protocols use formats defined in XML Schema (or other technologies) that exploit the URIs from the ontology concepts.
3. Objects about which data is exchanged are identified using Globally Unique Identifiers."

This approach should ensure that, although exchanges between data producers and clients may make use of different XML formats, the items the metadata is about and the meaning of the data elements is common across all formats.

Focus on TAPIR data services

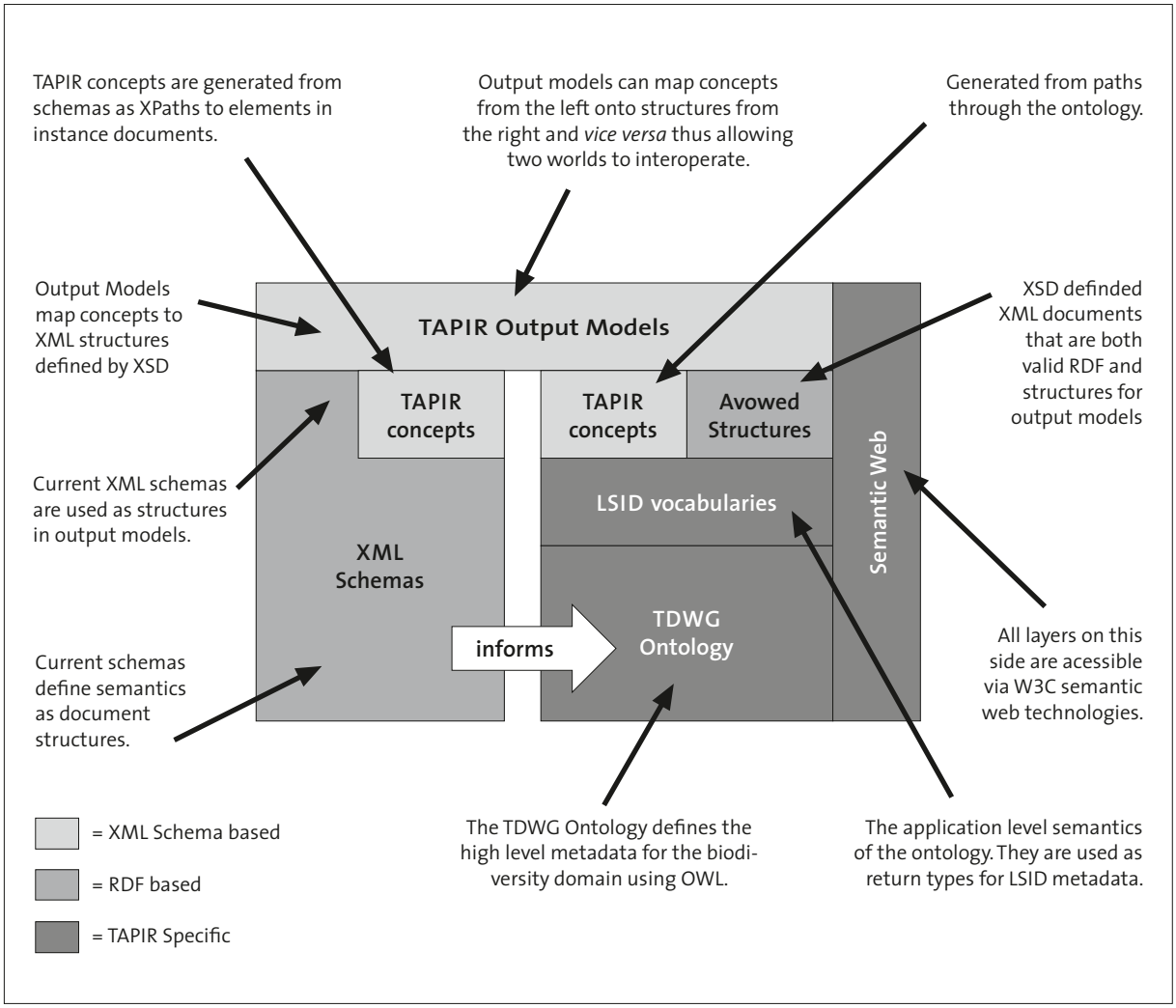
Secondly, the TDWG technical architecture is primarily, though not only, considered to be for data providers of the networks that use the TAPIR (TDWG Access Protocol for Information Retrieval) Web Service protocol for performing queries across distributed and heterogeneous data sources. TAPIR provides the means to query data suppliers based on conceptual schemas, query templates and output models that are usually defined by one or more federated networks.

When first developed, TAPIR was envisaged as a tool for unifying existing biodiversity data sharing networks that use the Global Biodiversity Information Facility (GBIF) accepted DiGIR and BioCASE protocols. However, TAPIR become such a generic product that its potential scope goes beyond biological observations and specimen collections,

also allowing for interoperability with geological, geospatial, ecological, climate, gene sequence and other data providers.

A suggested bridge between "two worlds"

In the TDWG Technical Roadmap 2007 and on http://wiki.tdwg.org/TAG a figure of the technical architecture is shown that confirms its focus on TAPIR data sources.



Source: TDWG Technical Architecture Subgroup Wiki, http://wiki.tdwg.org/TAG

The figure explains how a bridge between the "two worlds" of XML Schema based data provision and RDF/OWL based data integration could be realised. TAPIR output models (custom response types) would serve as the mapping point between concepts – on the left hand, concepts of RDF instance documents (generated from XML Schemas) and, on the right hand, concepts from the RDF/OWL based TDWG ontology and LSID metadata vocabularies (a more detailed explanation of this suggested approach is provided in TDWG TAG 2007).

The model assumes that content providers form an organised network where consumers pull data directly from data nodes using mutually agreed upon protocols. Other Semantic Web models such as, for example, used in the SPIRE project (Parr et al. 2006), build on the distributed provision of OWL documents that are indexed by semantic search engines like Swoogle. Documents relevant to a project are then captured and aggregated, and queried with SPARQL (the Semantic Web query language).

13.2 Ontology development and implementation by research projects

Many ongoing ontology developments

There are many ontologies that have been developed in research projects. Indeed, ontologies abound, however, many remain in an embryonic stage, because funding for the research project is drying out or there has not been the intention to go beyond a prototype ontology and serve a practical application of a "real world" user community.
The sections below briefly describe some noteworthy ontologies in Web Ontology Language (OWL) that are of interest to the fields of natural history and biodiversity. Some of them also have been used in a practial application.

OBO and OWL

The National Center for Biomedical Ontology's BioPortal 2.0 lists 111 ontologies that have been developed either in OBO format (representing the larger part) or Web Ontology Language (OWL-DL or OWL Full).
The OBO (Open Biomedical Ontologies) Flat File Format Specification, an ontology language originally designed for the Gene Ontology (GO), is widely used in the biomedical domain. The OBO Foundry is a community platform of OBO developers (http://www.obofoundry.org).
Also a number of ontologies that are relevant for the fields of natural history and biodiversity have been developed in OBO. For example, the ontology of the NCBI (National Center for Biotechnology Information) organismal classification has been developed in OBO (see: NCBI Taxonomy Browser). The classification uses a class hierarchy and includes terms for taxonomic ranks and a special relation type (has_rank) that links each taxonomic name to its appropriate rank term.
An other example is the Teleost Taxonomy Ontology (TTO) that is used in the Phenoscape project (http://phenoscape.org). Phenoscape is developing methods for comparing species that combine genomics and morphology. They use the TTO for taxonomic names to construct statements using terms from several ontologies that describe characters observed in the fish taxonomy literature. (Midford 2008)

Focus on OWL ontologies

In the biomedical sector, OWL has gained prominence through the development of several large ontologies such as the Biological Pathways Exchange (BioPAX) ontology, the GALEN ontology and the Foundational Model of Anatomy (FMA).
Recently there also have been efforts to establish an exact relationship between OBO and OWL and to develop applications that enable interoperability between OBO and Semantic Web tools and systems. (Davis et al 2007; Golbreich et al. 2007)
In our selection of ontologies we only include examples that have been developed in OWL. Some further examples may be found in a recent review of ontology development efforts (framework, domain-specific and other approaches) in the field of ecology. (Madin et al. 2008)
The examples below are arranged according to the level of application they show, from no identifiable application to experimental and on to working prototypic applications (the latter are described in more detail).

13.2.1 Ontogenesis Animal Behaviour and Animal Welfare ontologies

Ontogenesis

The Ontogensis project is a UK-based network of excellence to foster the creation, ontogeny and evolution of biological ontologies that has started in October 2006 and receives funding from the Engineering and Physical Sciences Research Council (EPSRC).

Ontologies

In the Ontogenesis project, so far two OWL ontologies have been developed for the description of biological research data: Animal Behaviour Ontology (ABO) and Animal Welfare Ontology (AWO).
The ontologies are available at http://ontogenesis.ontonet.org/moin/AnimalBehaviourOntologyDevelopment

Website

Ontogenesis network, http://www.ontonet.org

13.2.2 NESCent evolutionary informatics Comparative Data Analysis Ontology

Ontology purpose

Members of the NESCent (National Evolutionary Synthesis Center) evolutionary informatics working group have created the Comparative Data Analysis Ontology (CDAO) to facilitate the development of interoperable systems that support evolutionary comparative analysis.

Status of development

OWL-DL has been used to formalise key concepts and relations in evolutionary analysis, focusing on phylogenetic trees, character data, operational taxonomic units and evolutionary transitions.
The onotology has been subjected to some simple tests of representation and reasoning and is intended to be used in projects dedicated to establish interoperability of sequence family data resources.

Websites

Ontology page: http://www.evolutionaryontology.org
NESCent Evolutionary Informatics WG, https://www.nescent.org/wg_evoinfo/Main_Page

13.2.3 SEEK Extensible Observation Ontology

SEEK project brief

The Science Environment for Ecological Knowledge (SEEK) is a five year project funded by the National Science Foundation (USA) to create e-infrastructure for ecological, environmental, and biodiversity research. The motivation for this project is to remove problems that are encountered with accessibility and integration of large-scale biocomplexity data in the ecological sciences.
The SEEK participants are building EcoGrid, an integrated data grid of modular components for storage, sharing, access and analysis of a variety of ecological and biodiversity data. SEEK uses LSIDs to uniquely identify resources and store them on the EcoGrid. Furthermore, analytical tools are developed to allow an efficient use of the data stores. A middleware system using semantic technologies facilitates integration, reasoning over and synthesis of data and models used on EcoGrid. In particular, this system should be capable of determining whether relevant data and analytical components may be automatically transformed for use with a selected workflow.

SEEK ontologies

In the SEEK project OBOE, the Extensible Observation Ontology, has been developed using OWL-DL. This is a base ontology for generically describeing scientific observations and measurements. OBOE is now used to facilitate search and semi-automated integration of heterogeneous data of the Knowledge Network for Biocomplexity (KNB) repository (http://knb.ecoinformatics.org).

References

Bowers 2007; Madin et al. 2007; Schildhauer et al. 2008

Website

http://seek.ecoinformatics.org

13.2.4 BioImage system

BioImage – a Semantic Web image database

BioImage is an ontology-driven database for images of biological specimens. It has been developed by the Image BioInformatics Research Group of the Department of Zoology at the University of Oxford. Some development work for BioImage has been carried out in the framework of the EU FP5 project ORIEL (Online Research Information Environment for the Life Sciences; http://www.oriel.org).
Bioimage has been built using Jena and other Open Source components around an ImageStore ontology written in OWL-DL that describes all aspects of an image.

Features

BioImage simplifies manual metadata entry by dynamically creating from the underlying ontology simple Web form entry interfaces. If metadata already exists in digital

form, semi-automated entry is enabled. The metadata is saved in RDF format, a mechanism that eases migration to RDF data.
A semantically enhanced search interface allows for retrieval of relevant images. When interacting with Web services of other content providers, during the retrieval process textual descriptions of images are marked up on the fly with definitions of key terms.
The BioImage system also has been suggested for use as "a semantic data marshal" for laboratory information management and knowledge integration. The basis for this is that the system can handle different data types if an appropriate ontology is added and the data made available in RDF format.

References Catton et al. 2006; Shotton 2005

Website http://bioimage.ontonet.org/moin/FrontPage

13.2.5 Semantic WildNET

Semantic integration of ecological databases for biodiversity monitoring

Researchers from the University of Queensland, School of Information Technology and Electrical Engineering (Australia) have developed a biodiversity ontology in OWL and implemented a (prototype) system called Semantic WildNET.
Semantic WildNET applies Semantic Web/Grid technologies to integrate distributed ecological databases for purposes of biodiversity monitoring. It adds an ontology-based semantic search layer over the databases, enabling some automated reasoning. SPARQL, the query language for RDF, is combined with Google Maps to provide an intuitive mapping interface to query the integrated datasets.
Semantic WildNet at present provides a semantically-unified view of wildlife sighting data from the Environmental Protection Agency, species data from the Australian Museum and the National Herbarium; climate sensor data from the Bureau of Meteorology, and topographic maps from Geosciences Australia.

Reference Henderson, Khan and Hunter 2006

Website http://www.itee.uq.edu.au/~eresearch/projects/semanticwildnet/

13.2.6 SPIRE Evolutionary Trees and Natural History Ontology (ETHAN)

Context of development

SPIRE (Semantic Prototypes in Research Ecoinformatics) is a NSF funded project of several research groups investigating how Semantic Web applications can be used in the field of biodiversity. A research group at the University of Maryland (USA) has developed the Evolutionary Trees and Natural History Ontology (ETHAN).

Use case: Animal Diversity Web

ETHAN has been applied in a collaboration with the Animal Diversity Web (ADW). ADW is a large searchable online database of the University of Michigan's Museum of Zoology that holds descriptive texts, images of animal wildlife and museum specimens, sound recodings, and several hundred Quick Time Virtual Reality Movies that allow for exploring skulls in 3D.
ADW serves some 3000 Web pages of so called animal taxon accounts, mostly at the species-level. The backend is a relational database, TaxonDB (MySQL), that allows for taxon-based filtering of content. The taxonomic backbone has been constructed from a variety of sources including ITIS, Mammal Species of the world, EMBL reptile database and the Complete Checklist of the Birds of the World.

ETHAN

The Evolutionary Trees and Natural History Ontology (ETHAN) has been developed to provide a semantic layer on top the TaxonDB database. ETHAN actually combines two OWL ontologies:
The "Evolutionary Tree" is an OWL document of several hundred thousand scientific

names of species and higher taxonomic levels from the ADW TaxonDB which are represented in a simple class hierarchy. For example, Corvus corax (Northern raven) is a subclass of Corvus which itself is a subclass of Corvidae.
The "Natural History" part is a more complex OWL ontology that defines a set of behavioural and natural history concepts related to taxa as well as relationships between those concepts. It covers physical and reproductive description categories and quantitative measures such as body mass, metabolic rates, life spans, etc.
The categorical descriptors of habitats, reproductive behaviour and life history characteristics are represented as classes and class hierarchies. Their function is to facilitate organising taxa into groups that share a particular characteristic. Numerical measures are handled with annotation properties, which in OWL are associated only with a specified class.
The taxons from the ADW database are made a subclass of categorical desciptors. For example, Corvus corax is a subclass of "NearcticThing", "ThingWithSexualDimorphismSexesAlike" and other such descriptors.
ETHAN OWL taxon documents are created by associating categorical descriptors and measures to the appropriate scientific names of animals in the taxonomic hierarchy. Such documents provide the semantic layer for the taxon-related information in the ADW database. This information is retrieved from the database in XML format by the taxon name.
ADW's taxonomic backbone has been chosen to allow for immediate utility of the SPIRE project, however, it may be replaced by other taxonomic trees or phylogenetic structures. Easy replacability or merging with other RDF/OWL based resources was also the reason behind keeping the "Evolutionary Tree" separated from the "Natural History" part of ETHAN.

Available ADW OWL documents

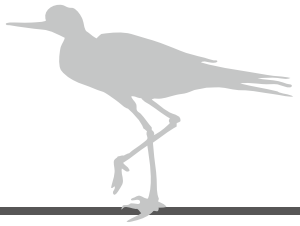
As one result of the collaboration with the ETHAN project, ADW since November 2006 has been providing OWL documents of all of its animal taxon accounts.
On the ADW website these are the "Information" pages for the taxons. At the bottom of these pages there is a "Get OWL" button which runs the transformation script that generates the OWL document. This allows for semantic search engine crawlers such as Swoogle to regularly capture and index these documents.
Since September 2008 the Animal Diversity Web also provides resources to the Encylopedia of Life project.

