Having the right connections: the LIMBER project

By Ken Miller, Information Development, UK Data Archive, University of Essex and Brian Matthews, Dept. of Information Technology, CLRC-Rutherford Appleton Laboratory.

Introduction

As with any journey, you have to make the right connections if you want to reach your desired destination. The goal in the LIMBER project is to facilitate cross-European data analysis independent of domain, resource, language and vocabulary. This paper describes the expertise, associations, standards and architecture underlying the project deliverables designed to achieve the project's ambitious aims.

The project

Limber (Language Independent Metadata Browsing of European Resources) is an EU (European Union) IST (Information Societies Technology) funded project that seeks to address the problems of linguistic and discipline boundaries, which, within a more integrated European environment, are becoming increasingly important. Decision-makers, researchers and journalists need to be provided with a broader, comparative picture of society across the continent; with the social science information often required to be correlated with information from domains such as environmental science, geography and health. This cross-discipline interoperability will be provided via a uniform metadata description. In addition, the provision of multilingual user interfaces and the controlled vocabulary of a multi-lingual thesaurus will make these datasets globally accessible in a range of end user natural languages.

The partners

The UK Central Laboratory of the Research Councils, Rutherford Appleton Laboratory (CLRC/RAL), act as co-ordinator of the project and as an innovator partner. As the UK office for W3C, the World Wide Web Consortium, CLRC/RAL has access to the latest developments in web technologies and recommendations that will be used in the project, in particular the Resource Description Framework (RDF) which will be implemented in the project tools. The Information Technology Department employs about 150 staff researching and developing IT standards and software, including considerable experience in EU linguistic and metadata projects.

The United Kingdom Social Science Data Archive (UKDA) houses the largest collection of accessible computer-readable data in the social sciences and humanities in the UK. The UKDA's mission is to preserve electronic data so that they continue to be available for use and to promote the wider and more informed use of those data in research and teaching. The UKDA has an international reputation among data archives for its technical developments and its involvement in several leading EU projects, such as NESSTAR, a pan European information and retrieval tool. The UKDA's thesaurus, HASSET (Humanities And Social Science Electronic Thesaurus), is available to all via the Web for the indexing of humanities and social science datasets.

The Norwegian Social Science Data Services (NSD) is currently one of the largest social science data archives and resource centres in the world. Besides being a gateway to rich data holdings, NSD also serves as a competence centre assisting researchers with respect to data gathering, questionnaire design, selection of software tools, social science data analysis and methodology. NSD also gives high priority to the development of software tools for data browsing and visualisation over the

Internet, as shown in the NESSTAR tools, an advanced gateway to the holdings and services of data archives and digital data libraries world-wide.

INTRASOFT International based in Luxembourg, and with its R&D Section in Athens, is a member of the Services and Informatics Expertise (SIX) Advisory Group. This independent European body, whose members are the leading European software and services companies, aims at increasing the role of the European software and services industry in the development of the Information Society. INTRASOFT believes that technologies whose target is to reduce the language barrier in web-based IT services will be of much demand in the near future. They have been involved in several EU linguistic and metadata projects as well as developing commercial software products.

The user group

The Council of European Social Science Data Archives (CESSDA) promotes the acquisition, archiving and distribution of electronic data for social science teaching and research in Europe. It encourages the exchange of data and technology and fosters the development of new organisations in sympathy with its aims. It associates and co-operates with other international organisations sharing similar objectives. In particular the German, French, Spanish and Greek Data Archives are involved in the evaluation of the multi-lingual thesaurus.

The community

The International Association for Social Science Information Service & Technology (IASSIST) is an organisation whose aim is to advance the interests of data professionals, to promote professional development, and to take an active role in the promotion of global exchange of information, experience and standards. With its global focus, IASSIST offers an opportunity to extend the professional activities and interests of its members beyond national boundaries, advancing issues to do with the development of social science information service and technology.

