EUROPEAN
COMMISSION

Community research

EU RESEARCH ON SOCIAL SCIENCES AND HUMANITIES

Multilingual Access to Data Infrastructures of the European Research Area

MADIERA

FINAL REPORT

EUR 23148

Interested in European research?

Research*eu is our monthly magazine keeping you in touch with main developments (results, programmes, events, etc.). It is available in English, French, German and Spanish. A free sample copy or free subscription can be obtained from:

European Commission
Directorate-General for Research
Communication Unit
B-1049 Brussels
Fax (32-2) 29-58220
E-mail: research-eu@ec.europa.eu
Internet: http://ec.europa.eu/research/research-eu

EUROPEAN COMMISSION

Directorate-General for Research
Directorate L — Science, economy and society
B-1049 Brussels
Fax (32-2) 2994462

http://ec.europa.eu/research/social-sciences
http://cordis.europa.eu/fp7/cooperation/socio-economic_en.html

EUROPEAN COMMISSION

EU RESEARCH ON SOCIAL SCIENCES AND HUMANITIES

Multilingual Access to Data Infrastructures of the European Research Area

MADIERA

Final report

HPSE-CT-2002-00139

Funded under the Key Action
'Improving the Socio-economic Knowledge Base' of FP5

DG Research
European Commission

Issued in
April 2006

Coordinator of project:
Norwegian Social Science Data Services (NSD)
Bergen, Norway
Atle Alvheim
www.madiera.net

Partners:
UK Data Archive (UKDA), Colchester, UK, Kevin Scürer
Danish Data Archive (DDA), Odense, DK, Anne Sofie Fink
Finnish Social Science Data Archive (FSD), Tampere, FI, Sami Borg
Nesstar Limited (NESSTAR), Bergen, NO, Jostein Ryssevik
Swiss Information and Data Archive Service for the Social Science (SIDOS), Neuchâtel, CH, Reto Hadorn
National Centre for Social Research (EKKE), EL, John Kallas
Zentralarchiv fűr Empirische Sozialforschung (ZA), Köln, DE, Rolf Uher

2007 Directorate-General for Research
Citizen and Governance in a knowledge-based society EUR 23148 EN

EUROPE DIRECT is a service to help you find answers to your questions about the European Union

Freephone number(*):
00 800 6 7 8 9 10 11

(*) Certain mobile telephone operators do not allow access to 00 800 numbers or these calls may be billed

A great deal of additional information on the European Union is available on the Internet.
It can be accessed through the Europa server (http://europa.eu).

Cataloguing data can be found at the end of this publication.

Luxembourg: Office for Official Publications of the European Communities, 2007

ISBN 978-92-79-07754-8

Printed in Belgium

PRINTED ON WHITE CHLORINE-FREE PAPER

Preface

Within the Fifth Community RTD Framework Programme of the European Union (1998–2002), the Key Action 'Improving the Socio-economic Knowledge Base' had broad and ambitious objectives, namely: to improve our understanding of the structural changes taking place in European society, to identify ways of managing these changes and to promote the active involvement of European citizens in shaping their own futures. A further important aim was to mobilise the research communities in the social sciences and humanities at the European level and to provide scientific support to policies at various levels, with particular attention to EU policy fields.

This Key Action had a total budget of EUR 155 million and was implemented through three Calls for proposals. As a result, 185 projects involving more than 1 600 research teams from 38 countries have been selected for funding and have started their research between 1999 and 2002.

Most of these projects are now finalised and results are systematically published in the form of a Final Report.

The calls have addressed different but interrelated research themes which have contributed to the objectives outlined above. These themes can be grouped under a certain number of areas of policy relevance, each of which are addressed by a significant number of projects from a variety of perspectives.

These areas are the following:

- ***Societal trends and structural change***

16 projects, total investment of EUR 14.6 million, 164 teams

- ***Quality of life of European citizens***

5 projects, total investment of EUR 6.4 million, 36 teams

- ***European socio-economic models and challenges***

9 projects, total investment of EUR 9.3 million, 91 teams

- ***Social cohesion, migration and welfare***

30 projects, total investment of EUR 28 million, 249 teams

- ***Employment and changes in work***

18 projects, total investment of EUR 17.5 million, 149 teams

- ***Gender, participation and quality of life***

13 projects, total investment of EUR 12.3 million, 97 teams

- ***Dynamics of knowledge, generation and use***

8 projects, total investment of EUR 6.1 million, 77 teams

- ***Education, training and new forms of learning***

14 projects, total investment of EUR 12.9 million, 105 teams

- ***Economic development and dynamics***

22 projects, total investment of EUR 15.3 million, 134 teams

- ***Governance, democracy and citizenship***

28 projects; total investment of EUR 25.5 million, 233 teams

- ***Challenges from European enlargement***

13 projects, total investment of EUR 12.8 million, 116 teams

- ***Infrastructures to build the European research area***

9 projects, total investment of EUR 15.4 million, 74 teams

This publication contains the final report of the project Multilingual Access to Data Infrastructures of the European Research Area, whose work has primarily contributed to the area 'The development of European infrastructures for comparative research in the social sciences and humanities'.

The report contains information about the main scientific findings of MADIERA and their policy implications. The research was carried out by eight teams over a period of 39 months, starting in December 2002.

The abstract and executive summary presented in this edition offer the reader an overview of the main scientific and policy conclusions, before the main body of the research provided in the other chapters of this report.

As the results of the projects financed under the Key Action become available to the scientific and policy communities, Priority 7 'Citizens and Governance in a knowledge based society' of the Sixth Framework Programme is building on the progress already made and aims at making a further contribution to the development of a European Research Area in the social sciences and the humanities.

I hope readers find the information in this publication both interesting and useful as well as clear evidence of the importance attached by the European Union to fostering research in the field of social sciences and the humanities.

J.-M. BAER,

Director

Table of contents

Abstract

Empirical comparative social research in Europe is hampered by a fragmentation of the scientific information space. Data and its derivatives, information and knowledge, are often scattered in space and divided by language and institutional barriers. The MADIERA data infrastructure is an answer to this. The MADIERA portal enables and promotes the development of a thoroughly comparative and cumulative research process that will be integrating and nurturing the entire European Research Area.

By using Nesstar and Semantic Web technology, the MADIERA project has developed a user friendly, highly functional and fully integrated web interface to data and resources distributed at various sites throughout the European Research Area.

The MADIERA portal provides access to an unprecedented quantity of social sciences quantitative datasets using an easy to use web interface. It harvests metadata from statistical datasets and variables published on the Semantic Web from all the largest European social sciences data archives, organizes and makes them available using a set of multilingual thesauri and taxonomies.

The expanded use of data will benefit and enhance comparative research; and the ability to harmonize datasets over time and geography will lead to significant improvement in our understanding of societies. Promoting comparative research is the core to developing a European Research Area. Increasing the availability of high-quality data is also a way of increasing the importance of secondary analysis in the social sciences. For that to become a reality the high-quality data needs high-quality documentation to accompany it and high-quality resource discovery tools to locate it.

Traditional national borders have to a large extent been done away with in the age of Internet. However, other borders like cultural and judicial borders still remain. To break language barriers MADIERA employs a multilingual thesaurus for automatic translation.

The key to realizing the benefits of modern computing technology and the Semantic Web is standardization. By adhering to standards we may let technology substitute for institution building. In a world of standards we also have the possibility to build an open but sustainable system, nurtured by the collective energy of the data and knowledge producing communities of the European Research Area.

The MADIERA portal applies and demonstrates the value of standards and the possibility to build decentralised infrastructures to make data available. The standardization of metadata, at the semantic, structural and syntactic level, facilitates interoperability

between systems, but also allows easier interpretation and better understanding of the substantive content.

The ability to see oneself compared to others is the key to the development of a European Research Area. MADIERA promotes a comparative perspective, secondary analysis of data and a continuous quality control of data through availability, open systems and intensive use. The portal promotes European integration by demonstrating practical solutions to how to build down borders without removing national distinctiveness.

I. EXECUTIVE SUMMARY

By using Nesstar and Semantic Web technology, the MADIERA project aims to provide a user friendly, highly functional and fully integrated web interface to data and resources distributed at various sites throughout the European Research Area.

The MADIERA portal now provides access to an unprecedented quantity of social sciences quantitative datasets using an easy to use web interface. It harvests metadata from statistical datasets and variables published on the Semantic Web from all the largest European social sciences data archives, organizes them using a set of multilingual thesauri and taxonomies and makes them available through a simple, responsive and highly customisable web interface.

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, she can then use the standard Nesstar web client to examine its complete metadata, apply statistical operations and download data.

The vision of the MADIERA project has been 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 final product, the MADIERA portal, is 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 collective resources of a selection of the existing 20+ social science data archives in Europe, with the potential for rapid expansion with the inclusion of new data-supplying points. The MADIERA infrastructure will, as the web itself, have the capacity to grow and diversify after the initial construction period. The main objective of the project has been to create an open but sustainable system, nurtured by the collective energy of the data and knowledge producing communities of the European Research Area.

Breaking these ideas down into more specific objectives, the MADIERA project has focused on the following specific goals:

- The development of an integrated and effective distributed social science portal to facilitate access to a range of data archives and their disparate resources.
- The employment of a multilingual thesaurus to break the language barriers to the discovery of key resources.

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

- Run an extensive program to add content, both at the data/information and knowledge levels.

- Carried out extensive training of data providers and users to inspire and encourage the continuous growth of the infrastructure developed tools and guides for the practical side of such work.

- Opened for the gradual integration of the emerging national infrastructures of the candidate countries into the European Research Area, by making available technical solutions and guiding material at low cost.

The MADIERA project addresses the need to facilitate cross-national social science and humanities research throughout the European Research Area. Since the middle of the 1990s, the social science data archives of Europe have, as a specific response to clearly formulated needs by the scientific user community, run a long-term focused project on developing a modern user friendly integrated web interface to their collective holdings, an interface that should be rich in functionality and content. 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 World Wide web period. Since the timid start, the ambitions have grown considerably and grown in parallel with the potential and the development of the Internet and 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 may be regarded as another important step in this concerted effort but before being integrated in this particular project, the development of the three main components, the metadata standard, the thesaurus and the software technology were all initiated as separate initiatives or projects.

The MADIERA portal is based on three main components:

- a common standard for data documentation, the DDI (Data Documentation Initiative);

- the comprehensive multilingual thesaurus, ELSST (European Language Social Science Thesaurus);

- the Nesstar technology for making data resources available on the web.

1. The metadata standard

Metadata, or simply 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, retrieve and interpret data. And of course metadata is necessary to drive the software that is needed to analyse the data, whereby data are converted from digits into information and knowledge.

Developed by a an international committee of data producers and data archives from Europe, USA and Canada, the **D**ata **D**ocumentation **I**nitiative (DDI) is a standard for documenting data. The objective was to build a generic standard for social science metadata expressed in a web-friendly framework allowing and encouraging exchange, integration and interoperation across resources from a broad range of providers.

The DDI metadata standard, implemented in XML, presents the structure and the possibilities. To make constructive use of this, some of the elements within that structure have to be defined by a set of potential values, an adding of semantics and knowledge content to the DDI structure. 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. Through CESSDA (Council of European Social Science Data Archives), the European social science data archives have 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.

2. Multilingual thesaurus

In terms of the creation of a valuable European-wide resource, the production of a 9-language version of the multilingual thesaurus ELSST (European Language Social Science Thesaurus) has been a major success of the MADIERA project. The way that that resource has then been applied in the MADIERA portal has surpassed all our expectations.

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 metadata standard 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. This thesaurus was originally based on the UKDA HASSET-thesaurus, the multi-language idea was developed within the EU-financed LIMBER project (Language Independent Metadata Browsing of European Resources) and has now been carried significantly forward within the MADIERA project. Such a thesaurus is a hierarchically arranged controlled vocabulary, which is used for indexing and retrieval purposes 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 were done properly a user would be able to specify his/her search criteria in any of the supported languages and get a list of hits independent of what language they are described in. The keywords assigned to the metadata from a multilingual thesaurus can be instantly translated back into the supported language of the user. Initial full translation of the returned resources might then be achieved by applying standard automated web-based translation services. We know that the quality of these 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 ELSST 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 opens 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.

At the beginning of the project ELSST was a four language multilingual thesaurus of approximately 1,355 concepts expressed in English, French, German and Spanish.

This was the final deliverable of another EU project, LIMBER (Language Independent Metadata Browsing of European Resources). This first version of ELSST had been developed from the UKDA monolingual thesaurus HASSET (Humanities And Social Science Electronic Thesaurus). The major 38 hierarchies of concepts had been reduced by excluding any country specific or organization specific terms to produce a more European-centered thesaurus. Although successful, the terms are still in the present version of ELSST, the true power of the thesaurus could not be demonstrated since the data archives of CESSDA that held data in the languages other than English had, at that time, not published their data using the DDI standard and Nesstar technology.

The ambitious infrastructure of the MADIERA portal with its underlying standards also requires the development of supporting software. Without easy-to-use and efficient software to support the creation of metadata and the publication process, it is difficult for archives to provide quality data content. The multilingual thesaurus is not only a key resource for the data location process, but also for the documentation and publication process. As part of the MADIERA project, an important supporting piece of software has been added to the Nesstar Publisher. The Publisher can import data from the common statistical packages and builds up a DDI-structured view of the data through a common CESSDA template. The Publisher now has the thesaurus built in to facilitate automatic insertion of keywords at study or variable level and will automatically classify and publish data according to the CESSDA topic classification.

Increasing the availability of high-quality data is a way of increasing the importance of secondary analysis in the social sciences. For that to become a reality the high-quality data needs high-quality documentation to accompany it and high-quality resource discovery tools to locate it; and that are what the ELSST thesaurus and the MADIERA portal delivers.

3. The technical platform

The MADIERA portal is built on Nesstar technology. Nesstar is a fully web-enabled technology providing powerful and advanced data analysis and visualization capabilities through a standard web-browser. It is a fully distributed technology providing integrated access to data stored on separate remote servers.

The Nesstar suite comprises the following components:

- Nesstar Server – providing the repository of data and metadata and services for their searching, access and analysis.

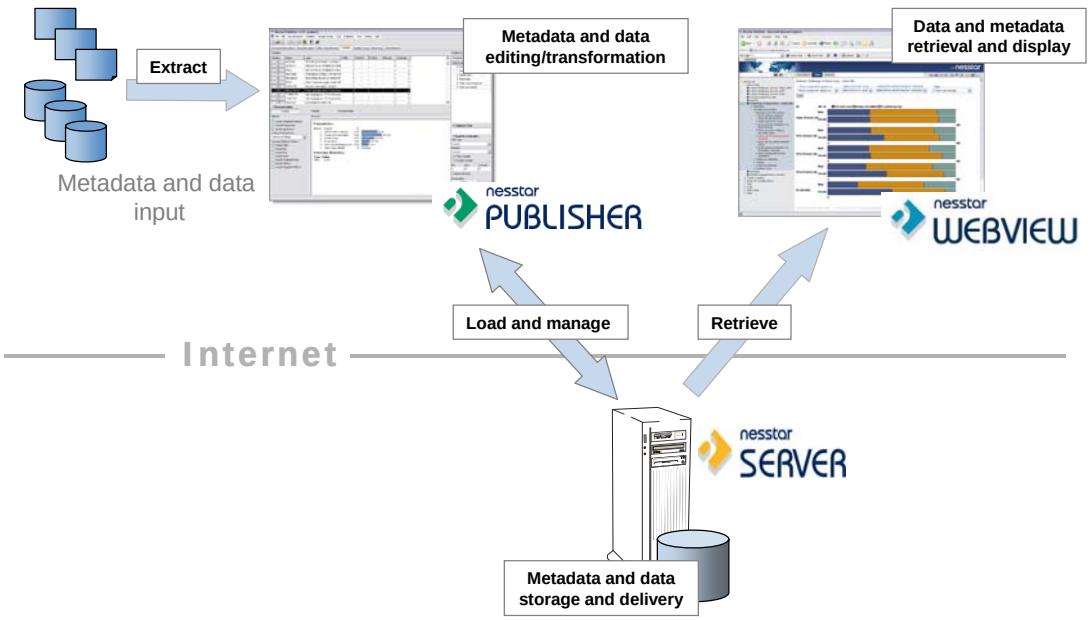
- Nesstar WebView – providing the web-based user interface to the system, with a variety of powerful visualisations on the data, including mapping and graphing.

- Nesstar Publisher – the desktop based extraction, transformation and loading tool for Nesstar, able to accept input from a wide range of formats, and provide facilities to add value to the data and metadata for loading on to Nesstar Server via Internet connectivity.

A completely new generation of the underlying Nesstar technology has been implemented and deployed in the portal. Nesstar 3.0 forms the basis of the MADIERA platform. This version includes extensions and improvements to all parts of the technology: server (Nesstar Server), end-user client (Nesstar WebView) as well as the data and metadata management tool (Nesstar Publisher).

The functionality of Nesstar covers four basic facets of the research process: resource location, metadata browsing, on-line analysis and data download. Within the MADIERA project these technologies have been refined to make the software even better suited as a tool for European comparative research. On top of Nesstar, the MADIERA portal platform has been designed and implemented and serves as a central resource to the MADIERA network.

Figure 1. The technical platform



4. User requirements and usability testing

The MADIERA project has been a user driven software development project. Users have been closely involved in the design from the beginning to the end of the project. The project has two main target groups: researcher with in the European social science community and data providers that wish to connect to the portal. During the project period there has been continuous contact with representatives for both groups.

User specification of needs, user testing of implemented solutions and analysis of user reactions and evaluations has been fundamental to the MADIERA project. The software tools developed under the portal have to offer functionality and solutions that are in accordance with expressed user needs. Since the MADIERA infrastructure is intended to help solve what often are non-routine problems over a broad range of types of users, it is important to ensure that the software tools developed is user friendly and appropriate for the target user groups.

A first step was to map out user requirements and analyse these in order to feed this information into the functional specification for the system architecture. The challenge was to collect information that would provide the project with a detailed understanding of work practices among potential users of the product. The methods applied had to accommodate the needs of independent users and those of users who were also publishers of data.What was needed was a sensitive method that was able to make the underlying logic in users' working-process explicit. The Contextual Design method was chosen.

The mapping was partly based on re-analysis of data from earlier projects, partly on a new user analysis for the MADIERA project. The material from the projects NESSTAR, LIMBER and FASTER were revisited and reanalyzed and the findings relevant to the MADIERA project were summed up and an interview guide was outlined. The review of existing data gave a valuable introduction to early discussions for development of the prototype

The more elaborate user analysis based on new data pointed at three different user profiles named 'the young researcher', 'the IT-curious researcher' and 'the traditional researcher'. Interviews with 14 potential users were carried out at all sites. The respondents were recruited from universities being researchers covering the field of empirical quantitative research. A user expert team which was set up in the beginning of the project with members from all project partners, carried out the interviews. The user requirements reported were made the basis for further development of the design and functionality of the portal.

The second step was to plan and carry out the usability testing. Usability is defined as:

> The extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use (ISO 9241-11:1998).

Usability testing has been carried out several times during the span of the project and at several stages. The testing proved valuable for the development of the design and functionality of the portal.

5. Specification and development of new functionality

Input for the functional specification came from the user analysis and the usability testing and also the project team's extensive experience with the development of earlier Nesstar products. The functional specification was developed through different stages. One of the original ideas of the MADIERA project was to develop some specific new technology to search for and locate data. Since the overarching aim is to develop a system that makes it easier to carry out comparative research, two specific problems were singled out to be added to standard search and browse technology:

- Specification and implementation of a geo-referencing system for social science data, to allow geographically based search for and location of resources.
- Development of a methodology for identification of comparable data.

In addition to that it was set as a goal to investigate and try to develop a standardized naming and identification system for social science data published on the web.

5.1. Implementation of geo-referencing functionality

As a part of the geo-referencing system a coordinate-based spatial search has been implemented as a feature of the MADIERA portal. The available first version of the demonstrator gives a user access to a map version of the NUTS system of statistical units, as a 3-level hierarchy (down to level 3 of the NUTS system). The user may mark an area of interest on the image of the map. The coordinates of the chosen area of interest will then be shipped to the search-module of the portal as search-parameters. These parameters can be used in isolation, or in combination with regular text-search parameters. Because the visual display is based on coordinates, the map-interface does have zoom and pan-capabilities, in addition to support of selection among available maps. The interface produces the list of available maps by connection to a OGC-

compliant web map server (WMS). The WMS also performs the actual rendering of the map coordinates onto a relevant image format (png. and jpg. are supported).

Additionally the portal's search-module treats searches for coordinates in a parallel way as traditional text-searches. This extension is vital for the system to work flexibly on different types of searches, because this is the component that retrieves the matching documents from the underlying servers and presents the portal with them. This open the possibility that location of data resources by strict spatial positioning, by picking codes or names from a menu/gazetteer or by searching the substantive content can be mixed freely by a user.

To establish a better map-basis for the spatial positioning and a realistic system of spatial units for a menu/gazetteer procedure, we have negotiated a contract with Eurographics, which allows us to use the material they have developed for the updated NUTS overview of 35 European countries. This means that we get the names, codes and coordinates down to level 3 or comparable for every country. The NUTS system is regarded as the most relevant nomenclature and frame of reference for such a practical procedure.

In the social sciences there is little tradition to document geographic location of data resources. Because of that it has become obvious that the project also have to suggest realistic publishing procedures for this type of content. This may be done in various ways, to some degree linked with the best-practices work on content provision outlined by WP6 and by expanding the dedicated Nesstar publishing tool. If we start from the NUTS nomenclature, we have a system of identifying codes and names at different levels of aggregation, linked to coordinates that outline the unit borders. This may be expanded with a standardized bounding box (i.e. the upper left hand corner and the lower right hand corner) and a (set of) standardized bounding polygons of every unit. In the data publishing process specification of geographic coverage could then be accomplished by selecting the unit code or name from a controlled vocabulary/menu or through clicking in a visualized map. Then several elements of the data documentation can be filled with the relevant information automatically, if we decide on geographic coverage, we get bounding box and bounding polygon for free.

The actual development of this publishing possibility is not regarded as an integrated part of the MADIERA project.

5.2. Identification of comparable data

Comparison is a relational and relative concept, it requires that there is a baseline defined and we are looking for other items that may be compared to that baseline. Usually that means an extension of an analytical dimension. We are either working from one data resource or some specific analytic task, and we are looking for other data resources or other pieces of information of a similar kind or similar content, to extend a dimension or expand on the content of our analyses.

The MADIERA intention was to investigate possibilities to supplement a data analytic process with a practical but useful and realistic procedure to look up relevant data while in the middle of the process itself. It is not to go out and search for comparative data in the traditional meaning of the term. This focus on the data-analytic process will have consequences for the order of priority between types of elements. The explicit prerequisite is data resources described according to the DDI metadata standard.