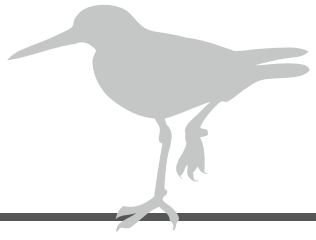
References

Parr et al. 2006 and 2008

Websites

SPIRE research group at at the University of Maryland, http://spire.umbc.edu/us/
Animal Diversity Web, http://animaldiversity.ummz.umich.edu





Part C:
Annexes and literature

14 Annex 1: Selected natural history and biodiversity metadata standards

There are many metadata standards used in the fields of natural history and biodiversity. Below we briefly describe two important standards that have been mentioned in this report several times, Darwin Core and ABCD (Access to Biodiversity Collections Data). Furthermore, we include the Ecological Metadata Language (EML) which is becoming increasingly popular.

14.1 Darwin Core

Brief description

Darwin Core is a metadata standard for describing the objects contained within natural history specimen collections and species observation databases. The Darwin Core (DwC) elements set consists of only 44 elements to simplify data interchange, however, it can be extended with additional elements. There are some standard DwC extensions available (Curatorial, Geospatial, Paleontology and Interaction Extensions), but, also elements of other metadata standards may be used to extend DwC that are appropriate for describing an organism occurrence.

Examples of use

Darwin Core is the single most used natural history and biodiversity data exchange standard in the world, exchanging over 140 million records from 3,000 datasets within the Global Biodiversity Information Facility (GBIF) network alone.
For example, DwC is used by the (US) National Biological Information Infrastructure (NBII) in conjunction with the Distributed Generic Information Retrieval (DiGR) protocol to harvest information from museum collections databases in the United States. This information is also made available through the GBIF portal. More than 36 million DwC-compliant specimen records have been provided to GBIF in this way.
The Avian Knowledge Network (AKN), http://www.avianknowledge.net, a network of North American institutions dedicated to the ecological study of bird populations, uses an extension of the Darwin Core schema (called Bird Monitoring Data Exchange). AKN nodes have contributed so far over 50 million observation records, mainly generated through broad-scale surveys.
The Ocean Biogeographic Information System (IOBIS), http://iobis.org, provides access to 16 million records of 102,000 species from 441 databases using DwC.

Websites

TDWG: DarwinCore Group – DwC, http://www.tdwg.org/activities/darwincore/
Darwin Core Wiki: http://wiki.tdwg.org/twiki/bin/view/DarwinCore/WebHome

14.2 ABCD (Access to Biodiversity Collections Data)

Brief description

ABCD is a comprehensive standard that contains about 700 elements for describing specimen, observation and other primary biodiversity data in great detail.
ABCD supports all of the information included in Darwin Core but aims to serve more complex requirements of occurrence and other descriptions. Whereas Darwin Core has a flat structure of elements, ABCD has a hierarchical structure that supports repeating elements and complex types.
ABCD was developed as a standard by a CODATA/TDWG task group with major input from the BioCASE (Biological Collection Access Service for Europe) network (11/2001-01/2005) and ENHSIN, the European Natural History Specimen Information Network (01/2000-12/2003). It was formally accepted as a standard by the Taxonomic Databases Working Group in 2005.

Examples of use

ABCD is used for data transmission in the BioCASE network, the European transnational network of biological collections of all kinds. It also has been accepted by the Global

Biodiversity Information Facility (GBIF) together with the BioCASE data transmission protocol (as an alternative to the DiGIR protocol).
Hence, ABCD data can be shared with GBIF but also, for example, Bioversity International, Atlas of Living Australia and many other networks.

Websites http://www.biocase.org
http://www.tdwg.org/activities/abcd/

14.3 Ecological Metadata Language (EML)

Brief description The Ecological Metadata Language (EML) is a metadata specification for use with ecological data. It includes elements intended to capture information on the taxonomic and geographic scope of a data set, and on any methods which went into the data capture.
EML has been developed in an open content, community oriented project and is currently maintained by the Knowledge Network for Biocomplexity (KNB).
EML is implemented as a series of XML documents that can by used in a modular and extensible manner to document ecological data. Each EML module is designed to describe one logical part of the total metadata that should be included with any ecological dataset.

Examples of use EML has been adopted in 2003 by the US Long Term Ecological Research Network as the official standard of the LTER Network and their Network Information System.
Other major organisations and projects such as the Global Biodiversity Information Facility (GBIF) and the Atlas of Living Australia (ALA) consider to use EML as the preferred metadata specification for ecological data. The expressiveness, modularity, extensibility, supporting software and community uptake of EML is clearly recognised. (GBIF / Tuama 2008)

Morpho Morpho is a dedicated open source software program for creating and managing EML (XML) data packages.

Websites http://knb.ecoinformatics.org/software/eml/
Detailed further information on EML is provided at: http://knb.ecoinformatics.org/software/eml/eml-2.0.1/eml-faq.html
Morpho, http://knb.ecoinformatics.org/software/morpho/

15 Annex 2: Environmental and biodiversity thesauri available in SKOS format

15.1 General Multilingual Environmental Thesaurus (GEMET)

GEMET has been developed from about 1995 onwards by the European Topic Centre on Catalogue of Data Sources (ETC/CDS) under contract to the European Environment Agency (EEA), and is currently managed by the European Environment Information and Observation Network (EIONET).

A core of general terminology for the environment

GEMET has been conceived as a general thesaurus aimed to provide a core of terminology for the environment. Specific thesauri and descriptor systems (e.g. for nature conservation) are not included, but were taken into account with regard to their general structure and upper level terminology.
GEMET currently is available in 26 languages.
GEMET contains over 6,000 descriptors which have been "arranged in a classification scheme made of three super-groups, 30 groups plus 5 accessory, instrumental groups. Each descriptor has been arranged in a hierarchical structure headed by a Top Term. The level of poly-hierarchy, i.e. the allocation of a descriptor to more than one group, has been kept to a minimum. Further, to allow a thematic retrieval of terms thematically related but scattered in different groups, a set of 40 themes have been agreed upon with the EEA and each descriptor has been assigned to as many themes as necessary. Thus, the user can access the thesaurus through the group-hierarchical list, through the thematic list or through the alphabetical list. As a complement to the hierarchical 'vertical' relations, an exhaustive series of strong 'horizontal' relations between terms (RT, Related Terms) have been introduced." (GEMET: About GEMET [2001], 2008)
In general, GEMET follows the ISO norms on monolingual and multilingual thesauri, however, the "group" and "theme" constructs are non-standard thesaurus constructs. This means that in order to express them in RDF a schema extension had to be made.

Availability

GEMET is freely available in several formats: It can be browsed and searched on-line, accessed through Web services. and downloaded as XML (RDF/SKOS) files.
For each of the different language versions of GEMET there is an XML file available for download. These files share the same markup structure and element names, only the element contents change with the language. The XML files are available from http://www.eionet.europa.eu/gemet/rdf
GEMET Web services: GEMET's data is exposed through the Web for remote applications using XML (RDF/SKOS), HTTP and XML/RPC. The Web service API for XML/RPC and HTTP is currently undergoing a change; for the proposal see: Zope/Plone products for EEA: Proposal for a new GEMET webservice API, https://svn.eionet.europa.eu/projects/Zope/wiki/GEMETWebServiceAPI

Website

http://www.eionet.europa.eu/gemet

15.2 CSA/NBII Biocomplexity Thesaurus Web Services

The Biocomplexity Thesaurus was developed through a partnership between the (US) National Biological Information Infrastructure (NBII) and CSA (Cambridge Scientific Abstracts) and launched in May 2003.

A merging of six thesauri

To create the Biocomplexity Thesaurus, the terminology of six thesauri has been merged, vetted and reconciled. These thesauri include the initial CERES/NBII Thesaurus (California Environmental Resources Evaluation System) and five CSA thesauri for the fields of Life Sciences, Aquatic Sciences and Fisheries, Sociology, Ecotourism Sciences and Pollution. In this work, the MultiTes 8.0 thesaurus development software package was used. (cf. ASIS&T 2003)

The Biocomplexity Thesaurus is a living resource that is updated regularly based on the decision of a thesaurus working group that reviews suggested additions and modifications (e.g. from the NBII nodes).

Supportive functions within NBII

The thesaurus supports NBII information services in a number of ways, for example, it provides subject metadata for resource indexing, drives the selection of literature citations from the CSA Internet Database Service, and aids searching within the My.NBII Portal (Intranet) which uses Plumtree technology. (cf. Zolly 2004)

Publicly available term search facililty

Moreover the Thesaurus is freely accessible for term searching at http://thesaurus.nbii.gov. The lookup tool performs automatic stemming for prefixes and suffixes, and the thesaurus can be "rotated" to examine facets of a particular concept.

Thesaurus Web services

For application developers, the CSA/NBII Biocomplexity Thesaurus is available via a Web services (SOAP) interface. Using the services an external application can query the thesaurus for matching terms, retrieve all related terms, or retrieve only terms related in specific ways (e.g. broader terms only).
The use of SKOS is currently in demonstration phase. Two demonstration Web service clients are offered, one for the NBII Thesaurus and one for the EIONET Multilingual Thesaurus as provided via NBII.
Developer resources offered comprise the SWAD Europe SKOS Service API, sample client source code for the demonstrators, and a customised SKOS JavaDoc. The latter adds to the SKOS Core API a convenience method getConceptResultsByKeyword, which speeds up search results that are returned from the web service.
The resources mentioned above are to be found under "Web Service" at http://thesaurus.nbii.gov/portal/server.pt

Website

http://nbii-thesaurus.ornl.gov/thesaurus/

15.3 CAIN Invasive Species Management Thesaurus

Brief description

The Invasive Species Management Thesaurus has been published by the Information Center for the Environment of the University of California, Davis.
The Centre hosts the website of the California Information Node (CAIN) of the National Biological Information Infrastructure (NBII) and provides access to a variety of data and information on areas such as biodiversity, invasive species, land use, and water quality.
The Invasive Species Management Thesaurus is a small thesaurus of 121 terms that is available in English, Spanish and Portuguese. The thesaurus has a rather flat structure (only one sub-level) and is most detailed with regard to types of habitats and species.
The thesaurus is available in SKOS format from the CAIN website.

Website

http://cain.ice.ucdavis.edu/thesauri/ismt/

16 Annex 3: Natural history and biodiversity organisations, projects and resources

16.1 Selected major European natural history and biodiversity organisations and projects

Below we include descriptions of some major European organisations and projects mentioned in the report that merit to be highlighted and described in appropriate detail.

16.1.1 Consortium of European Taxonomy Facilities (CETAF)

Brief description of the consortium

CETAF, the Consortium of European Taxonomy Facilities, was founded in 1996 by ten of the largest European taxonomic institutions (natural history museums, botanic gardens and other biological collections) to promote scientific research and access to European collections. Today CETAF represents 28 members and is the voice for taxonomy and systematic biology in Europe.

According to its self-definition, "CETAF strives to maximise the benefits that its member institutions can provide for the sustainable use of biodiversity in Europe and elsewhere in the world; coordinate work around the field of taxonomy with other institutions, and improve Europe's capacity to fulfil its commitments and obligations in taxonomy under European and international initiatives such as the Global Biodiversity Information Facility (GBIF) and the Global Taxonomic Initiative (GTI) as well as conventions (especially the CBD)."

CETAF also has initiated large European research and e-infrastructure projects such as SYNTHESYS (EU-FP6, integrated infrastructure initiative, 2004-2009) and EDIT (EU-FP6, network of excellence, 2006-2011).

Website http://www.cetaf.org

16.1.2 European Distributed Institute of Taxonomy (EDIT)

Project brief

EDIT, the European Distributed Institute of Taxonomy, is a EU-FP6 Network of Excellence project (03/2006-02/2011) that brings together an international consortium of 28 institutions. EDIT institutions represent around 30% of the world's taxonomic collections that are at the forefront of the development of state-of-the-art databases, information networks, and large-scale and specialised instrumental facilities (e.g. remote microscopy, DNA barcoding, etc.). The project aims to reduce fragmentation of European taxonomic research within the European Research Area and create a virtual centre, which will increase both the scientific basis and capacity for biodiversity conservation.

EDIT's eight work packages comprise: Coordination and Management, Integrating and Reshaping the Expert and Expertise Basis, Integrating Research Strategies and Liaison to Users of Taxonomy, Internet Platform for Cybertaxonomy: Tools, Sharing, Networking and Integration, Unifying Revisionary Taxonomy on the web, Applying Taxonomy to Conservation, and Training and Public Awareness.

An overview of the set of tools and services that are developed and implemented in the EDIT Platform for Cybertaxonomy is given in Müller et al. 2008. One example is the Virtual Taxonomic Library (ViTaL) that aims to leverage the discovery and accessibility of taxonomically relevant literature. Also Scratchpads are a component of this platform (see section 10.2).

Recently, EDIT has issued a scientific vision for the future of taxonomy in the next 10 to 20 years. The document emphasises: "Although an ever expanding repertoire of theoretical and practical tools is available to taxonomists (...), there will have to be substantial, even radical, changes in how taxonomy is done and its supporting infrastructure

operated, to exploit these opportunities to the full. 'Business as usual', even if scaled up, is simply not an option." (EDIT 2008)

Website http://www.e-taxonomy.eu

16.1.3 Synthesis of Systematic Resources (SYNTHESYS)

Project brief SYNTHESYS is an Integrated Infrastructure project initated by CETAF and funded under the EU FP6 (02/2004-07/2009; SYNTHESES II is expected to start in September 2009 and to run 5 years).
The current project provides transnational access grants to 20 CETAF natural history museums and botanical gardens, facilitates the creation of a virtual museum service, sets standards for collection management and databases, promotes best practice by offering training and workshops, and provides guidelines for the care, storage and conservation of collections.

Website http://www.synthesys.info

16.1.4 Biological Collection Access Service for Europe (BioCASE)

Development of unified access to Europe's biological databases The BioCASE network was established by an EU-funded project (11/2001-01/2005) that prepared unified access to distributed and heterogeneous European collection and observational databases. The project promoted using open-source, system-independent software and open data standards and protocols.
During the BioCASE project a network was formed by partners from 31 countries who in a first step provided meta-information on thousands of biological collections and then, in a second step, established a unit-level data access network (i.e. data of individual specimen or observation records).

BioCASE technologies Technologies developed by the project include the BioCASE protocol and the BioCASE provider software. These technologies make it possible to connect arbitrarily structured databases to the BioCASE network and the Global Biodiversity Information Facility (GBIF has accepted the BioCASE provider package for unit data as part of its standard services). The BioCASE network uses the ABCD standard for data transmission.

Supporting projects The development of BioCASE has been supported by several other projects, of which ENHSIN, the European Natural History Specimen Information Network (EU-FP5, 01/2000-12/2003) and ENBI, the European Network for Biodiversity Information (EU-FP5, 03/2002-02/2006) are two more recent ones.
Currently BioCASE is supported by SYNTHESYS and EDIT (see separate entries). EDIT aims to integrate the BioCASE portal into the EDIT Internet Platform for Cybertaxonomy.
In March 2008, the BioCASE portal was launched that both contributes to and builds upon the global efforts in biodiversity informatics led by the Global Biodiversity Information Facility (GBIF).