The Data Documentation Initiative (DDI) is an effort to establish an international criterion and methodology for the content, presentation, transport, and preservation of "metadata" about datasets in the social and behavioural sciences. With the achievements of the DDI, metadata can now be created in a uniform, highly structured format that is easily and precisely searchable on the Web, that lends itself well to simultaneous use of multiple datasets, and that will significantly improve the content and usability of metadata.

The Networked Knowledge Organization Systems/Services organisation (NKOS) is devoted to the discussion of the functional and data model for enabling knowledge organisation systems, such as classification systems, thesauri, gazetteers, and ontologies, as networked interactive information services to support the description and retrieval of diverse information resources through the Internet.

The SCHEMAS project is funded as part of the IST Programme, a theme of the EU's Fifth Framework. It aims to provide a forum for metadata schema designers involved in projects under the IST Programme and national initiatives in Europe. The main objectives are to provide information for schema implementers about the status and proper use of new and emerging metadata standards, as well as promoting good-practice guidelines for adapting multiple standards or metadata modules for local use in customised schemas.

The Institute for Learning and Research Technology (ILRT) is a centre of excellence in the development and use of Information and Communication Technology (ICT) to support learning and research. The ILRT focuses on a suite of projects and services in the development and use of ICT, supplemented by complementary training, advisory and consultancy activities. The ILRT has also developed a RDFschema for thesauri.

The Resource Description Framework (RDF) is a general framework for describing metadata about Web accessible resources being developed by the World Wide Web Consortium (W3C). This framework is intended to provide a simple model for user communities to define their own metadata descriptions, which can then be interpreted throughout the Web, especially via automated processors. RDF is based on a simple graph model capturing the resources on the Web and the relationships between them, in a flexible and extensible manner. This is realised using an XML format to compatible with other XML developments and tools. Thus by using RDF each user community can describe the properties of it own resources and combine them with the descriptions from other communities in a uniform manner. Thus RDF will allow the use of metadata across domains.

The dream (Figure 1)

Consider a data user sitting at his client PC, which has a user preference file written to an international standard and marked up in RDF. This user preference file contains information such as:- preferred language, second language, domain preferences, geographic preferences, metadata search elements, metadata hit list elements, ranking criteria and the IP address of the common gateway.