In MADIERA we initiate a search for comparable data from very specific information and ask for:

- additional examples of the same specific information, or

- additional examples of the same specific information where we are invoking more general information as a criterion, similar table for the same topic, the same topical group, the same or different universes, timepoints, etc

To develop a useful procedure, it is necessary to concentrate on the *concept* or *keyword* elements in the documentation of the datasets. The data archives have tested the publishing procedure implemented in Nesstar Publisher. The archives tagged up concepts at variable and variable group level for sets of studies and used the Nesstar Publisher version with the thesaurus ELSST included. The publishing software makes an analysis of the textual documentation linked to a variable, usually the question text, and invoking the vocabularies of the thesaurus, suggests potential keywords to concentrate and standardize the content-carrying part of the documentation. The conclusion is that this is an efficient, rational and more than anything, it is a standardized way of structuring the substantive content of data resources.

Whenever we are exploring/analyzing data resources, it will be possible to check or look up keywords connected with specific variables. By clicking directly on the keyword, a second search will be performed, with the keyword as search term.

5.3. The idea of a hyperlinked information-space

The MADIERA portal promotes an *extended* metadata concept where not only descriptions of the data are relevant information, but also various types of knowledge products deriving from their use. It is implying a *dynamic* concept where metadata is seen as a collection of information that is developed and enriched all the way through the life cycle of the dataset and not something that can be created and published once and for all. The perspective is leading 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 could be provided by others as an ongoing activity lasting for many years after the data themselves have left the production line, and the use of the data becomes an element that stimulates further use and creates wider relevance.

This allows for interactive knowledge products through inclusions of live tables and graphs into publications so that the reader is able to interact with the tables and use them as an entry point to the underlying data. Readers should be able to re-run an original analysis or add comments or results from alternative analyses to the metadata of a study.

The problem is to record and accumulate the dynamic information generated in the data use process, link it and make it relative to an identifiable starting point. Users should be allowed to feed their experiences back into the repository made up by the study. This is a technical, identification and an authorization problem.

Standards are essential for the functionality of MADIERA. To develop the dynamic (meta)-data concept, first and foremost a naming and identification recommendation for social science data resources is necessary. Without a consistent naming system that gives us better possibilities to identify resources a dynamic data resource concept may be difficult to develop.

The technological platform of MADIERA is well suited to support this dynamic growth of such hyperlinked information spaces. Our scenario implies that we are going backwards from a knowledge product to the underlying data resources and establish a connection. We may store a bookmark to the data with the knowledge product or store a link as a reference with the data resource, as part of the metadata. This requires an exact identification to get back to the originally used version of the data resource.

In accordance with the recommendations for DDI 3.0, MADIERA have recommended an identification system consisting of three logical components:

- original publisher, the metadata producer, the "owner", the authority;

- the instance, the actual data resource we access on the web;

- versions, the recorded history of the instance, as a configuration of module statuses.

When using data from a Nesstar server, it is at present possible to create both client-side and server-side bookmarks. The latter could be regarded as private workspaces on the server for every user and every data instance where it is possible for a user to store comments or identifiable bookmarks to actions on a specific data resource. Such a workspace is private in the meaning that it is password protected.

To allow users to add comments and hyperlinks to datasets on a server it is possible to set up an open version of this technology, along with every dataset there could be a "notepad", where users leave comments and hyperlinks.

Use of this functionality will in the first version be controlled by the following practical arrangements: The data publisher has an editorial right to edit whatever is put as comments with a dataset and access to datasets are controlled via the specific access rules and user registration defined by a data publisher. This means that a publisher usually has control over users possibilities to access data resources and can deny such access if systems are misused.

6. The MADIERA portal

The MADIERA portal provides access to almost 3000 studies from several European countries at three different levels: study, sections and variables. The studies cover most of the areas within the social sciences, politics, employment, culture, economics, social stratification, health etc. Using the portal, 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) she can then use the standard Nesstar web client to examine its complete metadata, apply statistical operations and download data.

The portal has several functions that are crucial to linking European data resources and providing unified access to social science data archives:

- provide a Yahoo-style overview of the data resources of the entire network using the MADIERA classification system to organize the resources;

- provide a home for the MADIERA multilingual thesaurus;

- provide a central metadata index that will support more efficient Google-style searching across servers/archives;

- provide a MADIERA registry service whereby new servers/archives can be dynamically added to the network.

The portal is implemented as a metadata harvester that automatically will upload metadata from the individual servers of the MADIERA network through the standard Nesstar API. The metadata are added to a central index that will provide lightening fast searching across a high number of servers. The search facility is powered by the MADIERA thesaurus to increase precision.

The portal's main technical features are that it makes available its full functionality through a HTTP/REST interface that can be easily accessed from any computer language. In addition it has a streamlined and efficient internal architecture. It provides a highly customisable user interface.

The main use cases supported by the MADIERA portal are finding studies published by any of the participating data archives by:

- browsing the multilingual CESSDA classification or the multilingual ELSST thesaurus;

- searching using a flexible search language that supports free text or field search, logical connectives, fuzzy and stem searches;

- finding studies relative to a certain geographical area using a graphical interface (based on the European NUTS classification).

Additionally it is possible to:

- browse the multilingual thesaurus structure (synonyms, related terms, and equivalent terms in other languages);

- view keywords associated with a study automatically translated in any of the supported languages (currently nine);

- capture terms used in user searches and not included in the supported thesauri.

The portal is administered via a system with a web interface, which makes it possible to add new archives/servers, remove existing archives/servers and to refresh the harvested contents of an archive, i.e. the index.

The widespread adoption of the DDI and the publishing of marked-up datasets available via the MADIERA portal will vastly improve access to a range of varied resources. The expanded use of data will greatly benefit and enhance comparative research; and the ability to harmonize datasets over time and geography will lead to significant improvement in our understanding of societies. Promoting comparative research is the core to developing a European Research Area. Increasing the availability of high-quality data is also a way of increasing the importance of secondary analysis in the social sciences. For that to become a reality the high-quality data needs high-quality documentation to accompany it and high-quality resource discovery tools to locate it. The Data Documentation Initiative (DDI), the European Languages Social Science Thesaurus (ELSST) and the MADIERA portal delivers that.

7. Policy implications

The empirical comparative social research in Europe has been hampered by a fragmentation of the scientific information space. Data and its derivatives, information and knowledge, are often scattered in space and divided by language and institutional barriers. Consequently, too much research is based on data from single nations, carried out by single-nation teams of researchers and communicated to single-nation audiences. The MADIERA infrastructure is an answer to this state of affairs. The portal enables the development of a thoroughly comparative and cumulative research process that will be integrating and nurturing the entire European Research Area.

The MADIERA reasoning has stressed that there is a major difference between how the answers to these challenges have been formulated up until now, in terms of centralization and establishment of large-scale European-wide institutions, and the potential that modern communication technology now offers.

The portal uses the emerging information technologies to encourage communication, sharing and collaboration across spatially dispersed but scientifically related communities. The MADIERA infrastructure connects existing and well functioning providers of content and services and tries to meet the demands of their users. The portal can be seen as a 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.

The infrastructure developed in this project offers better availability through a solution based on the Nesstar technology, a multilingual thesaurus and documentation standard. The researchers in the European Research Area can search for data and metadata by several methods. The first uses a simple text box for native language, user-entered search strings along with browsing trees, of the topic classification and concept keywords from the ELSST. In addition, browsing of the complete holdings of each participating archive is also possible. The default for the simple search is to find occurrences containing all the words entered across the all the metadata records. The system also supports exact phrase matching, Boolean and wildcard searches and can also be restricted to certain elements of the metadata. Help for these more advanced types of searches is provided in the portal.

In 2004 the report of the European Strategy Forum for Research Infrastructure (ESFRI) working group in the Humanities and Social Sciences recommended the establishment of a European Research Observatory which will build upon existing resources and both actively and systematically promote synergy and coherency (EROHS report, 2004). The MADIERA project is entirely in keeping with two of the main objectives which are meant to guide this initiative:

- The facilitation of access to and sharing of existing European and national data, thereby linking more efficiently and effectively data resources already available.

- The development of improved standards and documentation relating to existing European and national data in order to enhance the scientific quality of data and their potential for interoperability.

The main political potential of the MADIERA infrastructure lies in the following points:

- data is the most important component necessary for research;

- if data are stored and documented according to well-defined and generally accepted standards, we create well-structured possibilities;

- then we also make it possible to harvest the full potential of modern computer technology;

- if we develop possibilities, it is the responsibility of the single participant to make use of these possibilities.

Science does not empty data of its content. Social science is not mainly about accurate descriptions of the present, science is about relationships and processes in general. When we add a dataset or a variable to a collection, we add further relationships and promote and demonstrate the value of secondary use of data. The same data may be used for more than one purpose and the potential value may grow through additions of data or just use of the data, other researchers conclusions may be important data. Extended secondary use and integration of data and knowledge products will greatly enhance the possibilities for empirical social enquiry.

The key to realizing the benefits of modern computing technology and the Semantic Web is standardization. Standardization means to develop structured possibilities. By adhering to standards we may let technology substitute for institution building.In a world of standards we also have the possibility to build open-ended systems with the ability to grow, structures that are nurtured by the collective energy of the participants.

The MADIERA portal applies and demonstrates the value of standards and the possibility to build decentralised infrastructures to make data available.The standardization of metadata, at the semantic, structural and syntactic level, facilitates interoperability between systems, but also allows easier interpretation and better understanding of the substantive content.

A comparative perspective, the ability to see oneself compared to others is a key to the development of a European Research Area. MADIERA promotes a comparative perspective, secondary analysis of data and a continuous quality control of data through availability, open systems and intensive use.

Traditional formal national borders have to a large extent been done away with in the age of the Internet. However, other borders like cultural and judicial borders still remain. To break language barriers MADIERA employs a multilingual thesaurus for automatic translation. Building down such barriers is also to develop the ERA.An important remaining task is likewise to build down the judicial barriers, develop general policies and implement systems for controlled access to data.

The MADIERA portal vastly improves access to a range of varied data resources. The expanded use of data will benefit and enhance comparative research; and the ability to harmonize datasets over time and geography will lead to significant improvement in our understanding of societies. Promoting comparative research and a comparative perspective is the core to developing a European Research Area. Increasing the availability of high-quality data is also a way of increasing the importance of secondary analysis in the social sciences. For that to become a reality the high-quality data needs high-quality documentation to accompany it and high-quality resource discovery tools to locate it. The MADIERA portal demonstrates how this can be done as a distributed system, where participants themselves develop and maintain their own node in the system. Participation becomes an open possibility and an individual responsibility, freeing initiative and creating dynamics.Through the linking of data and knowledge products, making data available becomes a scientifically meriting activity; data takes on a greater relevance and importance. Through the new situation where data becomes an integrated part of knowledge products published on the web, more data will be made available.

The MADIERA portal promotes European integration by demonstrating practical solutions to how we can build down borders without removing national distinctiveness.

8. The consortium

The success of the project would not have been possible without the good cooperation between the eight partners and a collaborative effort. The most important requirement for the success of the new portal is that it should hold a sufficient amount of data and documentation of the data. All the partners have played important roles in filling the new portal with well-documented and interesting data.

The project had eight participants, five principal contractors and three assistant contractors from seven European countries. The project was led by the Norwegian Social Science Data Services (NSD), which has had many years experience in the preservation and dissemination of statistics.

The UK Data Archive (UKDA) has been responsible for coordinating the content side of the portal and also for managing the work with the thesaurus. The archive has long experience in this field after having led the development of two generations of multilingual thesauri.

NESSTAR Ltd has been responsible for the technology development and for the exploitation. In both fields the company have a lot of experience of developing and selling the Nesstar Software Suite.

Finnish Social Science Data Archive (FSD) has led on the dissemination of the project results and decided on the dissemination strategy. Additionally they have played an active part in the development of the thesaurus.

The Danish Data Archive (DDA) is part of the state archive and was able to bring a professional archival view to the development of web based systems. They co-ordinated the user analysis and had the responsibility for writing the reports.

The assistant contractors, Swiss Information and Data Archive Service for the Social Science (SIDOS), Greek Social Data Bank (EKKE) and Zentralarchiv fur Empirische Sozialforschung (ZA) enabled the wider contribution of data archives to inform the user analysis and provide a user forum for the validation. Additionally SIDOS and EKKE have translated the thesaurus into their national languages.

II. BACKGROUND AND OBJECTIVES OF THE PROJECT

The MADIERA application held as a starting point that data are the single most important component necessary for a science-based understanding of society. However, empirical comparative social research in Europe has been hampered by a fragmentation of the scientific information space. Data and its derivatives, information and knowledge, often are scattered in space and divided by language and institutional barriers. Consequently, too much research is based on data from single nations, carried out by single-nation teams of researchers and communicated to single-nation audiences. This state of affairs has prevented the development of a thoroughly comparative and cumulative research process that would be integrating and nurturing the entire European Research Area.

The MADIERA reasoning has stressed that there is a major difference between how the answers to these challenges have been formulated up until now, in terms of centralisation and establishment of large-scale European-wide institutions, and the potential that modern communication technology offers.

The answer is to focus on the power of emerging information technologies to encourage communication, sharing and collaboration across spatially dispersed but scientifically related communities. However, 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. What is required is a 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. Still we find that availability is 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.

If "sharing" is the most important single keyword characterizing a true grid, the key to realizing the benefits of both grid computing and the Semantic Web is

standardization.Standardization facilitates development or integration 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 requirement is standardization of metadata, at the semantic, structural and syntactic level, to facilitate interoperability.

Consequently the vision of the MADIERA project has been 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 final product, the MADIERA portal, is 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 collective resources of a selection of the existing 20+ social science data archives in Europe, with the potential for rapid expansion with the inclusion of new data-supplying points. The MADIERA infrastructure will, as the web itself, have the capacity to grow and diversify after the initial construction period. The main objective of the project has been to create an open but sustainable system, nurtured by the collective energy of the data and knowledge producing communities of the European Research Area. Breaking these ideas down into more specific objectives, the MADIERA project has focused on 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 their disparate resources.

2) The employment of a multilingual 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 empirical scientific material.

4) Run an extensive program to add content, both at the data/information and knowledge levels.

5) Carried out extensive training of data providers and users to inspire and encourage the continuous growth of the infrastructure developed tools and guides for the practical side of such work.

6) Opened for the gradual integration of the emerging national infrastructures of the candidate countries into the European Research Area, by making available technical solutions and guiding material at low cost.

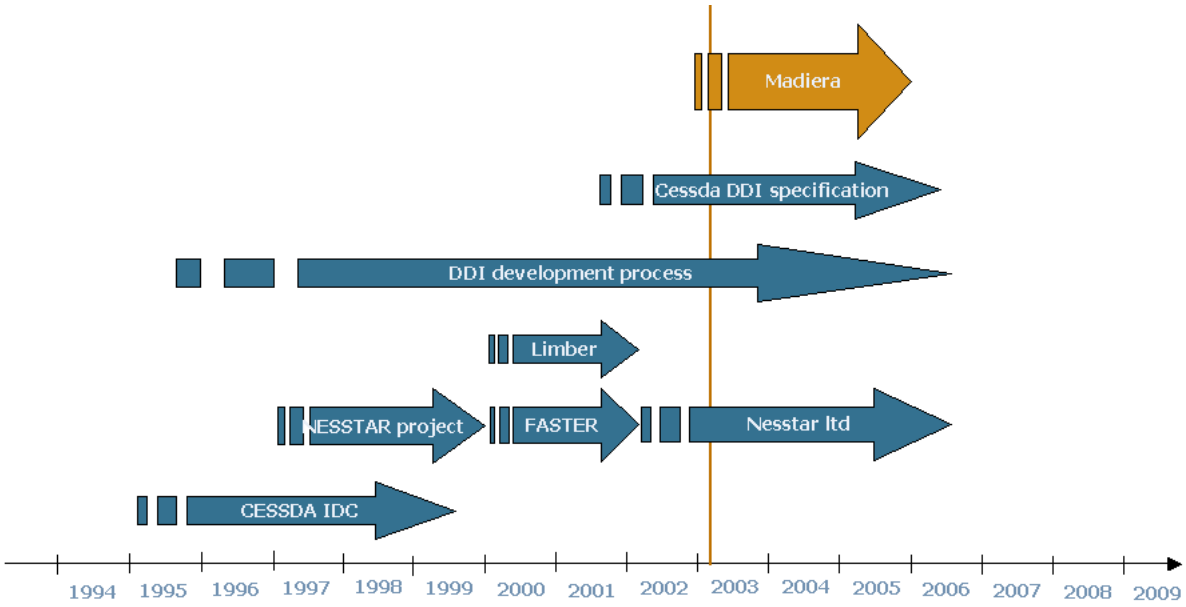
III. SCIENTIFIC DESCRIPTION OF PROJECT RESULTS AND METHODOLOGY

1. Implementation of a vision

Since the middle of the 1990s, the social science data archives of Europe have, as a specific response to clearly formulated needs by the scientific user community, run a long-term focused project on developing a modern user friendly integrated web interface to their collective holdings, an interface that should be rich in functionality and content. 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 World Wide web period. Since the timid start, the ambitions have grown considerably and grown in parallel with the potential and the development of the Internet and 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 may be regarded as another important step in this concerted effort but before being integrated in this particular project, the development of the three main components, the metadata standard, the thesaurus and the software technology were all initiated as separate initiatives or projects.

In this picture we can regard MADIERA as a modest and tentative implementation of the visions of the European social science communities, made possible through the financial and practical support of the EU research programs.

Figure 2. The building blocks of the MADIERA infrastructure



The MADIERA portal is based on three main components.

- A common standard for data documentation, the DDI (Data Documentation Initiative).
- The comprehensive multilingual thesaurus, ELSST (European Language Social Science Thesaurus).
- The Nesstar technology for making data resources available on the web.

These components will be presented in brief at first and then in depth later in this document.

1.1. 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 define 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 is 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 underway and almost completed on the development of version 3.0, aiming at a general ability to handle complex organized files.

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 his set of potential modes of data collection, or make his choices from some agreed upon standard set of modes of data collections. It would be a violation of the basic idea if we for this element put in a free-text description of our data collection process, disregarding the need to make it "machine actionable". The DDI 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 optional values, some syntactical rules and some clearly defined and normative best practice guiding the user, so that the material becomes machine actionable and not only human understandable. Throughout the DDI there are 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 studies or types of data. Generally, there is a need and a possibility to add the semantics and the knowledge content, the DDI itself represents only the structure, which may be differently applied by different scientific fields. 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 the data archives have been working to develop an European implementation of the DDI standard. At this point standardization means that there should be common agreement on lists of optional values (controlled vocabularies) for every element, there should be a common "template" and some generally agreed upon best practices. Some of the lists are simple some may be more complicated. The MADIERA project has taken up this challenge and has developed a common template, outlining the recommended best practices for documenting social science data that are intended for use in the European Common Research Area.

1.2. 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 metadata standard in a multi-language Europe, the thesaurus ELSST (European Language Social Science Thesaurus) stands out as the single most important component of the semantic and content-carrying kind mentioned above. This thesaurus was originally based on the UKDA HASSET-thesaurus; the multi-language idea was developed within the EU-financed LIMBER project (Language Independent Metadata Browsing of European Resources) and has now been carried significantly forward within the MADIERA project. Such a thesaurus is a hierarchically arranged controlled vocabulary, which is used for indexing and retrieval purposes 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 were done properly a user would be able to specify his/her search criteria in any of the supported languages and get a list of hits independent of what language they are described in. The keywords assigned to the metadata from a multilingual thesaurus can be instantly translated back into the supported language of the user. Initial full translation of the returned resources might then be achieved by applying standard automated web-based translation services. We know that the quality of these 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 ELSST 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 componentson the data input/data publishing side, as well as possibilities to browse and search more meaningfully on the data location and application side.

1.3. 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 data supplier side. Within the MADIERA project the intention has been to further refine the available technologies but to be explicitly aware of the data user side, to make the software even 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.

To give a comprehensive report on the outcomes of the MADIERA project, the scientific description of the project results and methodology will follow the outline below:

- a thorough description of the DDI metadata standard and the MADIERA implementation of the DDI standard;

- a description of the ELSST thesaurus itself and the very important administrative framework developed for the future maintenance and development of the thesaurus across the present 9 and potentially other European languages;

- a description of the core computer technology employed by the project;

- a description and discussion of the functionality specification and development work carried out within the project;

- a description of the user analysis work undertaken by the project;

- a description of the final MADIERA portal and the supporting documentation.

2. DDI and the MADIERA implementation

Metadata is data about data, the material that undertakes the complex explanatory task related to a core piece of data. Metadata converts naked data into information, the same way as analysis may further convert information into knowledge.

Metadata serves a variety of specific purposes, most of them may be found in the imaginary space delineated by concepts like "finding", "understanding", "assessing", "accessing" and "administering".

Finding: Precise metadata is the key to high precision resource discovery. A user is never searching for numbers, but for concepts or keywords measured and represented by numbers. Through catalogue information, study descriptions, question texts, definition of concepts or descriptions of sampling procedures etc, users are able to locate the collection of numbers that might fulfil their data needs.

Understanding: Metadata is giving meaning to numbers. Without human language descriptions of their various elements, data resources will manifest themselves as more or less meaningless collections of numbers to the end users. The metadata provides the bridges between the producers of data and their users and convey information that is essential for secondary analysts.

Assessing: Metadata is giving end-user a chance to assess the quality and relevance of a collection of numbers. By describing methodologies and procedures, as well as features related to the context of a particular study, end users are allowed to decide whether or not a data collection is meeting their professional or scientific standards.

Accessing: For several reasons data are not floating freely around. Scientific value of data resources may be matched by commercial value, making available high quality data is an expensive activity and data is often only available at a price. Another important restricting factor is the need to protect data privacy and guard against data misuse. Individual level data usually can only be made available if data privacy is protected.

Administering: Data resources have to be stored and maintained, wrapped in computer systems and administrative procedures. The common denominator of these concepts is the idea of "*sharing*". The evidence based knowledge production process is an activity with many groups of participants, each bringing different skills and resources to the table. It is also an activity that normally will be distributed in space as well as time.Metadata is therefore about communication. Metadata might be viewed as a structured conversation between the different persons, offices, organizations and

software processes working with a kernel, a dataset, all the way from the design process to the final users. The main purpose of this structured conversation is to make sure that all relevant information are passed on from one station to the next and that all participants have a chance to add their own relevant knowledge to this information exchange.

Within the academic sector social science data archives and data libraries have been established to provide researchers and students with data for secondary analysis. Some of these institutions have been in existence for 2-3 decades and house the largest collections of accessible computer-readable data in the social sciences in their respective countries. The primary goals of the archives and libraries have been to safeguard the data and to make them as easily accessible as possible for teaching and research independent of whether the users are able to pay for the services or not.

The social science data archives are rarely engaged in the collection of primary data. Neither are they themselves data users, they serve as professional brokers between various data providers and the academic community. Their holdings contain data from the public sector (statistical agencies, central government etc), the commercial sector (opinion and market research companies) and academic research. The archives do not only preserve data for future use but also add their own value to the collections:

- data received by the archives goes through a variety of checks and cleaning procedures to ensure their integrity;

- any system or software dependency is stripped away to make sure that data can be read at any time in the future;

- comprehensive computer-readable metadata are developed;

- data from various sources are often integrated and harmonised in order to produce easy-to-use information products (on-line databases, CD-ROMs etc.);

- data are catalogued and made accessible through electronic search and retrieval systems;

- in order to encourage the use of statistical data among students, teaching packages and interactive statistical laboratories, are developed.

