OECD CONFERENCE

ASSESSING THE FEASIBILITY OF MICRO-DATA ACCESS
LUXEMBOURG 26–27 OCTOBER 2006

Session 3: Micro-data in practice

A common interface to a semantic web for the social sciences

Submitted by Norwegian Social Science Data Services
Prepared by Atle Alvheim (atle.alvheim@nsd.uib.no)

Abstract: This paper describes an EU financed project, Madiera (Multi-lingual Access to Data Infrastructures in the European Research Area). The idea of the project is to implement the vision of a European Social Science data portal, summarized in the following points:

1. The development of an integrated and effective distributed social science portal to facilitate access to a range of data archives and disparate resources.
2. The employment of a multi-lingual thesaurus to break the language barriers to the discovery of key resources.
3. The development of specific add-ons to existing virtual data library technologies, in particular data location technologies and a metadata standard for scientific empirical material.
4. An extensive programme to add content, both at the data/information and knowledge levels.
5. Extensive training of data providers and users to encourage the continuous growth of the infrastructure

Introduction

> "It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts. **"**
>
> *Sir Arthur Conan Doyle, British mystery author & physician (1859–1930)*

This paper holds as its basic starting point that data are the single most important component necessary for a science-based understanding of society. However, empirical comparative research in Europe is hampered by a fragmentation of the scientific information space. Data and its derivatives, information and knowledge, are scattered in space and divided by language and institutional barriers. Consequently, too much research is based on data from a single nation, carried out by a single-nation team of researchers and communicated to a single-nation audience. This state of affairs is preventing the development of a comparative and cumulative research process integrating and nurturing the entire European Research Area.

Yesterday's answers to these challenges would probably have been formulated in terms of centralisation and establishment of large-scale European-wide institutions. Today's answers focus on the power of emerging information technologies to encourage communication, sharing and collaboration across spatially dispersed but scientifically related communities. Grid computing is increasingly being viewed as the next phase of distributed computing. Grid computing enables participants to share computing and information resources across organizational boundaries in a secure and efficient manner. More than anything, this suits the idea of collaborative and comparative scientific research. Grid computing enables research-oriented organizations to solve problems that were infeasible to solve due to computing-, data access and data integration constraints. In addition grids may reduce costs through automation and improved IT resource utilization, the expectation is that over time grid computing will enable a more flexible and efficient pan-European computing infrastructure. However, for most of the social sciences it is the data that is the critical resource, it is not the computer power.

This implies then that it is not solely or primarily a question of computer technology. Technology can only improve, strengthen and augment processes and activities that already are established. A virtual infrastructure will only make sense if it connects existing and well functioning providers of content and services and it will only survive if it is meeting the demands of their users. This leads over into the semantic web extension of the ordinary web, where information is given well-defined meaning.

Whilst data is not necessarily a scarce resource in Europe, they are not as available as they could be. Well-developed official statistical systems combined with a variety of both academically and commercially driven data gathering programs and activities are producing a wealth of data and information about various aspects of the European societies. Moreover, in the majority of European countries social science data archives have been established to secure the longer-term preservation of large parts of the available resources. These are institutions that do not to any significant degree collect data themselves, but are there mainly to preserve and make available for potential use what others may have collected. Across this field of producers we find that availability are severely hampered by technological, judicial, economic and retrieval-related factors. Data are locked in systems, fenced by (un)necessary rigorous rules, treated as an economic commodity, not being adequately documented and often not being intended for alternative use by original data collectors.

If "sharing" is the most important single keyword characterizing a true grid, the key to realizing the benefits of grid computing is *standardization.* Standardization facilitates development or inte-

gration of computer software so that the diverse resources that make up a modern computing environment can be discovered, accessed, allocated, monitored, and in general managed as single virtual systems – even when provided by different vendors or operated by different organizations.

The vision of the MADIERA project run by some of the European data archives was to develop an effective infrastructure for the European social science community by integrating data with other tools, resources and products of the research process. The aim was a fully operational Web-based infrastructure populated with a variety of data and resources from a selection of providers, a common integrated interface to the resources of the majority of the existing 20+ social science data archives in Europe including several newly established archives in the candidate countries. Furthermore, the infrastructure should, as the Web itself, have the capacity to grow and diversify even after an initial construction period. The main objective of the project is to create a sustainable system, nurtured by the collective energy of the data and knowledge producing communities of the European Research Area.