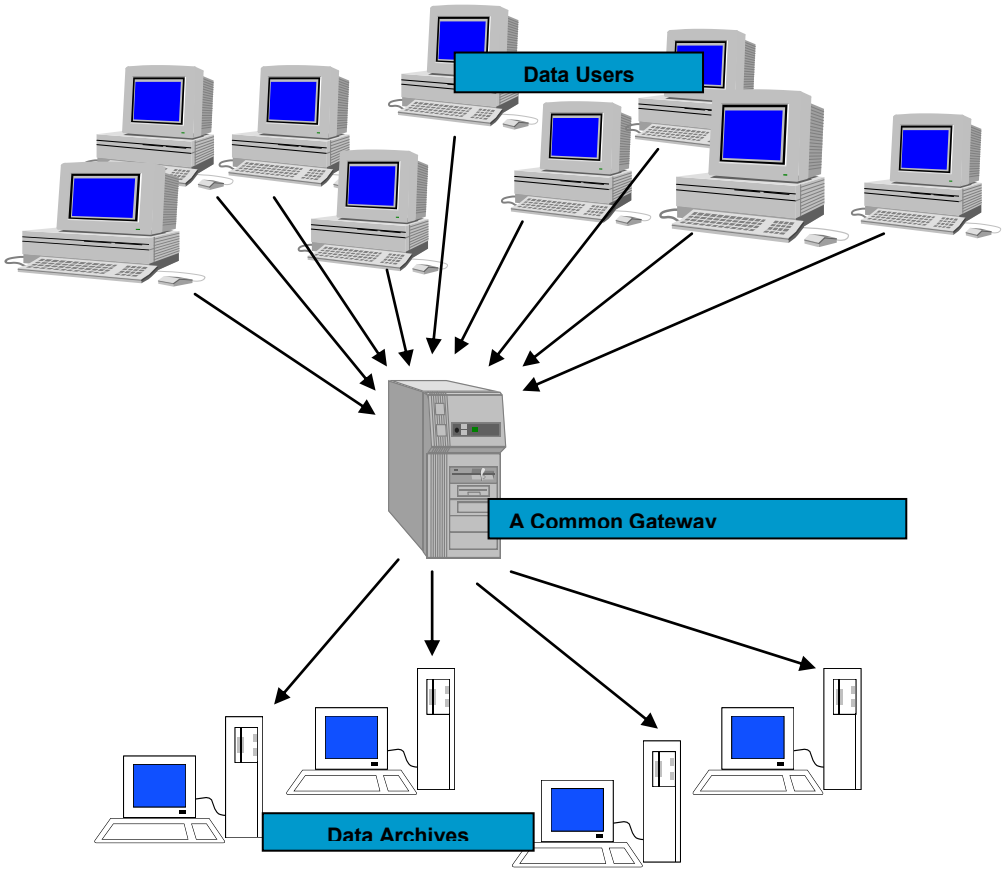

Figure 1. The Dream Gateway.

At the gateway there is the IP address of a resource directory, again written to an international standard and marked up in RDF, which can then be searched for resources that cover the user's domain and geographic preferences. The gateway would also have IP addresses of a thesauri directory and a metadata directory categorised by domain and marked up in RDF to international standards.

The user search interface is presented with search boxes and language as indicated in the user preference file and the domain and language specific thesaurus is made available for the user to select terms from the controlled vocabulary. The search is sent to the gateway as yet another RDF standard message.

At the gateway this search is distributed to the various resources identified in the resource directory. This directory also holds information as to which metadata standard, which language and which thesauri have been used at each resource. The thesauri directory holds information about mappings between same domain thesauri in different languages and between thesauri of different domains. The metadata directory holds information about mappings between the elements in each of the domain metadata standards. Hence the search entered by the user on elements from his domain metadata standard in his preferred language, using the controlled vocabulary of the domain thesauri, gets converted for resources being searched to reflect the appropriate language, thesauri and metadata.

The gateway, using the same resource, metadata and thesauri directories, then converts the information retrieved from each resource into the user's desired language and metadata elements and passes them back to the client. On the client the retrievals are displayed in a hit list using the elements and ranking defined in the user preference file.

The reality (Figure 2)

Unfortunately there is no international standard for user preferences, although there is a RDFschema for one in the NESSTAR system which could be expanded to cover the needs of LIMBER. There is also no international standard for thesauri content, although a RDFschema has been developed by ILRT, which LIMBER is modifying and extending. There is also no registry of thesauri and mappings, although NKOS hope to provide one. There is also no registry of metadata schemes and mappings, although the SCHEMAS project hope to provide one.