Due to the extensive refinements of the data sources, as well as a long-standing reputation of responsiveness to users' needs, non-academic users frequently request data and related services from the archives. This includes users from the public sector,

as well as from the mass media and private companies. To the extent that services to non-academic users do not run counter to the agreements with the data depositors, access is usually granted.

The characteristics of the user communities go a long way to explain the high priority that the archives have given to the development of metadata:

- users of archived data have rarely been engaged in the creation of a dataset;
- archived data will frequently be used for other research purposes than intended by the creators (secondary analysis);
- archived data will frequently be used many years after they were created;
- academic users are often comparing and combining data from a broad range of sources (across time and space).

This analysis underpins our "communication" perspective on metadata. Whereas creators and primary users of statistics might possess "undocumented" and informal knowledge, which will guide them in the analysis process, secondary users must rely on the amount of formal metadata that travels along with the data in order to exploit their full potential. For this reason it might be said that social science data are only made accessible through their metadata. Without human language description of their various elements, data resources will manifest themselves as more or less meaningless collections of numbers to the end users. The metadata provides the bridges between the producers of data and their users and convey information that is essential for secondary analysts.

Over the years many initiatives have been taken within the data archive movement to create metadata standards. None of these have, however, reached the level of acceptance that is needed for a standard to be successful. The majority of social science data archives have documented their holdings according to a standard study description agreed in the mid 1970's by an international committee of data archivists. Unfortunately many local "dialects" of this standard have evolved and the archives have adapted their metadata holdings to fit the requirements of different storage and retrieval systems. As a consequence the level of standardisation across archives has been rather low.

In order to improve this situation, an international committee, the Data Documentation Initiative (DDI) was established in 1995, to create a universally supported metadata standard for the social science community. The committee was initiated and organised by the Inter-University Consortium for Political and Social Research (ICPSR). The members were coming from social science data archives and libraries in USA, Canada and Europe

and from major producers of statistical data (like the US Bureau of the Census, the US Bureau of Labour statistics, Statistics Canada and Health Canada).

The original aim of the Data Documentation Initiative was to replace the old-fashioned and obsolete standard study description with a more modern and web-aware format. The first version of the new standard was consequently expressed as an SGML DTD. In 1997 it was translated to XML where it have stayed since. This was just a few months after the World Wide web Consortium (W3C) released the very first working draft for this new language which according to the visions of the creators would add a new dimension to web-publishing, especially related to resource discovery and metadata.

The first version of DDI (1.0) was released in the spring 2000, and gave a comprehensive standard for documentation of freestanding single survey files, i.e. it made available the traditional codebook in a web-friendly environment. Version 2.0 of the standard was released in 2003, now with the ability to handle aggregate data and complex tables. Parallel with the MADIERA project there has been work underway which is now almost completed on the development of version 3.0, aiming at a general ability to handle complex organized files. This is particular important for the MADIERA focus on comparative research and comparative data.

The DDI standard is a voluminous system; to describe all aspects of relevance for secondary users of data requires a lot of detail. To help users find their way in this multitude the Nesstar software operates with a *template* concept as a guide. The total DDI standard is built into the data publishing software, then a template might be defined as a tailor-made selection of elements for a specific project. For the MADIERA portal, the MADIERA/CESSDA template is defined as a recommendation for a common denominator for the European data archives.

The document/table (partially shown below) summarizes the total DDI, selects elements for the MADIERA template, compares with other major initiatives and standards (Dublin Core), identifies where there is a need for a controlled vocabulary and marks elements of specific interest for the data location technology (the search procedures).

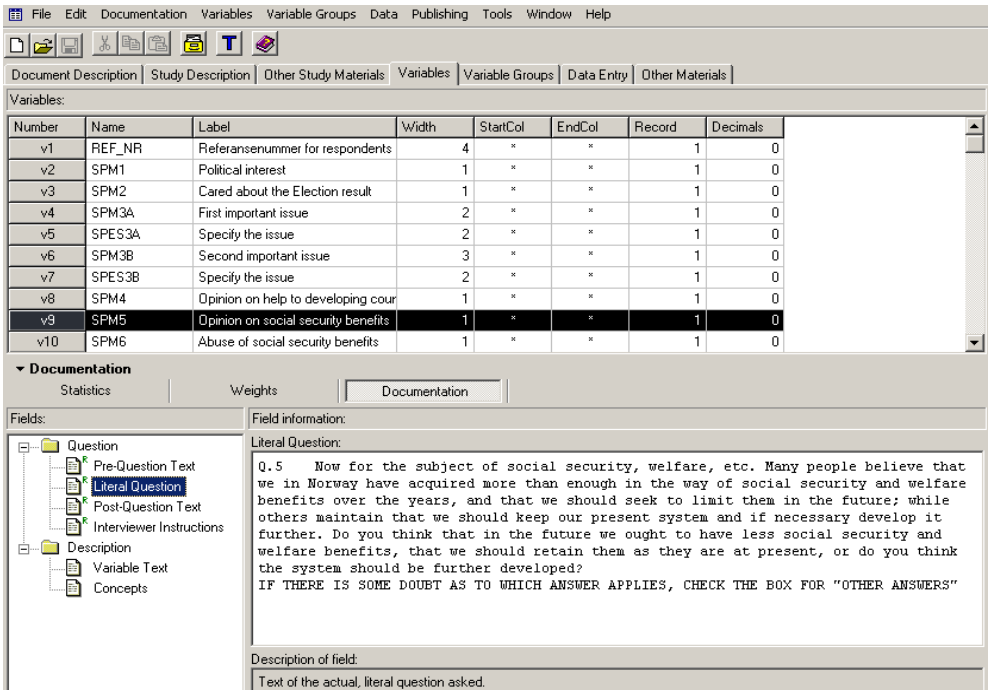
Figure 3. DDI elements in the MADIERA template

DDI Codebook , with possible CESSDA template	Recommendation CDG	Recommendation ICPSR	Dublin Core	Suggestion, Controlled vocabulary	Elements recommended for MADIERA search			Comments
					Study	Vargrp	Variable	
0.0 codeBook								
1.0 docDscr*								
1.1 citation?								
1.1.1 titlStmt								
1.1.1.1 titl	Mandatory	Mandatory	Title					
1.1.1.2 subTitl*								
1.1.1.3 altTitl*								
1.1.1.4 parTitl*	Recommended for 2.language							Systematically used to double language titles
1.1.1.5 IDNo*	Mandatory	Recommended	Identifier					**Identification number**
1.1.2 rspStmt?								
1.1.2.1 AuthEnty*		Recommended	Creator					

The Nesstar data publishing software allows a potential user to define her own template. However, the table indicated above represents the most thoroughly developed implementation of DDI and as such is a good recommendation for a best practice. For a comparatively oriented project like MADIERA it is of utmost importance that also the implementation of the standard is common across data suppliers.

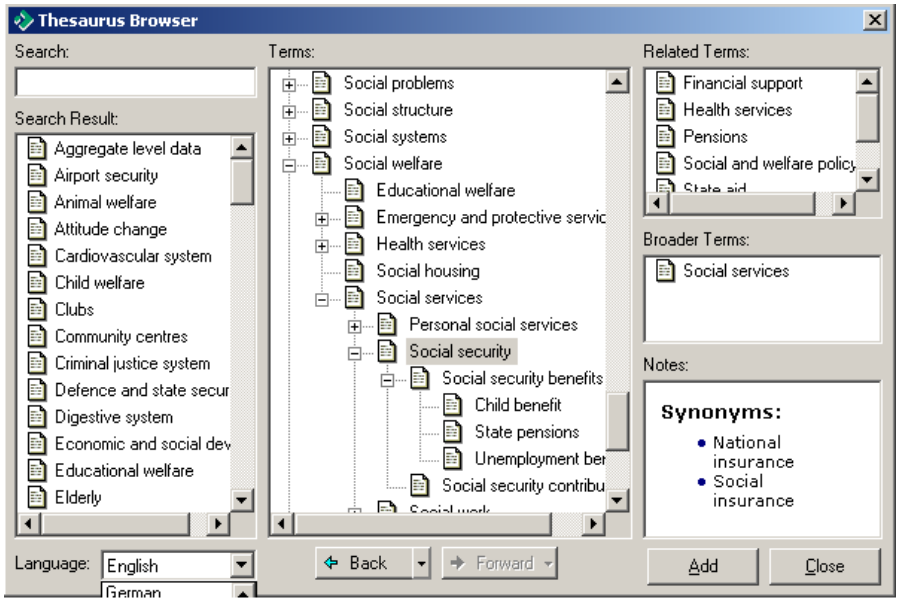
The DDI is a tool to create good descriptions of data for a broad spectre of uses. To develop an efficient and standardized data location technology on top of that, it will be of great help if we can concentrate the substantive meaning or content of longer texts into single or a few keywords. That would make searching more focused and precise. If we develop a procedure to insert keywords to summarize text, it should be used universally, across data publishers to make outcomes comparable. In the DDI, there are specifically two points where explanatory text could be easier handled by search procedures if summarized as keywords or concepts. One is where the subject of a *study* is presented as an abstract; the other is where each *variable* is documented with a question formulation. The Nesstar Publisher employs the ELSST thesaurus to allow insertion of keywords in such a procedure.

Figure 4. Documentation of a dataset in Nesstar Publisher



For a question text like the one shown above, it is obvious that it would be more efficient both to search and to index if we could substitute it with one or a few keywords. For that purpose, the multilingual ELSST thesaurus is used as an advanced controlled vocabulary.If we want to summarize the question text as keywords, we invoke the thesaurus by clicking on "Concepts", a quick scan of the text is carried out and the program returns a suggestion for the most relevant keyword.

Figure 5. Nesstar Publisher employs ELSST thesaurus



It is now possible to select and "add" keywords to the concept element for every variable by selecting from the menus.

As part of the MADIERA project, the interface shown above has been developed for the data documentation and publishing software, so that when a data-matrix is imported from one source and documented with question texts, etc from another source, the thesaurus can be invoked directly and keywords inserted appropriately on the fly.

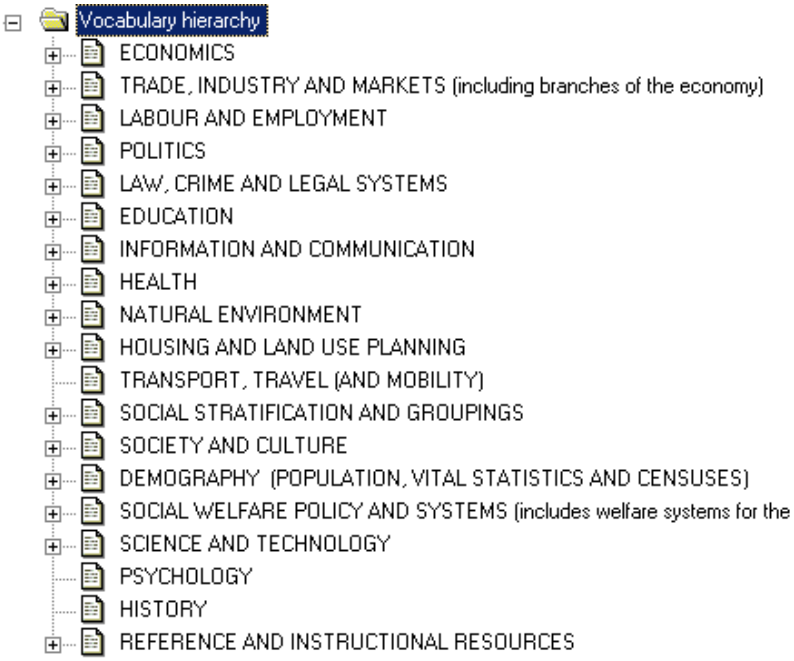
If the purpose is to find data, we might contrast two different methods: The "Google" style means to search through pre-indexed material, in practice to look up data by a pre-arranged central index. In contrast, the "Yahoo" style would be to browse by broader pre-defined categories in a clearer hierarchical structure. To insert keywords as described above is a procedure that betters the speed and precision of the "Google" style search.

The DDI element 2.2.1.2. topcClass (topical classification) serves the same purpose for the "Yahoo" style browsing. In the template this is a recommended element.

For the MADIERA project a common empirically based two-level topical classification has been developed. This specific classification is defined in the common template as a controlled vocabulary, and this allows for insertion of topical groupings in much the same way as the thesaurus allows for insertion of keywords.

The classification consist of the 19 main groups listed below, spanning a total of 83 subgroups:

Figure 6. MADIERA's topical classification

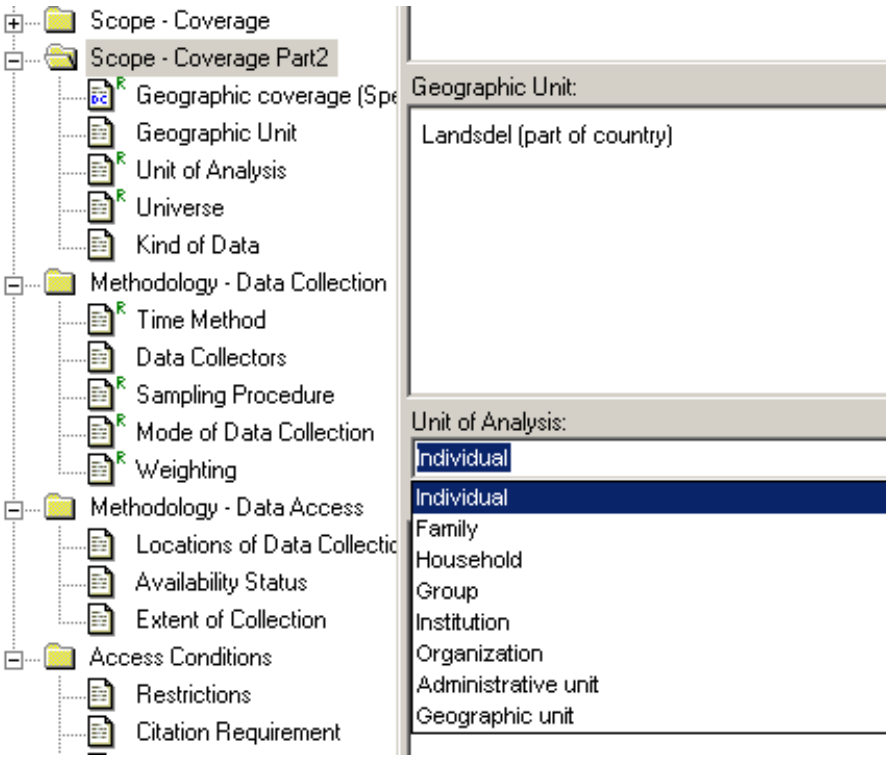


Nesstar Publisher is designed so that almost any clearly defined standard/standardized vocabulary can be linked into a documentation process via a URL. In addition to the above demonstrated examples this will be very useful for geographic specifications, since the procedure can allow for easy insertion of aggregate unit codes, names and coordinates, without such a procedure it would be very difficult to standardize such data.

The DDI element 2.2.3.8. *anlyUnit* demonstrates the need for a more elementary type of controlled vocabulary. In DDI documentary texts, the element is explained as "Basic unit of analysis or observation that the file describes: individuals, families/households, groups, institutions/organizations, administrative units, etc. The 'unit' attribute is included to permit the development of a controlled vocabulary for this element."

In the common MADIERA template there is included such a vocabulary. This is maybe not the ultimate and absolute final vocabulary; only extensive use will teach us the need for adjustments. But the positive side is that we by simple means obtain standardized use of such an element. It may for instance give us the possibility to search for datasets with household as the natural unit of analysis, since this piece of documentation is a recommended element in the template.

Figure 7. MADIERA' template includes controlled vocabularies



In the MADIERA template there are included controlled vocabularies for *Unit of analysis*, *Type of data*, *Time-method* of the data, *Sampling Procedure*, *Mode of Data Collection* and *Availability Status*. This could be regarded as a first efficient step towards a

harmonization of both procedures and actual metadata content at the different data-archives in Europe.

3. Multilingual thesaurus

In terms of the creation of a valuable European-wide resource, the production of a 9-language version of the multilingual thesaurus ELSST (European Language Social Science Thesaurus) has been a major success of the MADIERA project. The way that that resource has then been applied in the MADIERA portal has surpassed all our expectations. But perhaps more important is the fact that a robust system for the maintenance and upkeep of ELSST has been developed, as opposed to the promised prototype in the project plan, which means that the valuable work achieved during this project will not be lost. In fact it is envisaged that ELSST will grow not only by addition of concepts but by addition of languages as well. Already interest has been shown in providing Russian, Italian, Romanian, Austrian, Hungarian, Portuguese and Polish versions of ELSST.

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 metadata standard 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. This thesaurus was originally based on the UKDA HASSET-thesaurus; the multi-language idea was developed within the EU-financed LIMBER project (Language Independent Metadata Browsing of European Resources) and has now been carried significantly forward within the MADIERA project. Such a thesaurus is a hierarchically arranged controlled vocabulary, which is used for indexing and retrieval purposes 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 a list of hits independent of what language they are described in. The keywords assigned to the metadata from a multilingual thesaurus can be instantly translated back into the supported language of the user. Initial full translation of the returned resources might then be achieved by applying standard automated web-based translation services. We know that the quality of these 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 ELSST 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.

At the beginning of the project ELSST was a 4 language multilingual thesaurus of approximately 1,355 concepts expressed in English, French, German and Spanish.

This was the final deliverable of another EU IST project, LIMBER (Language Independent Metadata Browsing of European Resources). This first version of ELSST had been developed from the UKDA monolingual thesaurus HASSET (Humanities And Social Science Electronic Thesaurus). The major 38 hierarchies of concepts had been reduced by excluding any country specific or organization specific terms to produce a more European-centered thesaurus. Although successful, the terms are still in the present version of ELSST, the true power of the thesaurus could not be demonstrated since the data archives of CESSDA that held data in the languages other than English had, at that time, not published their data using the DDI standard and Nesstar technology.

Hence, the choice of mainly Scandinavian languages for this project, since the archives from those countries had been early adopters of both the DDI and Nesstar. Stage one was for the new languages of Finnish, Danish, Norwegian and Greek to provide terms for the concepts already in ELSST. Stage two was to bring in smaller but essential hierarchies from HASSET and for these concepts to be translated into all languages. The third and final stage was to bring in the remaining HASSET concepts and again provide terms to express these in all languages. Although there were no specific funds within the

MADIERA project for the continuation of the French, Spanish and German translation, it was hoped that separate funds would be found.

The major disappointment in the MADIERA project was that separate funds for the continuation of the German and Spanish versions of ELSST were not forthcoming. These funds however are still being sought, as it is the initial mass translation that cannot be adsorbed into daily running costs of most CESSDA archives.

However, one major and unexpected bonus for the MADIERA project was that the Swedish Data Archive did find extra resources and was able to complete all three stages, not only increasing the number of languages to 9 but also being able to provide DDI metadata describing data held in a Nesstar server.

3.1. Stage One

The translations of the existing ELSST concepts were staggered so that each archive could have an initial period where specific questions and problems could be addressed by the UKDA thesaurus team for each language. It also allowed time for the Thesaurus Construction and Translation Guidelines to be prepared for those archives, unlike the Finish Data Archive, that did not have their own thesaurus. The final versions of the Thesaurus Construction and Translation Guidelines were submitted to the EC as deliverable D5.1.

The user guide outlines the purpose, coverage and construction of the thesaurus along with the standards and conventions used. It also contains instructions, with examples, of how to search and navigate the web version of ELSST. The translation guidelines outline the procedures and standards adhered to in the creation of ELSST. In addition the guidelines contain a listing of translation sources, for all languages.

The hierarchies from ELSST were sent out in a format which seemed to suit the translators; namely a single listing with synonyms and scope notes integrated. The process of answering queries from the separate translation teams revolved around the meaning of terms, structural changes, the addition of scope notes and synonyms and the logging of suggested new terms. Hence it has proved necessary to create an administrative process so that the queries could be logged, the resolution recorded, and the changes published to the other partners.

By November 2003, the Finnish stage one version of ELSST had been completed. The Danish was completed by February 2004, the Greek by May 2004, with the Swedish and Norwegian completing stage one in September 2004. This was 6 months behind schedule, but work had already started on the stage two translations. However even at

this early stage it was envisaged that an extension might be required to complete the full thesaurus to the desired standard.

At this stage the UKDA massed in the translated concepts for each language, then did rigorous consistency checks. The inconsistencies were sent to the translators for resolution and the corrections input again by UKDA staff.

Stage one also include the setting up and running of an Administrative Workshop to discuss the best way to manage and maintain ELSST during and after the project.

The Workshop was held at the Butterfly Hotel in Colchester, Essex on the 10/11th June 2003. A draft agenda and a document listing the aims, objectives and hopeful outcomes was circulated to delegates, along with specific questions that required answers or at least strong proposals.

The workshop was extremely successful and resulted in agreement on several immediate and long-term issues. An initial draft report was made available on the internal MADIERA web site. Issues raised also meant revisions to D5.1 the User Guide and Translation Guidelines; new versions of these documents were also made available.

For the effective running of the project, the workshop delegates agreed the need for an improved method of logging queries and suggestions.Hence a database was designed and a prototype web interface developed by UKDA programmers during this stage with a live beta version being made available in August 2003.

The workshop report formed the basis of the specification for a prototype thesaurus management system. The system would be in place at the end of the MADIERA project to ensure the efficient maintenance and European-wide use of the multilingual thesaurus.

Running parallel to the above processes was the preparation of new hierarchies from the HASSET thesaurus for stage two of the translation work.

3.2. Stage Two

The Administrative Workshop held in June 2003, also agreed on the constitution and responsibilities of the ELSST management team during stages two and three of the project. The 5 member team from different partner archives were to be responsible for a)Reviewing candidate terms and new hierarchies; b)Setting up of specialist workgroups for specific problem areas; c)Ensuring voting on changes and resulting actions are carried out; d)Alterations in structure of the thesaurus; e)Release of official versions; f)Ensuring

consistency across languages; g)Translation guidelines and User Manual and other documentation and h)Seeking funding from CESSDA for ongoing maintenance.

In this stage the structure of existing HASSET hierarchies were tidied up to make them more consistent and ready for reduction for European usage. This involved the same process as employed in the LIMBER project of reduction by excluding any country specific or organization specific terms. Another part of the process included work identifying potentially ambiguous terms and providing scope notes for these in advance of the hierarchy selection process. During the previous project LIMBER, translation work was delayed whilst term ambiguity was clarified. Doing this work in advance meant that the translation ran more smoothly without individual translators misunderstanding the meanings of these terms.

The creation of the database to log suggestions and comments on thesaurus terms was in place and ready for use from October 2003. The software greatly increased the efficiency of the management team for both this and the final stage. Another EU project, Metadater, also adopted this software for the management of their workpackages.

