SEARCHING AND BROWSING MULTIPLE SUBJECT GATEWAYS IN THE RENARDUS SERVICE

Michael Day[1], Traugott Koch[2], Heike Neuroth[3]
[1]*UKOLN, University of Bath*, [2]*NetLab, Lund University Libraries*, [3]*Goettingen State and University Library*

This paper describes the main features of the Renardus cross-search and cross-browse service. Renardus is made up of a number of participating subject gateway services. Cross searching is based on the Z39.50 protocol. A review of data models in use by partner services helped define a minimum set of Dublin Core-based metadata elements that could be utilised as a common model for the Renardus service. This provides the basic infrastructure for interoperability between all participating gateways. The Renardus Service also uses classification mapping to enable subject browsing across all gateways with the Dewey Decimal Classification (DDC). The paper outlines the use of classification systems by Renardus partner gateways, the general mapping approaches taken by the project, the definition of mapping relationships, and technical solutions. There follows a description of how the mapping information is used within Renardus and several features that have been implemented to aid end-user navigation in deep subject-browsing structures.

Key words: information retrieval, metadata, classification mapping, interoperability

1 INTRODUCTION

Renardus (IST-1999-10562) was a research project funded between 2000 and 2002 by the European Commission as part of the Information Society Technologies (IST) programme. Project partners included national libraries, research centres and subject gateway services from Denmark, Finland, Germany, the Netherlands, Sweden and the UK; the project being co-ordinated by the National Library of the Netherlands (Heery, *et al*, 2001). The project developed a pilot Web-based broker service that enabled searching and subject-based browsing across a range of distributed subject gateway services (http://www.renardus.org/). The pilot service developed by the project has now migrated to an operational Renardus service hosted by the Goettingen State and University Library (SUB), managed by participating gateways through the Renardus Consortium (Huxley, 2002; Huxley, *et al.*, 2003).

Subject gateways are services that provide access to Internet resources that have been reviewed, selected and described by subject specialists. The exact selection criteria largely depend on the perceived usage base of the gateway, but typically include factors relating to the content and presentation of the resource and the integrity of the information and site provider (e.g., http://www.sosig.ac.uk/desire/ecrit.html). Subject gateways are almost always based on the manual creation of descriptive metadata and usually provide end users with both search and

[1] Corresponding author: Michael Day, UKOLN, University of Bath, Bath BA2 7AY, United Kingdom. E-mail: m.day@ukoln.ac.uk

subject-browse facilities. The existence of rich metadata means that gateways can offer more sophisticated search options than other Web indexes. The application of subject classification schemes means that gateway services often provide hierarchical browse structures for browsing (Koch, 2000). As the Internet itself is constantly evolving, subject gateways also need robust collection development policies that include the regular checking and updating of resources included in the database.

The Renardus project developed a pilot broker system that enabled searching and browsing across a number of distributed subject gateways through the use of a common metadata profile and by the mapping all locally-used classification schemes to a common scheme. This paper will explore some of the issues faced by the project in developing the pilot broker.

2 RENARDUS CROSS-SEARCHING

Content providers for the Renardus project were subject gateway services from Finland, Germany, the Netherlands, Sweden and the UK. These varied widely in their nature and details of technical implementation. Some services were focused on relatively limited subject ranges, e.g. both the German Centre for Documentation and Information in Agriculture's dainet service (http://www.dainet.de/) and Nova University's NovaGATE (Price, 2000) covered just agriculture, forestry and related subjects. The Goettingen State and University Library provided four gateways covering mathematics, the earth sciences, and Anglo-American history and literature (Fischer & Neuroth, 2000). Other services were more comprehensive in scope, e.g. the Finnish Virtual Library (http://www.jyu.fi/library/virtuaalikirjasto/) and DutchESS (Peereboom, 2000). One of the partner gateways, the UK Resource Discovery Network (http://www.rdn.ac.uk/), was itself a growing federation of gateway services with its own requirements for interoperability (e.g., Dempsey, 2000; Powell, 2001).