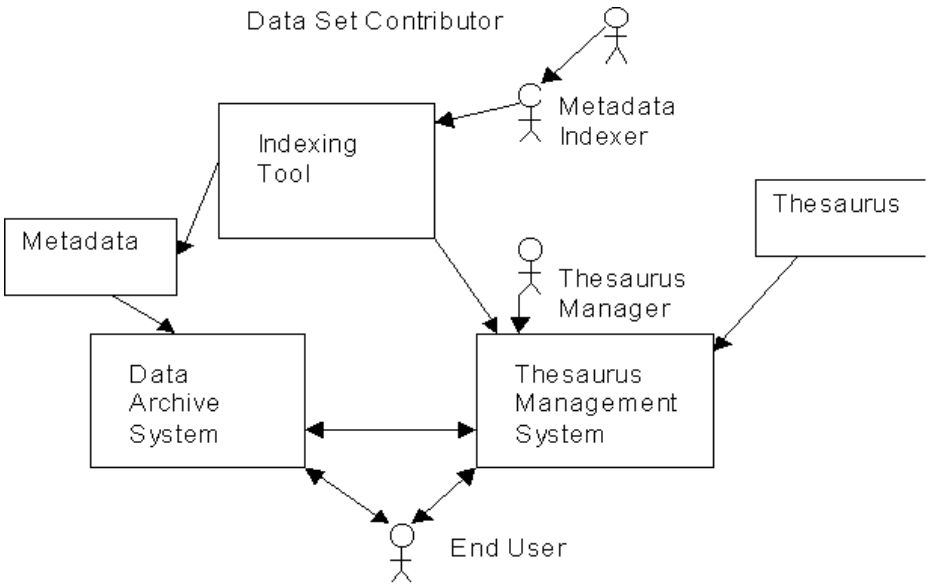

Figure 2. Overall Architecture of the LIMBER Multilingual Data Archive System

There are several schemes that could be used as part of a resource directory, namely Z39.50 explain, RSS (Rich Site Schema) and Dublin Core but none of these exactly meet the needs of LIMBER. However most metadata schemes do map to the Dublin Core, so initially there could be inter-operability between domains at this basic 15-element level.

LIMBER will use the multi-lingual ELSST (European Language Social Science Thesaurus), derived and translated from HASSET. Since copyright of HASSET, recognised as an excellent domain specific thesaurus, resides with the UKDA conversion to the LIMBER modified ILRT RDF format will be possible. It is hoped that other domain specific thesauri will also convert to this format and make their resource readily available. Incidentally the examples given in the published version of the ILRT RDFschema for thesauri all come from HASSET.

Tools developed in LIMBER will work with any thesaurus marked up in the LIMBER RDF format, and the semi-automatic indexing tool will apply keywords from these thesauri to any metadata record marked up in either XML or RDF. Initially all interfaces will have the choice of being displayed in French, German or Spanish, although the architecture and underlying format will allow further languages in the future.

Therefore, even without these international standards being available, LIMBER will still be able to provide multi-lingual interfaces to thesaurus aided searching across domains, using thesauri conforming to the LIMBER RDFschema and retrieving metadata mapped to the Dublin Core with assigned keywords translated back to the user's native language. The underlying metadata having been semi-automatically indexed by terms from the conforming thesauri. The information passed to search engines will allow selection of display elements in the hit list and ranking of items retrieved. Not quite the full dream - but a reality that allows that dream to come true.