During this stage several meetings took place between the partners to help develop a more coherent strategy for the development of the thesaurus features of both the MADIERA publisher and client software. They resulted in realistic targets being set for the MADIERA content provision and metadata workshops in June 2004 and the version release of the MADIERA software in September 2004.

The combined workshop on content provision and content metadata was held at the University of Essex 22-23 June 2004 and was extremely successful. Especially in the evaluation of the publishing software developed during the MADIERA project that incorporated the latest version of ELSST. Here semi-automatic assignment from ELSST to the relevant section of the DDI metadata is achieved by analysing question text and the categorised answers.

The first hierarchies for this stage were sent out for translation in all languages at the end of August 2004. In all a further 37 new hierarchies were distributed by November 2004. At this stage ELSST consisted of 86 hierarchies and 2119 preferred terms.By January 2005 the complete Danish, Greek, Swedish and Finnish translations of the hierarchies had been returned.As in stage one consistency checks were carried out by the UKDA and discrepancies returned to the submitting archive. UKDA staff once again input these final amendments.

At this stage it was made clear that the German and Spanish version of ELSST would not be completed within the MADIERA project, since external funding had not been found. The French translation, although funded, was unable to keep to the strict timetable laid down in the project. It had been hoped that the French translation would be performed using the prototype web interface to the ELSST database tables. However it proved that the software was not robust enough and required modification.

However by the May 2005 IASSIST (International Association of Social Science Information Service and Technology) conference in Edinburgh ELSST had 6 complete languages covering social science archives which held large amounts of data. The Norwegian version was completed in time for the Evaluation Workshop held at the UKDA June 27-28th 2005 where the usage of ELSST in the MADIERA portal was evaluated. The functionality to be evaluated had been decided at a series of meetings between the UKDA and Nesstar Ltd.

Also part of this workpage was the translation of the smaller hierarchy of terms from the CESSDA topic classification. It was extremely rewarding that all 9 languages, including French, German, Spanish and Swedish were completed by May 2005, since that greatly contributed to the success of the poster session at IASSIST and the wide interest generated.

Running parallel to the above processes was the preparation of new hierarchies from the HASSET thesaurus for stage three of the translation work and the creation of the final 2.2 version of ELSST. Due to the late start to this stage and the fact that the timetable included the summer months when annual holidays are traditionally taken, the partners asked that the project coordinators request an extension to the project. This was so this stage of the work could be completed to the same high quality of the other two stages.

3.3. Stage Three

In this stage the structure of the remaining HASSET hierarchies were tidied up to make them more consistent and ready for reduction for European usage. This involved the same management team and process as employed in stage two.

A paper on the MADIERA project was presented at the First International Conference on e-Social Science, held at the University of Manchester, June 22-24th. It was part of a metadata workshop which showed how the MADIERA project was using the DDI standard to build infrastructures and resources that were a prerequisite for an e-Social Science GRID.

The organization and great success of the Evaluation Workshops held at the UKDA June 27-28th 2005 was in part due to the meetings between UKDA and Nesstar Ltd along with the project-wide teleconferences. The event itself focused on the use of ELSST and the CESSDA topic classification within the MADIERA portal as search and browsing tools. The information gained from the exercise was used to improve the prototype version of the MADIERA portal. After the workshop a MADIERA project meeting was held to discuss and timetable the final stages of the project.

By September 2005 the preparation and dissemination of the final hierarchies of the last development stage of the multilingual thesaurus, ELSST version 2.2, had been completed.The translation, mass input, consistency checks and final amendments would be carried over into the 3 month extension that was granted to the project, thus allowing a longer time for quality translation from the other partners.

Also by September 2005 translation of ELSST into French, via the new web interface, was started. To begin with this involved the terms from Stage two of the project, but resources were made available that would take the work pass the extended end of the MADIERA project, with a full French version predicted to be available at the end of April 2006. Such an intense use of the interface meant that several refinements were made during the extension period and a robust version was released which is now being used by every partner.

The new web interface allows partners to update their language version of the multilingual thesaurus. For each language it is possible to edit preferred terms and add, delete or edit synonyms and scope notes. This meant that once the UKDA had entered the mass of remaining terms and carried out final consistency checks, responsibility for amending the entries to resolve the errors found was passed on to the individual partners. This further work on the underlying programs has also ensured that the maintenance and upkeep of the thesaurus will continue once the MADIERA project ends.

At the end of the project ELSST contained 3,209 concepts, more than double the initial number of 1,355. All of these had been expressed in English, Finnish, Danish, Greek, Swedish and Norwegian terms. The French still had approximately 1,000 terms left to translate. Although the German and Spanish had not progressed since the LIMBER project, the overall combination of this version of ELSST and the underlying data available in Nesstar servers meant that the functionality of the MADIERA portal was not impaired.

A paper on the MADIERA project was presented at the Association for Survey Computing International Conference on Survey Research Methods – Maximising Data Value.

The CESSDA Expert Seminar held in Madrid Spain at the Spanish Data Archive (CIS) focused on how the deliverables of the ELSST thesaurus and the MADIERA portal could be carried forward after the project has ended. It is envisaged that both the portal and the thesaurus will become CESSDA-wide resources and part of the new planned web site.

A similar technical meeting was held at NSD, in Bergen, in February 2006 to discuss the final refinements that were possible to the Final Report version of the MADIERA portal and what developments should continue after the end of the project and how these could be organized and resourced.

3.4. Conclusion

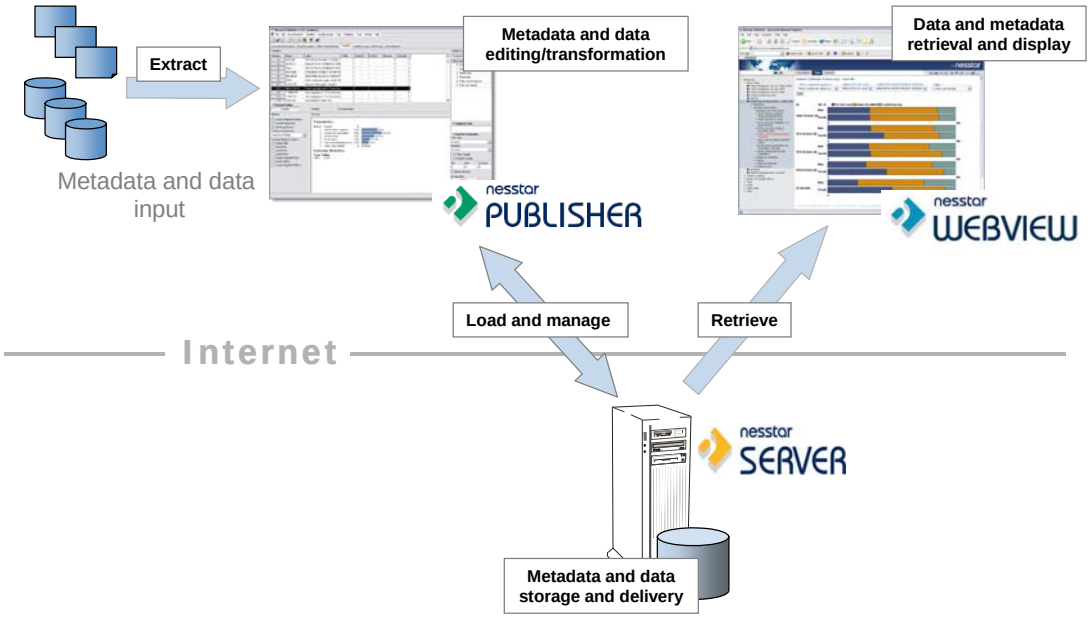
The ambitious infrastructure of the MADIERA portal with its underlying standards also requires the development of supporting software. Without easy-to-use and efficient software to support the creation of metadata and the publication process, it is difficult for archives to provide quality data content. The multilingual thesaurus is not only a key resource for the data location process, but also for the documentation and publication process. As part of the MADIERA project, an important supporting piece of software has been added to the Nesstar Publisher.The Publisher can import data from the common statistical packages and builds up a DDI-structured view of the data through a common CESSDA template. The Publisher now has the thesaurus built in to facilitate automatic insertion of keywords at study or variable level and will automatically classify and publish data according to the CESSDA topic classification.

Increasing the availability of high-quality data is a way of increasing the importance of secondary analysis in the social sciences. For that to become a reality the high-quality data needs high-quality documentation to accompany it and high-quality resource discovery tools to locate it; and that are what the ELSST thesaurus and the MADIERA portal delivers.

4. The technical platform

The MADIERA portal is built on Nesstar technology, which has been refined during the project period. A completely new generation of the underlying Nesstar technology has been implemented and deployed in the portal. Nesstar 3.0 forms the basis of the MADIERA platform. This version of Nesstar includes extensions and improvements to all parts of the technology: server, end-user client (Nesstar WebView) as well as the data and metadata management tool (Nesstar Publisher). This section gives a description the core computer technology employed by the project and of how the parts of the architecture are related:

Figure 8. The technical platform



On top of Nesstar, the MADIERA portal platform – a new technology serving as a central resource to the MADIERA network has been designed and implemented. A description of the portal platform will follow after the presentation of the Nesstar software.

4.1. Nesstar Server

The Nesstar server provides the basic building blocks of the MADIERA distributed network. It plays a similar role to MADIERA as a standard web-server do to the web in general and has a series of specialised functionality for distributed publishing and dissemination of statistical resources. The Nesstar server provides a robust, efficient and scalable platform for storing and serving statistical data.The Server builds on standard web technology and protocols providing universal access and seamless integration with the broader web services of the data archives.

The following functionality has been added:

Extensions to the metadata model: The metadata object model has been extended to support the complete set of metadata elements of the MADIERA DDI template.

Variable level searching: To support efficient location of potentially comparable variables, variable level searching has been implemented.

Performance: The overall performance of the Nesstar server has been radically improved to support the amount of data and the traffic of the MADIERA network.

Simplified installation and configuration: The installation and configuration tools of the server have been radically improved to make it easier to establish and maintain the MADIERA network.

Server side bookmarks: Server side storage of bookmarks has been implemented. This is a crucial building block of the "hyperlinked information space" concept as well as the agent technology.

Improved and generalised GIS interface: A more generalised and standardized interface to various mapping (GIS) systems has been developed. This is a prerequisite for a full implementation of the MADIERA geo-referencing functionality.

Support for derived variables: A system has been implemented, whereby derived variables (recoded or computed) can be held on a Nesstar server across user session. This is a crucial building block in all functionality related to comparable variables. With this functionality, variables can be harmonized according to the specifications of the user and stored and shared on a Nesstar server.

4.2. Nesstar WebView

WebView is the general end-user client of Nesstar and the MADIERA network, providing advanced and powerful data analysis capabilities through a standard web-browser. With WebView you can locate data using browse- and search-methods, display and browse metadata, analyse and visualize data, export tables and graphs to various tools like Word, Excel etc. and download data to a variety of formats (like SPSS, SAS etc.)The following functionality has been added:

Overall design: The overall design and look & feel of the end-user client has been dramatically improved, largely based on input from the MADIERA usability process.

Searching: The search dialogs has been improved to meet a more complete set of requirements regarding searches across multiple servers/archives.

Browse-list: The main navigation tree has been improved to make it easier to include information from a series of servers.

Metadata display: The way Nesstar displays metadata has been improved to meet the requirements of the data archives.

Subsetting: A more robust sub-setting dialog has been implemented. Sub-setting and downloading of data is crucial to the services of the data archives.

Codebook creation: A facility that creates a complete electronic code-book of datasets stored on a Nesstar server has been added.

Improved bookmarking: The Nesstar bookmarking functionality has been improved and extended. This is crucial to the "hyperlinked information space" concept.

4.3. Nesstar Publisher

Nesstar Publisher is the metadata and data preparation tool of the MADIERA platform. It enables the archives to extract data and metadata from a variety of formats and back-office systems and to prepare the information for publishing to the MADIERA network. It is also the tool that facilitates the creation of DDI-compatible metadata and provides the facility to index resources according the MADIERA multilingual thesaurus. The key role of the Nesstar Publisher is to enable and encourage the development of high quality metadata supporting the standards and best practices of the MADIERA network. The Nesstar Publisher stands out as a very significant piece of software in its own right. Since the project initiation this component of the whole system has increased substantially in importance. The Publisher not only make possible a flexible transformation of data stored as system files under most major statistical packages to DDI-structured XML-based Nesstar files or file systems, it also contains an internal version of the multilingual thesaurus. This makes possible a partly automatic, but still user supervised insertion of generalised keywords at various points in the content-carrying parts of the metadata. This will have a substantial positive effect for data-locating procedures that have been developed. The following functionality has been added:

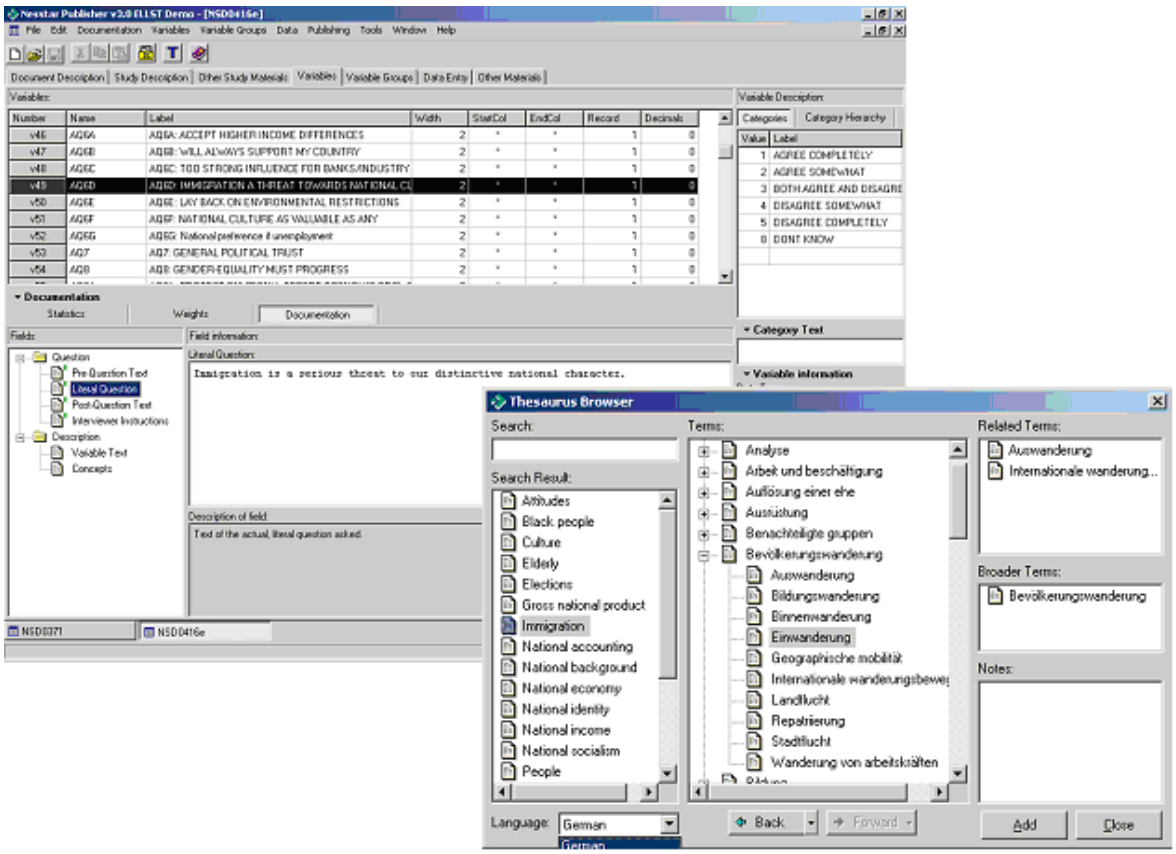
Remote publishing: The Publisher can now publish data and metadata over the web. This is simplifying the data publishing process of the involved archives.

Server and resource management: The Publisher has also got a series of server and resource management functions that make it easier to manage the content on a Nesstar server. Of special importance is a new component to manage the browse-list (or the MADIERA subject classification tree).A dataset that are classified according to the MADIERA standard will now automatically be published to the right catalogue(s). This is enforcing standardization across archives and ultimately making it easier for the end-users to locate relevant data across archives.

Improved metadata template support: The metadata template support has been radically improved. The agreed MADIERA "best practice" metadata template can now easily be distributed and shared across archives to enforce metadata standardization. It will also

be easier to make and propagate changes to the agreed standard and to create local archive-specific extensions.

Figure 9. Nesstar Publisher: Metadata template



Integrated thesaurus support: A module that supports the MADIERA multilingual thesaurus has been added to the Publisher. This is a module that makes it easy to add keywords from the thesaurus at the study, variable group or even variable level. The latter is of utmost importance, as it will dramatically increase the chances of identifying comparable variables across languages. To make it quick and easy to add keywords at variable level, a facility has been added whereby the question text is analysed and relevant keywords proposed by the system. In the current implementation this is based on a very simple text-query mechanism. Adding more advanced text indexing algorithms based on a learning material would dramatically improve the precision of the mechanism. The thesaurus support in the Nesstar Publisher is using an efficient combination of remote/local access. The Publisher is holding a local copy of the thesaurus, but will check the server for updates every time the thesaurus tool is opened for the first time in a session.

*Global variable repository:*A new functionality whereby an archive can establish a global variable repository (holding variable descriptions across datasets) has been added. This

is another mechanism that is implemented to make it easier to create high quality metadata and to improve standardization with the ultimate goal of making it easier to identify comparable variables.

Publishing of none-data resources: Functionality to publish documents and reports to a Nesstar server has been added. The none-data resources can be described using theDublin Core metadata standard and can be linked by internal references to data resources. The Nesstar metadata model supports the Dublin Core standard as well as the extensions to Dublin Core made for the e-GMS standard used by the e-Government initiative in the UK. Nesstar has also integrated major parts of the ISO11179 metadata registry standard. Publishing of none-data resources is one of the building blocks of "the hyperlinked information spaces" that are central to the MADIERA project.

4.4. More about the technology

Nesstar is a fully web-enabled technology providing powerful and advanced data analysis and visualization capabilities through a standard web-browser. This avoids expensive installation and maintenance of software components on the end-user's desktop. It is also a fully distributed technology providing integrated access to data stored on separate remote servers.

Unlike most data publishing and analysis systems, Nesstar supports micro-data (rectangular and hierarchical) as well as aggregated data (multi-dimensional tables or cubes). As most organizations possess both types of data, this not only removes the need for parallel investment in two lines of technologies, but also enables efficient integration of data across this divide. Examples of micro-data are survey-data and census data at the individual level. In addition to aggregated and disaggregated statistics, a Nesstar system can also be used to store and deliver other digital information products, like documents, pictures, maps etc. These additional information products can be described by metadata and made searchable along with the statistical resources. It is also possible to create links between data and knowledge products allowing the user to drill down from fact sheets and reports to more detailed data.

The Nesstar platform provides the shortest possible route to the web for the data within an organization. The Nesstar Publisher can import data from most known statistical packages, data production systems, file formats and databases.

Nesstar technology is as easily interfaced on the output side. Nesstar can export data to most known desktop analysis tools (such as statistical packages and spreadsheets), as well as being able to export tables, graphs and maps to standard office and authoring

tools. The ability to meld into the existing technological environment of the enterprise is one of the uniquely powerful properties of the Nesstar platform.

Nesstar WebView offers a truly unique and intuitive interface to statistical data allowing data users and analysts to focus on content and the creation of knowledge rather than techniques. Standard OLAP features (like drill-down/roll-up and slicing and dicing of multi-dimensional tables) are combined with a repertoire of statistical methods to produce a highly interactive, user-friendly and visually attractive interface to the data.

The Nesstar Server builds on standard web technology and protocols providing universal access and seamless integration with the broader intra-, extra- or internet services of the organization. Robust and scalable components from top to bottom secure remarkable performance even in situations with high volumes of data, high number of users and inexpensive hardware. A specialised statistical engine provides lightening fast tabulations and statistical analysis of datasets of any size, outperforming standard statistical packages like SPSS and SAS.

Nesstar has an integrated mapping module providing standard thematic mapping based on ESRI Shape files. Other mapping modules can be plugged in to extend this functionality.The Nesstar technology builds on rich metadata standards supporting intelligent knowledge management and retrieval. Of special importance is the support for the DDI (Data Documentation Initiative) standard which over the last couple of years has gained increasingly rapid acceptance amongst data producers and users world-wide. Nesstar is the only complete implementation of this standard and is currently recognised as the only viable technical solution for organizations migrating to the DDI.

Internally Nesstar servers store metadata in a relational database. However, on the input side as well as the output side metadata are exchanged as XML. The Nesstar Publisher can import metadata in XML format (or a variety of other formats) and will deliver metadata to the server as XML. Nesstar is as such fully compliant with the UK e-GIF standard for data and metadata transfer.

The metadata foundation is a key to Nesstar. Metadata are used:

- to support efficient and high precision resource location;
- to provide easy-to-use navigation structures;
- to provide the user with all the information needed to understand how to use a data resource and to assess it’s quality;

- to enrich any output from the system (screen display, print or export) with relevant information about the resource that the output is derived from;

- to provide logical descriptions of data that can be used by the software components to automate processes or guide the user;

- to facility efficient content management.

Nesstar is a highly customisable technology that can be easily modified to meet specific look and feel requirements or to integrate with other services and technologies on the web.

The Nesstar Server is built according to Sun's Java 2 Enterprise Edition (J2EE) framework with a J2EE compliant application server fronted by a web server and a web-client application. Persistent storage of metadata is supplied by a back-end relational database system.

A standard Nesstar system ships with the MySQL database integrated. MySQL can easily be replaced by MS SQLserver or Oracle which both have been tested with the system.

Nesstar is built on the J2EE compatible JBoss Application server and the Tomcat web-server. The WebView engine is heavily based on Cocoon and Velocity from Apache.

Nesstar services can be accessed through any JavaScript enabled browser. The entire server is implemented in Java, except for the statistical engine and the cube engine that are implemented in C++ for efficiency reasons.

The Nesstar server is currently available for Windows (2000 or XP), UNIX (Solaris) and Linux operating systems. The Nesstar Publisher runs on any Windows based computer.

4.5. Security and Access Control

Nesstar comes with an integrated access control system that will allow the data publisher to control the access to the data with high precision and security. The system supports high granularity protection of resources where even individual variables or metadata elements can be assigned specific rules separate from the dataset in which they are located.

The system is role-based, allowing the data owner to define the concrete access conditions of different categories of users.Authentication can either be based on user-id/passwords or third party access management services like the UK-based Athens system. The software comes with a built in user-database and a web-based tool to

manage users and roles. The internal user-database can be replaced by any existing SQL-based user management system that might already exist within the organization. The access control system of Nesstar has been further improved to meet the specific requirements of the MADIERA project.

5. User requirements and usability testing

The MADIERA project has been a user driven software development project. This means that users were closely involved in the design from the beginning to the end of the project. The project has two main target groups: researcher with in the European social science community and data providers that wish to connect to the portal. During the project period there has been continuous contact with representatives for both groups.