Website http://www.biocase.org

16.1.5 Pan-European Species directories Infrastructure (PESI)

Project brief PESI is a three-year research infrastructure project (05/2008-05/2011) funded under the EU-FP7 Capacities Work Programme. PESI is coordinated by the Zoological Museum Amsterdam and involves 40 partner organisations from 26 countries.
PESI will coordinate the integration and synchronisation of the European taxonomic information systems (species names directories) in Europe into a joint e-infrastructure

that leverages the management of biodiversity in Europe. More specifically, it aims to integrate the three main all-taxon registers in Europe, the Euro+Med PlantBase, the European Register of Marine Species, and Fauna Europaea.
A description of the wider scope of the project is provided on the PESI website)

Website http://www.eu-nomen.eu/pesi

Related websites European Register of Marine Species, http://www.marbef.org/data/erms.php
Euro+Med PlantBase, http://www.emplantbase.org
Fauna Europaea, http://www.faunaeur.org

16.1.6 LifeWatch

Project brief LifeWatch is a three year project (02/2008-01/2011) funded under the EU FP7 for preparing a European Research Infrastructure for global biodiversity research. This infrastructure has been identified by the European Strategy Forum on Research Infrastructures (ESFRI) to be supported by the Member States of the European Union. Currently 19 countries have expressed interest in the initiative.
LifeWatch investigates and prepares the required infrastructure for global biodiversity research, linking biodiversity data from ecological monitoring in marine and terrestrial environments to data in physical collections such as natural history museums and botanical gardens. The infrastructure should give users access to large data sets from different levels of biodiversity - genetic, population, species and ecosystem - together with analytical and modelling tools. A first cost estimate for building and maintaining the Research Infrastructure is € 1.5 billion over 25 years.
(For some background information see Berendsohn and Gebhardt 2008; Van Waeyenberge 2008)

Website http://www.lifewatch.eu

16.2 List of natural history and biodiversity organisations, projects and resources mentioned

Amphibia Web
http://amphibiaweb.org
Animal Diversity Web (ADW)
http://animaldiversity.ummz.umich.edu
AntWeb, Hymenoptera Name Server
http://www.antweb.org
AquaRing
http://www.aquaringweb.eu
Aquarium of Genoa
http://www.acquariodigenova.it
ARKive – Images of Life on Earth
http://www.arkive.org
Assembling the Tree of Life (AToL)
http://atol.sdsc.edu
Atlas of Living Australia (ALA)
www.ala.org.au
Avian Knowledge Network (AKN)
http://www.avianknowledge.net
Avibase – the world bird database
http://www.bsc-eoc.org/avibase/avibase.jsp
BioCASE – Biological Collection Access Service for Europe
http://www.biocase.org

Biodiversity Collections Index (BCI)
http://www.biodiversitycollectionsindex.org
Biodiversity Heritage Library (BHL)
http://www.biodiversitylibrary.org
BioImage project, Image BioInformatics Research Group, Department of Zoology, University of Oxford
http://bioimage.ontonet.org/moin/FrontPage
Biologia Centrali-Americana (BCA) electronic
http://www.sil.si.edu/digitalcollections/bca/
BioNET International
http://www.bionet-intl.org
Biorepositories.org
http://biorepositories.org
Birding.com
http://www.birding.com
CABI Bioscience
http://www.cabi.org
CAIN Invasive Species Management Thesaurus
http://cain.ice.ucdavis.edu/thesauri/ismt/
Catalogue of Life (CoL)
http://www.catalogueoflife.org
Consortium of European Taxonomy Facilities (CETAF)
http://www.cetaf.org
Convention on Biological Diversity
http://www.biodiv.org
Cornell Lab of Ornithology, Birds of North America (BNA)
http://bna.birds.cornell.edu/bna
Creating a Taxonomic e-Science (CATE)
http://www.cate-project.org
CSA/NBII Biocomplexity Thesaurus
http://nbii-thesaurus.ornl.gov/thesaurus/
Digital Morphology
http://digimorph.org
Digitaltaxonomy
http://digitaltaxonomy.infobio.net
DiscoverLife
http://www.discoverlife.org
EMBL reptile database
http://www.reptile-database.org
ENBI – European Network for Biodiversity Information: Digital Imaging of Biological Type Specimens.
A Manual of Best Practice. Häuser, C.L. et al., Stuttgart 2005
http://circa.gbif.net/Public/irc/enbi/comm/library?l=/enbi_reports/haeuser_digital/_EN_1.0_&a=i
Encyclopedia of Life (EOL)
http://www.eol.org
Encyclopedia of Life (EOL), LifeDesks
http://lifedesk.eol.org
Erudite Recorded Botanical Information Synthesizer (HERBIS)
http://www.herbis.org
EUNIS Habitat Classification, European Environment Agency
http://eunis.eea.europa.eu/habitats.jsp
Euro+Med PlantBase
http://www.emplantbase.org
European Distributed Institute of Taxonomy (EDIT)
http://www.e-taxonomy.eu
European Natural History Specimen Information Network (ENHSIN)
http://www.nhm.ac.uk/research-curation/research/projects/enhsin/index.html
European Network for Biodiversity Information (ENBI)
http://www.enbi.info

European Network for Science Centres and Museums (ECSITE)
http://www.ecsite.net
European Register of Marine Species
http://www.marbef.org/data/erms.php
European Virtual Anthropology Network (EVAN)
http://evan.at
Fauna Europaea
http://www.faunaeur.org
Field Guide: Birds of the World
http://www.flickr.com/groups/birdguide/
Field Museum of Natural History
http://www.fieldmuseum.org
Fishbase
http://www.fishbase.org
Genbank
http://www.ncbi.nlm.nih.gov/Genbank/
General Multilingual Environmental Thesaurus (GEMET)
http://www.eionet.europa.eu/gemet
Global Biodiversity Information Facility (GBIF)
http://www.gbif.org
Global Biodiversity Information Facility (GBIF), list of 27 global taxonomic databases
http://www.gbif.org/links/taxo
Global Biodiversity Information Facility (GBIF), Seed Money programme (DIGIT, ECAT)
http://www.gbif.org/prog
Global Biodiversity Information Facility (GBIF): Training Manual 1: Digitisation of Natural History Collections Data. Version 1.0. Copenhagen, 2008
http://www.gbif.org/GBIF_org/GBIF_Publications/trainingmanual1/index_html
Global Taxonomic Initiative (GTI)
http://www.cbd.int/gti/
Index Fungorum
http://www.indexfungorum.org
Index Herbariorum
http://sciweb.nybg.org/science2/IndexHerbariorum.asp
Insect and Spider Collections of the World (ISCW)
http://hbs.bishopmuseum.org/codens/
Integrated Open Taxonomic Access (INOTAXA)
http://www.inotaxa.org
Integrated Taxonomic Information system (ITIS),
http://www.itis.gov
International Plant Names Index (IPNI)
http://www.ipni.org
International Union for Conservation of Nature and Natural Resources (IUCN), Red List of Threatened Species
http://www.iucnredlist.org
iSpecies
http://ispecies.org
Knowledge Network for Biocomplexity (KNB)
http://knb.ecoinformatics.org
LifeWatch
http://www.lifewatch.eu
Lithuanian Sea Museum
http://www.muziejai.lt/Klaipeda/juru_muziejus.en.htm
Long Term Ecological Research Network
http://www.lternet.edu
Mammal Species of the world
http://vertebrates.si.edu/mammals/msw/
Marine Biological Laboratory
http://www.mbl.edu

Missouri Botanical Garden
http://www.mobot.org
Morphbank
http://www.morphbank.net
Morphster
http://www.morphster.org
National Center for Biomedical Ontology, BioPortal 2.0
http://bioportal.bioontology.org/ontologies
National Center for Biotechnology Information (NCBI), Taxonomy Browser
http://www.ncbi.nlm.nih.gov/Taxonomy/
Natural History Museum (NHM)
http://www.nhm.ac.uk
Nausicaa – the French National Sea Experience Centre
http://www.nausicaa.fr
NESCent (National Evolutionary Synthesis Center), Evolutionary Informatics WG
https://www.nescent.org/wg_evoinfo/Main_Page
Northern Temperate Lakes - Long Term Ecological Research Network (USA)
http://lter.limnology.wisc.edu/
OBO (Open Biomedical Ontologies) Foundry
http://www.obofoundry.org
Online Research Information Environment for the Life Sciences (ORIEL)
http://www.oriel.org
Ontogenesis
http://www.ontonet.org
Pan-European Species directories Infrastructure (PESI)
http://www.eu-nomen.eu/pesi/
Peabody Museum of Natural History at Yale University
http://www.peabody.yale.edu
Phenoscape project
http://phenoscape.org
Plazi.org
http://plazi.org
Rotterdam Zoo
http://www.rotterdamzoo.nl
Royal Belgian Institute of Natural Sciences
http://www.naturalsciences.be
Royal Botanic Gardens, Kew
http://www.kew.org
Royal Museum for Central Africa (RMCA), Belgium
http://www.africamuseum.be
Science Environment for Ecological Knowledge (SEEK)
http://seek.ecoinformatics.org
Scratchpads
http://scratchpads.eu
Semantic Prototypes in Research Ecoinformatics (SPIRE)
http://spire.umbc.edu/us/
Semantic WildNET
http://www.itee.uq.edu.au/~eresearch/projects/semanticwildnet/
Smithonian National Museum of Natural History
www.mnh.si.edu
Species 2000 programme
http://www.sp2000.org
Synthesis of Systematic Resources (SYNTHESYS)
http://www.synthesys.info
Taxonomic Database Working Group (TDWG), database: Biodiversity Information Projects of the World
http://www.tdwg.org/biodiv-projects/projects-database
Taxonomic Database Working Group (TDWG), Technical Architecture Group

http://wiki.tdwg.org/TAG

Taxonomic Search Engine (TSE)

http://darwin.zoology.gla.ac.uk/~rpage/portal/

Tree of Life

http://tolweb.org/tree/

TreeBase

http://www.treebase.org

uBio – Universal Biological Indexer and Organizer

http://www.ubio.org

University of Texas UTCT Data Archive

http://utct.tacc.utexas.edu

World Ocean Network

http://www.worldoceannetwork.org

ZipcodeZoo

http://zipcodezoo.com

ZooBank

http://www.zoobank.org

Zoological Museum of the University of Amsterdam (ZMA), bird collection, 3D images of type specimens

http://ip30.eti.uva.nl/zma3d/

17 Annex 4: Cultural heritage organisations, projects and resources

17.1 Selected projects related to the EDL initiative

This section documents some selected projects that are related to the European Digital Library initiative. Some of them are recently started projects under the eContentplus programme, some precursors of these projects, which have developed an important stock of knowledge, tools and expertise to build on (e.g. with regard to multi-lingual access to library resources).
The newer large projects such as Athena and EuropeanaLocal are expected to prepare more institutions to collaborate with and contribute content to the European Digital Library. According to available presentations of these projects, they also intend to prepare participating institutions to contribute available thesauri, classification schemes or other knowledge organisation systems in SKOS format.
The projects included below of course are not all projects that relate in some way or other to the European Digital library initiative. Other such projects are among the ones that have been funded under the 2005, 2006 and 2007 calls of the *e*Content*plus* programme, in the areas of digital libraries and cultural and scientific/scholarly content (see Literature: eContentplus Programme: Projects)
Moreover, there are several related research and technological development projects that have been funded under the European Union's 6th and 7th Framework Programmes for Research and Technological Development (see Literature: European Commission, unit: Cultural Heritage and Technology Enhanced Learning, DigiCult).

17.1.1 Europeana

Project brief

Europeana, originally known as EDLnet (European Digital Library Network), is an *e*Content*plus* project (07/2007-06/2009) that realises a European Digital Library (EDL) prototype website which was officially launched on the 20th of November 2008.
The project is run by a core team based in the National Library of the Netherlands, Koninklijke Bibliotheek. It builds on the project management and technical expertise developed by The European Library (TEL), which is a service of the Conference of European National Librarians.
Overseeing the Europeana project is the EDL Foundation, which includes major European cultural heritage associations.

Project philosophy

The Europeana project among other objectives has been entrusted to find consensual technical solutions to interoperability issues of the European Digital Library (EDL).
Europeana supports the development of solutions to the interoperability of cultural and scientific heritage content held by European libraries, archives, museums and audiovisual collections in the context of the European Digital Library initiative. It is fully considered that no solution can be imposed from above and progress can only be made by consent. Also the Conference of European National Librarians (CENL) had to develop a clear collaborative framework for its members, defining how the members relate to each other in the context of their shared European online platform (called, The European Library – TEL, see below).
Such clarity may currently not exist between other types of libraries, museums, archives and audiovisual collections nor with the relevant associations representing these organisations. However, if a shared understanding is found among institutions from these domains, a technical dialogue can be established to find common solutions to interoperability.

Europeana content

The Europeana website gives users access to some 2 million digital objects, including film material, audio recordings, photographs, historic maps, books, manuscripts and

archival records; the intention is to by 2010 reach a volume of well over 6 million digital objects. The interface is intended to be multilingual, initially in French, English and German, but further languages should be included after the launch of the website.

Relevance to STERNA The STERNA partnership among other objectives aims to provide content/ metadata to the emerging European Digital Library, which is expected to build on the results of the Europeana project.
The Europeana project defines the technological roadmap for the European Digital Library (see section 2.2). The roadmap suggests SKOS as method of choice to create a data layer ready for semantic query methods. This includes that content holders will have to provide their controlled vocabularies in SKOS.

References Europeana project deliverables are available at: http://www.europeana.eu/outcomes.php
The deliverables D2.2 and D2.5 provide technical requirements for content providers to have their data integrated into Europeana.

Website http://www.europeana.eu

17.1.2 The European Library (TEL)

Project brief The European Library (TEL) is an online portal that provides access to electronic resources of most national libraries of Europe which cooperate in the Conference of European National Librarians (CENL). An important basis of TEL has been CENL's GABRIEL (GAteway and BRIdge to Europe's National Libraries) service that was integrated in TEL in 2005.
The TEL platform became the starting point for developing the envisioned European Digital Library, which is now showcased by Europeana. In view of making TEL an important organisational ground of the future European Digital Library, a number of already completed or ongoing projects have received funding from the European Commision: TEL-ME-MORE, EDLproject (see below) and TELplus (also included below).

Relevance to STERNA TEL is of general interest as an important organisational ground of the future European Digital Library and centre of a cluster of supporting projects. Some specific results of these projects are of interest to STERNA (see below).

Website http://www.theeuropeanlibrary.org

17.1.3 EDLproject

Project brief The *e*Content*plus* EDLproject (09/2006-02/2008) supported TEL to incorporate collection records of nine national libraries within the European Union/European Free Trade Association, thereby extending the grasp of the future European Digital Library. A technological focus point of EDLproject was the enhancement of multilingual capabilities of TEL's user portal.

Relevance to STERNA The EDL technological roadmap suggests to make use of domain-specific Dublin Core application profiles. A report of the EDLproject provides an interesting assessment of the metadata interoperability of TEL and discusses how such interoperability between museums, archives, audio-visual archives and libraries could be approached. The report draws on the work of the Metadata Sub-group of the European Commission's i2010 Interoperability Expert Group and consultations with the projects DISMARC (music archives) and VideoActive (historic TV content). (Chambers 2007)

Website http://www.edlproject.eu

17.1.4 TELplus

Project brief TELplus is an *e*Content*plus* project (09/2007-11/2009) that aims to strengthen, extend and improve the services of The European Library (TEL). Specifically it focuses on capturing through OCR the content of more than 20 million text pages in many languages, and on making library data OAI compliant and harvestable.

Relevance to STERNA In the TELplus project there is ongoing work on multi-lingual and semantic approaches under their work package 3: Improving Access. Multi-lingual subject access is explored building on experiences of the MACS (Multilingual Access to Subjects) project, that developed manually an alignment between parts of three library vocabularies: LCSH (English), Rameau (French) and SWD (German). The multilingual search system developed by MACS exploits equivalence links created among the three vocabularies.
TELplus wants to investigate how automated techniques can be applied to multi-lingual cases similar to the one explored by MACS. With respect to semantic access, strategies that are considered comprise converting vocabularies to SKOS and identifying semantic correspondences between subjects (semantic alignment). (cf. Isaac 2007b)
Also of interest are practical TELplus experiences with Optical Character Recognition methods and making library data OAI compliant.

Websites TELplus, http://www.theeuropeanlibrary.org/telplus/
MACS, https://macs.vub.ac.be/pub/

17.1.5 MICHAEL and MICHAELplus

Project brief The MICHAEL and MICHAELplus (Multilingual Inventory of Cultural Heritage in Europe, 2004-2008) projects were funded under the European Commission's eTen programme to develop a multilingual inventory service for digital resources from the cultural sector across Europe, in particular, resources related to national cultural portals. The MICHAEL European portal, launched in December 2006, allows users to search, browse and examine descriptions of resources held in institutions from across Europe.
Technical results of the projects include the MICHAEL data model for multilingual digital cultural heritage inventories, an open source technical platform for national instances (built on Apache Tomcat, Cocoon, XtoGen, XML etc.), and interoperability protocols for national instances to contribute data to the European service.
The MICHAEL platform supports interoperability on the schema, record and repository levels. Scalability is achieved through schema mapping techniques and metadata can be harvested using the OAI-PMH. The end-user can make cross-lingual queries to all the archives through the controlled vocabularies embedded in the platform.