Breaking this vision down into more specific objectives, we can outline the following specific goals:

1. The development of an integrated and effective distributed social science portal to facilitate access to a range of data archives and disparate resources.
2. The employment of a multi-lingual thesaurus to break the language barriers to the discovery of key resources.
3. The development of specific add-ons to existing virtual data library technologies, in particular data location technologies and a metadata standard for scientific empirical material.
4. An extensive programme to add content, both at the data/information and knowledge levels.
5. Extensive training of data providers and users to encourage the continuous growth of the infrastructure
6. The gradual integration of the emerging national infrastructures of the candidate countries into the European Research Area.

The main components

The present MADIERA portal is based on three main components.

- A common standard for data documentation, the *DDI*
- The comprehensive multilingual thesaurus, *ELSST*
- The *Nesstar* technology for making data resources available on the web

The metadata standard

Metadata, or the slightly more limited and operationalized term documentation is a mandatory part of the material that is necessary to study and describe society. Metadata serve several purposes, they constitute the instruments, describe the structural complexities of a data resource and convey the content and the meaning that is necessary to use, locate and find, to retrieve and interpret data. And of course metadata is necessary to drive the software that is needed to analyse, which is the process whereby we convert data from digits into information and knowledge. It provides information on the setting and the administrative metadata specifies the administrative rules and regulations that defines possibilities for access and use. Since there may be distance between producers and users of data, metadata make up the bridge, and this becomes particularly

true for the data archives, where the job to a large extent is to bring data from producers on one side to users on the other side.

The **D**ata **D**ocumentation **I**nitiative (DDI) is a standard for documenting data developed by an international committee of data producers and data archives from Europe, USA and Canada. The objective of this work is to build a generic standard for social science metadata expressed in a Web-friendly framework (implemented in XML) allowing and encouraging exchange, integration and interoperation across resources from a broad range of providers. In that respect there was a one to one overlap between the DDI standard and the MADIERA project. The first version of DDI (1.0) was released in the spring 2000, and gave a comprehensive standard for documentation of free-standing single survey files. The latest full version, *version 2.0* of the standard was released in 2003, now with the ability to handle aggregate data and complex tables. Currently there is work nearly completed on a version 3.0, aiming at a general ability to handle complex organized files, comparative data, geographic data and time-related data and instruments.

The DDI metadata standard, supplied with a tag-library and implemented in XML, presents the structure and the possibilities. To put it into actual use there is a need to supply it with a lot of detailed definitions, operationalisations of elements. One example: The DDI has an element 2.3.1.6. *Mode of Data Collection*. To make constructive use of this, the user have to define her set of potential modes of data collection, or make her choices from some agreed upon standard set of modes of data collections. The element is there to *allow* description of an important piece of meta-information, but in addition to that there have to be some list of possible values and some clearly defined and normative best practice guiding the user. Throughout the DDI there is a lot of elements that have to be specified this way. Any user could define his more or less personal implementation of the DDI, maybe the first and most important best practice is to agree on a recommendation of which elements to use for which types of data. Generally, there is a need and a possibility to add the semantics and the knowledge content, the DDI itself represents the structure, which may be differently applied by different scientific fields. But the vocabularies, the ontologies and the thesauri used to specify the content allow us to add machine-understandable and Web-accessible semantics to DDI-described data. The European social science data archives (through their common organization CESSDA) have over the last 30 years and over the last 10 years of the age of the Internet in particular been heavily involved in this kind of specification work, in practice working to develop a European implementation of the DDI standard. At this point standardization means that there should be common agreement on lists of optional values for every element, there should be a common "template" and some generally agreed upon best practice. A first version of this common template is now available.

A multilingual thesaurus

Language barriers are major obstacles to efficient resource location and utilization across the European Research Area. This is specially so for comparative research that normally requires data and resources from more than one language community. Apart from a handful of significant comparative data collections that are available in several languages, the majority of sources describing European societies are only documented in one language (typically the language of the country from which the data derives). Translation into one or more additional European languages has in most cases not been carried out, due to the costs involved.