User specification of needs, user testing of implemented solutions and analysis of user reactions and evaluations has been fundamental to the MADIERA project. The software tools developed under the portal have to offer functionality and solutions that are in accordance with expressed user needs. Since the MADIERA infrastructure is intended to help solve what often are non-routine problems over a broad range of types of users, it is important to ensure that the software tools developed is user friendly and appropriate for the target user groups.

A first step was map out user requirements and analyse these in order to feed this information into the functional specification for the system architecture. The challenge in user analysis was to collect information that would provide the project with a detailed understanding of work practices among potential users of the product. The methods applied had to accommodate the needs of independent users and those of users who were also publishers of data.What was needed was a sensitive method that was able to make the underlying logic in users' working-process explicit. The Contextual Design method was chosen.

The mapping was partly based on re-analysis of data from earlier projects, partly on a new user analysis for the MADIERA project. The material from the projects NESSTAR, LIMBER and FASTER were revisited and reanalyzed and the findings relevant to the MADIERA project were summed up and an interview guide was outlined. The review of existing data gave a valuable introduction to early discussions for development of the prototype

The more elaborate user analysis based on new data pointed at three different user profiles named 'the young researcher', 'the IT-curious researcher' and 'the traditional researcher'. Interviews with 14 potential users were carried out at all sites. The

respondents were recruited from universities being researchers covering the field of empirical quantitative research. A user expert team which was set up in the beginning of the project with members from all project partners, carried out the interviews. The user requirements reported were made the basis for further development of the design and functionality of the portal.

The second step was to plan and carry out the usability testing. Usability is defined as:

> The extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use. (ISO 9241-11:1998)

Usability testing has been carried out several times during the span of the project and at several stages. The testing proved valuable for the development of the design and functionality of the portal. The target user group for the MADIERA portal is researchers within the social sciences as well as data providers. Whereas the interview about user requirements were carried out among the first group, the usability tests have been targeted at data providers. However these data providers, mainly data archives, are daily in close contact with the European social science community. The two first usability tests were carried out according to the method "heuristic evaluation" which is a method developed by Jacob Nielsen (http://www.useit.com/). The third and the fourth test used a structured questionnaire to test usability in order to adapt to concrete wishes from the development team

The goal of the first testing was to identify possible usability problems associated with the preliminary version of the MADIERA portal, the prototype. The first formal usability test report was send out in November 2003. This was a test of the functionality of the Madeira user interface carried out by the User Expert Team. The testers were instructed to identify usability problems and rate them according to severity. Five tests were carried out according to the principles of heuristic evaluation. Altogether the testers reported 34 distinct usability problems. The severity of the problems ranges from being major problems, which are important to fix, to problems of a cosmetic nature. Overall, major usability problems are few and far between.It appears from the means of the severity ratings that there are no "usability catastrophes". This reflects the fact that all the system's main functionalities such as running diverse analyses, downloading data and saving tables seem to work without any major problems. Overall, the severity scores of the usability problems indicate that usability problems of the MADIERA prototype are modest in kind and in number.

The second usability test was carried out among this last group at the Content Provision Workshop at the University of Essex in June 2004. The testing focused on the new functionality within Nesstar Publisher. Since the project initiation Nesstar Publisher has increased substantially in importance. The publisher not only make possible a flexible transformation of data stored as system files under most major statistical packages to DDI-structured XML-based Nesstar files or file systems, it also contains an internal version of the multilingual thesaurus. This makes possible a partly automatic, but still user supervised insertion of generalised keywords at various points in the content-carrying parts of the metadata. This will have a substantial positive effect for data-locating procedures that may be developed. For that reason the testing of the Data Publisher, which also employs a common template for data documentation work developed for the CESSDA archives became the main topic for workshop

The third usability testing was carried out during the workshop in June 2005 where all the CESSDA members were invited. At this event the MADIERA portal was presented for several of the European Social science data archives. They are potential providers of data to the portal – they are also very important links to the research community and data users in their countries.Since one of the central aims of the project is that the MADIERA infrastructure will become the new CESSDA portal, extensive information activity have been aimed at this community. The success of MADIERA will to a considerable degree be dependent upon how the project deals with the future needs of the European data archives. It is registered as an initial success that CESSDA has decided to make the MADIERA portal its new integrated catalogue.

The workshop focused on evaluating a more robust version of the MADIERA portal and the latest Nesstar software. The test focused on the newest functionality; and in particular the use of the multilingual thesaurus ELSST and the geo-referencing tools.Both in the hands on evaluation of the software and in evaluating the portal, the discussions were regarded as extremely useful.

However, this time the testing team used a structured questionnaire to test usability in order to adapt to concrete wishes from the development team. The test had four parts aimed at searching for information, retrieval of datasets using the CESSDA Classification scheme (browsing), using the ELSST_FREE_TEXT/ELSST (browsing), google style searching for datasets compared to browsing respectively. The participants were requested to carry out specific assignments in order to test various functionalities of the portal. Alongside participants were asked to take elaborate notes describing their successes and failures.The main conclusion from the workshop was general satisfaction with the MADIERA portal among the 24 participants.

The fourth and final test was also carried out among users working at a data archive. They evaluated the final version of the portal using the same questionnaire as in the third test in order to make a comparison possible.

The observations and comments they made during the test made it clear that the MADIERA portal was offering the user community a very good and useful way of finding studies. There are many ways of finding a study as the data is indexed in several ways. The users were impressed by the short response time.The users likewise rather quickly discovered the many ways of sorting the result list, the importance of using the number of hits to be displayed and how you can have different outputs by using the 'mouse over' function on the study, section or variable labels.

In short: there is much information to be found within a very short time and the use of the thesaurus when searching in your own language was found to be very useful and impressive.

6. Specification and development of new functionality

Input for the functional specification came from the user input and also the project team's extensive experience with the development of earlier Nesstar products. The functional specification was developed through different stages focusing on the technical development to be developed through the project. One of the original ideas of the MADIERA project was to develop some specific new technology to search for and locate data. Since the overarching aim is to develop a system that makes it easier to carry out comparative research, two specific problems were singled out to be added to standard search and browse technology:

- Specification and implementation of a geo-referencing system for social science data, to allow geographically based search for and location of resources.
- Development of a methodology for identification of comparable data.

In addition to that it was set as a goal to investigate and try to develop a standardized.

- Naming and identification system for social science data published on the net.

This was regarded as a potentially very important requirement for "the hyperlinked information space" concept that the project aimed to fill with content.

6.1. Implementation of a geo-referencing system

Most social science data have a spatial reference; in some way the data are located in a space. This spatial reference will often be of interest to researchers, either to find data or to judge their relevance. Such a spatial reference may be recorded as the name of an unit, a statistical code for a unit or one or more (x,y) coordinates. And social science data are not all the same, the spatial reference could be given both for the dataset as such, i.e. *the coverage* of the dataset, or for the individual *units* of the dataset, the *location* of units. This means that the information could be stored in the pure documentation or be part of the data matrix as such. There are clear differences between aggregate statistics, where the units usually have identifying codes and often also name in the datafile, while sampled individual level data do not carry such information at the record level because of data protection, etc.

Using available spatial information, we might try to answer different types of questions:

- find the coverage of a data collection;

- find the geographic type of data;

- availability of (geographic) levels in the data;

- availability of data for a specific location (or level).

To look up data by spatial reference, we need to analyze the characteristics of the spatial references and develop a meaningful user interface. Which data in a well-described dataset is useful and necessary when we are trying to locate data by geographic location or coverage, and what are the characteristics of an interface that make this easy and intuitive understandable for a user. The three first points above are not part of our problem, they start from a given dataset. It is the fourth question that is of interest to us the way we delineate our task. This will involve both "horizontal" positioning in space and "vertical" level. Sometimes we are looking very specific at geographic location, sometimes we are looking for the levels that data are available for.

For MADIERA, DDI is the given metadata standard. If we focus on the *study* level, the DDI standard defines (implicitly) the *most important unit* (2.2.3.3) by singling out *country* as a close to mandatory piece of information, the *hierarchy* of geographic levels (2.2.3.4) and the *lowest unit* of interest (2.2.3.5).In addition a *specific position* is defined as the geographic bounding box (the smallest square box that covers the area) and/or the geographic bounding polygon (the smallest polygon). These last two positioning

means are defined in terms of x,y coordinates and allows for different ways of setting up coordinate systems.

To go further down beyond the study level to the data level could pose some practical problems, for individual level/sampled data it requires that we have the possibility to read through the file to pick up necessary data and that is not always possible, in particular because it could mean a violation of privacy/data protection. Such problems often means that the actual data are not published or made available over the net, only the metadata part is published.

The DDI elements described are of two distinctly different kinds. 2.2.3.3 Country is a very specific and discrete piece of information. The same goes for 2.2.3.6/7, the bounding box or bounding polygon. 2.2.3.4 geoCover however is defined so that the top level is specific (i.e. the country of Denmark) while the other subelements by the nature of its reference back to the top level become relative, i.e. are general levels, like county, municipality, areas, etc.

The DDI allow that the *coverage* of a dataset could be by several levels in a hierarchical system and each level can be linked to an external resource, i.e. a standard or system for coding. In addition we may have several variables pointing to different codings at the same level or that only one variable is used to specify several different levels, e.g. one variable may give both county and municipality in a concatenated code. The code for NUTS (Nomenclature of Territorial Units for Statistics/Nomenclature des Unités Territoriales Statistiques) units specifies 4 levels, country and NUTS 1, 2 and 3, which means that 4 levels are stored in one variable. This is usually the case with hierarchical codes and the attributes for the geoCover element reflect such complexity. If we list lower levels of a hierarchy it will to a large extent be useless information until we are able to point to variables in the datafile where the actual information is stored and to an external resource that explains the coding. Presently the available attributes specifies the *geographic type* (polygon, line, point) of the geographic element, the *vocabulary* or *standard* used in the codification and via geoRef points to the actual variable in the datafile. We see that the vocabulary only give the "headings", the general information about a level. We would have to go down and read the file to find the specific information. From the geoCover element we would know if the dataset holds information about towns, but we have to read the file to see if the town of Odense is represented.

The three types of spatial data representation, name, code or coordinate differ. These are three sides of the same coin that may function to supplement each other, summing up to a larger whole because each represents something unique. The *name* is discrete and

usually very specific, the *code* allows in addition for specification of levels and sequences and the *map* position units both absolute and relative to each other. In addition a map visualizes, if we don't know a name or code, or if we don't know the correct spelling, then we can select based on the map. A map is a more general help-tool while a list helps us go more specific.

At the data level, the DDI-element *4.3 var* is used to specify information about each variable in the dataset. The attribute *geog* is a yes|no attribute indicating whether the variable relays geographic information, while geoVocab records the coding scheme used in this variable.

Element 4.3.23 geoMap is used to point (using a URI) to an external (coordinate based) map that may be used to display the geography in question. The *levelno* attribute indicates the level of the geographic hierarchy relayed in the map while the *mapformat* attribute indicates the format of the map. It is possible to store many geographic levels in one file, and the more complicated we make a *mapformat*, the richer and more complicated we can make the match between levels in the data files and levels in the map. For the MADIERA project we have concluded that we would not gain much by developing these kinds of complexities. Contrary, for our limited data location purposes we decided to stay at the study level and disregard the information at the file level. Researchers interested in finding data will most likely do well with the information available at the study level.

MADIERA users access this information via an interface. With well-documented data there is a potential to make such an interface very sophisticated. It could be a table/menu or simpler than that: an outline map. A table/menu possibility could be developed more or less dynamically if the geoCover element points to external resources, i.e. gazetteers or controlled vocabularies. In the ultimate version of a map possibility, we could have a top level where we zoom in or select part of the world, Americas vs Europe, etc

It is possible to search for names/a text (i.e. Nation or geoCover (highest level) from a menu based on a gazetteer like ISO 3166 (the standard for naming of countries).However, this does not give more than what already is in Nesstar, it will not make use of the specific geographic starting point. And cross-national gazetteers are not easy to find. NUTS might be regarded as one, but represents a limited area, ISO 3166 (The standard for coding countries) is another.

To make use of hierarchical coding systems it probably has to be in combination with a map or carthographic picture, not for technical reasons but because a map-image conveys more intuitive information.

To visualize as more than a picture we started from a coordinate file, a map outlining the NUTS system. To generalize the procedure the coordinates just have to be standardized to decimal latitude/longitude coding to be used for larger areas.

If we as a specific example think of an interview in a Eurobarometer, carried out in Gent, Belgium (and we add an imagined geographic reference variable V005), then the meta-data would look approximately as follows:

2.2.3.3 Nation	Belgium
2.2.3.4 geoCover	Belgium,
(Vlaams Gewest)	NUTS1(geoType=polygon,geoClass=NUTS,geoRef=V005)
(Prov.Oost-Vlaandern)	NUTS2(geoType=polygon,geoClass=NUTS,geoRef=V005)
(Arrondisement Gent)	NUTS3(geoType=polygon,geoClass=NUTS,geoRef=V005)
geoClass could be expressed as a URI	
V005 will have the value "BE234" for this unit	
2.3.3.5 geogUnit	NUTS3
geoBndBox	
2.3.3.6.1 westBL	West Bounding Longitude Value
2.3.3.6.2 eastBL	East Bounding Longitude Value
2.3.3.6.3 southBL	South Bounding Longitude Value
2.3.3.6.4 northBL	North Bounding Longitude Value

boundPoly	
2.3.3.7.1 polygon1	(point1, point2, point3, point4+)
point1	
gringLat	
gringLon	
point2……	
var	(geog Y,geoVocab=NUTS)
4.3.24	geoMap (…URI=http://…, mapformat=..., levelno=NUTS3)

Within the MADIERA project we tailored this for resource location purposes:

- availability of data resources for a specific geographic data item, i.e. a point, a string or a polygon (Data for Belgium, or for Gent, Belgium);

- availability of data at a specific geographic level (Data at the NUTS3 level).As the discussion of the complexities of data has shown, it seemed over-ambitious within the MADIERA limits of time and resources to try to build a demonstrator for a complete procedure. The problem of *levels* is present in several of the DDI elements. Some of the information needed to develop a search procedure is available directly as part of the metadata while other important parts are only available indirectly from external resources. These external resources have to be produced and maintained. A full-fledged procedure also have to mix use of pure geographic information (which may be static and more easy to set up for demonstration purposes) and use of other types of information, standards for specification of names, codes and levels (which may be more complicated to develop).If geography is recorded as variables in the data files,such variables are coded according to standards that have to be available and actionable in a machine-readable version. A further factor of practical importance is that for social science data this have to be coupled with a publishing procedure, because presently we have fairly little data already published that make use of mapping and coordinates to establish the location in space.

By the term *external resource* is meant that whatever is shared by multiple datasets and is in a common structure should be kept in separate files outside the actual metadata of a file and only referenced or invoked when needed. That would be the best solution for consistent maintenance. It would also be the best solution for development of a generic

procedure, a procedure that is expandable through external initiatives. The geoClass specification given in element 2.2.3.4, the geoVocab under 4.3 and geoMap under 4.3.24 are all examples of external resources that may potentially be invoked for use. However, it is way beyond the potential of a project like MADIERA to actually *develop* and *maintain* those within the present time and resources limits. But they will in the future be needed for an elaborated and flexible procedure.

In MADIERA we focused on dataset "coverage". Although the aim was a rich but generic procedure, we in this first phase disregarded the potential lower levels of coverage recorded in the meta-data. This was caused by the fact that the geoCover element have to make extensive use of the geoClass attribute, and a more sophisticated procedure have to make use of this attribute. We also focused on the pure geographic search, disregarding the alternative possibilities presented by free text search against designated gazetteers or search or look-up by codes against vocabularies or standards. This was again caused by the fact that the geoClass and geoVocab attributes have to present the data needed, and within the limits of the MADIERA project we could only describe the potential represented by external datasets of that kind.

The ambition was to develop a "demonstrator" based on a reasonable background map. The examples developed focus on Europe down to NUTS level 2 to be able to demonstrate the potential of the procedure.

The data in the elements 2.2.3.6 geoBndBox and 2.2.3.7 boundPoly elements are intended for geographic search. It could be possible to search both for polygons that only partly overlap with a drawn rectangle and for polygons that lies totally within. This can be regulated as a user option.

A geographic search requires a background map for the search interface. Such a map could be rather crude, i.e. Europe at the country level. It could as well be rather detailed, outlining units at a low level of aggregation. Or it could be that the map in the interface was possible to navigate, with possibilities to zoom in and zoom out. Europe by country would as a first start look like the illustration below. In such a map it is possible to mark areas of interest in at least two different ways;

- clicking on a country (a polygon) or

- drawing a rectangle.

Clicking on a polygon would be to send *one point* (one set of x,y coordinates) to the search procedure. This is the same as drawing a rectangle of no or zero extension, which have some far-reaching consequences. A geographic search procedure could search for

such a single point, against the DDI-elements geoBndBox and boundPoly. The problem is that there will be hits for all datasets with geographic coverage at this *or higher levels*. With only one point indicating Switzerland, we will find all datasets with coverage Switzerland, but also all datasets with coverage Europe, etc, because there is no delimitation of the search. Such a technique can only be used to exclude datasets with geographic coverage at the same or lower levels of a hierarchy. It is possible to expand on this way of marking a single point or a "no-extension" polygon to search for data, but then we have to employ additional information. Since we do not know the extension of the spatial search, we could for instance use identification code or unit name to exactly identify spatial units of interest. Such an adjustment requires that identification code or name is available along with the coordinates, as an identifier in the file with the map. Clicking on a country (Belgium) would send one identification code (BE) along with one set of x,y coordinates to a search procedure that searches against the appropriate elements in the metadata of a study, the geoBndBox, boundPoly or the geoClass external resource. This is a likely further step for refinement of the procedures presented here.

Figure 10. Performing a spatial search

A simpler way of supplying a spatial search procedure with the demarcation needed to get a more precise search is to draw a rectangle in the map. Then we send two (x,y) points to the search procedure, i.e. the upper left hand corner and the lower right hand corner. This facilitates a precise search. We may think of two varieties of this procedure:

- only search for spatial units with coverage totally within the limits of the rectangle;
- search for spatial units that overlaps with the rectangle

As a first step MADIERA singled out point 1. However, since many spatial units may have irregular shapes, it could be a reasonable extension to say that we look for units where a certain percentage of the surface is within, let's say that 90% of the surface should be within.

The concept "comparative dataset" is not yet defined in the DDI. However, since it is possible to have more than one boundPoly element specified, that may be regarded as a

preliminary and very simple definition of a comparative dataset. Defining the extension of a spatial search via the drawing of a rectangle give us a crude possibility also to look up such studies. But because of all potential "noise", a procedure that intends to find comparative studies probably is more in need of further refinements than a procedure developed to look up single units.

The map of countries in Europe is now used as the starting point for both the search and a publishing procedure. In the data-publishing situation, clicking on the appropriate country could insert the needed coordinates directly into the geoBndBox and the boundPoly sub elements. If this was a comparative dataset, it should be possible to click on more than one country (select more than one unit) and separate boundPoly polygons should be generated. This last point tells us that if at all needed, the data in the geoBndBox element should be generated automatically on the basis of the polygons, if this was the preferred procedure.

Although the publishing procedure is outside the immediate concerns here, it is important to stress that a map-based interface should not be the only way of inserting coordinates into the metadata elements GeoBndBox or boundPoly during publishing. Often it will be more convenient to look up such data for spatial units indirectly via a name list or a code system.

Figure 11. A coordinate point in a map

If we drill down a map with a somewhat finer granularity, it will immediately create the impression that the lower level units have some importance, which usually would mean that there are data available for these units. If the unit of interest is the Vlaams Gewest region in Belgium (at NUTS1 level) then the top level coverage may either be Belgium or Vlaams Gewest. If a user of this technology want to search for data for this region and draws a rectangle as shown, he will find units with geographic coverage for this region only, and not find datasets with geographic coverage Belgium that holds data for all three regions in Belgium.That might be counter-intuitive and may be a candidate for further refinements when time and resources allow.

Refinements could either start from a search with no extension (a point), which would find datasets for both Vlaams Gewest and Belgium. The coordinate point has to be identified by a unit *name* or a *code* (here BE2) that would be unique, and also establish the level. For datasets with coverage Belgium, the geoClass attribute would tell if data are available at the level of region (NUTS1) and the specific unit Vlaams Gewest.

It is an alternative, but probably more complicated procedure to just use the rectangle in the map to pick up the identifications of units and then go about searching dataset metadata for names or levels.

On the basis of the above discussion, the project have implemented:

A. A coordinate-based spatial search.

This has been implemented as a feature of the MADIERA portal. The available first version of the demonstrator gives a user access to a map version of the NUTS system of statistical units, as a 3-level hierarchy (down to level 3 of the NUTS system). The user may mark an area of interest on the image of the map. The coordinates of the chosen area of interest will then be shipped to the search-module of the portal as search-parameters. These parameters can be used in isolation, or in combination with regular text-search parameters. Because the visual display is based on coordinates, the map-

interface does have zoom and pan-capabilities, in addition to support of selection among available maps. The interface produces the list of available maps by connection to a OGC-compliant web map server (WMS). The WMS also performs the actual rendering of the map coordinates onto a relevant image format (png and jpg are supported).

B. The Portal's search-module treats searches for coordinates in a parallel way as traditional text-searches.

This extension is vital for the system to work flexibly on different types of searches, because this is the component that retrieves the matching documents from the underlying servers and presents the portal with them. This open the possibility that location of data resources by strict spatial positioning, by picking codes or names from a menu/gazetteer or by searching the substantive content can be mixed freely by a user.

To establish a better map-basis for the spatial positioning and a realistic system of spatial units for a menu/gazetteer procedure, we have negotiated a contract with Eurographics, which allows us to use the material they have developed for the updated NUTS overview of 35 European countries. This means that we get both the names, codes and coordinates down to level 3 or comparable for every country. The NUTS system is regarded as the most relevant nomenclature and frame of reference for such a practical procedure.

In the social sciences there is little tradition to document geographic location of data resources. Because of that it has become obvious that the project also have to suggest realistic publishing procedures for this type of content. This may be done in various ways, to some degree linked with the best-practices work on content provision outlined by WP6 and by expanding the dedicated Nesstar publishing tool. If we start from the NUTS nomenclature, we have a system of identifying codes and names at different levels of aggregation, linked to coordinates that outline the unit borders. This may be expanded with a standardized bounding box (i.e. the upper left hand corner and the lower right hand corner) and a (set of) standardized bounding polygons of every unit. In the data publishing process specification of geographic coverage could then be accomplished by selecting the unit code or name from a controlled vocabulary/menu or through clicking in a visualized map. Then several elements of the data documentation can be filled with the relevant information automatically, if we decide on geographic coverage, we get bounding box and bounding polygon for free.

The actual development of this publishing possibility is not regarded as an integrated part of the MADIERA project.