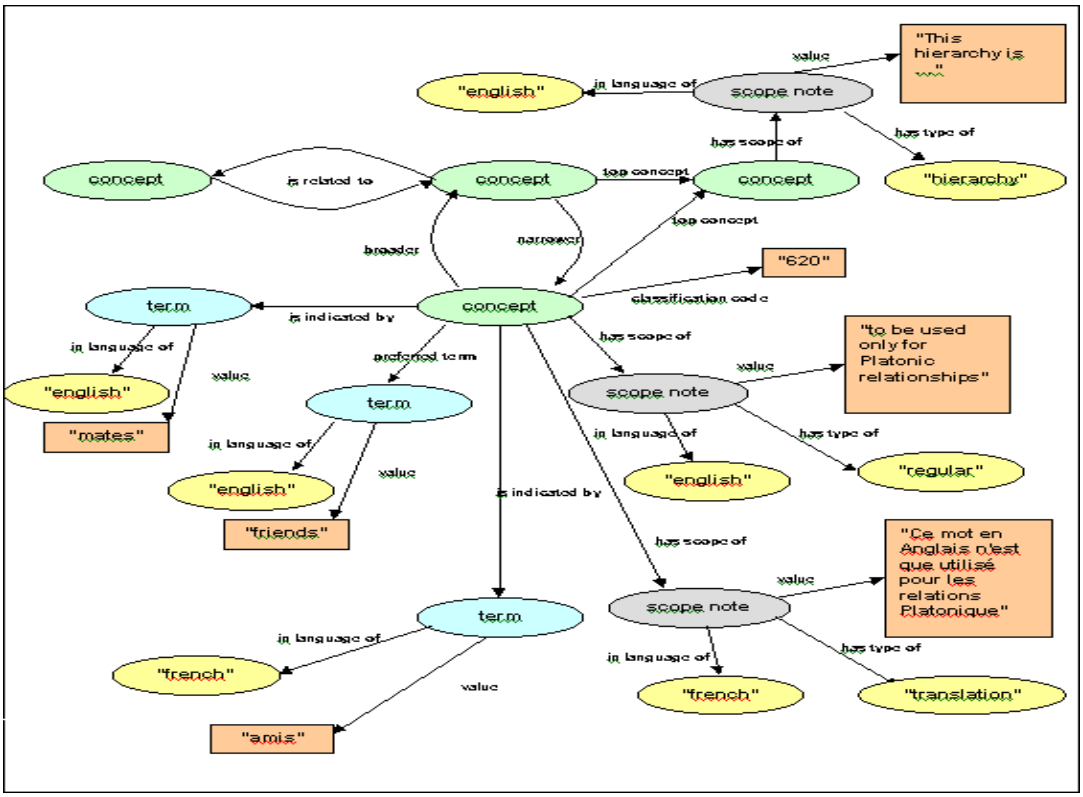

Figure 3. A RDF instance diagram for the multilingual *Friend* concept.

The thesaurus (Figure 3)

The ELSST thesaurus will be created from the present UKDA HASSET. This will involve reducing the present hierarchies so that all cultural and institutional specificity are removed. On the assumption that other domain specific thesauri will be available in the same RDF format certain hierarchies will be dropped, as they can be included at a later date by mapping to an existing thesaurus. Certain areas not covered at present, such as methodology, will be added to the new thesaurus.

The resulting broad-based social science thesaurus will be suitable for use by any resource in the social science domain. Because of the time limitations of the project, a target of 1,500 preferred terms from the minimum of 20 hierarchies has been set. The thesaurus will also include all synonyms to these terms and all top terms of hierarchies in the existing HASSET that either map to existing thesauri or which, although not in the major 20 hierarchies, would have been present if resources were available. Each hierarchy will be sent to the CESSDA archives for evaluation of coverage and usefulness.

As each hierarchy is reduced it will be translated, initially into French, German and Spanish, although several CESSDA archives have also expressed a desire to allocate independent resources to translate into their own languages. Although it is hoped that, at this broad level, one-to-one equivalence will be possible for the vast majority of terms, the format will allow for non-equivalence and different structures in each language. Extensive use of scope notes will resolve ambiguities, translation assumptions and subject coverage of hierarchies. The translated hierarchies will be sent to the appropriate archives of CESSDA for evaluation and addition of language specific synonyms.

The resulting four language, multi-lingual thesaurus will then be converted into the LIMBER RDFschema format ready for use in the management, indexing, browsing and search tools developed in LIMBER.

The metadata

The present metadata standard used by the CESSDA members is the DDI XML (eXtensible Mark-up Language) codebook DTD (Document Type Definition). It is specifically defined to describe flat single survey data files. However, problems arise when different types of data need to be described, when series data need to be linked, when different output formats are required and when references to different metadata standards need to be made. All these would require major changes to the DTD, thus making metadata already created invalid against the new format.

LIMBER proposes to investigate the possibility of changing the DDI codebook to a more modular and extensible format such as RDF or XMLschema. The project plans to develop a high level object orientated conceptual model that could be translated into whichever format becomes international excepted. This would allow for wider coverage, cross-domain linkages, fix present inconsistencies and give the standard a formal development and upgrade path, which would ensure that earlier versions still compiled with the latest release of the standard.

The multi-lingual interface

All screens and drop-down menus will be available in German, French, Spanish and English to begin with, but defined in a standard format that can easily be translated to other languages in the future. All metadata retrieved will, where possible, be displayed in either the user's first or second choice language. This will be achieved via the creation of metadata element heading files in all four languages, translation of keywords from the thesaurus and display of the correct language variance within the metadata itself. The query itself will be translated via the multi-lingual thesauri, unless the resources searched are all using the same thesaurus, in which case the search would be carried out on the standard notation common to all languages.