Table 1. The Renardus Application Profile

Element/Refinement	Scheme	Obligation
Title		Mandatory
Title.Alternative		Optional
Creator		Recommended
Creator	LastName, FirstName	Recommended
Description		Mandatory
Subject		Mandatory
Subject	Renardus-DDC	Mandatory
Subject	DDC, LCC, LCSH, MeSH, UDC	Recommended
Identifier	URI	Mandatory
Language	ISO 639-2	Recommended
Type		Recommended
Type	DCMI Type Vocabulary	Recommended
Country	ISO 3166-1	Recommended

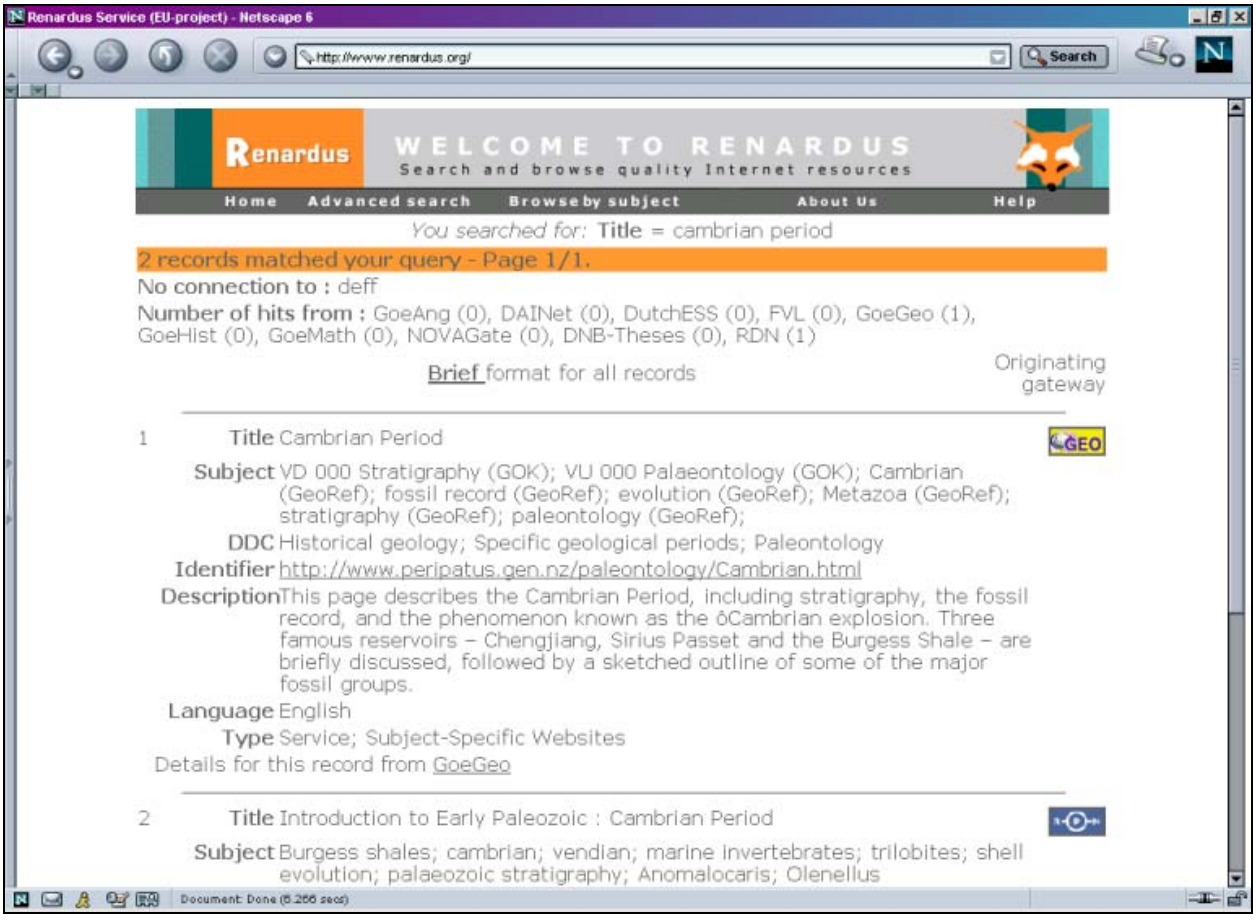
Source: Neuroth & Koch (2001)

As may be expected, many of these gateways had developed (or adopted) their own technical solutions. Metadata standards also varied, although most of these converged on implementations

of the Dublin Core metadata element set developed and maintained by the Dublin Core Metadata Initiative (DCMI). The objective of Renardus was to provide integrated access to the Internet resource catalogues provided by all participating gateways. Its first challenge was to devise a technical basis for doing so. After a review of broker models used to integrate access to distributed and heterogeneous information resources, the project chose to base the Renardus cross-search facility on the ANSI/NISO Z39.50 protocol and developed a project specific profile.

