



European Strategy Forum
on Research Infrastructures

ROADMAP FOR EUROPEAN RESEARCH INFRASTRUCTURES

REPORT OF THE

SOCIAL SCIENCES AND HUMANITIES

ROADMAP WORKING GROUP

VER. 4 SEPTEMBER 2006

Foreword

This document is the report of the Social Sciences and Humanities Roadmap Working Group (SSH RWG) to ESFRI recommending new or upgraded pan-European Research Infrastructures (RIs), to be considered for the first Roadmap for European Research Infrastructures.

Two Expert Groups (EG) were proposed by the SSH RWG and established by ESFRI to assess infrastructural needs for the humanities and social sciences and to recommend new or upgraded RIs in the areas. The members were selected on the basis of their expertise, including in science policy development and their international reputation, from nominations by ESFRI delegations on the basis of their known capabilities. The terms of reference of the working groups ensured that the members acted in a personal capacity, and that conflicts of interest should be declared and dealt with.

The EGs conclusions are based on proposals received via the exercise of List of Opportunities and a mapping of potential new (or major upgrade) pan-European Research Infrastructures for ESFRI consideration. The SSH RWG conducted the mapping to supplement the proposals received via the List of Opportunities to have a broader basis for the review. Based on this material the EGs proposed seven RIs to meet the infrastructural needs within the domain for the next 10 – 20 years. In order to create scientifically and conceptually stronger RIs some of those are compounded from several received proposals. This has been done in close dialogue with the proposers. The reports of the two EGs are incorporated in this document.

Rapporteurs for the seven proposed RIs were invited to present their projects for the SSH RWG members in the SSH RWG meeting in Brussels April 19. Based on the presentations and the reports from the two EGs the final decision on which RIs to be recommended for the ESFRI Roadmap was taken by the SSH RWG in the following meeting in Brussels May 15.

The SSH RWG recommendation consists of six RIs of pan-European nature. Although they may differ from each other in many respects they have three distinct features.

- Their aim is in the end to provide the scientific communities with high quality data for research purposes
- They do all consist of distributed systems involving institutions and initiatives in a number of European countries
- They build to some extent on institutions or services the through their existence has proven their excellence and high quality

The EGs identified eight RI initiatives as embryonic according to the criteria set by ESFRI. These RIs are including in this report. In the SSH RWG's final decision one of the projects recommended by one of the EGs was reclassified from mature to embryonic. The total number of proposals recommended by the SSH RWG as embryonic is therefore nine. The SSH RWG recommends that none of the proposals reviewed should be rejected from future considerations.

Bjørn Henriksen
Chair SSH RWG

Content

1	EXECUTIVE SUMMARY	4
1.1	STRATEGIC GOALS.....	4
1.2	EUROPEAN SOCIAL SURVEY - ESS.....	5
1.3	SURVEY OF HEALTH, AGEING AND RETIREMENT IN EUROPE - SHARE	5
1.4	COUNCIL OF EUROPEAN SOCIAL SCIENCE DATA ARCHIVES - CESSDA	5
1.5	COMMON LANGUAGE RESOURCES AND TECHNOLOGY INFRASTRUCTURE – CLARIN	5
1.6	EUROPEAN RESEARCH OBSERVATORY FOR THE HUMANITIES AND THE SOCIAL SCIENCES - EROHS	6
1.7	DIGITAL RESEARCH INFRASTRUCTURE FOR THE ARTS AND HUMANITIES - DARIAH.....	6
1.8	RECOMMENDATIONS ABOUT THE ROADMAP PROCESS IN THE FUTURE	6
2	SCIENTIFIC LANDSCAPE OF SOCIAL SCIENCES AND HUMANITIES.....	7
2.1	STATUS	7
2.2	INFRASTRUCTURE NEEDS	7
2.3	LONG-TERM STRATEGIC GOALS FOR RESEARCH INFRASTRUCTURES	9
3	SOCIAL SCIENCES AND HUMANITIES ROADMAP WORKING GROUP - SSH RWG.....	11
3.1	MEETINGS HELD BY THE SSH RWG.....	11
3.2	MEMBERS OF SSH RWG	11
3.3	LIST OF OPPORTUNITIES	12
3.4	PRESENTATIONS OF RESEARCH INFRASTRUCTURE PROJECTS	12
3.5	EXPERT GROUPS	12
3.6	MAPPING OF POTENTIAL NEW (OR MAJOR UPGRADE) PAN-EUROPEAN RESEARCH INFRASTRUCTURES FOR ESFRI CONSIDERATION	13
4	METHODOLOGY USED BY EXPERT GROUPS	14
4.1	EUROPEAN CULTURAL HERITAGE EXPERT GROUP – ECH EG	14
4.2	EUROPEAN RESEARCH OBSERVATORY FOR THE HUMANITIES AND SOCIAL SCIENCES (EROHS) EXPERT GROUP – EROHS EG	17
5	RECOMMENDED PROPOSALS FOR NEW AND UPGRADED RESEARCH INFRASTRUCTURES.....	21
5.1	RECOMMENDATIONS	21
5.2	EUROPEAN SOCIAL SURVEY – ESS.....	22
5.3	SURVEY OF HEALTH, AGEING AND RETIREMENT IN EUROPE – SHARE	31
5.4	COUNCIL OF EUROPEAN SOCIAL SCIENCE DATA ARCHIVES – CESSDA	40
5.5	COMMON LANGUAGE RESOURCES AND TECHNOLOGY INFRASTRUCTURE – CLARIN	58
5.6	EUROPEAN RESEARCH OBSERVATORY FOR THE HUMANITIES AND THE SOCIAL SCIENCES – EROHS	73
5.7	DIGITAL RESEARCH INFRASTRUCTURE FOR THE ARTS AND HUMANITIES – DARIAH.....	85