The architecture (Figure 2)

The LIMBER system is designed as three distinct stand alone products: a multi-lingual thesaurus management tool, a user browsing interface and a semi-automatic indexing tool. However, the architecture underlying these will also allow existing search engines to use the LIMBER products as a plug-in to enable multi-lingual cross-domain searching and relevance feedback.

The following calls have been identified as those necessary in any API (application program interface).

GetSynonyms, a function that returns all the synonyms of a preferred term from a thesaurus, used for free-text searching.

GetCode, a function that returns the notation code of a preferred term from a thesaurus, used to perform controlled vocabulary searches over resources using the same thesaurus to index their metadata records.

GetLanguageX, a function that returns the equivalent term in language X of a preferred term from a thesaurus identified in the user's natural language, used to perform multi-lingual free-text searches and provide relevance feedback to the metadata retrieved.

GetRelated, a function that returns all the immediate relationship terms (narrower, broader, related) of a preferred term from a thesaurus, used to display alternative search strategies.

GetScopeNote, a function that returns the scope notes assigned to a preferred term from a thesaurus, used to explain ambiguity, translation assumptions and scope of hierarchies.

GetContaining, a function that returns all preferred terms and synonyms that contain words from a user entered string, used when a controlled vocabulary search is attempted on a term not found in a thesaurus. This function has the possibility of two further variations which involve the use of either stemming or truncation techniques.

GetHierarchy, a function that lists all narrower terms to a preferred term that is the top term of a particular hierarchy, used for browsing a thesaurus for possible search terms.

The tools

Of the tools to be developed in the LIMBER project, the one that poses the biggest challenge is the semi-automatic indexing tool. The objective is to assign keywords from a thesaurus to specific elements within any metadata records that conform to an XML or RDF standard. The metadata record will be parsed and the concepts within each element extracted. The extracted concepts will then be matched against the domain specific thesaurus and the most likely preferred terms listed for possible inclusion as index terms.

Those concepts that cannot be matched will be listed as possible additions to the thesaurus. If selected, the term will be automatically included in the metadata record in the format defined by the XML or RDF schema.

This tool will also learn from the metadata that it indexes. This learning process can be accelerated in certain instances where keywords have already been assigned manually; in the case of the UKDA, 4,000 metadata records have already been assigned keywords from HASSET.

Conclusion

The LIMBER project does have very ambitious goals. However, we feel confident given the connections that have been made through our partners, the user group and other institutions working in similar fields, that these goals are attainable. The adoption of the standards and architecture chosen will hopefully result in a positive demonstration of a unique and powerful tool for resource discovery across domain and language.

References

Project Deliverables: -

Wilson, M.D. LIMBER Annexe 1 "Description of Work"
Miller, K.P. LIMBER D1 "Requirements Definition"
Ramfos, A. and Malamateniou, F. LIMBER D2 "System Architecture"

Project Web Site: -

LIMBER: http://venus.cis.rl.ac.uk/limber/

User Group Web Site: -

CESSDA (Council of European Social Science Data Archives):
http://www.nsd.uib.no/Cessda/

Community Web Sites: -

IASSIST (International Association for Social Science Information Service & Technology): http://datalib.library.ualberta.ca:80/iassist/
NKOS (Networked Knowledge Organization systems/Services):
http://www.alexandria.ucsb.edu/~lhill/nkos/index.html
ILRT (The Institute for Learning and Research Technology) http://www.ilrt.bris.ac.uk/
SCHEMAS (EU IST project): http://www.schemas-forum.org/

Standards Web Sites: -

DDI (Data Documentation Initiative) CodeBook DTD (Document Type Definition):
http://www.icpsr.umich.edu/DDI/codebook.html
RDF (Resource Description Framework): http://www.w3.org/RDF/
XML (eXtensible Mark-up Language): http://www.w3.org/XML/
Dublin Core: http://purl.org/DC/

Related Web Sites:-

NESSTAR (Networked European Social Science Tools and Resources):
http://www.nesstar.org/
HASSET (Humanities and Social Science Electronic Thesaurus):
http://biron.essex.ac.uk/searching/zhasset.html