The project first undertook a detailed survey of all the data models in use by partner gateways and used the results to agree a minimum set of Dublin Core-based metadata elements that could be utilised as a common model (exchange format) for the Renardus service. The Dublin Core metadata element set was chosen because it offered a good prospect for interoperability with future partners or services (Neuroth & Koch, 2001). The survey looked at the semantics and syntax of all metadata elements used by participating services and their obligation, i.e. whether elements were mandatory, recommended or optional. Specific issues for Renardus were the language of the metadata content and whether specific schemes were used for dates, language codes, subjects, etc. From the data collected, the project defined a core set of elements that Renardus could use to link the diverse metadata used by participating services. The resulting data model, an application profile of Dublin Core (Heery & Patel, 2000), is summarised in Table 1.

Figure 1. Renardus Search Results (Detailed Record View)



Participating gateways were then expected to map their existing metadata schemas to this application profile. The minimum set of metadata that they were expected to provide included the Dublin Core elements 'title,' 'description,' 'subject' and 'identifier,' with an additional 'subject' field with a Dewey Decimal Classification (DDC) code extracted from the classification mapping information developed for the cross-gateway browse service (see section 3, below). Other elements were not mandatory, but their presence would greatly assist with enhancing the search functionality offered by Renardus. Additional 'administrative' elements included a 'Full Record URL' that would lead the users of Renardus to the metadata record created by the participating gateway and a service identifier ('SBIG ID') that would indicate which gateway the record originated from. All participating services were also expected to provide a collection level description record. More details on the Renardus application profile are available in the paper by Neuroth & Koch (2001).

Each participating service had to set up a Renardus Z39.50-compliant server and import records from their database normalised according to the Renardus profile. Once set up, the Renardus broker service can then search across these services in response to user queries and retrieve records in detailed or brief format (Figure 1). The interface offers both a simple and advanced search. The simple version searches the elements 'title,' 'description,' 'subject' and the DDC caption in 'subject.' Advanced searching enables field specific searches of all of these elements, 'creator' and document 'type,' with filtering by the document 'type,' 'language' and 'country' elements. In the advanced search, users can also select which particular gateways they would like to search.

3 RENARDUS CROSS-BROWSING

One of the most important services offered by gateways is subject browsing, typically based on classification schemes. The advantages of using classification systems to support subject access and navigation, multilingual access, for broadening or narrowing searches, etc. have been described elsewhere (e.g., Koch & Day, 1997). The Renardus project, therefore, proposed that it should provide some kind of subject browsing across all participating services (Koch, *et al*., 2003). Older projects, like the EU-funded DESIRE project (http://www.desire.org/), had already investigated the use of classification schemes by gateways and experimented with automatic classification technologies (Koch & Vizine-Goetz, 1999). Renardus, by contrast, was concerned with investigating ways in which users could browse a single subject hierarchy giving access to the content of all partner gateways. The problem was that different gateway services use a wide range of classification schemes to provide access to resources. These included subsets of well-known universal schemes like the Dewey Decimal Classification or Universal Decimal Classification, but also some specialised systems designed for particular subject areas or schemes produced locally by the gateway itself. In order to achieve *consistent* browse access to the content of partner gateways, Renardus decided that all of the different classification systems used needed to be mapped to a common classification system that could be used as a common switching language and browsing structure. The scheme chosen was the Dewey Decimal Classification (DDC).

A universal scheme like DDC had important advantages over other candidate schemes for an application like cross browsing in a universal service like Renardus. One of the main advantages to Renardus was that its online availability (e.g., in the form of WebDewey) meant

that it could be integrated as a useful tool in the Renardus mapping process. Other advantages included the scheme's universal subject coverage, its global use, the large number of digital resources that had been classified using it, and the speed and frequency of updates, especially with regard to the content of digital resources. The basis for the use of the DDC by Renardus was a research agreement with the scheme's owner, OCLC Forest Press, part of OCLC Online Computer Library Center (http://www.oclc.org/dewey/). The license allowed the project to use the full DDC classification system to construct and offer the Renardus cross-browsing pages.