However, the language challenge can be attacked by other means than large-scale translations. In the practical implementation of the DDI in a multi-language Europe, the thesaurus ELSST stands out as the single most important component of the semantic and content-carrying kind mentioned above. The thesaurus was originally based on the UKDA HASSET-thesaurus, the multi-language idea was developed within the EU-financed LIMBER project (Language Inde-

pendent Metadata Browsing of European Resources) and has been carried further within the Madiera project. A thesaurus is a hierarchically arranged controlled vocabulary, which is used for indexing and retrieval in the field of information science. If comparative data resources can be efficiently identified across language barriers, the first hurdle is already passed. This can be achieved by the use of language-independent classifications of resources as well as language-independent and thesaurus-supported application of keywords and terms to the relevant parts of the metadata records. If this is done properly a user would be able to specify his/her search criteria in any of the supported languages and get hits independent of what language they are described in. The keywords assigned to the metadata from a multi-lingual thesaurus can be instantly translated back into the supported language of the user. Initial translation of the returned resources might then be achieved by applying standard automated Web-based translation services. The quality of such translation services still do not meet scientific standards, but they might be used as a first pass in order to decide whether the use of human-powered translation might be worthwhile. And the data-location and retrieval purpose is not dependent upon the full and optimal translation service.

The ELLST thesaurus at present covers core concepts in social science research and methodology for nine European languages, English, French, Spanish, German, Greek, Norwegian, Danish, Finnish and Swedish. The thesaurus open enormous possibilities for meaningful data classification and data retrieval across the language barriers of Europe. It allows for automatic insertion of keywords and automatic classification of text components on the data input/data publishing side, as well as possibilities to browse and search more meaningfully on the data location and application side.

The technological platform

The technological platform NESSTAR has been developed through the EU-financed NESSTAR (Networked Social Science Tools And Resources) and FASTER (Flexible Access to Statistical Tables for European Research) projects. It is a state-of-the-art suit of software tools developed to run real-life data services at data archives and other large organizations.

The functionality of NESSTAR at project initiation covered four basic facets of the research process: resource location, metadata browsing, on-line analysis and data download. However, even if NESSTAR have been developed for several different types of users, both data providers and data users, the focus of the functionality have been tilted somewhat towards the supplier side. Within the MADIERA project the intention is to further refine the available technologies but to be explicitly aware of the data user side, to make the software better suited as a tool for European comparative research. NESSTAR is developed to exploit the full potential of the DDI standard, it handles the data structures described, make use of the content-carrying parts to catalogue or locate data, contextual and managerial metadata are important as basis for controlled access to data and data is in the end delivered to internal or external analysis technology.

The implementation of a vision

Since the middle of the 1990s, the social science data archives of Europe have run a long-term focused project on developing a modern user friendly integrated Web interface to their collective holdings. It started off as an idea about an integrated common catalogue for all the major social science data archives in Europe, an idea of the early web period. Since the timid start, the ambitions have grown considerably and grown in parallel with the potential and the development of the web itself. The work has both capitalized on and sparked off other related projects.
In particular several EU-sponsored projects have been important building blocks in this long-term strategic plan. The MADIERA project is just the present last step in this concerted effort but before being integrated in this particular project, the development of the three main compo-

nents, the metadata standard, the thesaurus and the software technology were all initiated as separate initiatives or projects.

If we return to the 6-point list of more specific objectives, what is now the state of the development?

The development of an integrated and effective distributed social science portal to facilitate access to a range of data archives and disparate resources

This social science portal has during the course of the project undergone some development. Initially the project intended to link together a set of separate Nesstar servers, where the data archives that take part in the Madiera project published and maintained their data holdings. The documentation standard should be common, while language normally would be national. Nesstar servers operate on the base of the same principle as traditional Web servers: published resources are assigned an URL and are directly accessible through the HTTP protocol. When the resource URL is accessed using HTTP the server returns the RDF description of the corresponding statistical object (a dataset, a group of variable, a variable, etc.) In addition to direct URL access, Nesstar servers also provide a more sophisticated query language that allows the retrieval of objects that satisfy specific conditions. A common but distributed index would allow one server to access data from another server