Relevance to STERNA The XML-based Michael platform does not provide for semantic interoperability, however, there seem to be plans to upgrade the platform using Semantic Web technologies. Two methods have been considered for this: using SKOS vocabularies or applying ontology alignment techniques. (cf. Christaki et al. 2007)

References A concise description of the MICHAEL platform is to be found at http://www.michael-culture.gr/mpf/pub-mpf/about.html

Website MICHAEL European Service, http://www.michael-culture.org

17.1.6 Athena

Project brief Athena (Access to cultural heritage networks across Europe) is a*e*Content*plus* project (11/2008-10/2010) that builds on the achievements of the MINERVA (Ministerial Network for Valorising Activities in Digitisation), MINERVAplus and MINERVA-EC projects

as well as the MICHAEL projects. Additional technical work to MICHAEL includes to develop a set of plug-ins to be integrated within the EDL, facilitating access to, and re-use of, digital content of European cultural institutions.
Athena has partners from 22 European countries with a focus on museums.

Relevance to STERNA According to presentations of the project co-ordinator (cf. Caffo 2008a+b) the wide-ranging activities of Athena should ultimately enable any museum and other cultural institution wishing to share their data and get visibility through Europeana

- to map its metadata into domain-specific Dublin Core application profiles,
- publish existing terminologies and thesauri using SKOS and achieve semantic interoperability with the European Digitial Library,
- moreover the institutions should be enabled to describe their own content and services and make them discoverable using available MICHAEL inventory services.

Website http://www.athenaeurope.org

17.1.7 EuropeanaLocal

Project brief EuropeanaLocal (originally, EDLocal) is an *e*Content*plus* project (06/2008-05/2011). The project has a large partner network and aims to make accessible to Europeana over 20 million content items that are held by regional and local institutions across 27 countries. According to the project website, "EuropeanaLocal will work with the EDL Foundation to establish simple, efficient and sustainable processes through which local and regional institutions can easily make their content available to Europeana during and after the project. It will adopt and promote the use of Europeana's infrastructures, tools and standards, as specifications emerge – especially OAI-PMH repositories and Europeana Metadata Application Profiles initially, but moving forward to semantic web technologies later."

Relevance to STERNA EuropeanaLocal aims to allow European regional and local museums, archives and libraries to participate in the EDL initiative. Work with technical partners on the regional and local level will focus on conversion of metadata and controlled vocabulary and implementation of OAI-PMH repositories. An interesting outcome of the project may be a EuropeanaLocal prototype service, i.e. a service specifically adapted to the needs of regional and local institutions.

Website http://www.europeanalocal.eu

17.2 List of cultural heritage organisations, projects and resources mentioned

AHRC ICT Methods Network, UK
http://www.methodsnetwork.ac.uk
Art & Architecture Thesaurus (AAT), Getty Research Institute
http://www.getty.edu/research/tools/vocabulary/
Artchive
http://artchive.com
ATHENA – Access to cultural heritage networks across Europe
http://www.athenaeurope.org
BELIEF – Bringing Europe's Electronic Infrastructures to Expanding Frontiers
http://www.beliefproject.org
Bibliopolis
http://www.bibliopolis.nl
BRICKS – Building Resources for Integrated Cultural Knowledge Services
http://www.brickscommunity.org
Cantabria Cultural Heritage ontology
http://www.cidoc2008.gr/cidoc/Documents/papers/drfile.2008-06-18.1772912112

CASPAR – Cultural, Artistic and Scientific Knowledge Preservation, for Access and Retrieval
http://www.casparpreserves.eu
CIDOC Conceptual Reference Model (CRM) / ISO 21127:2006 – A reference ontology for the interchange of cultural heritage information • http://cidoc.ics.forth.gr
Conference of European National Librarians (CENL)
http://www.cenl.org
CONTRAPUNCTUS – Preservation and Unification of New and Existing Braille Music Digital Sources for a New Access Methodology • http://www.punctus.org
CulturaItalia
http://www.culturaitalia.it
DELOS – A Network of Excellence on Digital Libraries
http://delos-noe.iei.pi.cnr.it
DigiCULT Forum
http://www.digicult.info
Digitaal Erfgoed Nederland (Digital Heritage Netherlands)
http://www.den.nl
DILIGENT – A Digital Library Infrastructure on Grid Enabled Technology
http://www.diligentproject.org
DPE – Digital Preservation Europe
http://www.digitalpreservationeurope.eu
DRIVER – Digital Repository Infrastructure Vision for European Research
http://www.sherpa.ac.uk/projects/driver.htm
EASAIER – Enabling Access to Sound Archives through Integration, Enrichment and Retrieval
http://ww.easaier.org
EDLnet – European Digital Library Network
http://www.europeanlibrary.org
EDLproject
http://www.edlproject.eu
English Heritage, Centre for Archaeology
http://www.english-heritage.org.uk
English Heritage, National Monuments Record Thesauri
http://thesaurus.english-heritage.org.uk
ENRICH – European Networking Resources and Information concerning Cultural Heritage
http://enrich.manuscriptorium.com/
EPOCH – Excellence in Processing Open Cultural Heritage
http://www.epoch-net.org
Europeana
http://www.europeana.eu
EuropeanaLocal
http://www.europeanalocal.eu
FACET
http://www.comp.glam.ac.uk/~FACET
French National Library, Mandragore collection
http://mandragore.bnf.fr/html/accueil.html
Getty Research Institute
http://www.getty.edu/research/
Iconclass
http://www.iconclass.nl
IMPACT – Improving Access to Text
http://www.impact-project.eu
Instituut Collectie Nederland (Netherlands Institute for Cultural Heritage)
http://www.icn.nl
MACS – Multilingual Access to Subjects
https://macs.vub.ac.be
MDA Archaeological Objects Thesaurus
http://www.mda.org.uk/archobj/archcon.htm
MDA, Collections Trust

http://www.mda.org.uk

MEMORIES – Design of an Audio Semantic Indexation System Allowing Information Retrieval for the Access to Archive Content • http://www.memories-project.eu

MICHAEL and MICHAELplus – Multilingual Inventory of Cultural Heritage in Europe
http://www.michael-culture.org

MINERVA / MINERVAPlus / MINERVA EC – Ministerial Network for Valorising Activities in digitisation
http://www.minervaeurope.org

MultiMATCH – Multilingual/Multimedia Access to Cultural Heritage
http://www.multimatch.eu

MultimediaN E-Culture project
http://e-culture.multimedian.nl

Museo24
http://www.museo24.fi

MuseumFinland – Finnish Museums on the Semantic Web, research project
http://www.seco.tkk.fi/applications/museumfinland/

MuseumFinland portal
http://www.museosuomi.fi

National Board of Antiquities, Finland
http://www.nba.fi

National Library of the Netherlands, Koninklijke Bibliotheek
http://www.kb.nl

OCLC (Online Computer Library Center) terminology services project
http://www.oclc.org/research/projects/termservices/

PLANETS – Preservation and Long-term Access to our Cultural and Scientific Heritage
http://www.planets-project.eu

PrestoSpace – Preservation towards storage and access. Standardised Practices for Audio-visual Contents in Europe
http://prestospace.org

Reference Network Architecture (RNA) project
http://www.rnaproject.org

Rijksbureau voor Kunsthistorische Documentatie (RKD), Netherlands
http://website.rkd.nl

Rijksmuseum, Amsterdam (Aria Masterpieces collection)
http://www.rijksmuseum.nl

SCULPTEUR – Semantic and content-based multimedia exploitation for European benefit
http://www.sculpteurweb.org

STAR – Semantic Technologies for Archaeological Resources
http://hypermedia.research.glam.ac.uk/kos/star

STITCH – Semantic Interoperability to access Cultural Heritage
http://www.cs.vu.nl/STITCH/

TEL – The European Library
http://www.theeuropeanlibrary.org

TEL-ME-MORE
http://www.theeuropeanlibrary.org/portal/organisation/cooperation/archive/telmemor/

TELplus
http://www.theeuropeanlibrary.org/telplus/

TNT – The Neanderthal Tools
http://www.the-neanderthal-tools.org

Tropenmuseum, Amsterdam
http://www.tropenmuseum.nl

UNESCO Thesaurus
http://www2.ulcc.ac.uk/unesco/

Union List of Artist Names (ULAN), Getty Research Institute
http://www.getty.edu/research/tools/vocabulary/

Visual Resources Association (VRA) standard
http://www.vraweb.org

Volkenkunde
http://www.volkenkunde.nl

18 Literature

Agosti, Donat (2008): Access to Biodiversity Information: From Printed to Semantic, Enhanced e-Publications. NDAP International Conference, Taipei, March 19, 2008 http://www.ndap.org.tw/96AnnualExhibition/InternationalConference/files/20080319_taipei_agosti.pdf

AHRC ICT Methods Network (2008): Case study 13: STAR – Semantic Technologies for Archaeological Resources http://www.methodsnetwork.ac.uk/resources/casestudy13.html

Aitchison, J., Gilchrist, A. and David, D. (2000): Thesaurus construction and use: a practical manual. London: Aslib IMI

Antbase (2008): Launch of Plazi.org (February 27, 2008) http://antbase.blogspot.com/2008/02/launch-of-plaziorg.html

Antoniou, Grigoris and Van Harmelen, Frank (2004): A Semantic Web Primer. Cambridge, Mass.: MIT Press 2004

AOL/Marketwire (2008): Cognition Creates World's Largest Semantic Map of the English Language With More Than 10 Million Semantic Connections (September 16, 2008) http://money.aol.com/news/articles/_a/bbdp/cognition-creates-worlds-largest/176372

Arnold, David and Geser, Guntram (2008): EPOCH Research Agenda for the Applications of ICT to Cultural Heritage. Full Report, May 2008 • http://public-repository.epoch-net.org/publications/RES_AGENDA/research_agenda.pdf

ASIS&T – American Society for Information Science & Technology (2003): Biocomplexity Thesaurus Launched. Bulletin of the American Society for Information Science & Technology, October/November 2003 http://findarticles.com/p/articles/mi_qa3991/is_/ai_n9343023

Berendsohn, Walter G. and Gebhardt, Marie (2008): LifeWatch – e-Science and Technology Infrastructure for Biodiversity Research. Proceedings of TDWG 2008 • http://www.tdwg.org/proceedings/article/view/372

Berners-Lee, Tim (1998a): Interpretation and Semantics on the Semantic Web http://www.w3.org/DesignIssues/Interpretation.html

Berners-Lee, Tim (1998b): Semantic Web Road Map http://www.w3.org/DesignIssues/Semantic.html

Berners-Lee, Tim (2000): Semantic Web. Presentation at XML2000 http://www.w3.org/2000/Talks/1206-xml2k-tbl/Overview.html

Berners-Lee, Tim (2003): WWW past & future. Presentation at the Royal Society http://www.w3.org/2003/Talks/0922-rsoc-tbl/

Berners-Lee, Tim (2005): Web for real people. Presentation at WWW conference 2005 http://www.w3.org/2005/Talks/0511-keynote-tbl/

Berners-Lee, Tim, Hendler, James and Lassila, Ora (2001): The Semantic Web. In: Scientific American, May 2001, p. 29-37 http://www.sciam.com/article.cfm?id=the-semantic-web

Binding, C., May, K. and Tudhope D. (2008): Semantic Interoperability in Archaeological Datasets: Data Mapping and Extraction via the CIDOC CRM, pp. 280-290, in: Christensen-Dalsgaard, Birte et al. (eds.): Research and Advanced Technology for Digital Libraries, 12th European Conference, ECDL 2008, Aarhus, Denmark, September 14-19, 2008. Proceedings. Lecture Notes in Computer Science 5173, Springer 2008.

Binding, C., Tudhope, D. and Vlachidis, A. (2008): STAR – Semantic Technologies for Archaeological Resources. Cost Action 21: Towntology, 3rd Workshop: Construction of multilingual ontologies for Urban Civil Engineering projects, 20 October 2008, University of Zaragoza, Spain http://iaaa.cps.unizar.es/towntology/presentations/vlachidis_mapping.pdf

Binding, Ceri (2008) STAR - Semantic Technologies for Archaeological Resources. ISKO UK Workshop "SKOS – Sharing Vocabularies on the Web via Simple Knowledge Organisation System", University College London, July 21, 2008 http://hypermedia.research.glam.ac.uk/kos/star/http://www.iskouk.org/presentations/STAR_UCL_20080721a.pdf

Binding, Ceri and Tudhope, Douglas (2004): KOS at your Service: Programmatic Access to Knowledge Organisation Systems. Journal of Digital Information, 4(4), (2004) • http://journals.tdl.org/jodi/article/view/jodi-124/109

Biodiversity Information Standards (TDWG), database Biodiversity Information Projects of the World http://www.tdwg.org/biodiv-projects/projects-database

BioNET International, "the global network for taxonomy" http://www.bionet-intl.org

BirdLife International (2004): State of the world's birds 2004: indicators for our changing world http://www.biodiversityinfo.org/sowb/userfiles/docs/SOWB2004_en.pdf

BirdLife International (2008): State of the world's birds 2008: indicators for our changing world http://www.biodiversityinfo.org/sowb/userfiles/docs/SOWB2008_en.pdf

Bowers, Shawn (2008): Ontology Frameworks for Modeling Observational Data Semantics