As has been said before, the classification solutions that had been adopted by the gateways participating in Renardus were very heterogeneous. In order to help prepare the mapping effort, it was necessary to conduct a detailed review of the schemes in use by partner gateways. An analysis showed, for example, that several gateways used specialised subject schemes with deep structure. For example, one gateway had 800 thematic classes structured in five levels. Other subject structures were not so extensive, with one or two levels of hierarchy and between 18 and 60 classes that would require mapping.

3.1 The mapping process

Some practical principles were required to maintain consistency in the mappings and to ensure that the resulting Renardus browse interface was balanced. Firstly, it was agreed that mapping relationships would be expressed between a pair of classes and not between a DDC class and individual resources. Secondly, the mapping was to be carried out in one direction only, from the DDC to the local classification (the gateway's local browsing system). In order to help establish a balanced Renardus service during the development stage, it was suggested that gateways should finish mapping the top level of a local browse hierarchy, before moving progressively down through other levels. While the ultimate goal was to map DDC to all local classes, priority was given, however, to mapping the most frequently used classes in the local gateway.

In producing these initial guidelines, the project was aware that there were a large number of issues that required discussion. These issues included the specifics of how the DDC should be used to create a browse structure and how the mappings should be displayed in the Renardus browse interface. Other problematic issues included the depth of the mapping (on both sides), how Renardus should treat local classes that contained both generalities and specialities, the exclusion of non-topical classes (e.g. auxiliary tables). It was recognised that some of the subject areas that provide the main focus of a gateway could be located deep within the DDC hierarchy. It was also not clear how the project would solve the conflict between the compact structures that are often used in specialised subject classifications and the 'shattering' of the same discipline within universal systems. For example, engineering is expressed in 800 classes within the specialised Ei Thesaurus (http://www.ei.org/) but dispersed in around 2,300 categories in the DDC. Another problematic issue was the influence of the degree of subject overlap between the Renardus participants on the mapping practice. It remained to be seen what would be the best trade-off between consistency, accuracy and usability in the Renardus cross-browsing service.

3.2 Mapping relationships

Many other mapping projects, (e.g. those involved in conversions between two classification systems for use in OPACS or union catalogues) had limited themselves to the establishment of simple connections between pairs of classes. These projects are often unspecific concerning defining the character and degree of the indicated equivalence. However, the structures and levels of detail, the vocabularies, languages and cultural contexts of the locally applied classification systems used by Renardus gateways and the DDC are very different. Renardus, therefore, assumed that a simple equivalence between the content of two classes would be unusual.

For the Renardus subject browsing pages, it was felt that users needed to be advised that certain links from a DDC class, point to a class in a local gateway containing broader or narrower areas of content, or showing major or minor overlaps with the DDC class. This was especially true, as there would quite often be multiple links to classes found within a number of different gateways. One such link might be fully equivalent; another might show a minor overlap. The need for a more detailed specification of the degree of equivalence was even greater when the mapping was used in the Renardus advanced search feature. Using this, a result list could be ranked according to the degree of relationship between the individual resource's local class and the DDC class used for searching.

In order to deal with this problem, Renardus defined five distinct mapping relationships. The local class is deemed to be either fully equivalent, a narrower or broader equivalent, or has a major or minor overlap when compared with the DDC class. These relationships are influenced by the possible relationships between sets in set theory and can be illustrated via Venn diagrams. 'Fully equivalent' means that the subject content of the local page that one is linked to, is generally the same as the subject indicated on the Renardus browsing page. A 'narrower equivalent' indicates that the subject content of the local page is a true subset of the browsing page, whereas "broader equivalent" reflects the opposite, where the local page contains the entire subject content of the Renardus browsing page. A "major overlap" exists when the content of the local page represents a large part of the browsing page plus other related subjects. Conversely, "minor overlap" indicates some equivalence to part of the browsing page but that the class may also include other related subjects. Renardus maps in one direction only, from the DDC to the local classification(s). The three types of equivalence relationship require that one of the two classes is a true subset of the other, i.e. that it cannot also be mapped to another part of the classification scheme. Full equivalence is the intermediate situation where both classes are basically 100% equivalent. The two overlapping relationships require that parts of both classes clearly do not belong to the subject content of the other class. Thus certain logical rules apply which would permit a formal quality control of the mapping process.