6.2. Identification of comparable data

Comparison is a relational and relative concept, it requires that there is a baseline defined and we are looking for other items that may be compared to that baseline. Usually that means an extension of an analytical dimension. We are either working from one *data resource* or some specific *analytic task*, and we are looking for other data resources or other pieces of information of a *similar kind* or *similar content*, to extend a *dimension* or *expand on the content* of our analyses.

The MADIERA intention was to investigate possibilities to supplement a data analytic process with a practical but useful and realistic procedure to look up relevant data while in the middle of the process itself. It is not to go out and search for comparative data in the traditional meaning of the term. This focus on the data-analytic process will have consequences for the order of priority between types of elements. The explicit prerequisite is data resources described according to the DDI metadata standard.

In the practical research situation we distinguish between two stages. In the first stage we try to locate information resources or studies of interest for a problem area. Under the MADIERA portal we may think in terms of at least three ways of establishing this *baseline*, or a first phase towards the starting point of the actual analytic work.

a) a search;

b) a "drill-down approach";

c) through a geographic search/delineation.

Metadata or documentation of the substantive content of a data resource is rarely very precisely specified, it is intended basically for human interpretation, and much of the documentation work at this level is done by archive professionals, not the people that originally designed the data collection instruments. In the language of the database community there is heterogeneity, semantic mismatch between schemas. This may include differences between national languages. The practical matching problems encountered require some possibilities to harmonize inconsistencies. This generates a need for structuring in the documentation-/publication process. The more order there is introduced through the documentation process across potential resource publishers, the better the practical possibilities to compare. Classifications are an important ingredient on which all statistical systems build, classifications are difficult to develop for broad substantive content however. The closest we come to a generalization of the classification concept for fairly general substantive content of a study is a common *thesaurus*. It is more relevant to think in terms of a thesaurus than in terms of an

ontology (defined as an explicit specification of a conceptualisation), as a thesaurus is a more general tool covering a broader part of a scientific field. In the NESSTAR publication tool (Nesstar Publisher) the ELSST thesaurus is used as such a structuring and defining tool to insert *keywords* or *concepts* at study (add keywords to abstracts), variable group and variable level (concepts). An additional kind of structuring is reached through controlled vocabularies and a common topical classification of studies, available through the common template for documentation work. But controlled vocabularies are typically used only where there is a specified and fairly short list of options.

One MADIERA conclusion is that the data-publishing instrument should employ the same external resources (thesaurus, controlled vocabularies, methodological categories) through the data documentation and publication process as the MADIERA portal should have available for the resource location process.

When searching, there are two possible types of hit lists, either a list of *studies* or a set of *variables.*

1) If the starting point/baseline is a dataset/study, we can look for comparable data on many elements of the meta-information given for the dataset, with an equal focus on information elements at study level and at variable level, but variables as such are not explicitly defined as the information focus.

2) If the starting point/baseline is in the middle of some analytic work, then it is more likely thatsome very specific piece of information will be the starting point to look up comparable data and then our focus would be more specifically at the variable level.

We may distinguish these two scenarios as a top-down and a bottom-up scenario. For MADIERA we have concluded that it is most relevant to start from the bottom-up version. This is a perspective where we work from a specific analytic situation, and the most immediate problem is to start from the variables of the actual present analysis being carried out. If we look at a univariate frequency listing, then we are interested in additional examples of this same variable or the *concept* or *topic* represented by the variable, in other datasets, at different time-points, for other countries, etcIf our analysis happens to be a crosstabulation, then we are interested in comparable tables, that is crosstables using the same variables, etc

Most methodological and "practical" information on a dataset is located at the dataset level, in section 2 of the DDI DTD. The substantive information is more complex, and is found both at study- (subject/abstract) and variable group- and variable level, and will to

some degree be of relevance at all three levels. At the variable group level it will be possible to organise variables under meaningfully labled subjects. The substantive information may be expected to be less precise and it is more difficult to establish procedures and criteria to measure comparability. Methodological information is possible to specify as controlled vocabularies, and it is important that common use of vocabularies through a common templates is advocated.To employ ELSST to look up synonyms or alternative search criteria is relevant for substantive oriented searches, for methodological searches it is more a question of matching on elements of controlled vocabularies.

In MADIERA we initiate a search for comparable data from very specific information and ask for:

- additional examples of the same specific information, or

- additional examples of the same specific information where we are invoking more general information as a criterion, similar table for the same topic, the same topical group, the same or different universes, timepoints, etc

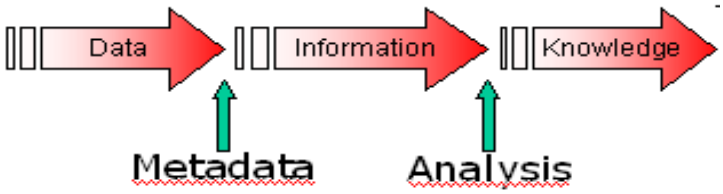
To develop a useful procedure, it is necessary to concentrate on the *concept* or *keyword* elements in the documentation of the datasets. The data archives have tested the publishing procedure implemented in Nesstar Publisher. The archives tagged up concepts at variable and variable group level for sets of studies and used the Nesstar Publisher version with the thesaurus ELSST included. The publishing software makes an analysis of the textual documentation linked to a variable, usually the question text, and invoking the vocabularies of the thesaurus, suggests potential keywords to concentrate and standardize the content-carrying part of the documentation. The conclusion is that this is an efficient, rational and more than anything, it is a standardized way of structuring the substantive content of data resources.

Whenever we are exploring/analyzing data resources, it will be possible to check or look up keywords connected with specific variables. By clicking directly on the keyword, a second search will be performed, with the keyword as search term.

6.3. Naming/identifying data resources made available for empirical research

Metadata converts data into information and analysis converts information into knowledge. However, data may be used many times, and there is a need that metadata be expanded by information about data use.

Figure 12. Metadata converts data into information and analysis converts information into knowledge



One aim is to develop interactive knowledge products through inclusions of live tables and graphs into publications so that the reader is able to interact with the tables and use them as an entry point to the underlying data and the contextual information for any analysis. Using such functionality, a reader should be able to re-run an original analysis of the author or to carry out a comparative analysis using alternative sources (for example from the same or a different country). It should be possible to add comments or results from alternative analyses to the metadata of a study. This requires a technical framework that allows

1) Hyperlinks from the metadata of a study to reports and publications displaying analytical results based on the data. Data should point to use.

2) For communication purposes, the current e-mail/web addresses to relevant researchers, support staff, departments etc. should be part of the metadata supporting a data resource. Data could point to (potential) users.

If we regard documented data resources as stable entities that are described once and for all, then this would not be technically difficult to do; we might just fill our documentation with hyperlinks. However, this will among other things soon present us with an update requirement. Production and publication of data onto the web comes before use, and the problem becomes how to link data and its use as a dynamic entity on the web. In the MADIERA context we need to create a feedback system to the body of metadata allowing a user to add to the collective memory of a data source, creating the notion of data resources as dynamic entities or dynamic repositories that keep a log of its own use.

The problem is to record and accumulate the dynamic information generated in the data use process, link it and make it relative to an identifiable starting point.Users should be allowed to feed their experiences back into the repository made up by the study. This is a technical, identification and an authorization problem. Who should be allowed to do what, and who has the authority or "ownership" to data once they are published to the web?

Such a perspective leads to a greatly *extended* metadata concept where not only descriptions of the data are relevant information, but also various types of knowledge products deriving from their use. It is implying a *dynamic* concept where metadata is seen as a collection of information that is developed and enriched all the way through the life cycle of the dataset and not something that can be created and published once and for all. The perspective is leading 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 could be provided by others as an ongoing activity lasting for many years after the data themselves have left the production line, and the use of the data becomes an element that stimulates further use and creates wider relevance.

Standards are essential for the functionality of MADIERA. To further develop the dynamic data concept, first and foremost a naming and identification recommendation for social science data resources is necessary. Without a consistent naming system that gives us better possibilities to identify resources a dynamic data resource concept may be difficult to develop.

There is, at the moment, no naming or identification system for data that is comparable to the systems used for written publications. For instance: How should identical versions of the same dataset, stored at two different data archives, be named and identified? And how should slightly different versions of the same dataset be named and identified independent of where the data are stored? Is it at all necessary that datasets be uniquely identified? Datasets documented under DDI carries along a very elaborate description of themselves internally as part of the metadata. For what purposes do they need to be uniquely identified? Is there a distinction between technical requirements and more substantive data use? Presently there are no answers to these questions, so it was set as one of the tasks of the MADIERA project.

The technological platform of MADIERA is well suited to support this dynamic growth of a hyperlinked information space. By hyperlinked information space is meant a framework for bringing live data into online texts, as well as linking on-line scientific texts into the metadata body of a data material. The former is achieved by traditional hyperlinks and

bookmarking technology where specific tables or other analytic results in published material on the web carries a hyperlink back to the original dataset used. If this link is activated, the dataset is automatically opened for use with the correct variables, reproducing the analysis used as the starting point and immediately being ready for further analytic processing. The other problem, using a data resource as a repository that collects and stores information about its own use could benefit from a naming convention, which makes it possible to name, bookmark and hyperlink all relevant resources in the MADIERA repositories. At least that was the assumption of the MADIERA application.However, we see that both scenarios imply that we are going backwards from a knowledge product to a data resource, and establishing and recording the connection between the two components is not the major problem. The real difference will be that in the first situation we store the bookmark with the knowledge product, while in the second situation we store the link as a reference with the data resource, as part of the metadata. The identification need has something to do with the implicit need we have to find the latest or richest or best or most correct version of a dataset, be it if we start from the data as such and are looking for use, or if we return to the data from a use situation. With a modular DDI 3.0, we need a mechanism that guarantees that we go back to the originally used version.

To uniquely identify a specific data resource published on the web is complex. We have to distinguish the starting-point, the original published data from the process of dynamic changes over time, there is an identifiable first version of a data resource and then comes a process that needs to be versioned. A data resource however consist of both data and metadata, these two components are rarely produced at the same time and in the same place, giving further problems for how to identify. If we think of the data production as the starting point, then there will naturally be a very long metadata development history to record and version. Metadata describing a study is in the upcoming DDI version 3.0 split over several modules, the "data" as such tend to stay much more stable than the metadata, which is the real dynamic component.

Data archives cannot just copy the systems used in the library world, the ISBN system. Archives and libraries are not the same and research data are stores in archives. Libraries identify unique objects, while archives deal with the unique and its relationships. The concern is not only in identifying the object but placing it in context. Archives attempt to record the intellectual, physical, and temporal provenance of the object.

DDI version 2.0 was focused on the library concept. It thinks in terms of single objects, a data file, and creates the ultimate bibliographic record for that object. If, and only if,

there is an accepted title for the source document, and anyone creating an XML instance at another location for the same file uses the same source and cites it accurately, can anyone tell if instance A and instance B are related. There is presently no effective means of writing a "generic" XML instance that can be used on multiple data files of matching structure and differing coverage.This problem, created by the hierarchical nature of the source, study and instance elements will come up in several varieties, where the potentially different branching developments of studies may be the most immediately relevant for MADIERA.

In the archive, the idea of a study is based on the life-cycle model up through the production of data products and their record subsets. A study as a data collection process consisting of the conception of the study, its description, the data collection instrument and process, data cleaning, data entry and storage, and the products resulting from the collection of the data. A study may be part of a series, for example the Eurobarometer series. While all of these parts may not be included in specific DDI or other metadata structured formats, the metadata standard should ideally provide an option for identifying and pointing to them regardless of their format. The DDI aims at a structured description for materials that are needed for machine processing of the search process, data identification, data selection and data manipulation.

The idea of a DDI instance is based on archival/library holdings and the requirements of data discovery and processing. Under version 2.0 of DDI, an instance is a file that has to be republished entirely when there is a new version, it would be possible to develop a version or edition identification system. A version identifier however would carry little or no information on the nature of a change/update. Under the proposed version 3.0 a DDI instance consists at the most general level of an instance module, which lists the related modules for a particular occurrence of a datafile or dataset. This wrapper may refer to modules that are commonly held between two or more datafiles/sets and what is enclosed may vary by institution and/or system. An instance module is specific and product based. It can contain multiple physical data file descriptions that are part of the same product. It could contain only one physical file of a given multi-file product if this is what is in the holdings of a particular library/archive. It should not contain more than one study. For processing purposes an archive/library may choose to create separate instances for each physical data file, but is not required to do so. The flexibility created by the modularity open possibilities to record the nature of changes much more accurately.

In DDI 3.0 there is a logical hierarchy where the backbone goes from the general instance module via a StudyUnit module (that operates at a conceptual level), a

DataCollection module (instruments and processes), a LogicalStructure module (content and structure), a PhysicalStructure module (data layout) and a PhysicalInstance (actual file description) module. In addition there are several other supplementary modules, one of which is the Archive module that gives the details of the organization holding the specific instance.

The overall instance and each module carries an identifier, with a version number and a version log. In total, the identifier consists of three parts, the identifier, versioning information, and authority. If archive ZA copies a study (Eurobarometer 60) to DDA, ZA was the authority. An archive can only version what it owns.

The DDI instance sent to DDA would minimum include a StudyUnit, a DataCollection, a LogicalStructure, a DataStructure, and a PhysicalInstance module, each with the ZA identifier. If DDA publish the instance as a DDA study this would most likely mean that they rename the data file itself to reflect the DDA numbering system and its location. The authority now shifts to DDA with the DDA authority code in the identifier. Since the individual modules also have identifiers of the same format, the instance has by inclusion the original StudyUnit, DataCollection etc...but with the DDA PhysicalInstance module which includes the original ZA PhysicalInstance PLUS the differences.

There may be many varieties on this topic, that may be solved along the same simple lines. An instance is a configuration of modules and the instance module at the top stores this map. The original conceptualisation and data collection description is regarded as persistent, not to any significant degree influenced by later changes. It is the PhysicalInstance that usually are modified or re-stored in other locations.

If we pick up the components we have distinguished so far, we have the following:

Although it is difficult to establish a data versioning authority, we need to identify some origin/original resource, where the organised identification starts with the publishing process, not the actual data production process. Everything starts somewhere, and of course there is an initial data collection, a very first step before any branching. This original data-component has to be uniquely identified, when it is put on the web. The practical situation will be that such starting points are not absolute, that data in some way could be derived from other sources by different data producers and could potentially be linked backwards into prehistoric times. This may mean that a data collection process give rise to more than one object on the web, without being linked into the same study. This would probably be fairly rare for survey type data, while there are more examples from official published statistics where this is the model, in particular historical data.

However, data resources are stored in data-archives of many varieties and data archives tend to put their own stamp on the products on their shelves, basically because they often contribute the systematic metadata and publish the collections on Internet. If the starting point is a resource archived at NSD, we would identify it as NSD…. even if NSD is only the creator of the metadata part. If it is a resource archived at UKDA, it could be UKDA…., and viceversaFor a DDI instance this means to establish a metadata authority and an identifiable starting point for a new data collection *on the web*. We know that Eurobarometers are collected by one authority (EU) which will continue to hold the copyright and the IPR to the data, although the data are made available for research by another institution (ZA). The whole collection of metadata will document this. But ZA is the prime authority for the web version. In addition the data are distributed to many national archives that may tailor the metadata to their national user communities. This would usually add more XML instances on the web, and the original study identifier is logged in the identifier system to link them. In reality, the dynamic data resource often starts to build up after there has been some initial branching from the original data collection. The original in such a scheme is identified as the ZA-version, the first one published to the internet, and later versions point back to that.

But, whatever the identification of the origin, it is supported with one more piece of information, that describes the history of a data resource, the versioning.

Some changes are *minor*, they do not affect the use of the associated data.

Other changes are *major*, significant additions or alterations that affect the use of the data or the analytical results.

We now have 3 logical components:

- original publisher, metadata producer, the "owner", the authority;
- the instance, the "wrapper"-defined view of the actual data resource we access on the internet;
- versions, the recorded history of the instance, as a configuration of module status.

There are some problems left here, but with the low constraints on publishing on the internet it is difficult to come around them. In the above scheme we might think of ZA4321 as a Eurobarometer study, with a specific instance, 1.0 If the greek archive make a greek translation, that will usually be published to a different server, under a new authority and becomes a second instance because it is the practical authority, not the study that has changed. However, the versioning of the greek version would point back

to the ZA version and there is the possibility to distinguish. When further instances come along in Finland or Spain we have a possibility to go upwards, and if applications are good enough we could even go sideways

Within an instance both the pure data and the metadata could have its own identity. The data is identified in the physical instance module which is linked on a 1:1 basis to a physical structure module which in term is linked on a 1:1 basis to a logical structure module. This will uniquely link the data file to its original concept and identifier. If the data is stored in an XML structure which carries metadata within it, the link may be bi-directional. The data is "versioned" when it is corrected. Dynamic data that is updated through the addition of records but no change in logical structure should retain its base number with a subversion update that signals that the new version is triggered by a *data* update. This could be as direct as adding a data creation date to differentiate it from a different cause for versioning.

Table 1. Examples of Levels of Versioning

Module	Source	Type of change
Group	Originator	Addition of a new iteration in the series; correction of information content
Concept	Originator	Correction of information content; draft development
Data Collection	Originator	Correction of information content; draft development
Logical Structure	Originator/Archive	CORRECTION OF INFORMATION CONTENT; NEW OUTPUT PRODUCT ADDED
Physical Structure (of logical record)	Originator/Archive	Correction of information content; new physical structure added/subtracted
Physical Structure (file coverage)	Originator/Archive	Correction of information content; new sub-set of file created
Archive	Archive	Correction of information content; new modules added; collection information added/subtracted;
Wrapper	Archive	This structure implied a wrapper which indicates the modules contained in the instance. The version would change with each change in content

When using data from a Nesstar server, it is at present possible to create both client-side and server-side bookmarks. The latter could as well be regarded as a private workspace on the server for every user and every data instance where it is possible for a user to

store comments or identifiable bookmarks to actions on a specific data resource. Such a workspace is private in the meaning that it is password protected.

To allow users to add comments and hyperlinks to datasets on a server it is possible to set up an open version of this technology, along with every dataset there could be a "notepad", where users leave comments and hyperlinks.

Use of this functionality will in the first version be controlled by the following practical arrangements: The data publisher has an editorial right to edit whatever is put as comments with a dataset and access to datasets are controlled via the specific access rules and user registration defined by a data publisher. This means that a publisher usually has control over users possibilities to access data resources and can deny such access if systems are misused.

7. The MADIERA portal

The MADIERA portal provides access to almost 3000 studies at three different levels: study, sections and variables. The portal can presently be found via www.madiera.net

The portal has several functions that are crucial to linking European data resources and providing unified access to social science data archives:

- provide a Yahoo-style overview of the data resources of the entire network (using the MADIERA classification system to organize the resources);
- provide a home for the MADIERA multilingual thesaurus;
- provide a central metadata index that will support more efficient Google-style searching across servers/archives;
- provide a MADIERA registry service whereby new servers/archives can be dynamically added to the network (and old servers taken down without effecting the performance of the system).

The MADIERA portal consists of several building blocks:

- a metadata standard (DDI);
- a technological platform (Nesstar);
- a multilingual thesaurus (ELSST);
- a classification system for studies;

- a georeferencing system for social science data;

- a methodology for comparable data;

- a technologyfor linking data and knowledge products.

The portal is implemented as a metadata harvester that automatically will upload metadata from the individual servers of the MADIERA network through the standard Nesstar API. The metadata are added to a central index that will provide lightening fast searching across a high number of servers. The search facility is powered by the MADIERA thesaurus to increase precision.

The portal's main technical features are that it makes available its full functionality through a HTTP/REST interface that can be easily accessed from any computer language. In addition it has a streamlined and efficient internal architecture. It provides a highly customisable user interface.

The main use cases supported by the MADIERA portal are finding studies published by any of the participating data archives by:

- browsing the multilingual CESSDA classification or the multilingual ELSST thesaurus;

- searching using a flexible search language that supports free text or field search, logical connectives, fuzzy and stem searches;

- finding studies relative to a certain geographical area using a graphical interface (based on the European NUTS classification).

Additionally it is possible to

- browse the multilingual thesaurus structure (synonyms, related terms, and equivalent terms in other languages);

- view keywords associated with a study automatically translated in any of the supported languages (currently nine);

- capture terms used in user searches and not included in the supported thesauri.

The portal is administered via a system with a web interface, which makes it possible to add new archives/servers, remove existing archives/servers and to refresh the harvested contents of an archive, i.e. the index.

7.1. To search for data in the MADIERA portal

The MADIERA portal provides searching across and access to social science data held in several European archives. The metadata that describes these data can be searched by several methods and at different levels of detail. In addition to free text search a user may choose from several pre-organised subject lists.

In the DDI, the content of a study is described at 3 different levels, study, section and variable. At the study level there is both a summary topical classification and a set of keywords. The section level and the single variables may be characterized by a *concept* or by *keywords*.

The portal provides a simple text box for user-entered search strings.

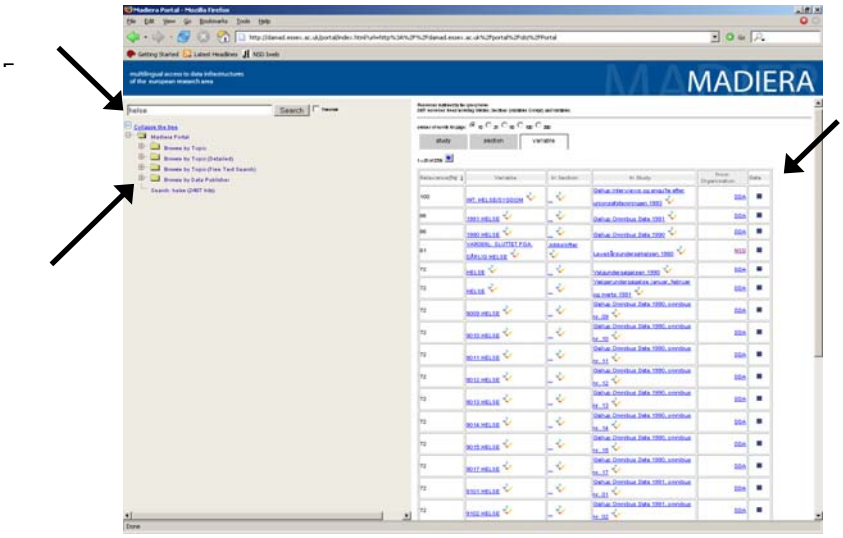
The default for the simple search is to find occurrences containing all the words entered across the whole metadata record.However, the search also supports exact phrase matching, Boolean and wildcard searches and can also be restricted to a selection of elements of the metadata.Help for these more advanced types of searches is provided in the portal. The search results are displayed at up to three levels; namely study, section and variable.

Such a search can be widened by ticking the translate check box before conducting a search. In this case the exact phrase or individual words, wherever possible, will be translated into terms from the other eight languages, using either ELSST or the CESSDA topic classification, and the search is performed using all languages.

Words that do not appear in ELSST are captured and can be displayed through an administration interface for possible inclusion into the thesaurus.