http://acdrupal.evergreen.edu/files/semanticweb/bowers-evergreen-2008.pdf
British Standards Institution (2005): BS 8723-2: Structured vocabularies for information retrieval – Guide – Part 2: Thesauri. London 2005
British Standards Institution (2006): BS 8723-3: Structured Vocabularies for Information Retrieval – Guide – Part 3. Vocabularies other than thesauri. London 2006
Butler, M.H. et al. (2004): Data conversion, extraction and record linkage using xml and rdf tools in project simile. Technical report, Digital Media Systems Lab, HP Labs Bristol http://www.hpl.hp.com/techreports/2004/HPL-2004-147.html
Byrne, Kate (2008a): Relational Databases and RDF. University of Edinburgh, School of Informatics, Multi-agent Semantic Web Systems Course, 8 February 2008 [113 slides] • http://homepages.inf.ed.ac.uk/s0233752/docs/reldb2rdf.pdf
Byrne, Kate (2008b): Relational Database to RDF Translation in the Cultural Heritage Domain. School of Informatics, University of Edinburgh, May 2008 • http://homepages.inf.ed.ac.uk/s0233752/docs/rdb2rdfForCH.pdf
Byrne, Kate (2008c): Having Triplets – Holding Cultural Data as RDF. In: Proceedings of the workshop on Information Access to Cultural Heritage (IACH 2008) at ECDL 2008, Denmark, Aarhus, 18 September 2008 http://ilps.science.uva.nl/IACH2008/papers/Byrne_RDF_IACH2008.pdf
Cabral, L., Domingue, J., Motta, E., Payne, T. and Hakimpour, F. (2004): Approaches to Semantic Web Services: An Overview and Comparisons. In: Proceedings of the 1st European Semantic Web Symposium (ESWS2004) http://kmi.open.ac.uk/technologies/irs/cabralESWS04.pdf
Caffo, Rosella (2008a): ATHENA. Access to cultural heritage networks across Europe. MICHAEL: Perspectives on cultural sector resource discovery. Royal Institute of British Architects, London, 23 May 2008 http://www.ukoln.ac.uk/events/michael-may-2008/presentations/r-caffo.ppt
Caffo, Rossella (2008b): ATHENA - Access to cultural heritage networks across Europe. Project presentation at Culture OnLine, 6-7 June 2008 • www.minervaeurope.org/events/Caffo_EVA%20Florence_18_April_2008.ppt
Catapano, Terry and Weitzman, Anna (2007): Progress in making literature easily accessible: schemas and marking up. TaxonX / Goldengate & taXMLit / INOTAXA. TDWG Annual Meeting, October 19, 2007 http://wiki.tdwg.org/twiki/pub/Literature/WebHome/Catapano_Weitzman_Markup_Final.pdf
CATE (2007): An LSID Resolution Service for CATE http://www.tdwg.org/uploads/media/LSID_Resolution_Service_For_CATE.pdf
Catton, C., Sparks, S., Shotton, D.M. (2006): The ImageStore Ontology and the Bioimage Database: Semantic Web Tools for Biological Research Images. Jena User Conference, Bristol, 11-12 May 2006 http://jena.hpl.hp.com/juc2006/proceedings/catton/paper.pdf
CETAF – Consortium of European Taxonomic Facilities (2004): Biodiversity and Europe: The contribution of Taxonomy and the European Taxonomic Facilities. Position Paper • http://www.cetaf.org/Maramos.pdf
Champin, Pierre-Antoine (2001): RDF Tutorial http://www710.univ-lyon1.fr/~champin/rdf-tutorial/rdf-tutorial.html
Chavan, V. and Krishnan, S. (2003): Natural history collections: A call for national information infrastructure. In: Current Science, Vol. 84, issue 1, January 2003 • http://www.ias.ac.in/currsci/jan102003/34.pdf
Christaki, Anna et al. (2007): Achieving Interoperability in the MichaelPlus Project http://www.delos.info/files/pdf/DELOS%20Multimatch%202007/Papers/8tzouvaras.pdf
CIDOC-CRM website http://cidoc.ics.forth.gr
Cousins, Jill and Siebinga, Sjoerd (2008): Introduction to Europeana prototype1. Presentation held at the Europeana/EDLnet conference: "Users expect the interoperable", Koninklijke Bibliotheek, The Hague, 23-24 June 2008 http://dev.europeana.eu/public_documents/Intro_to_demo_of_Europeana_prototype1.pps
Cripps, Paul et al. (2004): Ontological Modelling of the work of the Centre for Archaeology, CIDOC CRM Technical Paper http://cidoc.ics.forth.gr/docs/Ontological_Modelling_Project_Report_%20Sep2004.pdf
Cui, Hong (2008): Converting taxonomic descriptions to new digital formats. In: Biodiversity Informatics, 5, 2008, pp. 20-40 • https://journals.ku.edu/index.php/jbi/article/view/46/1551
D'Andrea, Andrea and Niccolucci, Franco (2008): Mapping, Embedding and Extending: Pathways to Semantic Interoperability. The Case of Numismatic Collections. Proceedings of the First International Workshop, SIEDL 2008, Tenerife, June 2, 2008, pp. 63-75 • http://image.ntua.gr/swamm2006/SIEDLproceedings.pdf
DAP 2005 – Dynamic Action Plan for the EU co-ordination of digitisation of cultural and scientific content. UK Presidency of the EU 2005, MLA – Museums-Libraries-Archives, DCMS, November 2005 http://www.minervaplus.ru/docums/dap-e.htm
Davies, C.E., Moss, D., Hill, M.O. (2004): EUNIS Habitat Classification. Revised 2004. Report to the European Environment Agency, European Topic Centre on Nature Protection and Biodiversity, October 2004 http://eunis.eea.europa.eu/upload/EUNIS_2004_report.pdf

Davies, Rob (2008): EuropeanaLocal – its role in improving access to Europe's cultural heritage through the European Digital Library. Workshop on Information Access to Cultural Heritage (IACH 2008) at ECDL 2008, Aarhus, 18 September 2008 • http://www.edlocal.eu/eng/Publications/Papers-and-documents and http://ilps.science.uva.nl/IACH2008/papers/Davies_EuropeanaLocal_IACH2008.pdf

Davis, Melissa J. et al. (2007): Integrating hierarchical controlled vocabulary with OWL ontology: A case study from the domain of molecular interactions. 6th Asia Pacific Bioinformatics Conference (APBC07), Kyoto, January 14-17, 2008, http://www.itee.uq.edu.au/~eresearch/papers/2007/APBC07.pdf

DCMI – Dublin Core Metadata Initiative (2008): Expressing DC metadata using RDF (DCMI Recommendation 2008-01-14) • http://dublincore.org/documents/dc-rdf/, see also the notes on this recommendation at http://dublincore.org/documents/dc-rdf-notes/

DCMI – Dublin Core Metadata Initiative
http://dublincore.org

de By, Rolf A. (2002): Recent proposals for specifically distinct bird species
http://home.planet.nl/~by000012/SM/Split/NewSplits.html

Dekkers, Makx (2001): Application Profiles, or how to Mix and Match Metadata Schemas. In: Cultivate international, issue 3 • http://www.cultivate-int.org/issue3/schemas/

DELOS (2005): Semantic Interoperability in Digital Library Systems. Project deliverable D5.3.1. Authors: M. Patel, T. Koch, M. Doerr, C. Tsinaraki. June 2005 • http://delos-wp5.ukoln.ac.uk/project-outcomes/SI-in-DLs/

DELOS: A Network of Excellence on Digital Libraries: Cluster on Knowledge Extraction and Semantic Interoperability
http://deloswp5.ukoln.ac.uk/

Denny, Michael (2002): Ontology Building: A Survey of Editing Tools (November 6, 2002)
http://www.xml.com/pub/a/2002/11/06/ontologies.html

Denny, Michael (2004): Ontology Tools Survey, Revisited (July 14, 2004)
http://www.xml.com/pub/a/2004/07/14/onto.html

Dextre Clarke, Stella G. (2007): Evolution towards ISO 25964: an international standard with guidelines for thesauri and other types of controlled vocabulary. In: IWP – Information Wissenschaft & Praxis, 2007, issue 8, pp. 441-444 http://www.agi-imc.de/isearch/dgi_publications.nsf/93387c5d893ee67bc12572590061a297/76884b709abadb72c12573a30067118b?OpenDocument

Digital Morphology project, University of Texas at Austin
http://digimorph.org

Doerr, M., Ore, Ch.-E., Stead, S. (2007): The CIDOC Conceptual Reference Model – A New Standard for Knowledge Sharing. 26th International Conference on Conceptual Modeling (ER 2007), Auckland, New Zealand. Conferences in Research and Practice in Information Technology (CRPIT), Vol. 83, Grundy, J. et al (eds.) http://cidoc.ics.forth.gr/docs/CRM_Tutorial_ER2007.pdf

Doerr, Martin (2001): Semantic Problems of Thesaurus Mapping. In: Journal of Digital Information, Vol. 1, issue 8
http://jodi.tamu.edu/Articles/v01/i08/Doerr/

Drenth, Bert Degenhart (2008): Using web services for terminology control. 2008 Annual Conference of CIDOC, Athens, September 15 – 18, 2008 • http://www.cidoc2008.gr/cidoc/Documents/papers/drfile.2008-06-18.6811555833

EC 2002 – Commission of the European Communities, DG Information Society: The DigiCULT Report. Technological Landscapes for tomorrow's cultural economy – Unlocking the value of cultural heritage. Authors: G. Geser and A. Mulrenin. Luxembourg • Available for download at: http://www.digicult.info/pages/report.php

EC 2005 – Commission of the European Communities: i2010: digital libraries. Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, COM(2005) 465 final, Brussels, 30.9.2005 • http://ec.europa.eu/information_society/activities/digital_libraries/doc/communication/en_comm_digital_libraries.pdf

EC 2006a – Commission of the European Communities: Digitisation and online accessibility of cultural material and digital preservation. Commission Recommendation of 24 August 2006, Official Journal of the European Union (2006/585/EC): L236/28, 31.8.2006
http://europa.eu.int/information_society/activities/digital_libraries/doc/recommendation/recommendation/en.pdf

EC 2006b – Commission Staff Working Document. Commission Recommendation on the digitisation and online accessibility of cultural material and digital preservation. Impact Assessment. Brussels, SEC(2006) 1075, 24.08.2006 • http://ec.europa.eu/information_society/activities/digital_libraries/doc/recommendation/impact_assessment/en.pdf

EC 2006c – Commission of the European Communities, DG Information Society and Media: i2010 Digital Libraries. Luxembourg: Office for Official Publications of the European Communities
http://ec.europa.eu/information_society/activities/digital_libraries/doc/brochures/dl_brochure_2006.pdf

EC 2006d – Commission of the European Communities, website: i2010: Digital Libraries Initiative

http://ec.europa.eu/information_society/activities/digital_libraries/index_en.htm
EC 2007 – Commission of the European Communities: Scientific information in the digital age: access, dissemination and preservation. Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, COM(2007) 56 final, Brussels, 14.2.2007 http://ec.europa.eu/research/science-society/document_library/pdf_06/communication-022007_en.pdf
eContentplus Programme: Projects (selected for funding under the calls for 2005-2007) http://ec.europa.eu/information_society/activities/econtentplus/projects/index_en.htm
EDIT – European Distributed Institute of Taxonomy (2008): Taxonomy in Europe in the 21st century. Report to the Board of Directors, European Distributed Institute of Taxonomy. http://ww2.bgbm.org/EditDocumentRepository/Taxonomy21report.pdf
EDLnet (2007): Initial Semantic and Technical Interoperability Requirements. Report by M. Dekkers, S. Gradmann, C. Meghini, N. Aloia, C. Concordia. Project deliverable D 2.2, 17 December 2007 • http://www.europeana.eu/public_documents/EDLnet_D2_2_Initial_Semantic_and_Technical_Interoperability_Requirements_final.pdf
EDLnet (2008): Europeana Outline Functional Specification. For development of an operational European Digital Library. Report by M. Dekkers, S. Gradmann, C. Meghini. Project deliverable D 2.5. Public Draft Version, 1.2., 20 August 2008 • http://www.europeana.eu/public_documents/EDLnet_D2.5_Outline_Functional_Specifications20080820_version_1.2_commentfree.pdf
EDLproject / Sally Chambers (2007): Towards Metadata Interoperability between Archives, Audio-Visual Archives, Museums and Libraries: What can we learn from The European Library metadata interoperability model? (August 31, 2007) http://www.edlproject.eu/downloads/EDLproject_D1.1_metadata_interoperability_report_v1.4.pdf
ENBI – European Network for Biodiversity Information / Häuser, Christoph L. et al. (eds., 2005): Digital Imaging of Biological Type Specimens. A Manual of Best Practice. Stuttgart http://circa.gbif.net/Public/irc/enbi/comm/library?l=/enbi_reports/haeuser_digital/_EN_1.0_&a=i
Europa.eu (2006): European Commission steps up efforts to put Europe's memory on the Web via a "European Digital Library" (IP/06/253), Brussels, 2 March 2006 • http://europa.eu/rapid/pressReleasesAction.do?reference=IP/06/253&format=HTML&aged=0&language=EN&guiLanguage=en
European Commission, unit: Cultural Heritage and Technology Enhanced Learning, DigiCult, FP7 projects for the 'digital libraries' objective (call 1) • http://cordis.europa.eu/fp7/ict/telearn-digicult/digicult-call1_en.html
European Commission, unit: Cultural Heritage and Technology Enhanced Learning, DigiCult: Research topics and projects, http://cordis.europa.eu/fp7/ict/telearn-digicult/digicult-projects_en.html
Europeana (2008): Specification for the Metadata Elements for the Europeana Prototype, V2.0, 28/08/2008 http://dev.europeana.eu/public_documents/Specification_for_metadata_elements_in_the_Europeana_prototype.pdf)
Euzenat, Jérôme et al. (2007): Results of the Ontology Alignment Evaluation Initiative 2007. The Second International Workshop on Ontology Matching, collocated with the 6th International Semantic Web Conference ISWC-2007, Busan, Korea, November 11, 2007 • http://www.dit.unitn.it/~p2p/OM-2007/0-o-oaei2007.pdf
Flowers, R. W. (2008): Taxononomy's unexamined impediment. In: EDIT newsletter #9, June 2008, pp. 8-9 http://www.e-taxonomy.eu/files/newsletter9.pdf
Foulonneau, Muriel (2004): Collaborer pour de nouveaux services culturels en ligne: le protocole OAI http://www.culture.gouv.fr/culture/mrt/numerisation/fr/technique/documents/guide_oai.pdf
Foulonneau, Muriel (ed., 2003): Open Archives Initiative – Protocol For Metadata Harvesting. Practices of cultural heritage actors, September 2003 • http://www.oaforum.org/otherfiles/oaf_d48_cser3_foullonneau.pdf
Gauch, Susan (2003): BDEI: Biodiversity Information Organization using Taxonomy (BIOT). Proc. of the National Conference on Digital Government Research, Los Angeles, CA, May 20- 28, 2002, pp. 169-174 http://ittc.ku.edu/~sgauch/papers/BIOT2003.doc
GBIF – Global Biodiversity Information Facility (2007): 2007-2008 Request for proposals: GBIF seed mony for content development • http://www.gbif.org/GBIF_org/documents/seedrfp
GBIF – Global Biodiversity Information Facility (2008a): Response by the Global Biodiversity Information Facility (GBIF). UK House of Lords Science & Technology Committee. Call for Evidence: Systematics and Taxonomy (05/03/2008), http://www.parliament.uk/documents/upload/stSTGBIF.pdf
GBIF – Global Biodiversity Information Facility (2008b): GBIF Training Manual 1: Digitisation of Natural History Collections Data. Version 1.0. Copenhagen http://www.gbif.org/GBIF_org/GBIF_Publications/trainingmanual1/index_html
GEMET (2008): About GEMET [2001] http://www.eionet.europa.eu/gemet/about?langcode=en
Gerbracht, Jeff and Kelling, Steve (2008): The Species Profile Model from an Avian Perspective. Proceedings of TDWG 2008 • http://www.tdwg.org/proceedings/article/view/393