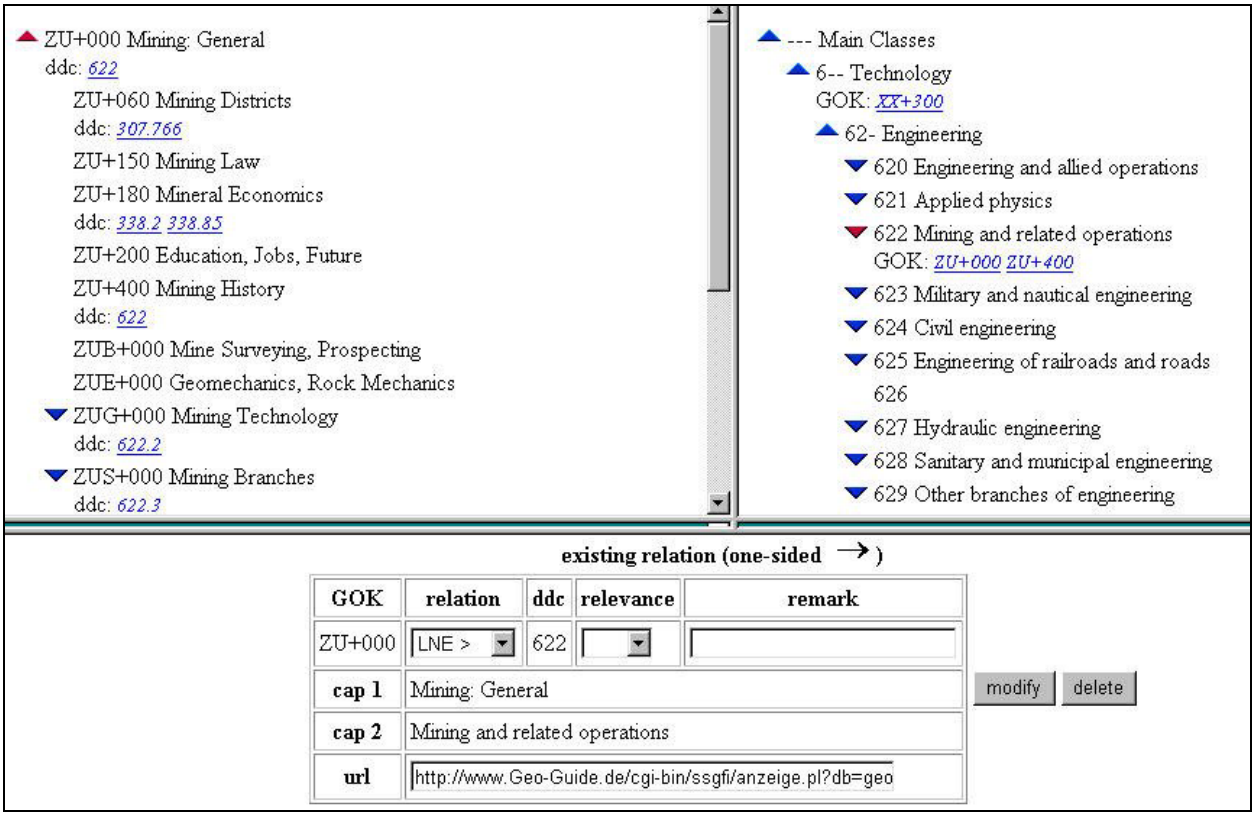
3.3 Technical solutions and tools

The main sources that are used for the classification mapping effort were the local classification systems and the enhanced DDC as presented by WebDewey. To support the practical effort, Renardus adapted a mapping tool developed by the German CARMEN project (http://www.bibliothek.uni-regensburg.de/projects/carmen12/). This tool is Web-based and requires the open-source database software mySQL, an Apache Web server, JavaScript, and PHP

scripts at the server side. The classification scheme information and mapping information were stored on different servers, partly to fulfil the obligations of the OCLC license. Each gateway participating in the mapping effort needed to provide a machine-readable version of their classification scheme (or schemes) that could be used by the mapping tool.

The user interface of the mapping tool (Figure 2) consists of three windows: one for the local target classification, another for displaying and navigating the source classification (the DDC). The third window receives and displays the mapping information, including relationships and notes. Mapping relationships are displayed as links in both classification windows. The tool was adapted to create and store the mapping information in a mySQL database in a syntax specified by Renardus. This information can be imported using Perl scripts into the main Renardus system in order to create the mapping links on the subject browsing pages and can also be used by each gateway's normalisation scripts so that they can generate a DDC mapping for each resource in the local gateway's Renardus database (one of the mandatory fields in the Renardus application profile).