The browse by topic option employs the 2-level CESSDA topical classification scheme, which is displayed in each of the 9 languages. Clicking on an element from this listing in any language performs a search for that exact term across all 9 languages, but only searches against the specific study-level metadata element named"topic classification"

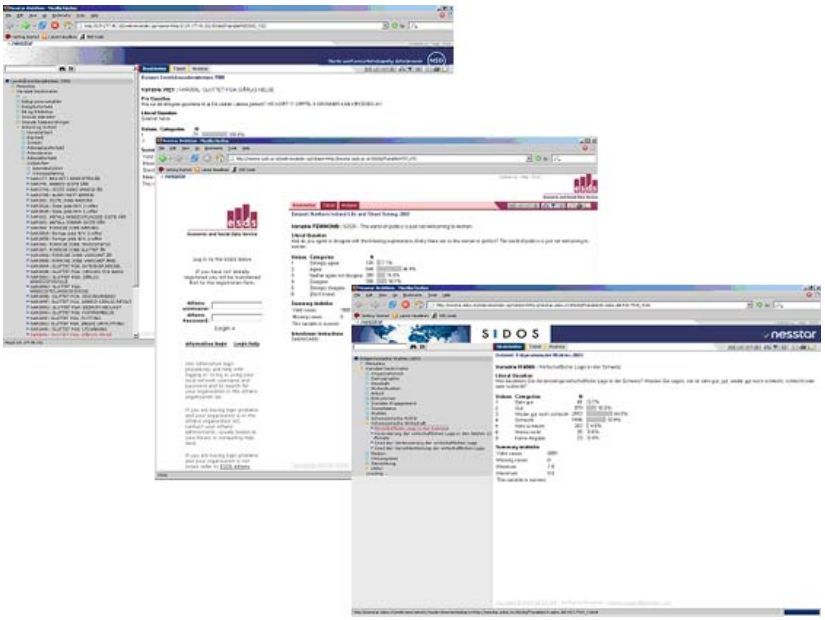
Figure 13. Performing a search in the portal



The topical classification tag is only defined for studies, not for sections or variables. Hence the search results are displayed only at study level. The more detailed subject browsing employs the whole keyword battery that ELSST represent, in any of the 9 languages. The search is conducted against concepts assigned at the study-level "keyword" element or the section-level or variable-level "concept" elements of the metadata. Clicking on a search element in the list in whatever language performs a search for that exact concept plus synonyms, in all languages, as defined in the thesaurus.As well as displaying the hits from the search, related terms from ELSST are listed as hyperlinks along with the number of hits that would result from a search carried out on these different concepts.

Clicking on an item from the hit list will open a data resource in a separate browser window. The data is accessed at the server where they are stored. Further exploration, analysis of the actual data, sub-setting or downloading can then occur. A user get access to the documentation right away and can browse all the metadata, but in order to access the actual data, the user need to log in. Access rules can vary between archives and over different types of data and the user have to apply for access permission from the respective archive

Figure 14. Data displayed at three different archives website



7.2. System architecture

The MADIERA portal is a Semantic Web portal that provides an easy-to-use, multilingual, single point of access to a wide array of high quality statistical datasets published by some of the major European social sciences data archives.

In particular this portal has the following characteristics:

- it makes available its full functionality through a REST interface that can be easily accessed from any computer language;
- it organizes information simply but effectively using a set of multilingual thesauri and classifications;
- it has a streamlined and efficient internal architecture;
- it provides a highly customisable user interface.

The social science data archives that are part of the MADIERA project publish their statistical holdings on the Semantic Web using Nesstar servers. These Nesstar servers operate on the same principle of traditional web servers; namely that 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. This object can be a complete dataset or a group of variables or single variable within a dataset. 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.

The MADIERA portal operates as a web search engine by browsing and querying the Nesstar 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 base 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 are then matched with a set of multi- and monolingual thesauri and classifications to illustrate different ways and technologies for searching for/locating data resources:

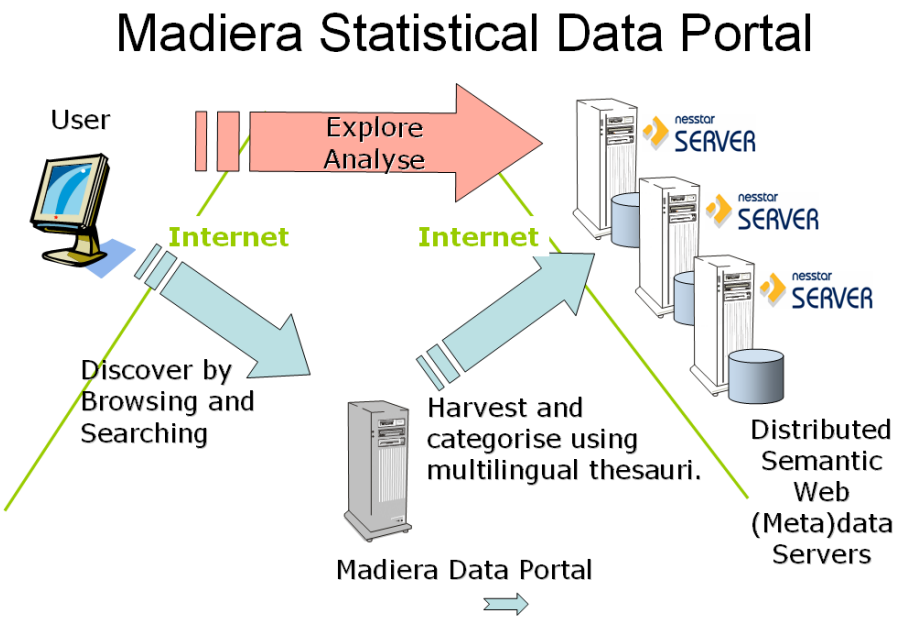
- currently the CESSDA classification of data sets;

- the European Language Social Science Thesaurus, both available in a number of European languages;

- and the NUTS system for classification of dataset geographic coverage).

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".

Figure 15. The architecture of the MADIERA portal



As with Google and Yahoo the user discovers the available "data" resources by browsing and searching indexes held on the MADIERA portal. Once discovered, the exploring and analysis of the actual data is performed via direct communication with the individual "data" resource.

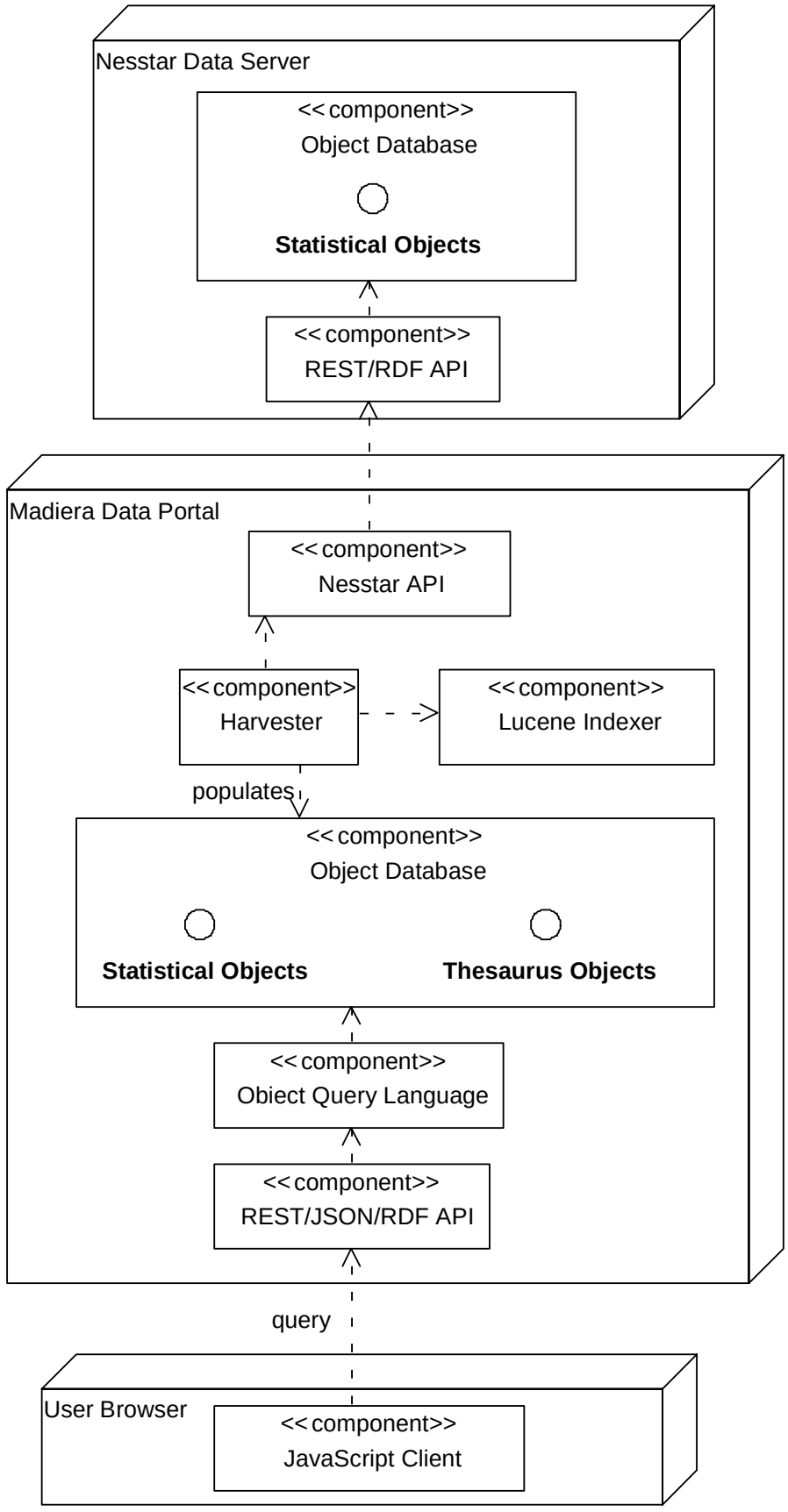
The object database that contains the statistical and thesauri objects can be queried using a flexible object-oriented query language. The query language has been optimised to efficiently support high-latency client-server applications by allowing more than one logical query to be performed at the same time.

The user clients query the portal using an HTTP/REST interface. The result of the queries is logically a set of subject-predicate-object triples that can be returned in either RDF or in JSON (a JavaScript-compatible data notation [3]) format. Given the extreme simplicity and standardized nature of the portal API we expect that it will be reused in wider applications (meta-portals) and that third parties will be able to quickly develop customized clients in their favourite computer language.

The default MADIERA client is web-based. Contrary to most web applications where the user interface is generated on the server side, the MADIERA client is written in HTML/JavaScript and executes completely in the user browser. This makes for a very responsive and highly customisable client. The complete separation of server and client reduces the complexity of the system and makes sure that the full functionality of the portal is available though the REST API.

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.

Figure 16. MADIERA portal system architecture



IV. CONCLUSIONS AND POLICY IMPLICATIONS

Throughout this document we have propagated the standpoints that data is the most important component necessary for good social science research. To foster god research, we need to ease the access to high quality data.

Another basic starting point is the belief that the promotion of a comparative perspective will be fruitful for the development of a European Research Area. We understand ourselves better when we can compare with others. Still there are many barriers to efficient comparative research, barriers that can be built down through the development of more efficient means and technologies.

The MADIERA project has in practices implemented the idea that this can be developed as a decentralized and distributed system. Through elaborate use of standards it is possible to develop decentralized open-ended systems, systems with the capacity to grow almost without limits. Behind this lies a strong confidence in the principles and ideas of the Semantic Web.

The MADIERA project has achieved its goals and met the expected outcome. The MADIERA portal, which effectively demonstrates all the goals of the project, can be found at via www.madiera.net and is a significant contribution to the European Research Area infrastructure for the social sciences and humanities.

The project has succeeded in developing an effective infrastructure for the European social science community by integrating data with other tools, resources and products of the research process. The MADIERA portal is a fully operational web-based infrastructure populated with a variety of data and resources from a selection of providers, it makes up a common integrated interface to the collective resources of several of the social science data archives in Europe. It has a formidable potential for expansion with the inclusion of new data-supplying points.

The MADIERA infrastructure has the capacity to grow and diversify after the initial construction period. It has reached the main objective which has been to create an open but sustainable system, nurtured by the collective energy of the data and knowledge producing communities of the European Research Area.

1. Results and transnational relevance

1.1. An integrated and effective distributed social science portal

The MADIERA portal provides access to almost 3000 studies from several European countries at three different levels: study, sections and variables. The studies cover most of the areas within the social sciences, among others, politics, labour and employment, culture, economics, social stratification, health etc. Using the portal, 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) she can then use the standard Nesstar web client to examine its complete metadata, apply statistical operations and download data.

The portal has several functions that are crucial to linking European data resources and providing unified access to social science data archives:

- provide a Yahoo-style overview of the data resources of the entire network (using the MADIERA classification system to organize the resources);

- provide a home for the MADIERA multilingual thesaurus;

- provide a central metadata index that will support more efficient Google-style searching across servers/archives;

- provide a MADIERA registry service whereby new servers/archives can be dynamically added to the network (and old servers taken down without effecting the performance of the system).

The fragmentation of the scientific information space has been a major obstacle for empirical comparative social research in Europe. The MADIERA infrastructure not only builds down many of the former barriers through innovative and effective new technology, but also promotes comparative and integrative perspectives.Where formerly data and derived knowledge products were kept distinctly apart, there may now be a much closer integration. This will underline the importance of data access, data use and critical reading. Access to data becomes a component in the development of democracy. The MADIERA infrastructure enables the development of a thoroughly comparative and cumulative research process that would be integrating and nurturing the entire European Research Area.

The portal uses the emerging information technologies to encourage communication, sharing and collaboration across spatially dispersed but scientifically related communities. The MADIERA infrastructure connects existing and well functioning providers of content and services and tries to meet the demands of their users. The portal can be seen as a Semantic Web extension of the ordinary web, where information is given well-defined meaning.

Data is not necessarily a scarce resource in Europe. However, there are large differences in the traditions and possibilities to make data available for scientific use. 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. Through making data generally more available and demonstrating the importance of data as basis fro development, the infrastructure developed in this project underscores the importance of data and empirical scientific work.

1.2. A multilingual thesaurus to break the language barriers

The portal offers a working implementation of the extensible and distributed multilingual web-based infrastructure for the European social science research community.

Social science archives stores and make available for secondary usedatasets from many studies by government and academia. The datasets are described by metadata, and to efficiently syntecize the substantive content the metadata standard make use of terms defined in a common multilingual thesaurus. The project has developed multilingual tools to support simpler user access to the data stored in different archives across Europe and to integrate it with data from other domains. This makes the information available as a common warehouse for all of Europe, for users to retrieve and communicate in the language(s) of their choice, as a basis for further research, policy making and planning by individuals, companies and government organisations

The thesaurus, the European Languages Social Science Thesaurus (ELSST) employed in the portal is translated into nine languages and contains 3,209 concepts.

To support the application of ELSST to the portal, guidelines for the use of the thesaurus have been written and a prototype of a common indexing management tool linked to the

thesaurus has been created. These developments provide the potential for data archives to index data resources consistently and homogeneously and permits researchers across the European Research Area to browse for and locate distributed resources in their native language.

1.3. The development of specific add-ons to existing virtual data library technologies

The project has developed data location technologies and participated in the development of a metadata standard for empirical scientific material. These developments recognize and directly address the need to facilitate cross-national social science and humanities research throughout the European Research Area. The project team has contributed its expertise in the field of metadata standards development through it work with the DDI committee, which is now about to release extensions to accommodate needs directly identified as part of this project, e.g. metadata for comparative data, for time-series data, for aggregate data and geographic data.

The widespread adoption of the DDI and the publishing of marked-up datasets available via the MADIERA portal will vastly improve access to a range of varied resources. Expanded use will greatly enhance comparative research; the ability to harmonize datasets over time and geography will lead to significant improvement in our understanding of societies. Increasing the availability of high-quality data is a way of increasing the importance of secondary analysis in the social sciences. For that to become a reality the high-quality data needs high-quality documentation to accompany it and high-quality resource discovery tools to locate it. The expansions of the DDI standard have to a considerable degree enhanced the potential of the MADIERA portal to handle complex organized data and aggregate statistics.

In order to encourage more comparative research in Europe, two functionalities have been added to Nesstar's standard search and browse technology:

- Specification and implementation of a geo-referencing system for social science data, to allow geographically based search for and location of resources.
- Development of a methodology for identification of comparable data.

1.3.1. Implementation of a geo-referencing system

a) A coordinate-based spatial search. This has been implemented as a feature of the MADIERA portal. The available first version of the demonstrator gives a user access to a map version of the NUTS system of statistical units, as a 3-level hierarchy (down to level 3 of the NUTS system). The user may mark an area of interest on the image of the map. The coordinates of the chosen area of interest will then be shipped to the search-module of the portal as search-parameters. These parameters can be used in isolation, or in combination with regular text-search parameters. Because the visual display is based on coordinates, the map-interface does have zoom and pan-capabilities, in addition to support of selection among available maps. The interface produces the list of available maps by connection to a OGC-compliant web map server (WMS). The WMS also performs the actual rendering of the map coordinates onto a relevant image format (png and jpg are supported).

b) The Portal's search-module treats searches for coordinates in a parallel way as traditional text-searches. This extension is vital for the system to work flexibly on different types of searches, because this is the component that retrieves the matching documents from the underlying servers and presents the portal with them. This open the possibility that location of data resources by strict spatial positioning, by picking codes or names from a menu/gazetteer or by searching the substantive content can be mixed freely by a user.

1.3.2. Identification of comparable data

In MADIERA we initiate a search for comparable data from very specific information and ask for:

- additional examples of the same specific information, or

- additional examples of the same specific information where we are invoking more general information as a criteria, similar table for the same topic, the same topical group, the same or different universes, timepoints, etc

To develop a useful procedure, it was necessary to concentrate on the *concept* or *keyword* elements in the documentation of the datasets. The data archives have tested the publishing procedure implemented in Nesstar Publisher. The archives tagged up concepts at variable and variable group level for sets of studies and used the Nesstar Publisher version with the thesaurus ELSST included. The publishing software makes an

analysis of the textual documentation linked to a variable, usually the question text, and invoking the vocabularies of the thesaurus, suggests potential keywords to concentrate and standardize the content-carrying part of the documentation. The conclusion is that this is an efficient, rational and more than anything, it is a standardized way of structuring the substantive content of data resources.

Whenever we are exploring/analyzing data resources, it will be possible to check or look up keywords connected with specific variables. By clicking directly on the keyword, a second search will be performed, with the keyword as search term.

1.4. An extensive program to add content, both at the data/information and knowledge levels.

Through several workshops an extensive user guides the project has carried out extensive training of data providers and users to inspire and encourage the continuous growth of the infrastructure developed tools and guides for the practical side of such work. The potential benefit to the data archiving community can now be realised as a result of the development and application of the MADIERA content publishing tool. This tool is accompanied by explanatory material which has been tried and tested both by those archives which contributed directly to the project and by sister archives within the CESSDA community. The work to date has enabled the publication of information relating to nearly 3000 datasets covering most areas within the social sciences and as a result of this activity, the building blocks are now in place to enable researchers within the European Research Area to browse and discover details of these rich but distributed resources using a single entry point.

1.5. The portal is open for the gradual integration of the emerging national infrastructures of the candidate countries into the European Research Area

It is simple for more data providers to join the portal because the technical solutions and guiding material are made available at low cost. The future potential for the outputs of this project has been clearly demonstrated by the interest of the non-participating CESSDA archives in the work. Representatives from the majority of the CESSDA archives attended the project requirements and evaluation workshops and thereby contributed actively and directly to the requirements and evaluation exercises. Several countries have expressed a desire to undertake translation as soon as funding is available. Several archives are in the process of building up Nesstar servers that might seamlessly be integrated into the MADIERA portal.

In 2004 the report of the European Strategy Forum for Research Infrastructure (ESFRI) working group in the Humanities and Social Sciences recommended the establishment of a European Research Observatory which will build upon existing resources and both actively and systematically promote synergy and coherency (EROHS report, 2004). This initiative would be guided by four main objectives:

- The facilitation of access to and sharing of existing European and national data, thereby linking more efficiently and effectively data resources already available.
- The development of improved standards and documentation relating to existing European and national data in order to enhance the scientific quality of data and their potential for interoperability.
- The generation of new and genuinely European data. This will involve both the collection of new data and the digitisation of materials, which are currently non-computerised.
- The provision of research training programmes for the next generation of researchers.

The MADIERA project is entirely in keeping with the first two objectives and as the EROHS report says: "Taken together these first two objectives represent a huge step forward for the humanities and social sciences" (2004:p15) whereas the two second objectives may be viewed as natural next steps and as a continuation of the MADIERA project.

The MADIERA project was initiated in order to meet some of the shortcomings in the European social scientific research infrastructure. The project has aimed to solve some of the problems which hamper pan-European research and which are described in the EROHS report. By building an integrated European social science infrastructure, the portal addresses some of the major gaps and deficiencies identified by the EROHS report through the facilitation of access to and sharing of existing European and national data; the development of improved standards and documentation, and by enabling the linking of cross-national data. The MADIERA portal brings together data, which are stored within confines of the nation-states, and enables the researchers to compare findings across countries. Through the MADIERA portal, data are more available for secondary analysis then ever: 3000 data sets that can be searched for in nine different languages. All datasets available are documented in DDI and datasets can be compared across borders because they are documented through agreed metadata standards. Standardisation facilitates development of middleware so that the diverse resources that make up the

MADIERA portal can be discovered, accessed, allocated, monitored, and in general managed as single virtual systems – even when provided by different vendors or operated by different organisations.

2. The European collaborative effort

The success of the project would not have been possible without the good cooperation between the eight partners and a collaborative effort. The most important requirement to the success of the new portal is that it should hold a sufficient amount of data and documentation of the data. All the partners have played important roles in filling the new portal with well-documented and interesting data.

The project had eight participants, five principal contractors and three assistant contractors from seven European countries. The project was led by the Norwegian Social Science Data Services (NSD), which has had many years experience in the preservation and dissemination of statistics.

The UK Data Archive has been responsible for coordinating the content side of the portal and also for managing the work with the thesaurus. The archive has long experience in this field after having led the development of two generations of multilingual thesauri.

NESSTAR Ltd. has been responsible for the technology development and for the exploitation. In both fields the company have a lot of experience of developing and selling the Nesstar Software Suite.

Finnish Social Science Data Archive (FSD) has led on the dissemination of the project results and decided on the dissemination strategy. Additionally they have played an active part in the development of the thesaurus.

The Danish Data Archive is part of the state archive and was able to bring a professional archival view to the development of web based systems. They co-ordinated the user analysis and had the responsibility for writing the reports.

The assistant contractors, Swiss Information and Data Archive Service for the Social Science (SIDOS), Greek Social Data Bank (EKKE) and Zentralarchiv fur Empirische Sozialforschung (ZA) enabled the wider contribution of data archives to inform the user analysis and provide a user forum for the validation. Additionally SIDOS and EKKE have translated the thesaurus into their national languages.