German National GTI Focal Point: How many taxonomists are there?
http://www.gti-kontaktstelle.de/english/taxonomy_E.html
Geser, Guntram (2003): A Cultural Heritage Semantic Web Example & Primer. In: G. Geser (ed.): DigiCULT Thematic Issue 3: Towards a Semantic Web for Heritage Resources. Salzburg, May 2003, S. 26-36
http://www.digicult.info/pages/Themiss.php
Giunchiglia, F., Shvaiko, P., and Yatskevich, M. (2005): Semantic Schema Matching. University of Trento, Dept. Information and Communication Technology, Technical Report # DIT-05-014, March 2005
http://eprints.biblio.unitn.it/archive/00000748/01/014.pdf
Global Taxonomic Initiative (GTI)
http://www.cbd.int/gti/
Godfray, H.C.J., Clark, B.R., Kitching, I.J., Mayo, S.J., Scoble, M.J. (2007): The Web and the structure of taxonomy. In: Systematic Biology, 56 (6): 943-955.
Golbreich, Christine et al. (2007): OBO and OWL: Leveraging Semantic Web Technologies for the Life Sciences. 6th International Semantic Web Conference 2007, Busan, Korea, 11-15 November 2007
http://iswc2007.semanticweb.org/papers/169.pdf
González, M., Bianchi, S. and Vercelli, G. (2008): Semantic framework for complex knowledge domains. International Semantic Web Conference 2008 (ISWC 2008), Karlsruhe, Germany, 26-30 October 2008
http://ftp.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-401/iswc2008pd_submission_17.pdf
González, Marta (2008a): Formalised AQUARING Domain Ontologies. Project deliverable 3.2, final version, 2 January 2008 • http://www.aquaringweb.eu/documents/AqR-DELIV3.2-v1.0.pdf
González, Marta (2008b): Aquaring WP3. Metadata and Semantic Resources. Presentation at project review, Luxemburg, 22 January 2008 • http://www.aquaringweb.eu/documents/AqR-Lux-Review-3-RBTK-WP3.ppt
Good, B.M. and Wilkinson, M.D. (2006): The Life Sciences Semantic Web is Full of Creeps! Briefings in Bioinformatics 2006 7(3):275-286 • http://bib.oxfordjournals.org/cgi/content/full/7/3/275?ck=nck#T1
Gradmann, Stefan (2007a): Interoperability of Digital Libraries. Report on the work of the EC working group on DL interoperability. Presentation at the seminar "Disclosure and Preservation. Fostering European Culture in the Digital Landscape", Lisbon, 7-8 September 2007
http://bnd.bn.pt/seminario-conhecer-preservar/doc/Stefan%20Gradmann.pdf
Gradmann, Stefan (2007b): Interoperable Information Space – moving towards the European Digital Library. Presentation at Second DELOS Conference on Digital Libraries, Tirrenia, Pisa, Italy, 5-7 December 2007
http://www.delos.info/index.php?option=com_content&task=view&id=602&Itemid=334
Gradmann, Stefan (2008): Making Europeana Interoperable (the) Six Most Challenging Issues. DFL3@ECDL2008, Århus, 18 September 2008
http://www.delos.info/files/pdf/DLFoundations2008/5_GradmannDLFoundations08.pdf
Gruber, Tom (1995): What is an Ontology?
http://www-ksl.stanford.edu/kst/what-is-an-ontology.html
Gruber, Tom (2007): Ontology • http://tomgruber.org/writing/ontology-definition-2007.htm (to appear in the Encyclopedia of Database Systems, Ling Liu and M. Tamer Özsu (eds.), Springer 2008)
Guarino, Nicola (1998): Formal ontology and information systems. In: Proc. of the 1st International Conference on Formal Ontologies in Information Systems (FOIS'98), Trento, Italy, 6-8 June 1998. Amsterdam, IOS Press, pp. 3-15
http://www.loa-cnr.it/Papers/FOIS98.pdf
Guarino, Nicola (2002): Ontology-Driven Conceptual Modelling, part 1-3
http://ontology.ip.rm.cnr.it/Tutorials/
Gwinn, Nancy E. and Rinaldo, Constance (2008): The Biodiversity Heritage Library: Sharing biodiversity literature with the world. World Library and Information Congress, 74th IFLA General Conference and Council, Québec, Canada, 10-14 August 2008 • http://www.ifla.org/IV/ifla74/papers/109-Gwinn-en.pdf
Heery, Rachel (2004): Metadata Futures: Steps Toward Semantic Interoperability. Metadata in Practice. Eds. Diane I. Hillmann and Elaine L. Westbrooks, 257-271. Chicago: American Library Association.
Heery, Rachel and Patel, Manjula (2000): Application Profiles: Mixing and Matching Metadata Schemas. Ariadne [Online], no. 25 • http://www.ariadne.ac.uk/issue25/app-profiles/
Heidorn, P. Bryan and Wei, Qin (2008a): Automatic Metadata Extraction from Museum Specimen Labels. Proceedings of the International Conference on Dublin Core and Metadata Applications 2008
http://www.ideals.uiuc.edu/bitstream/2142/9138/2/HeidornDC2008.pdf
Heidorn, P. Bryan and Wei, Qin (2008b): Automatic Metadata Extraction (Darwin Core) From Museum Specimen Labels
http://dc2008.de/wp-content/uploads/2008/09/heidorndc2008b.pdf
Heikka, Juhani et al. (2006): The Museum24 project: New channel into the local history – Cultural Heritage on the Semantic Web • http://www.seco.tkk.fi/events/2006/2006-05-04-websemantique/presentations/friday-1520-

szasz-museo24_paris_final.pdf
Henderson, M., Khan, I. and Hunter, J. (2006): Semantic WildNET: An Ontology based Biogeographical System http://www.itee.uq.edu.au/~eresearch/projects/ecoportalqld/papers/SemWildNET.pdf
Hernández, Francisca (2007): Case Study: An Ontology of Cantabria's Cultural Heritage. W3C Semantic Web Use Cases and Case Studies (May 2007) • http://www.w3.org/2001/sw/sweo/public/UseCases/FoundationBotin/
Hernández, Francisca et al. (2007): Semantic Approach on Cultural Heritage Domain, pp. 105-106, Aroyo, L., Hyvönen, E. and van Ossenbruggen, J. (2007): Cultural Heritage on the Semantic Web. Workshop 9 of the 6th International Semantic Web Conference, Korea, 2007 • http://www.cs.vu.nl/~laroyo/CH-SW/ISWC-wp9-proceedings.pdf
Hernández, Francisca et al. (2008): Building a cultural heritage ontology for Cantabria. 2008 Annual Conference of CIDOC, Athens, September 15 – 18, 2008 http://www.cidoc2008.gr/cidoc/Documents/papers/drfile.2008-06-18.1772912112
Hildebrand, M., van Ossenbruggen, J., Hardman, L (2006): /facet: A Browser for Heterogeneous Semantic Web Repositories. In: International Semantic Web Conference (ISWC2006), pp. 272-285 available from: http://ftp.cwi.nl/CWIreports/INS/INS-E0604.pdf
Hillmann, D.I., Naomi, D. and Phipps, J. (2004): Improving Metadata Quality: Augmentation and Recombination. DC-2004 International Conference on Dublin Core and Metadata Applications, 11-14 October 2004, Shanghai, China http://www.cs.cornell.edu/naomi/DC2004/MetadataAugmentation--DC2004.pdf
Hilse, Hans-Werner and Kothe, Jochen (2006): Implementing Persistent Identifiers. Overview of concepts, guidelines and recommendations. Consortium of European Research Libraries, European Commission on Preservation and Access (ECPA), November2006 • http://nbn-resolving.de/urn:nbn:de:gbv:7-isbn-90-6984-508-3-8
Hodge, Gail (2000): Systems of Knowledge Organization for Digital Libraries: Beyond Traditional Authority Files. Published by the Digital Library Federation, Council on Library and Information Resources, Washington, DC. April 2000. http://www.clir.org/pubs/reports/pub91/pub91.pdf
Hunter, Jane (2002): Combining the CIDOC CRM and MPEG-7 to Describe Multimedia in Museums. In: Museums on the Web 2002, Boston, April 2002 • http://www.archimuse.com/mw2002/papers/hunter/hunter.html
Hunter, Jane and Lagoze, Carl (2002): Combining RDF and XML Schemas to Enhance Interoperability Between Metadata Application Profiles • http://archive.dstc.edu.au/RDU/staff/jane-hunter/www10/paper.html (04-04-2003).
Hyvönen, Eero et al. (2002a): Semantic Interoperability on the Web: Case Finnish Museums Online. Towards the Semantic Web and Web Services. Proceedings of the XML Finland 2002 Conference http://www.cs.helsinki.fi/u/eahyvone/xmlfinland2002/ProceedingsXML2002-final.pdf
Hyvönen, Eero et al. (2002b): Cultural Semantic Interoperability on the Web: Case Finnish Museums Online http://iswc2002.semanticweb.org/posters/hyvonen_a4.pdf
Hyvönen, Eero et al. (2004a): A Content Creation Process for the Semantic Web. Proceeding of OntoLex 2004: Ontologies and Lexical Resources in Distributed Environments, Lisbon, Portugal http://www.seco.tkk.fi/publications/2004/hyvonen-salminen-et-al-a-content-creation-process-2004.pdf
Hyvönen, Eero et al. (2004b): Finnish Museums on the Semantic Web: The User's Perspective on MuseumFinland. In: D. Bearman and J. Trant (eds.): Museums and the Web 2004: Proceedings. Toronto: Archives & Museum Informatics, 2004 • http://www.archimuse.com/mw2004/papers/hyvonen/hyvonen.html
Hyvönen, Eero et al. (2005): MuseumFinland - Finnish Museums on the Semantic Web. Journal of Web Semantics, vol. 3, no. 2 • http://www.seco.tkk.fi/publications/2005/hyvonen-makela-et-al-museumfinland-finnish-2005.pdf
Hyvönen, Eero et al. (2008): Building a national semantic web ontology and ontology service infrastructure the Finnonto approach. In: Proceedings of the 5th European Semantic Web Conference (ESWC 2008), 1-5 June 2008 http://www.seco.tkk.fi/publications/2008/hyvonen-et-al-building-2008.pdf
IMPACT: Improving Access to Text (FP7-ICT project) http://www.impact-project.eu/tools-and-applications/enhancement-enrichment-ee/
International Organization for Standardization; ISO 5964-1985: Documentation – Guidelines for the Establishment and Development of Multilingual Thesauri. Geneva: ISO 1985
International Organization for Standardization; ISO 2788-1986: Documentation – Guidelines for the Establishment and Development of Monolingual Thesauri. 2nd ed. Geneva: ISO 1986
Isaac, Antoine (2007a): Accessing Cultural Heritage Collections using Semantic Web Techniques. STITCH Project. Presentation at Book & Digital Media Master, March 2nd, 2007 http://www.few.vu.nl/~aisaac/talks/Isaac-Talk-BDMaster.pps
Isaac, Antoine (2007b): Controlled Vocabularies in TELPlus. EDLProject Workshop, 22-23 November 2007 http://www.edlproject.eu/workshop/downloads/Isaac-Contr_vocab_TELPlus-EDL.ppt
Isaac, Antoine (2008): On practical aspects of enhancing semantic interoperability using SKOS and KOS alignment. ISKO UK Meeting, July 21, London • http://www.iskouk.org/presentations/isaac_21072008.pdf
Isaac, Antoine et al. (2007): The value of usage scenarios for thesaurus alignment in Cultural Heritage context

(GTT – Brinkman case), pp. 25-39, in: Aroyo, L., Hyvönen, E. and van Ossenbruggen, J. (2007): Cultural Heritage on the Semantic Web. Workshop 9 of the 6th International Semantic Web Conference, Korea, 2007 http://www.cs.vu.nl/~laroyo/CH-SW/ISWC-wp9-proceedings.pdf

Isaac, Antoine et al. (2008): Putting ontology alignment in context: usage scenarios, deployment and evaluation in a library case • http://www.eswc2008.org/final-pdfs-for-web-site/oa-1.pdf

Isaksen, Leif (2008): The TRANSLATION Framework for Archaeological Excavation Data: Transparent Negotiation and Sharing of Local Application Terminologies, Instances and Ontologies (PhD progress report). University of Southampton, School of Electronics and Computer Science http://leifuss.files.wordpress.com/2008/08/translationframework.pdf

IVOA - International Virtual Observatory Alliance (2008): Vocabularies in the Virtual Observatory. Version 1.15. IVOA Proposed Recommendation, September 12, 2008 http://www.ivoa.net/Documents/PR/Semantics/Vocabularies-20080912.html

Jacob, Elin K. (2003): Ontologies and the Semantic Web. In: Bulletin of the American Society for Information Science & Technology, April/May 2003, available at FindArticles.com http://findarticles.com/p/articles/mi_qa3991/is_200304/ai_n9235530

Jacob, Elin K. (2004): Classification and categorization: a difference that makes a difference. In: Library Trends, Winter 2004, available at FindArticles.com • http://findarticles.com/p/articles/mi_m1387/is_3_52/ai_n6080402

Jupp, S., Bechhofer, S. and Stevens, R. (2008): SKOS with OWL: Don't be Full-ish! OWLED 2008 workshop, co-located with the International Semantic Web Conference (ISWC), Karslruhe, Germany. October 26-27, 2008 http://www.webont.org/owled/2008/papers/owled2008eu_submission_22.pdf

Kennedy, J. at al (2006): TDWG Core Ontology (October 2006). Presentation at the TDWG 2006 Annual Meeting http://tdwg2006.tdwg.org/fileadmin/2006meeting/slides/Kennedy_TdwgOntology_abs0013.ppt

Kim, Ke Chung and Byrne, B. Loren (2006): Biodiversity loss and the taxonomic bottleneck: emerging biodiversity science. In: Ecological Research, Vol.21, 6 / November 2006, pp. 794-810 http://www.environment.psu.edu/publications/reports/kim_fig/appendix4.pdf

Koch, Traugott: Controlled vocabularies, thesauri and classification systems available in the WWW http://www.mpdl.mpg.de/staff/tkoch/publ/subject-help.html

Kollias, Stefanos and Cousins, Jill (eds.): Semantic Interoperability in the European Digital Library. In: Proceedings of the First International Workshop, SIEDL 2008, Tenerife, June 2, 2008 http://image.ntua.gr/swamm2006/SIEDLproceedings.pdf

Kondylakis, H., Doerr, M., and Plexousakis, D. (2006): Mapping language for information integration. Technical report, ICS-FORTH • http://www.ics.forth.gr/isl/publications/paperlink/Mapping_TR385_December06.pdf.

Koning, D., Sarkar, I.N., Moritz, T. (2005): TaxonGrab: Extracting Taxonomic Names From Text. In: Biodiversity Informatics, 2, 2005, pp. 79-82 • https://journals.ku.edu/index.php/jbi/article/view/17/9

Krishtalka, Leonard and Humphrey, Philip S. (2000): Can Natural History Museums Capture the Future? BioScience, July 2000 / Vol. 50 No. 7, pp. 611-617 http://www.uprm.edu/biology/profs/chinea/UIP-MAPR/PLANTA/krishtalka_e2000.pdf

Kroski, Ellyssa (2005): The hive mind: folksonomies and user-based tagging (12 July 2005) http://infotangle.blogsome.com/2005/12/07/the-hive-mind-folksonomies-and-user-based-tagging/

Lampe, Karl-Heinz (2006): CIDOC CRM Knowledge Mapping in Biodiversity. Semantic Interoperability for e-Research in the Sciences, Arts and Humanities, Imperial College, London, 30 March 2006 http://cidoc.ics.forth.gr/workshops/london_workshop/Lampe.pdf

Lampe, Karl-Heinz and Krause, Siegfried (2008): How CIDOC-CRM supports interoperability? http://www.europeana.eu/public_documents/Keynote_HowCIDOC-CRM_supports_interoperability_Seigfried_Karuse_and_Karl_Lampe.pps

Leary, Patrick R. et al. (2007): uBioRSS: Tracking taxonomic literature using RSS. In: Bioinformatics, 23(11):1 434-1436 http://bioinformatics.oxfordjournals.org/cgi/content/full/23/11/1434

Lepage, Denis (2008): Domain-centric observation networks: Experiences gained from the Avian Knowledge Network. TDWG 2008 conference • http://www.tdwg.org/proceedings/article/view/410/0 http://www.tdwg.org/fileadmin/2008conference/slides/Lepage_19_05_AKN.ppt

Lourdi, Irene and Papatheodorou, Christos (2008): Semantic integration of collection level information: A crosswalk between CIDOC/CRM & Dublin Core Collection Application Profile. 2008 Annual Conference of CIDOC, Athens, September 15 – 18, 2008 • http://www.cidoc2008.gr/cidoc/Documents/papers/drfile.2008-06-18.3928994098

Lourdi, Irene et al. (2007): Integrating Dublin Core metadata for cultural heritage collections using ontologies. Proceedings of the International Conference on Dublin Core and Metadata Applications, 2007 http://www.dcmipubs.org/ojs/index.php/pubs/article/viewFile/16/11

Lowndes, M. (2006): An introduction to the Semantic Web for Museums. In: J. Trant and D. Bearman (eds.):