Figure 2. Renardus Mapping Tool Interface



The enhanced DDC was delivered by OCLC in several XML encoded data files with a XML DTD, tag/attribute information and additional information about hierarchy. It contained 25,500 main schedule entries (notations) and 35,700 different records. Using these files, an initial complete hierarchical set of web pages can be generated allowing a user to navigate through the DDC structure. It was decided, however, that completely empty branches in the lower part of the DDC hierarchy could be removed from the display, assuming they were not required to assist as transitional steps during browsing.

3.4 The cross-browsing feature in Renardus

The Renardus pilot uses the DDC mapping information to support two functions, i.e. to create the cross-browse service and to provide additional information for the advanced search feature. The aim of the Renardus cross browse is to allow users to navigate through the subject hierarchies of the DDC classification and, on finding something suitable, to let them 'jump' from a chosen class to related classes in the local subject gateways. The project called this type of navigation 'browse and jump.' The system specifies the different equivalences and degrees of overlap in the user interface, enabling the user to visualise the resources in the context of their local browsing structures and to continue browsing there (Figure 3).

Figure 3. Renardus Browsing Interface

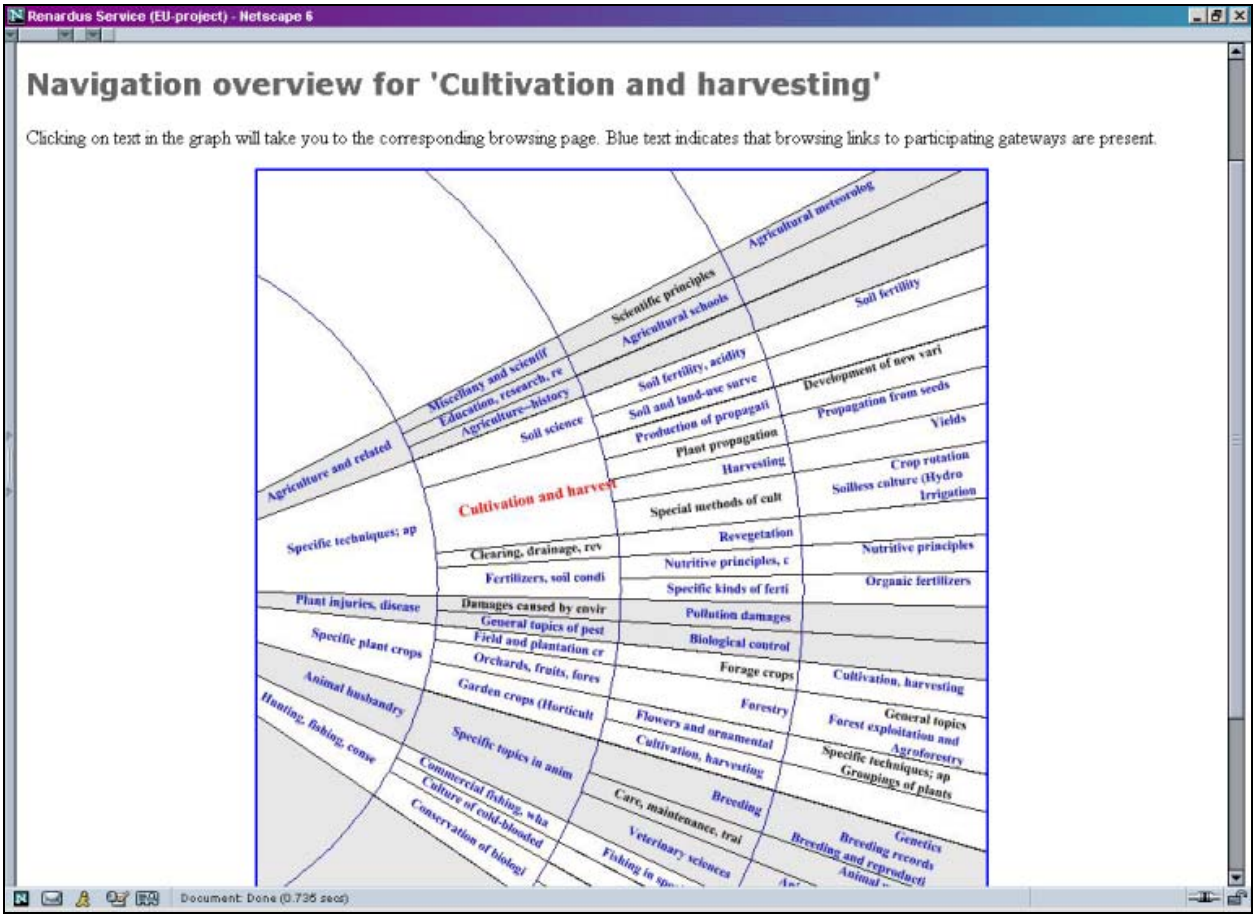
The upper part of every browse page displays the available categories in the actual section of the hierarchy, with links to all levels above and one level below for users to follow. The lower half of the browsing pages shows one or more links to related resource collections. The local classification caption, the local classification code and the icon of the gateway that the user would 'jump' to when clicking on the link, are also displayed. The related collections are presented in a ranked order according to the recorded mapping relationship: fully equivalent