3. Future needs for research

The MADIERA infrastructure is an effort to make more and better documented research-relevant data available for potential users. Through the project we have focused on three central components, the necessary metadata standard, the language problem and efficient content description, and the technical system needed to make data available on the web. The project have reached the goals set more than three years ago and proved that through setting up a portal giving access over the web to almost 3000 datasets over 9 different servers.

However, the fundamental ideas and the implemented solutions also represent a *formidable potential for expansion and additional functionality*. There are many complexities to this picture that are not yet covered. Some of these are loose ends we may label as management problems while others are genuine research-related problems that require further research efforts.

The management type problems are linked to the maintenance of the portal as it now stands. Addition of new archives to the portal will mean that, as well as adding a new server to the system, the portal screens will have to be amended to reflect the new language and at least the topical classification will need to be translated for the new country. Once the portal begins to be used in earnest, there will be a need for ongoing maintenance and support and further programming will be likely as users request more refinements and bugs or unnecessary restrictions may become apparent.

It is anticipated that the CESSDA partners will have resources to undertake this level of work, however, resources will need to be found for continuing maintenance. The Directors of CESSDA are currently reviewing the terms, definition and criteria for membership of the organisation to broaden participation and place the organisation and its associated Infrastructure on a more secure financial footing.

It is expected that the greater integration of processes over time and the type of data service that the portal represent will bring additional incentives to preserve data and increase their usability. In this way the whole data and metadata production process becomes more transparent, efficient and economical. With an increase in well-documented data available for secondary analysis and social indicator research, the portal will contribute to best practice development in survey resource sharing and data distribution. The portal will inspire next generation processing and analysis of huge amounts of data in order to increase empirical evidence and knowledge about European and global socio-economic developments.

This naturally then exposes the lacunas and loose ends.

We have pointed out that good metadata serves many purposes. For the MADIERA portal, the metadata is there to facilitate resource discovery, communication and data dissemination. There are other aspects. There are quite practical questions like the ability to handle other and more complex types of data. The DDI is by any standard a tremendous success, but success in one area also demonstrates the potential in other areas and it is no reason why we should stop short at this point.

The inspiration of the MADIERA portal might be subdivided in at least 7 topics for further work:

1) Data availability

2) Metadata development, standards and applications

3) Data discovery, browsing and investigation

4) Data dissemination

5) Data security

6) Data harmonization

7) Data preservation

For data *availability*: We have held that data is not necessarily a scarce resource in Europe, but that is a loose assumption. We do not really know what the situation are across Europe in terms of data collection, culture of data sharing and the availability of the data there is for research purposes. Who collect what data for what purposes in various national settings? To document the benefits of research there is a need to document data availability. This means both collection and sharing. With better information on the real availability of data, it would be possible to work more focused on fostering a culture of data sharing between fields of use, and it would be possible to promote technology for data documentation and data publication towards shared use.

For *metadata standards*: The important metadata standards in use all start from a fairly simple data model. Such simplified models not only exclude types of data but also exclude many important research questions. There is a need to define and model complex social science data types. The DDI is specially developed for web-applicability, other metadata standards are also coloured by their primary purpose or justification. There is a constant need to think of expansion or integration of standards, the present state of the art more than anything demonstrates that it is possible to develop such

tools. What are needed are ambitions, financial support and focused work. The DDI has developed from describing sampled survey type data to the ability to describe aggregate or cubed data. In version 3.0 the aim is to be able to describe instruments, geographic data, better handle aggregate data and complex sample type data. To allow for this, modularity has become essential. But complexity is not only external to the file concept, what we may picture as the ability to define complex studies or projects above the file. Complexity also goes internal of the file, instruments differ in complexity, questions and variables may make up complicated structures to measure complicated concepts. These questions have to be addressed, both in terms of better logical descriptions and possibilities to develop more appropriate applications.

For *data discovery*: Documentation of the substantive content of data resources is rarely precisely specified, it is often intended for human interpretation, and much of the documentation work at this level is done by archive professionals, not the people that originally designed the data collection instruments. This generates needs for structuring of the documentation-/publication process. The more order there is introduced through the documentation process across potential resource publishers, the better the practical possibilities to compare and to look up through searching. Classification is an important ingredient on which all statistical systems build, but classifications are difficult to develop for broad substantive content. The closest we come to a generalization of the classification concept for general substantive content is a common *thesaurus*. In the MADIERA portal the ELSST thesaurus is used as such a structuring and defining tool both to insert *keywords* or *concepts* in study descriptions and to deliver search terms for browsing or look-up of data. Additional kinds of structuring are reached through controlled vocabularies and a common topical classification of studies, available through the common template for documentation work. There is a constant need to develop these tools further, in particular to follow up on the ELSST thesaurus so that more languages are incorporated when new servers are hooked up to the portal, or new terms inserted when there is an empirically demonstrated need for that.

On *data dissemination*: The MADIERA portal employs the Nesstar software to locate, access, browse and download data. Such software is a dynamic tool, it is *middleware*

that is the glue that links together data resources, metadata standards, researchers, technical context and methodological technology in a constantly moving interplay to develop better knowledge. The Nesstar software is based on and closely interwoven with the DDI standard. When components or contexts change, this important type of "glue" also have to adjust. An infrastructure is also a network of bilateral links, dependent upon resources and maybe more than anything, dependent upon information, know-how and a

constant analysis of what possibilities there are, what trends are emerging, what potential are opened for tapping. The data dissemination process benefits greatly from the availability of such a software suite. The software not only is a vehicle in itself, but it is also a home and a testing ground for related tools like the thesauruses, vocabularies, taxonomies and ontologies developed. There is a constant need to keep up the dynamics of this research supporting functions.

On *data security*: Data security is an essential obligation on data producers, disseminators and users towards the data subjects. Yet, the availability of quality data is critical to social science research. This means that access to data have to be controlled. Scientific value of data in itself may create a need for protection, data collectors must have a right to embargo the empirical material collected. Then there are various aspects of the fact that scientific value of resources may be matched by commercial value, making available high quality data is an expensive activity and data can often only be available at a price. And of course another important restricting factor is the need to protect data privacy and guard against data misuse. Individual level data usually can only be made available if data privacy is protected.

For a general data access-point like the MADIERA portal this generates a series of problems:

- There is a need for a European-wide data access agreement; participants have to agree on or to a policy of data access.
- Then there will be a need to agree on, implement and adopt a single cross-national registration and authentication system. For several types of follow-up activities we need to know who is using what for which purposes.
- The point above mainly goes for sampled individual level data. For aggregate type data the privacy needs could be protected via an elaborate date disclosure procedure. However, data quality depends on its completeness and accuracy, which could be compromised by data security procedures. As such, there could be an unavoidable tension between the users' need for detailed data and the necessary confidentiality restrictions. Still there will be the need to document the use of data, via user logging or pre-registration.

This points towards a major research and development task, the development of a cross-national registration and authentication system for users of social science data, a system that both promotes the basic idea that data should be easily available and protects the data against unauthorized use.

On *data harmonization*: The main problem for presentcomparative social science research is the fragmentation of the data base. Data collected in various countries for a common comparative study often use distinct classifications for key variables, in order to keep sufficiently close to the daily experience of the respondents. When researchers come to harmonising data collected for uncoordinated surveys, the gap between the various classifications used may prove a barrier to success. MADIERA has brought us a long step forward towards improvement of this situation but more needs to be done to permit an increase in the amount of comparative research using data from distributed sources by developing simpler means of identifying and harmonising potentially comparable data from multiple datasets across geography or time. We have presented metadata as a structured conversation between the researchers, institutions and software processes that are working with a kernel, a dataset, all the way from the design process to the final use. The main purpose of this structured conversation is to make sure that all relevant information are passed on from one station to the next and that all participants have a chance to add their own relevant knowledge to this information exchange.Then it becomes self-evident that it is as important that information can be compared, is measured on comparable scales.

Specifically, further work is needed in the following areas:

- Improving substantive content organization and data resource discovery potential across a distributed network of collections by further development of underlying Semantic Web technologies (standard metadata, thesauri and ontologies). By describing methodologies and procedures, as well as features related to the context of studies, end users are allowed to decide whether or not a data collection is meeting their professional or scientific standards.

- Increasing ICT-resource sharing and extend access to social science data by reducing the technical and administrative barriers to accessing data through the development of middleware layers specifically designed to provide an interface for social science researchers to access the necessary research resources;

On *data preservation:* A data archive focus on data dissemination and data maintenance/preservation as the two major tasks. Whilst the Internet has become the groundbreaking channel for data dissemination, it is certainly an unsafe chest for preservation. The gap between the sophisticated Internet dissemination tools and the current technologies of preservation has increased, affording a significant advantage to the former. There is a need for focused work to reduce this gap by upgrading the contents and quality of the preserved data.

V. DISSEMINATION AND EXPLOITATION OF RESULTS

1. Strategy for dissemination

The aim of the workpackage on dissemination has been to disseminate the results of the project to the wider research community to both raise awareness and use and to encourage the creation of everything from full knowledge products to simple enhancements to the metadata. The MADIERA project has two main target groups; researcher within the European social science community and data providers that wish to connect to the portal. In the period, there has been continuous contact with representatives for both groups. The strategy has been threefold:

Firstly, the most important requirement to the success of the new portal is that during, and especially at the end of the project, the portal should hold a *sufficient amount of data and documentation of the data*. As we wanted to have the new portal filled with interesting data, the national data archives have played a significant role central in this task. Since one of the central aims of the project is that the MADIERA infrastructure will become the new CESSDA portal, extensive information activity have been aimed at this community and a special CESSDA-oriented strategy has been one of the top priorities of the project. It is registered as an initial success that CESSDA has decided to make the MADIERA portal its new integrated catalogue.

But also archives outside the CESSDA community and other data providers should be welcomed and supported to start producing data documentation in the DDI format and to set up a server to be linked with the new database. The project has been presented at several conferences, most importantly at the annual IASSIST conferences in 2003, 2004 and 2005. The project and the MADIERA portal will also be represented at the conference in May 2006. IASSIST is an international organisation of professionals working with information technology and data services to support research and teaching in social sciences. The members' work in a variety of settings, including data archives, statistical agencies, research centres, libraries, academic departments, government departments, and noon-profit organisations. This conference provided a opportunity to target potential users of the MADIERA system, to disseminate information about the project and collect future requirements from the community.

Secondly, the promotion of the project and its objectives should be as *collaborative* as possible, and it should be accomplished *by dividing the task into several different subtasks*. One aspect of this is to first establish a site for the prototype of the portal, and then expand it to the desired magnitude. Both national and international research

information providers have participated in the promotion work in order to get a broad coverage to the information on the new products and services.

In order to ensure that the results are exploited as widely as possible and that the system is used for data publishing beyond the initial social science archive community there have been arranged three successful exploitation event/workshops where the MADIERA portal has been presented for several of the European Social science data archives. They are potential providers of data to the portal – they are also very important links to the research community and data users in their countries.

The workshops have focused on evaluating different versions of the MADIERA portal and the latest Nesstar software. Both in the hands on evaluation of the software and in evaluating the portal, the discussions have been regarded as extremely useful.

Thirdly, when it comes to user demands, the project should offer *tailor-made information for different key user groups.* This means concentrating on those features of the project that are of particular interest to that user group, tailoring the information to their needs and selecting a suitable forum. Different brochures and other material have been distributed. In addition a website for the project has been set up

Figure 17. Channels for dissemination: MADIERA website + brochure to providers

2. Results coming out of the project

Table 2. List of project results

Result	Partner involved	Exploitation intention
1. New 'CESSDA' portal	Consortium	Individual archives, research institutes, end users
2. Multilingual thesaurus operation	UKDA, FSD, DDA, EKKE, SIDOS, NSD	Individual archives, libraries, end users (researchers, policy makers, others)
3. Additional options to core Nesstar technologies	Nesstar Ltd. and NSD	Users of the Nesstar server technologies and end users
4. Metadata structures for resource repositories and enhancements to the DDI	Public domain	Individual archives, research institutes, libraries and other information publishers

The set up of the portal ensures that it has the capacity to grow and diversify even after the initial construction period. We are very pleased that the members of CESSDA (Council of European Social Science Data Archives) have decided that the MADIERA portal shall serve as the basis for the new CESSDA Common Catalogue, a common integrated European data catalogue for all the data archives. We have managed to create a sustainable system, and it looks very likely that it in the future will be nurtured by the collective energy of the data and knowledge producing communities of the European Research Area. That the MADIERA portal has become the new 'CESSDA' portal, called "C-CAT" is one of the most high profile and exciting results to come out of the project. Expansion into central and Eastern Europe has seen CESSDA grow and these countries look to the partners represented in the consortium to guide and lead their developments. It has been decided by the CESSDA community that this will become the 'official' interface to the community and will be used heavily. Additionally it will be a major step towards creating a European identity to the data collections and make a significant contribution to development, perception and functionality for the European research area in the social sciences and beyond. In addition it will serve as an exemplar of the type of integrated, yet distributed and locally managed, infrastructure.

The multilingual thesaurus is based on two generations of earlier work at the UK Data Archive. Firstly there was the HASSET thesaurus and then, under the auspices of the LIMBER project, the ELSST thesaurus. The new development will mean that an improved version will be created and translated into several languages. When the project achieves the first goal of a fully operational and functional web of information, then it is intended to widen the exploitation of the thesaurus and the UKDA will discuss with the other

partners the best way of taking this forward. One option will be the licensing of the technology to Nesstar Ltd. who are set up to realise wider exploitation outside the academic community.

Nesstar Ltd and NSD have developed enhancements to the underlying Nesstar technologies. These have been made available to the project partners at no cost in order to ensure the full participation of all of the European archives. Organisations outside this immediate network will be encouraged to join the network and the negotiations with them have been carried out during the project.

The exploitation of the metadata will be as wide as possible and this IPR have been placed in the public domain. All stakeholders have benefited from as wide a take-up of metadata standards as possible. It facilitates interoperability and long-term data preservation and the partners are all committed to continuing their international leadership and delivery of improved metadata models and implementations in the wider community. In particular the metadata developments will be presented as valuable input to the DDI committee.

VI. REFERENCES AND BIBLIOGRAPHY

References

Alvheim, A. and Ryssevik, J. "MADIERA, Multilingual Access to Data Infrastructures of the European Research Area" Paper presented at the First International Conference on e-Social Science, Manchester, UK, June 2005.

Assini, Pasqualino. "Objectifying the web the 'light' way: an RDF-based framework for the description of web objects". Poster presented at the Tenth International World Wide web Conference, May 1-5 2001, Hong Kong.

Assini, Pasqualino. "NESSTAR: A Semantic web Application for Statistical Data and Metadata" in Real World Semantic web Applications, Vipul Kashyap and Leon Shklar eds. IOS Press, Amsterdam, 2002, ISBN 1 58603 306 9, pagg. 173-183.

Blank, G. and Rasmussen, K.B. "The Data Documentation Initiative: The Value and Significance of a Worldwide Standard." Social Science Computer Review 22, no. 3 (August 2004): 307-318.

"Blueprint for the European Research Observatory for the Humanities and Social Sciences EROHS", Report Compiled for the European Strategy Forum for Research Infrastructure (ESFRI) by the Ad Hoc Working Group on Research Infrastructure in the Humanities and Social Sciences (RISSH), Copenhagen, May 2004.

ISO 9241-11:1998: Ergonomic requirements for office work with visual display terminals (VDTs) -- Part 11: Guidance on usability.

MADIERA portal http://purl.oclc.org/NET/MADIERA_portal

MADIERA project http://www.MADIERA.net

Miller, K. and Matthews, B.Having the right connections: the LIMBER project; Journal of Digital Information (2001).

Miller, K. and Vardigan, M. "How Initiative Benefits the Research Community - the Data Documentation Initiative" Paper presented at the First International Conference on e-Social Science, Manchester, UK, June 2005.

Nesstar Ltd web Site http://www.nesstar.com

Norwegian Social Science Data Services. "Providing Global Access to Distributed Data Through Metadata Standardisation: The Parallel Stories of NESSTAR and the DDI." Working Paper No. 10, UN/ECE Work Session on Statistical Metadata, Geneva, Switzerland, September 22-24, 1999.

Ryssevik, J. "Bazaar Style Metadata in the Age of the web - An 'Open Source' Approach To Metadata Development." Working Paper No. 4, UN/ECE Work Session on Statistical Metadata, Washington, DC, November 28-30, 2000.

Ryssevik, J. and Musgrave, S. "The Social Science Dream Machine." Social Science Computer Review 19, no. 2 (summer 2001): 163-174.

http://www.useit.com/useit.com: Jakob Nielsen's website.

VII. ANNEXES

1. List of participants

Coordinator

Atle Alvheim
Norwegian Social Science Data Services (NSD)

Harald Hårfagres gate 29

5006 Bergen

Norway

Tel: +47 55 58 21 17

Fax: +47 55 58 96 50

Email: nsd@nsd.uib.no

Contractors

Norwegian Social Science Data Services (NSD) - www.nsd.uib.no

UK Data Archive (UKDA) - www.data-archive.ac.uk

Danish Data Archive (DDA) - www.dda.dk

Finnish Social Science Data Archive (FSD) - www.fsd.uta.fi

Nesstar Limited - www.nesstar.com

Swiss Information and Data Archive Service for the Social Sciences (SIDOS) - www.sidos.ch

Greek Social Data Bank (EKKE) - www.ekke.gr

Zentralarchiv für Empirische Sozialforschung (ZA) - www.gesis.org/ZA

2. Deliverables

Table 3. List of deliverables

No	Deliverable title	Completed
D1.1	Project Initiation Document	X
D3.1	Functional Specification and Design	X
D5.1	Guidelines for Thesaurus construction and translation	X
D1.2	Quality Assurance Plan (to be peer reviewed)	X
D2.1	User Analysis Report	X
D3.2	MADIERA Prototype	X
D7.1	Dissemination Plan	X
D2.2	Usability test – MADIERA Prototype	X
D3.3	MADIERA Beta Version	X
D4.1	Recommendation – Geo-referencing system	X
D6.1	Guidelines – Content provision and access control	X
D2.3	Usability test – MADIERA Beta version	X
D4.2	Methodology – identification of comparable elements	X
D3.4	MADIERA Version 1.0	X
D4.3	Naming and identification recommendation	X
D5.2	Report on administrative mechanisms for thesaurus maintenance	X
D6.2	User guides and training packs for content provision	X
D6.3	First version of hyper-linked information space demonstrator	X
D6.4	Data archive content provision workshop	X
D6.5	Workshop on content metadata (CDG/DDI)	X
D7.3	User guides and training packs for adding specific "knowledge products"	X
D8.1	Draft Technological Implementation Plan	X
D8.2	Workshops for non-archive data providers	X
D2.4	Usability test - MADIERA Version 1	X
D5.3	Extended multilingual thesauri	X

D6.6	Hyperlinked Information-Space Demonstrator Version 2	Partly implemented *)
D4.4	Package of revised recommendations	Abandoned- the recommendations can be found in D4.1, D4.2 and D4.3
D5.4	Evaluation Workshops	**X**
D1.3	Final Report	**X**
D2.5	Final Usability test report	**X**
D3.5	MADIERA Version 1.1	**X**
D5.5	Additional thesaurus hierarchies	**X**
D8.3	Final Technological Implementation Plan	**X**
D7.2	On-going dissemination events	**X**

*) This technology is under development. Hyperlinking directly into datasets it is demonstrated in other portals NSD has developed, but it has not been implemented in the MADIERA portal. We hope to implement during 2006.

3. Presentations

Kick-off meeting of the MetaDater project (FP5 project), in Cologne, 7-8 January 2003, Bjarne Øymyr, NSD.

Kick-off meeting for the Co-ordinators for projects funded under the Third Call of the Key Action in Brussels, 13-14 March 2003, Bjarne Øymyr, NSD.

IASSIST Conference in Ottawa, 26-29 May 2003, Ken Miller, UK Data Archive.

The project was presented at the CESSDA meeting in Dublin, 1-2 April 2003, Bjarne Øymyr, NSD.

Two presentations were held at the IASSIST conference in Wisconsin 25-28 May 2004:

- A general introduction of the MADIERA project: MADIERA: "A European Infrastructure for Web-based Data Dissemination: An Overview", Atle Alvheim NSD.
- The NESSTAR Publisher and its thesaurus functionality: "No Longer Lost in Translation", Ken Miller, UKDA.

"Practical Semantic web Portal Design", WWW2005 10-14 May 2005 Chiba, Japan. Developer's Day - Session on 'Semantic web, Theory and Practice', Titto Assini, Nesstar.

"ELSST we forget, the MADIERA portal", Demos and Posters held at the IASSIST in Edinburgh 23-27 May 2005. A demonstration of the progress made within the MADIERA project with the multilingual thesaurus ELSST and its use in the Nesstar system, Kenneth Miller, UKDA. The session proved to be very successful and generated wide interest.

"Practical Semantic Portal Design: The MADIERA Data Portal" Demos and Posters of the 2nd European Semantic web Conference (ESWC 2005), Heraklion, Greece, 29. May - 1. June 2005, Titto Assini, Nesstar.

MADIERA - Multilingual Access to Data Infrastructures of the European Research Area – Ken Miller, UKDA A paper on the MADIERA project was presented at the First International Conference on e-Social Science, held at the University of Manchester, 22-24 June 2005. It was part of a metadata workshop which linked the DDI (Data Documentation Initiative), a metadata standard for social science data, with EU projects MADIERA and MetaDater.

Workshop/Explotiation event, University of Essex, Great Brittain, 27-28 June 2005:

- Introduction to the MADIERA portal- Ken Miller, UKDA;

- Georeferences - extending search for data – Ørnulf Risnes, NSD.

ASC International Conference on Survey Research Methods – Maximising Data Value, Buckinghamshire, 15-16 September 2005, Ken Miller

Evaluation Workshop, EKKE Athens, 29-30 September 2005, Ken Miller and Atle Alvheim were invited to be evaluators at a workshop organized by EKKE in Athens. This was to evaluate the deliverables from a Greek national project "Node for Secondary Processing" which had similar aims to the MADIERA project of promoting comparative research.

The history and the future of the MADIERA portal 26-28 October 2005 on a CESSDA Expert meeting in Madrid, Atle Alvheim, NSD

The MADIERA portal - Weaving the web of Euroepan Social Science, Atle Alvheim. The project was presented at a poster session at a conference organised by the European Commission 12-13 December 2005: Social sciences and humanities in Europe: New challenges, new opportunities.

4. List of tables and figures

European Commission

EUR 23148 — EU RESEARCH ON SOCIAL SCIENCES AND HUMANITIES — Multilingual Access to Data Infrastructures of the European Research Area - MADIERA

Luxembourg: Office for Official Publications of the European Communities

2007 — 116 pp. — 21,0 x 29,7 cm

ISBN 978-92-79-07754-8

How to obtain EU publications

Our priced publications are available from EU Bookshop (http://bookshop.europa.eu/), where you can place an order with the sales agent of your choice.

The Publications Office has a worldwide network of sales agents. You can obtain their contact details by sending a fax to (352) 29 29-42758.

KI-NA-23148-EN-N

ISBN 978-92-79-07754-8