Museums and the Web 2006: Proceedings, Toronto: Archives & Museum Informatics, published March 1, 2006 at http://www.archimuse.com/mw2006/papers/lowndes/lowndes.html
Lund Principles and Action Plan (2001)
http://cordis.europa.eu/ist/digicult/lund-principles.htm
Lyal, Christopher and Weitzman, Anna L. (2008a): Releasing the content of taxonomic papers: solutions to access and data mining. In: Proceedings of the BNCOD 2008 Workshop: Biodiversity Informatics: challenges in modelling and managing biodiversity knowledge. Cardiff University, UK, 10th July 2008
http://biodiversity.cs.cf.ac.uk/bncod/LyalAndWeitzman.pdf
Lyal, Chris and Weitzman, Anna L. (2008b): Precision in accessing the descriptions, keys and other contents of digitized taxonomic literature: the INOTAXA prototype. Proceedings of TDWG 2008, abstract
http://www.tdwg.org/proceedings/article/view/415
Madin, Joshua et al. (2007): An ontology for describing and synthesizing ecological observation data.
In: Ecological Informatics, Vol. 2, Issue 3, October 2007, pp. 279-296
Madin, Joshua et al. (2008): Advancing ecological research with ontologies. In: Trends in Ecology & Evolution, Vol., 23, Issue 3, March 2008, pp. 159-168
Mäkelä, Eetu (2007): MuseumFinland – Finnish Museums on the Semantic Web. "Intelligent Access to Digital Heritage" conference, Tallinn, October 18-19, 2007
http://conference2007.kul.ee/failid/Makela_museumfinland-18_10_2007.pdf
Mallet, Jim (2004): Taxonomic inflation
http://www.ucl.ac.uk/taxome/jim/Sp/taxinfl.html
MARC21 formats for authority data and classification data
http://www.loc.gov/marc/
Marine Metadata Interoperability
http://marinemetadata.org
Masci, M.E., Buonazia, I. and Merlitti, D. (2007): The project of the Itallian culture portal. A standard based model for interoperability amongst cultural heritage data sources. XXI International CIPA Symposium, Athens, Greece, October, 1-6, 2007 • http://cipa.icomos.org/fileadmin/papers/Athens2007/FP096.pdf
Mathes, Adam (2004): Folksonomies - Cooperative Classification and Communication Through Shared Metadata. Report, Graduate School of Library and Information Science, Illinois Urbana-Champaign, 2004
http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html
May, K., Binding, C., Tudhope D. (2008): A STAR is born: some emerging Semantic Technologies for Archaeological Resources. Proceedings Computer Applications and Quantitative Methods in Archaeology (CAA2008), Budapest.
May, Keith (2006): Integrating Cultural and Scientific Heritage: Archaeological Ontological Modelling for the Field and the Lab. CIDOC CRM SIG Workshop, Heraklion (2006)
http://cidoc.ics.forth.gr/workshops/heraklion_october_2006/may.pdf
Maynard, Diana et al. (2007): Benchmarking of annotation tools. Knowledge Web project deliverable D1.2.2.1.3, October 2007 • http://www.kaiec.org/fileadmin/publications/Maynard07Benchmarking.pdf
McCallum, S.H. (2005): MARCXML Sampler. World Library and Information Congress: 71th IFLA General Conference and Council, "Libraries - A voyage of discovery", Oslo, Norway, August 14 - 18 2005
http://www.ifla.org/IV/ifla71/papers/175e-McCallum.pdf
McGuiness, D. L. (2002): Ontologies come of age. In: Fensel, D.; Hendler, J.; Lieberman, H. and Wahlster, W. (eds.): Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential. Cambridge: MIT Press 2003,
http://www.ksl.stanford.edu/people/dlm/papers/ontologies-come-of-age-mit-press-(with-citation).htm
Mergen, P., Vanhee, H., Lacaille, A., Cael, G., Louette, M. (2008): RDF based Reference Network Architecture for a distributed Digital Library system. The Royal Museum for Central Africa's use case on African Bird information (STERNA project). Proceedings of TDWG 2008 • http://www.tdwg.org/proceedings/article/view/384
Merholz, Peter (2004). Metadata for the Masses. Adaptive Path, October 19, 2004
http://www.adaptivepath.com/publications/essays/archives/000361.php
MICHAEL-EU Dublin Core Application Profile
http://www.ukoln.ac.uk/metadata/michael/michael-eu/dcap/
Michener, William K. et al. (2007): A knowledge environment for the biodiversity and ecological sciences.
In: Journal of Intelligent Information Systems, volume 29, no.1, August 2007, pp. 111-126
Midford, Peter E. (2008): Taxonomic ontologies: Bridging phylogenetic and taxonomic history. Proceedings of TDWG 2008 • http://www.tdwg.org/proceedings/article/view/414
Mika, Peter (2005): Ontologies are us: A unified model of social networks and semantics. In: Proceedings of the 4th ISWC, LNCS 3729, Springer, 2005 • http://www.cs.vu.nl/~pmika/research/papers/ISWC-folksonomy.pdf
Mika, Peter (2007): Social Networks and the Semantic Web. Springer 2007

(Series: Semantic Web and Beyond, Vol.5)
Mikhalenko, Peter (2005): Introducing SKOS (June 22, 2005)
http://www.xml.com/pub/a/2005/06/22/skos.html
Miles, Alistair (2005): Simple knowledge organisation and the Semantic Web. Notes on the Bliss Classification Association Annual Lecture 2005 • http://isegserv.itd.rl.ac.uk/public/skos/press/Bliss2005/skos-bliss-bulletin-2006.pdf
Miles, Alistair (2006): Simple Knowledge Organisation and the Semantic Web. Short paper for the Bliss Classification Association's annual bulletin
http://isegserv.itd.rl.ac.uk/public/skos/press/Bliss2005/skos-bliss-bulletin-2006.pdf
Miles, Alistair (2007): Tutorial – Vocabularies. International Conference on Dublin Core and Metadata Applications, Singapore, 27-31 August 2007 • http://isegserv.itd.rl.ac.uk/public/ajm65/dc2007/tutorial.pdf
Miles, Alistair (2008): SKOS Issues Review (21 February 2008)
http://lists.w3.org/Archives/Public/public-swd-wg/2008Feb/0096.html; Draft - SWD F2F Day 2, 07 May 2008, http://www.w3.org/2008/05/07-swd-minutes.html#item05; and SWD issue tracking: Issue-77: SubjectIndexing, http://www.w3.org/2006/07/SWD/track/issues/77
Miles, Alistair / SWAD-Europe Thesaurus Activity (2001): RDF Encoding of Classification Schemes: An example encoding of the PACS scheme, with some recommendations for classification schemes in general
http://www.w3.org/2001/sw/Europe/reports/thes/8.5/
Miles, Alistair et al. (2005): SKOS: A language to describe simple knowledge structures for the web. XTech 2005: XML, the Web and beyond • http://www.idealliance.org/proceedings/xtech05/papers/03-04-01/
Miles, Alistair (2007): Collaboration in the Value Grid for Semantic Technologies. Proceedings of the UK e-Science All Hands Meeting 2007, Paper for the Workshop on Issues in Ontology Development and Use, September 10-13, 2007, http://www.allhands.org.uk/2007/proceedings/papers/854.pdf
MINERVA / MINERVAPlus / MINERVA EC: Ministerial Network for Valorising Activities in Digitisation: Coordinating digitisation in Europe: progress reports of the National Representatives Group, 2002-2007
http://www.minervaeurope.org/publications/globalreport.htm
Miranker, D., Bafna, S. and Humphries, J. (2006): Schema Driven Assignment and Implementation of Life Science Identifiers (LSIDs). University of Texas at Austin, Department of Computer Sciences. Technical Report TR-06-50, October 19, 2006 • http://www.morphster.org/papers/tr06-50.pdf
Morris, Robert A. (2008): A gentle stroll through the Species Profile Model. Proceedings of TDWG 2008
http://www.tdwg.org/proceedings/article/view/381
Müller, Andreas et al. (2008): EDIT Platform for Cybertaxonomy – An Overview. Proceedings of TDWG 2008
http://www.tdwg.org/proceedings/article/view/375
Nadeau, David and Sekine, Satoshi (2007): A survey of named entity recognition and classification. Paper published in the Journal of Linguisticae Investigationes 30:1, 2007 • http://nlp.cs.nyu.edu/sekine/papers/li07.pdf
National Information Standards Organization. ANSI/NISO Z39.19-1993. Guidelines for the Construction, Format and Management of Monolingual Thesauri. Bethesda, Maryland: NISO Press; 1993.
NCBI – National Center for Biotechnology Information: Taxonomy Browser
http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/
Nederbragt, Hans / Trezorix (2008): Introduction to the STERNA architecture
http://www.sterna-net.eu/images/stories/documents/sterna_architecture_01.pdf
Nußbaumer, P. and Haslhofer, B. (2007a): CIDOC CRM in Action – Experiences and Challenges.Poster for the 11th European Conference on Research and Advanced Technology for Digital Libraries (ECDL07), Budapest
http://www.cs.univie.ac.at/upload//550/papers/cidoc_crm_poster_ecdl2007.pdf
Nußbaumer, P. and Haslhofer, B. (2007b): Putting the CIDOC CRM into Practice – Experiences and Challenges. University of Vienna, Technical Report, September 2007
http://www.cs.univie.ac.at/publication.php?pid=2965
OAI – Open Archives Initiative
http://www.oaforum.org
OCLC Terminology services project website (2008)
http://www.oclc.org/research/projects/termservices/
Omelayenko, Borys (2008): Porting Cultural Repositories to the Semantic Web. In: Kollias, Stefanos and Cousins, Jill (eds.): Semantic Interoperability in the European Digital Library. Proceedings of the First International Workshop, SIEDL 2008, Tenerife, June 2, 2008, pp. 14-35 • http://image.ntua.gr/swamm2006/SIEDLproceedings.pdf
OMG – Object Management Group (2004): Life Sciences Identifiers Specification, v1.0 (formal/04-12-01), Dezember 2004 • http://www.omg.org/technology/documents/formal/life_sciences.htm (http://www.omg.org/docs/formal/04-12-01.pdf)
Ontology Matching, a rich and up-to-date information website

http://www.ontologymatching.org
Orme, E.R., Jones, A.C. and White, R.J. (2008): LSID Deployment in the Catalogue of Life. In: Proceedings of the BNCOD 2008 Workshop: Biodiversity Informatics: challenges in modelling and managing biodiversity knowledge. Cardiff University, UK, 10th July 2008 • http://biodiversity.cs.cf.ac.uk/bncod/OrmeJonesAndWhite.pdf
Page, R.D.M. (2005): A taxonomic search engine: Federating taxonomic databases using web services. In: BMC Bioinformatics 6(48) • http://www.biomedcentral.com/1471-2105/6/48.
Page, R.D.M. (2006): Taxonomic names, metadata, and the Semantic Web. In: Biodiversity Informatics, Vol 3., 2006, pp. 1-15 • https://journals.ku.edu/index.php/jbi/article/view/25/12
Page, R.D.M. (2008a): IAG review of BIG, May 1, 2008 http://blog.eol.org/category/biodiversity-informatics/
Page, R.D.M. (2008b): LSID tester, a tool for testing life science identifier resolution services. In: Source Code for Biology and Medicine 3(2) • http://www.scfbm.org/content/3/1/2
Panzer, Michael (2007): Towards the "webification" of controlled subject vocabulary. A case study involving the Dewey Decimal Classification. 6th European NKOS Workshop, Budapest, September 21, 2007 • http://www.comp.glam.ac.uk/pages/research/hypermedia/nkos/nkos2007/presentations/NKOS_2007_webification_2-Panzer.ppt
Parr, C.S., Sachs, J., Finin, T. (2008): Lessons learned from semantic web prototyping in ecology. Proceedings of TDWG 2008 http://www.tdwg.org/proceedings/article/view/411
Parr, Cynthia (2008): The Encyclopedia of Life: Status report on species pages, contributions, and curators. Proceedings of TDWG 2008 • http://www.tdwg.org/proceedings/article/view/424
Parr, Cynthia et al. (2006): ETHAN: the Evolutionary Trees and Natural History Ontology. Technical Report, Computer Science and Electrical Engineering, University of Maryland, November 1, 2006, http://ebiquity.umbc.edu/_file_directory_/papers/320.pdf
Phipps, J., Hillmann, D.I. and Paynter, G. (2005): Orchestrating Metadata Enhancement Services: Introducing Lenny. Proceedings of the International Conference on Dublin Core and Metadata Applications, Madrid, Spain http://arxiv.org/ftp/cs/papers/0501/0501083.pdf
Plazi.org (2008): Press release, January 20, 2008: Plazi.org – the digital repository for species descriptions http://plazi.org/?q=node/27
Powers, Shelley (2003): Practical RDF. Cambridge: O'Reilly 2003
Quintarelli, E., Resmini, A., and Rosati, L. (2007): Facetag: Integrating Bottom-up and Top-down Classification in a Social Tagging System. Research paper presented at IA Summit 2007, Las Vegas http://www.facetag.org/download/facetag-20070325.pdf
Raatikka, Vilho and Hyvönen, Eero (2002a): Ontology-based Semantic Metadata Validation. In: Towards the Semantic Web and Web Services. Proceedings of the XML Finland 2002 Conference http://www.cs.helsinki.fi/u/eahyvone/xmlfinland2002/ProceedingsXML2002-final.pdf
Raatikka, Vilho and Hyvönen, Eero (2002b): Semantic Interoperability on the Web: Case Finnish Museums Online. In: Towards the Semantic Web and Web Services. Proceedings of the XML Finland 2002 Conference http://www.cs.helsinki.fi/u/eahyvone/xmlfinland2002/ProceedingsXML2002-final.pdf
Remsen, D. and Lane, M. (2008): Taxonomically informed biodiversity informatics supports taxonomy. In: EDIT newsletter #9, June 2008, pp. 10-12 • http://www.e-taxonomy.eu/files/newsletter9.pdf
Ross, Seamus (2004): Progress from National Initiatives towards European Strategies for Digitisation, pp. 88-98, in: Towards a Continuum of Digital Heritage: Strategies for a European Area of Digital Cultural Resources, European Conference, Den Haag: Dutch Ministry of Education, Culture and Science, 15-16 September 2004 http://eprints.erpanet.org/103/01/sross_denhaag_dutch_paper.pdf
Rycroft, S., Roberts, D., Harman, K. and Smith, V. (2008): Small pieces loosely joined: Building scientific web communities with Scratchpads. In: Proceedings of TDWG 2008 • http://www.tdwg.org/proceedings/article/view/334
Sautter, G., Böhm, K., Agosti, D. (2006): A combining approach to Find All Taxon Names (FAT) in legacy biosystematics literature. Biodiversity Informatics, 3, 2006, pp. 46-58 https://journals.ku.edu/index.php/jbi/article/viewFile/34/19
Sautter, G., Böhm, K., Agosti, D. (2007a): A quantitative comparison of XML Schems for taxonomic publications. In: Biodiversity Informatics, 4, 2007, pp. 1-13 • https://journals.ku.edu/index.php/jbi/article/view/36/20
Sautter, G., Agosti, D., Böhm, K. (2007b): Semi-Automated XML Markup of Biosystematics Legacy Literature with the GoldenGATE Editor, in Pacific Symposium on Biocomputing 12:391-402(2007) http://psb.stanford.edu/psb-online/proceedings/psb07/sautter.pdf
Schildhauer, Mark et al. (2008): SONet (Scientific Observations Network) and OBOE (Extensible Observation Ontology): facilitating data interoperability within the environmental and ecological sciences through advanced semantic approaches. Proceedings of TDWG 2008 • http://www.tdwg.org/proceedings/article/view/434
Schopf, J. M. et al. (2008): Managing Biodiversity Knowledge in the Encyclopedia of Life. In: Proceedings of the BNCOD