The major problem with this initial architecture would be that servers for some of the intended search functionality become dependent upon each other. A better solution would be to think in terms of a portal where we lift the common index out: The Madiera Portal developed operates as a Web search engine by browsing and querying the Nesstar Data Servers to harvest the RDF descriptions of the available statistical objects. The Portal accesses the data servers using the Nesstar API, a Java library that automatically converts the RDF descriptions returned by the servers to corresponding Java objects and stores them in an in-memory object database. The objects so collected are then indexed on the basis of the contents of their title, keywords and abstract properties using the Lucene text indexer and search engine. Using the indexing terms extracted by Lucene the statistical objects could then be matched with a set of multi and monolingual thesauri and classifications. The portal automatically matches resources with the terms in all the languages supported by the thesauri so that, for example, a dataset with the keyword "GLOBAL WARMING" is also associated to the corresponding French term "RECHAUFFEMENT PLANETAIRE".

Using the Madiera Web client, European social researchers can easily locate data resources published by any of the participating data archives by either *browsing* one of the available thesauri in their preferred language or by performing an explicit *search*. Once a researcher has found a useful resource (e.g. a study or a statistical variable) s/he can then use the standard Nesstar Web client to examine its complete metadata, apply statistical operations and download data.

The employment of a multi-lingual thesaurus to break the language barriers to the discovery of key resources

The MADIERA portal employs a multilingual thesaurus to solve the problems that different languages create for the discovery of key resources. A thesaurus is a hierarchically arranged controlled vocabulary, which is used for indexing and retrieval purposes, activities related to the user side. However, one of the conclusions reached very early was that such an ambitious infrastructure will have difficulties unless the development of standards is coupled with supporting software. Without easy-to-use and efficient software to support the input side, the documentation and publication process, it will be difficult to reach the content provision goals of the project. The multilingual thesaurus is therefore not only a key resource for the data location process, but

also for the documentation and publication process. An important supporting piece of software is then the Nesstar Publisher. The Publisher can import data from the common statistical packages and builds up a DDI-structured view of the data through the common template. Then the Publisher have the thesaurus built in to facilitate automatic insertion of keywords at study or variable level and the Publisher will automatically classify and publish data according to the CESSDA topical classification. Another "semi-Madiera" addition to the Publisher is the possibility to publish texts, or non-data resources on a server, i.e. documents that may be invoked by internal references from ordinary data resources. Such documents typically will be documented according to the Dublin Core standard.

The development of specific add-ons to existing virtual data library technologies, in particular data location technologies and a metadata standard for scientific empirical material

We know that finding and getting access to appropriate resources are time-consuming and diverts focus away from the creative part of the research process. Google and other success stories tell us that efficient resource location is of enormous importance. It is especially hard to stay continuously updated on the availability of new resources. Making use of DDI and ELSST, MADIERA is addressing this problem from four different angles:

Standard keyword and free-text searching: Given the detailed structure of the DDI standard the user is allowed to search for data with greater precision, hopefully avoiding the overflow of hits that usually is the case when searching the Web. Researchers interested in particular subjects are now able to move beyond the keywords and abstracts that normally are included in on-line catalogues and search directly on variable descriptions, question texts etc. Likewise, using the DDI structure as a menu, searches can be conducted on concepts of the DDI structure, such as method of data collection (e.g. self-completion questionnaires) or sampling strategy (e.g. random, stratified, etc) The resulting hit list is either presented as a list of studies or a list of variables in context.

Browsing of structured subject-oriented catalogues: Apart from standard text searching (the Google style), browsing of subject hierarchies (the Yahoo style) might be seen as the other major method of navigating a complex information space. MADIERA offer several ways of browsing by subject categories. First, the CESSDA topical classification mentioned above is a hierarchical two-level browse-list standardised across the European data archives and it is defined as a component in the common documentation template used to input the documentation of a study. Next, the thesaurus functions the same way as a topical classification under this browsing perspective. In theory the procedure used to classify studies could work all the way down to individual variables, where subject hierarchies are created automatically from relevant descriptors in the metadata. Often one study contains content that spans several categories, for that reason variable groups and single variables are assigned keywords that may be used for automatic classification as well as text searching. We could imagine the user being allowed to create different or customised private views to the information space (sorted by data provider, time, geography subject, data type etc.) through selecting and organizing categories, branches and nodes that are of special interest. Customised views can even be shared among researchers working in related fields.
Development of this functionality will require some more work on the harvesting capabilities of the Madiera portal.