classes are displayed first and minor overlapping classes last, thus encouraging the user to explore first the collections that are closest in coverage to the chosen DDC class.

It is clear that very large browsing structures - like that represented by the full DDC - need to provide additional assistance to guide users. Investigations by the project did not find any 'tried-and-tested' solutions that Renardus could immediately apply. Therefore, some experimental navigation support features were implemented for practical evaluation.

A 'graphical navigation overview' link (Figure 4) is available on every browse page. It provides a visual 'fisheye' overview of all the available categories that surround a chosen subject term, normally at one level above and two levels below within the hierarchy. Colours are used to help display the selected class within its context and all other classes that contain mappings. This feature is intended to increase the speed of users' navigation of the browse structure and to provide an immediate subject overview. Clicking on categories within the graphical display shows the relevant Renardus browsing page for this subject. An experimental text-based version of the browsing overview is also available.

Figure 4. Renardus Graphical Navigation Overview



On all browsing pages (apart from the top level) a search box is provided to 'Find a different start-page for browsing.' Using this, several valid alternative browsing pages are usually displayed. This feature offers a short cut for users who know significant terms from a valid category elsewhere in the Renardus DDC structure. This may also be an option if users have

difficulty finding exactly where their main area of interest is hidden within the DDC hierarchy. From the alternative list, users can go to a selected browsing page or graphically explore the hierarchical environment of this subject for further navigation.

Renardus also offers a short cut to viewing individual resource descriptions from all related collections with the feature: 'Merge the resource-descriptions from all related collections listed here.' Users will be shown an integrated list of resources from all related collections listed on the page, presented in the usual Renardus search results display. The main disadvantage with this 'virtual browsing' is that users may lose context and the potential additional information available from exploring the local gateway's browse structure. The same kind of search can be carried out on the 'advanced search' page by selecting the "DDC Classification" element for the search.

As noted earlier, the DDC mapping information is also used in the Renardus 'advanced search' feature. While the general subject element allows searching on all local subject information (e.g. uncontrolled keywords, controlled keywords from thesauri and subject headings, classification captions and notations, etc.) the "DDC classification" element enables searches to be made of captions of the mapped DDC classes.

One of the advantages of using a scheme like DDC is that the existence of translations means that it would be possible to 'plug-in' non-English language versions of the classification to generate browsing interfaces in multiple languages. In the final year of the project, Renardus experimented with interfaces to the browsing facility using versions of DDC in French, Italian, German and Spanish.

4 CONCLUSIONS AND FUTURE WORK

Funding for the Renardus project ended in 2002. Since then the pilot developed by the project has evolved into a service run by the Renardus Consortium, a co-operative venture of participating gateways and technical partners. The number of participating services has declined slightly since the end of the project, but the number of records being searched has continued to grow. Renardus continues as an active demonstrator broker service and the consortium is prepared to collaborate with other gateways and in relevant research and development opportunities.

Some future work might include further enhancements to the browsing interfaces. While Renardus developed interesting ways of browsing in large subject structures, there remains much that is not known about their effectiveness. That said, however, recent analysis of Renardus usage logs suggests that "systematic browsing of large information systems with the help of classification hierarchies seems to be widely accepted by users, especially when there is graphical support" (Koch, *et al*., 2004).