2008 Workshop: Biodiversity Informatics: challenges in modelling and managing biodiversity knowledge. Cardiff University, UK, 10th July 2008 • http://biodiversity.cs.cf.ac.uk/bncod/SchopfEtAl.pdf
Schreiber, Guus et al. (2006): MultimediaN E-Culture Demonstrator. In: International Semantic Web Conference (ISWC2006), Athens, USA. Cruz et al. (eds.), LNCS Volume 4273, November 2006, pp. 951-958.
Shirky, Clay (2005): Ontology is Overrated: Categories, Links, and Tags http://www.shirky.com/writings/ontology_overrated.html
Shotton, David (2005): Using the Semantic Web to address problems inherent in biological information management. UK e-Science All Hands Meeting Nottingham, September 20th 2005 http://www.jisc.ac.uk/media/documents/programmes/eresearch/4shottonall_hands_meeting.pdf
Shreeves, S.L., Riley, J. and Milewicz, Liz (2006): Moving towards shareable metadata. In: First Monday, volume 11, number 8 (August 2006) • http://www.firstmonday.org/issues/issue11_8/shreeves/index.html
Si, Libo (2007): Encoding formats and consideration of requirements for terminology mapping. The 6th European Networked Knowledge Organization Systems (NKOS) Workshop, Budapest, September 21, 2007 http://www.comp.glam.ac.uk/pages/research/hypermedia/nkos/nkos2007/programme.html
Sinclair, P.A.S. et al. (2005): Concept browsing for multimedia retrieval in the SCULPTEUR project. In: Proceedings of The 2nd Annual European Semantic Web Conference, Heraklion, Crete http://www.acemedia.org/ESWC2005_MSW/papers/ESWC_2005_MMSW_Sinclair_SCULPTEUR.pdf
Sinha, Rashmi (2005): A cognitive analysis of tagging (or how the lower cognitive cost of tagging makes it popular), 27 September 2005 • http://rashmisinha.com/2005/09/27/a-cognitive-analysis-of-tagging/
SKOS Simple Knowledge Organization System - homepage http://www.w3.org/2004/02/skos/
Slavic, Aida (2005): Knowledge organization systems, network standards and semantic Web. University College London, School of Library, Archives and Information Studies http://dlist.sir.arizona.edu/1326/02/semweb_kos_EN_2.pdf
Smith, Dan and Szekely, Ben (2005): LSID best practices. A guide to deploying Life Science Identifiers (April 5, 2005) http://www-128.ibm.com/developerworks/opensource/library/os-lsidbp/
Smithonian National Museum of Natural History (2008): Birds collection search http://nhb-acsmith2.si.edu/emuwebvzbirdsweb/pages/nmnh/vz/QueryBirds.php
Soberon, J. and Peterson, A. T. (2004): Biodiversity informatics: managing and applying primary biodiversity data. Philosophical Transactions of the Royal Society of London B, 359:689-698
Sowa, John: Ontology http://users.bestweb.net/ ~sowa/ontology/index.htm
Specia, Lucia and Motta, Enrico (2007): Integrating Folksonomies with the Semantic Web. Proceedings of the 4th European Semantic Web Conference 2007, Innsbruck, Austria http://www.eswc2007.org/pdf/eswc07-specia.pdf
Speers, Larry (2005): E-Types – A New Resource for Taxonomic Research, pp. 13-18, in: ENBI / Häuser, C.L. et al. (2005): Digital Imaging of Biological Type Specimens. A Manual of Best Practice http://circa.gbif.net/Public/irc/enbi/comm/library?l=/enbi_reports/haeuser_digital/_EN_1.0_&a=i
Summers, Ed et al. (2008): LCSH, SKOS and Linked Data. Proceedings of the International Conference on Dublin Core and Metadata Applications 2008 • http://inkdroid.org/bzr/lcsh/docs/dc2008.pdf
SWAD-Europe Thesaurus Activity (2004): SKOS-Core Guidelines for Migration. Guidelines and case studies for generating RDF encodings of existing thesauri, • http://www.w3.org/2001/sw/Europe/reports/thes/1.0/migrate/
SWAD-Europe Thesaurus Activity http://www.w3.org/2001/sw/Europe/reports/thes/
SWAD-Europe Thesaurus links http://www.w3.org/2001/sw/Europe/reports/thes/thes_links.html
SWAD-Europe: Semantic Web Advanced Development for Europe (SWAD-E) project (FP5-IST, May 2002 to October 2004) http://www.w3.org/2001/sw/Europe/
Szász, Barnabás et al. (2006): Cultural Heritage on the Semantic Web – the Museum24 project. Symposium on "Digital Semantic Content across Cultures", Louvre, Paris, 4-5 May 2006 http://www.artio.net/download/museo24_louvre_article.pdf
TDWG (2008): TAPIR - TDWG Access Protocol for Information Retrieval. Protocol Specification - Version 1.0, 18 September 2008 • http://www.tdwg.org/dav/subgroups/tapir/1.0/docs/TAPIRSpecification_2008-09-18.html
TDWG / Pereira, R., Pyle, R. and Richards, K. (2008): LSID Authority Setup Guides http://www.tdwg.org/activities/guid/documents/lsid-setup-guides/
TDWG TAG (2006a): TDWG Technical Architecture Group meeting, eScience Institue, Edinburgh, UK, April 11-13, 2006 http://wiki.tdwg.org/twiki/bin/view/TAG/TagMeeting1Report

TDWG TAG (2006b): TDWG Core Ontology Meeting, eScience Institute, Edinburgh, UK, May 16-18, 2006
http://www.nesc.ac.uk/talks/687/coremeetingreport.pdf
TDWG TAG (2007): Technical Roadmap 2007, Technical Architecture Group, 27th August 2007
http://wiki.tdwg.org/twiki/pub/TAG/RoadMap2007/TAG_Roadmap_2007_final.pdf
TDWG TAG (2008): Technical Roadmap 2008. Technical Architecture Group, 15th October 2008
http://wiki.tdwg.org/twiki/pub/TAG/RoadMap2008/TDWG_TAG_Roadmap_2008.pdf
TDWG TAG Ontology Wiki pages
http://wiki.tdwg.org/twiki/bin/view/TAG/TDWGOntology
Tennant, Roy (2004): Bitter Harvest. Metadata Harvesting Issues, Problems, and Possible Solutions
http://library.acadiau.ca/access2004/presentations/tennant1.ppt
Tennis, Joseph T. (2006): Social tagging and the next steps for indexing. In: Proceedings of the 17th SIG Classification Research Workshop, 2006 • http://dlist.sir.arizona.edu/1726/01/sigcr-06tennis.pdf
Tordai, A., Omelayenko, B. and Schreiber, G. (2007): Thesaurus and metadata alignment for a semantic e-culture application. In Proceedings of the 4th International Conference on Knowledge capture (KCAP-2007), October 28–31, 2007, Whistler, British Columbia, Canada, pp. 199–200 • http://www.cs.vu.nl/~guus/papers/Tordai07a.pdf
Trezorix (2008): Sterna architecture. Overview of the proposed Sterna software architecture (2 February 2008)
http://www.rnaproject.org/media/34563/rna%20report%20-%20c%20sterna%20architecture.pdf
Tudhope, Douglas, Koch, Traugott. and Heery, Rachel (2006): Terminology Services and Technology: JISC State of the art review • http://www.jisc.ac.uk/media/documents/programmes/capital/terminology_services_and_technology_review_sep_06.pdf
Tudhope, D. and Binding, C. (2008): Making KOS Machine Understandable. Additional Report for DELOS Work Package 5, 29 February 2008
http://hypermedia.research.glam.ac.uk/media/files/documents/2008-07-05/Additional-report-wp5.pdf
Tudhope, Douglas (2006): A tentative typology of KOS: towards a KOS of KOS? The 5th European Networked Knowledge Organization Systems (NKOS) Workshop at the 10th ECDL Conference, Alicante, Spain, September 21, 2006,
http://www.comp.glam.ac.uk/pages/research/hypermedia/nkos/nkos2006/
Tudhope, Douglas et al. (2006): Query expansion via conceptual distance in thesaurus indexed collections. Journal of Documentation, 62 (4), pp. 509–533
http://hypermedia.research.glam.ac.uk/media/files/documents/2008-04-02/JDOCfinal-Tudhope.doc
Tudhope. D., Binding, C., May, K. (2008): Semantic interoperability issues from a case study in archaeology. In: Kollias, S. and Cousins, J. (eds.): Semantic Interoperability in the European Digital Library, Proceedings of the First International Workshop SIEDL 2008, Tenerife, pp. 88-99
http://hypermedia.research.glam.ac.uk/media/files/documents/2008-07-05/SIEDL08-Tudhope-v3.pdf
Tuominen, Jouni et al. (2008): ONKI-SKOS – Publishing and Utilizing Thesauri in the SemanticWeb
http://www.seco.tkk.fi/publications/2008/tuominen-et-al-onki-skos-2008.pdf
Uschold, Mike and Jasper, Robert (1999): A Framework for Understanding and Classifying Ontology Applications. In: Proceedings of the IJCAI-99 workshop on Ontologies and Problem-Solving Methods (KRR5), Stockholm, Sweden, August 2, 1999 • http://www.cs.man.ac.uk/~horrocks/Teaching/cs646/Papers/uschold99.pdf
van Assem, M. et al. (2004): A Method for Converting Thesauri to RDF/OWL. In: McIlraith, S.A. et al. (eds.): Proceedings of the Third International Semantic Web Conference (ISWC'04) Lecture Notes in Computer Science – 3298. Hiroshima, Japan. Springer, pp. 17–31, http://www.cs.vu.nl/~mark/papers/Assem04a.pdf, supplementary website: http://thesauri.cs.vu.nl/
van Assem, M. et al. (2006): A Method to Convert Thesauri to SKOS. Lecture Notes in Computer Science (Springer), volume 4011, pp. 95–109, http://www.cs.vu.nl/~mark/papers/Assem06b.pdf; supplementary website: http://thesauri.cs.vu.nl/eswc06/
van Gendt, M. et al. (2006): Semantic Web Techniques for Multiple Views on Heterogeneous Collections: a Case Study. 10th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2006), Alicante, Spain, 2006 • http://www.cs.vu.nl/STITCH/papers/STITCH-ECDL06.pdf
van Ossenbruggen, Jacco et al. (2007): Searching and Annotating Virtual Heritage Collections with Semantic-Web Techniques. In: Museums and the Web 2007, April 11-14, 2007
http://www.archimuse.com/mw2007/papers/ossenbruggen/ossenbruggen.html
Van Waeyenberge, Sandra (2008): LifeWatch: Moving Forward. In: EDIT newsletter #10, August 2008
http://www.e-taxonomy.eu/files/Newsletter10.pdf
Vatant, Bernard (2008): Wondering about either SKOS or Web Ontology Language (OWL)? Use both! ISKO UK Workshop "SKOS – Sharing Vocabularies on the Web via Simple Knowledge Organisation System", University College London, July 21, 2008 • http://www.iskouk.org/presentations/vatant_21072008.pdf
Veldhuijzen van Zanten, H., Van Spronsen, E. and Altenburg, R. (2005): 3D Imaging for a Virtual Museum. Bird Type

Specimens of the Zoological Museum Amsterdam, pp. 272-283, in: ENBI / Häuser, C.L. et al. (eds.): Digital Imaging of Biological Type Specimens. A Manual of Best Practice. Stuttgart 2005 http://circa.gbif.net/Public/irc/enbi/comm/library?l=/enbi_reports/haeuser_digital/_EN_1.0_&a=i
Vizine-Goetz D., Houghton A., Childress E. (2006): Web Services for Controlled Vocabularies. In: ASIS&T Bulletin, June/July 2006 • http://www.asist.org/Bulletin/Jun-06/vizine-goetz_houghton_childress.html
Voß, Jakob (2007): Tagging, Folksonomy & Co - Renaissance of Manual Indexing? Paper Submitted to the 10th International Symposium for Information Science, Cologne • http://arxiv.org/PS_cache/cs/pdf/0701/0701072v1.pdf
W3C / Burrueta, D. and Phipps, J. (2008): Best Practice Recipes for Publishing RDF Vocabularies. W3C Working Group Note 28 August 2008 • http://www.w3.org/TR/swbp-vocab-pub/
W3C / Isaac, A. and Summers, Ed (2008): SKOS Simple Knowledge Organization System Primer, W3C Working Draft 29 August 2008 • http://www.w3.org/TR/skos-primer/
W3C / Isaac, A., Phipps, J. and Rubin, D. (2007): SKOS Use Cases and Requirements. W3C Working Draft 16 May 2007 http://www.w3.org/TR/2007/WD-skos-ucr-20070516/
W3C / Miles, A. and Bechhofer, S. (2008): SKOS Simple Knowledge Organization System Reference, W3C Working Draft 29 August 2008 • http://www.w3.org/TR/skos-reference/
W3C / Sauermann, L. and Cyganiak, R. (2008): Cool URIs for the Semantic Web. W3C Interest Group Note 31 March 2008 http://www.w3.org/TR/cooluris/
W3C SWDWG – Semantic Web Deployment Working Group (2007): SWDWG Amsterdam F2F October 2007: Topic: SKOS Concept Semantics. Patterns for Working with SKOS and OWL (10 May 2007) http://purl.org/net/skos/2007/10/f2f/skos-owl-patterns.html
W3C: RDF Vocabulary Description Language schema (RDFS) http://www.w3.org/TR/rdfschema
W3C: Resource Description Framework (RDF) Model and Syntax Specification http://www.w3.org/TR/REC-rdf-syntax/
Wei, Q., Freeland, C. and Heidorn, P.B (2008): An Evaluation of Taxonomic Name Recognition (TNR) in the Biodiversity Heritage Library. In: Proceedings of TDWG, 2008 • http://www.tdwg.org/proceedings/article/view/380
Weinberger, David (2002): Small Pieces Loosely Joined: A Unified Theory of the Web. Perseus Publishing
Weinberger, David (2005): Trees vs. Leaves. JOHO – The Journal of the Hyperlinked Organization http://www.hyperorg.com/backissues/joho-jan28-05.html#leaves
Weitzman, Anna L. and Lyal, Christopher (2004): An XML schema for taxonomic literature – taXMLit http://www.sil.si.edu/digitalcollections/bca/documentation/taXMLitv1-3Intro.pdf
Weitzman, Anna L. and Lyal, Christopher (2005): INOTAXA – INtegrated Open TAXonomic Access and the "Biologia Centrali-Americana" • http://units.sla.org/division/dbio/events/conf_past/Baltimore/inotaxa.pdf
Wester, Jeroen and Nederbragt, Hans (2007): RNA-project: Using things like thesauri and taxonomies in real cases!, pp. 93-99, in: Aroyo, L., Hyvönen, E. and van Ossenbruggen, J. (2007): Cultural Heritage on the Semantic Web. Workshop 9 of the 6th International Semantic Web Conference, Korea, 2007 http://www.cs.vu.nl/~laroyo/CH-SW/ISWC-wp9-proceedings.pdf
Wheeler, Quentin D. (2007): Invertebrate systematics or spineless taxonomy? In: Zhang, Z.-Q. & Shear, W.A. (eds., 2007): Linnaeus Tercentenary: Progress in Invertebrate Taxonomy. Zootaxa 1668: 11-18 http://www.mapress.com/zootaxa/2007f/zt01668p018.pdf
Wheeler, Quentin D. (ed., 2008): The New Taxonomy. Systematics Association Special Volume Series. Boca Ratan, CRC Press.
Wielemaker, J., Hildebrand, M., van Ossenbruggen, J. (2007): Using Prolog as the fundament for applications on the Semantic Web. In: Heymans, S. et al. (eds.): Proceedings of the 2nd Workshop on Applications of Logic Programming and to the Web, Semantic Web and Semantic Web Services (Porto, Portugal, September 13, 2007). CEUR volume 287, pp. 84–98 • http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-287/paper_1.pdf
Zeng, M. and Chan, L. (2004): Trends and issues in establishing interoperability among knowledge organization systems. Journal of American Society for Information Science and Technology, 55(5), 377-395
Zeng, Marcia D. (2005): Standards for Controlled Vocabularies. 7th NKOS Workshop, JCDL2005, Denver http://nkos.slis.kent.edu/2005workshop/z3919.ppt
Zolly, Lisa (2004): The CSA/NBII Biocomplexity Thesaurus: Current Initiatives, Future Directions. CENDI Terminologies Workshop. Washington, DC, September 16, 2004 • http://www.cendi.gov/presentations/KOS_NBII_Zolly.ppt
Zthes specification for thesauri http://zthes.z3950.org

Disclaimer

Imprint

This report is a product of the STERNA project that is supported and partly funded by the *e*Content*plus* programme of the European Commission.

Author:
Guntram Geser, Salzburg Research

Reviewers:
Andreas Gruber, Salzburg Research
Patricia Mergen, Royal Museum for Central Africa

Graphics & layout:
Daniela Gnad, Salzburg Research

Images:
Image on cover and page 1: Courtesy of Halldor Eiriksson, Shutterstock
Images on pages 7, 17, 63 and 101: Courtesy of Teylers Museum
Image sources:
Page 7: Cornelis Nozeman, *Nederlandsche vogelen*. Amsterdam, J.C. Espp & Zoon, 1770
Page 17: Herman Schlegel et A.H. Verster de Wulverhorst, *Traité de Fauconnerie*. Leiden et Düsseldorf, Arnz & Comp, 1844
Page 63: John James Audubon, *The Birds of America*. London, R. Havell & Son, 1827-38
Page 101: Daniël Giraud Elliot, *A Monograph of the Pittidae, or Family of Ant-Thrushes*. New York, D. Appleton & Co, 1863

ISBN 978-3-902448-19-4

Printed in Austria, 2009

Project Coordinator

salzburg research

Salzburg Research Forschungsgesellschaft m.b.H.
Jakob Haringer Straße 5/3 | 5020 Salzburg, Austria
Phone: +43.662.2288-201 | Fax: +43.662.2288-222
Project website: http://www.sterna-net.eu

Project Management
Andrea M. Mulrenin
andrea.mulrenin@salzburgresearch.at

The STERNA Consortium and Contributors

The STERNA project is supported and partly funded by the *e*Content*plus* programme of the European Commission.

ISBN 978-3-902448-19-4