Geographical/map-based resource location: Location in space is an important social science variable and spatial location may be of interest both to search for and to judge the relevance of data. However, geographical information for social science data is also a complex matter and it takes many elements and attributes of the DDI to accommodate everything. The Madiera project intends to offer a geographical interface to the information space, exactly how to do it is still being tested. The user will be allowed to circle in or mark one or more area on the map and retrieve lists of

resources covering or deriving from the selected areas. Further filters narrowing the search may be applied when using this method. Since the system will be metadata driven it is important that data are well described and it will probably be just as important to develop the documentation input side (the mirror image side of the search procedure). This means that the documentation instrument also needs a cartographic visualisation where selection of geographic position automatically inserts coordinates and identifications in the metadata of a study.

Specialised search for comparative data: Identifying comparable or potentially comparable data is always a challenge, especially as we move beyond the group of major studies that are explicitly designed to be *comparative* across time or across nations. What this feature will offer is a way of identifying potentially comparable sources of data (datasets, groups of variables or single variables) across the entire distributed information space. The feature will establish "comparability" by analysing a range of metadata descriptors. The end-user will be given a high degree of flexibility in defining the relevant set of comparability criteria and the system will then identify potentially comparable data sources, i.e. studies that addresses the same problem areas, studies with the same questions, variables with the same keywords, for the same geographic areas, same sampling procedure, etc. Any decisions or harmonization will be left to the user.

An extensive programme to add content, both at the data/information and knowledge levels

We have stressed the importance of the connection and interdependence between grid computing, standardisation/standards development and accompanying software. To develop a programme to promote the publication of content, several inter-related activities have been initiated

- Work on the uptake and implementation of the DDI, the data archives of Europe have initiated the CESSDA DDI Group with a mandate of developing a common template for data documentation work
- Work on the Nesstar Publisher, the need to develop mirror image input functionality for the data location functionality the Madiera project is intended to develop
- Workshops to train personnel from relevant data archiving institutions on procedures for data documentation and data publishing

Extensive training of data providers and users to encourage the continuous growth of the infrastructure

This point implies extension in two ways. More/other data producers should contribute data, and in addition the output side user community, the data users and research society should contribute to the growth of the infrastructure. It is a complicating factor that the data liberation ideology runs counter to commercial interests. But this point also implies more than pure data publishing. The infrastructure is based of an extended meta-data concept where metadata is regarded as a dynamic communication node or a facilitator for the interchange of information and insight that is the driving force of a process. Metadata makes it possible to extract knowledge from numbers and to share this knowledge with others. Then the conversation around the data and the various layers of knowledge products that derive from this conversation should be fed back into the metadata. For a secondary user of a data source it is important to have access to relevant information about the data production, but it is also of immense value to get access to the knowledge of previous users, not only to avoid walking down analytical paths that are already fully explored, but also to learn from past experiences and to make it possible to add new approaches and new insight to the layers of already accumulated knowledge. This perspective leads to an *extended* metadata concept where not only descriptions of the data are relevant information, but also various types of knowledge products (formal as well as informal) deriving from the use of the data. It implies a *dynamic* concept where metadata is seen as a collection of information that is developed and enriched through the life cycle of the dataset and not something that can be

created and published once and for all. The perspective lead to a concept where a broad spectre of actors is seen as legitimate contributors to the metadata holdings. Whereas the core metadata are still developed by the data producers as part of the data production and publishing process, further layers of metadata will be provided by others as an ongoing activity lasting for many years after the data themselves have left the production line.

So far, the Nesstar Publisher is extended with a possibility to publish documents, articles, texts on a server in the same way as data resources are published, and these components can then be linked together both ways, each referencing the other. This make the dataset or data-collection a repository that not only bridges the gap between data producers and –users, but also functions as a bridge between different users, or between the basic data and the various derived knowledge products.

If we look at it from the knowledge repository viewpoint, the data use process can be chopped up in several discrete functions. The table below illustrates how the three main ingredients are thought to support each other in the Madiera infrastructure.

	Functions						
	Describe	Discover	Evaluate	Access	Analyze	Manage	Download
DDI	X	X	X	X	X	X	
ELSST	X	X	X				
NESSTAR		X		X	X	X	X