A wider area of continued development might be with relation to terminology services. These have been defined by Vizine-Goetz, *et al*. (2004) as Web Services that combine various types of knowledge organisation resources or vocabularies, "including authority files, subject heading systems, thesauri, Web taxonomies, and classification schemes." Such services developed by (or in association with) vocabulary owners could be used, for example, to provide up to date, authoritative and sustainable mapping information that could be used by services like Renardus for producing subject browse interfaces, for search query expansion, and much else.

There is much research work going on in this area at the moment, some of it related to Semantic Web developments, so it will be an interesting area to watch.

REFERENCES

Dempsey, L. (2000). The subject gateway: experiences and issues based on the emergence of the Resource Discovery Network. *Online Information Review*, 24(1), 8-23.

Fischer, T., & Neuroth, H. (2000). SSG-FI - special subject gateways to high quality Internet resources for scientific users. *Online Information Review*, 24(1), 64-68.

Heery, R., & Patel, M. (2000). Application profiles: mixing and matching metadata schemas. *Ariadne*, 25. Retrieved June 18, 2004, from: http://www.ariadne.ac.uk/issue25/app-profiles/

Heery, R., Carpenter, L., & Day, M. (2001). Renardus project developments and the wider digital library context. *D-Lib Magazine*, 7(4). Retrieved June 18, 2004, from: http://www.dlib.org/dlib/april01/heery/04heery.html

Huxley, L. (2002). Renardus: following the fox from project to service. In: *Research and Advanced Technology for Digital Technology: 6th European Conference, ECDL 2002, Rome, Italy, September 16-18, 2002* (pp. 218-229). Lecture Notes in Computer Science, 2458. Heidelberg: Springer-Verlag.

Huxley, L., Carpenter, L., & Peereboom, M. (2003). The Renardus broker service: collaborative frameworks and tools. *The Electronic Library*, 21(1), 39-48.

Koch, T. (2000). Quality-controlled subject gateways: definitions, typologies, empirical overview. *Online Information Review*, 24(1), 24-34.

Koch, T., & Day, M. (1997). *The role of classification schemes in Internet resource description and discovery*. DESIRE project. Retrieved June 18, 2004, from: http://www.ukoln.ac.uk/metadata/desire/classification/

Koch, T., & Vizine-Goetz, D. (1999). Automatic classification and content navigation support for Web services. In: *Annual Review of OCLC Research 1998*. Retrieved June 18, 2004, from: http://digitalarchive.oclc.org/da/ViewObject.jsp?objid=0000003489

Koch, T., Neuroth, H., & Day, M. (2003). Renardus: cross-browsing European subject gateways via a common classification system (DDC). In: I. C. McIlwaine (ed.), *Subject retrieval in a networked world* (pp. 25-33). UBCIM Publications, 25. Munich: K. G. Saur. Retrieved June 18, 2004, from: http://www.lub.lu.se/~traugott/ drafts/preifla-final.html

Koch, T., Ardö, A., & Golub, K. (2004). Browsing and searching behavior in the Renardus Web service: a study based on log analysis. Poster, Joint Conference on Digital Libraries 2004 (JCDL 2004), Tuscon, Ariz., USA, 7-11 June 2004. Retrieved June 18, 2004, from: http://www.it.lth.se/knowlib/publ.htm

Neuroth, H., & Koch, T. (2001). Metadata mapping and application profiles: approaches to providing the cross-searching of heterogeneous resources in the EU project Renardus. In: *DC-2001: International Conference on Dublin Core and Metadata Applications 2001, Tokyo, Japan, 24-26 October 2001*. Retrieved June 18, 2004, from: http://www.nii.ac.jp/dc2001/proceedings/product/paper-21.pdf

Peereboom, M. (2000). DutchESS: Dutch Electronic Subject Service - a Dutch national collaborative effort. *Online Information Review*, 24(1), 46-49.

Powell, A. (2001). An OAI approach to sharing subject gateway content. Poster, WWW10, the Tenth International World Wide Web Conference, Hong Kong, 1-5 May 2001. Retrieved June 18, 2004, from: http://www10.org/cdrom/posters/1097.pdf

Price, A. (2000). NOVAGate - a Nordic gateway to electronic resources in the forestry, veterinary and agricultural sciences. *Online Information Review*, 24(1), 69-73.

Vizine-Goetz, Hickey, C., Houghton, A., & Thompson, R. (2004). Vocabulary mapping for terminology services. *Journal of Digital Information*, 4(4). Retrieved June 18, 2004, from: http://jodi.ecs.soton.ac.uk/Articles/v04/i04/Vizine-Goetz/