1 EXECUTIVE SUMMARY

The Social Sciences and Humanities Roadmap Working Group (SSH RWG) proposed two Expert Groups (EG):

- European Cultural Heritage Expert Group – ECH EG, chair: Maurice Bric, Ireland
- European Research Observatory for the Humanities and Social Sciences (EROHS) Expert Group – EROHS EG, chair: Niels Ploug, Denmark

ESFRI later followed the recommendation and did establish the two groups. Both started to work in the autumn of 2005.

The ECH EG was established to identify and develop research infrastructures (RI) for the area. The group applied the criteria set by ESFRI to identify mature proposals. Three projects were recommended for the Social Sciences and Humanities Roadmap Working Group (SSH RWG) to be included in the Roadmap. Each of these represents distributed facilities. They are either based on existing European infrastructures that require major up-grades, and new functional nodes, or on national facilities, which should be joined to enhance pan-European distributed facilities. These infrastructures are:

- Common Language Resources and Technology Infrastructure - CLARIN
- Digital Research Infrastructure for the Arts and Humanities - DARIAH
- European Research Infrastructure for Conservation and Analysis - EURICA

The EROHS EG was given two interrelated tasks:

- To specify and substantiate the proposal for the setting up of EROHS
- To evaluate the proposals for new and upgraded pan-European research infrastructures in the social sciences

EROHS will cover both the Humanities and the Social Sciences. The Group was therefore set up with experts from both fields.

The EROHS EG recommended four infrastructures for the SSH RWG:

- European Social Survey - ESS
- Survey of Health, Ageing and Retirement in Europe - SHARE
- Council of European Social Science Data Archives - CESSDA
- European Research Observatory for the Humanities and Social Sciences - EROHS

Based on the recommendations from the two EGs and after thorough discussions in the SSH RWG the SSH RWG decided on its meeting on May 15, 2006, to recommend the following proposals for consideration for the ESFRI Roadmap:

- ESS
- SHARE
- CESSDA
- CLARIN
- EROHS
- DARIAH

1.1 STRATEGIC GOALS

The SSH RWG has identified six current and future different infrastructure needs for the two fields:

- Data collection
- Digitalization
- Interoperability of data
- Interoperability between fields and language
- Central access/location services
- Harmonization of data access policies

For the six infrastructure needs, the following strategic goals have been identified in order to address the needs:

- European Comparative Data and Modeling
- Data Integration and Language Tools
- Coordination and Enabling

The first strategic goal is linked to the need for European-wide data and the specification of best practice standards for European data collections. Among the recommended mature projects ESS and SHARE have been identified as necessary first steps since they are currently seen as best practice examples for comparative surveys or panel studies.

The second goal encompasses the interoperability of data and languages. CESSDA has been seen as an organization that is contributing substantially to a rapid European-wide data integration and CLARIN as a tool to overcome the language barriers.

The third goal, which is linked to EROHS and DARIAH, is directed towards the harmonization of data access policies, the standardization of digitalization processes as well as the interoperability between the social sciences and the humanities in general.

1.2 EUROPEAN SOCIAL SURVEY - ESS

The ESS infrastructure already exists, and a major upgrade is proposed. Between 25 and 28 European countries are now committed to the project, and have together with the EC secured funding for the ESS's biennial rounds 1 – 4. ESS has also been awarded separate infrastructure support from the Commission for the next five years (from 2006 - 2011). The infrastructure support is, however, explicitly not to be used to support the biennial rounds of the survey. Instead it is for outreach activities and methodological refinements only.

By including ESS in the proposals, the SSH RWG hopes to unify, regularise and secure the funding for the RI as a whole (the surveys, the outreach activities and the methodological work) over an extended period, though naturally with periodic reviews. A large and complex time series such as the ESS requires such continuity of funding, which is a prerequisite of appropriate planning. But a major upgrade would also help to fill debilitating gaps in the present programme of work – allowing much-needed new programmes of work on:

- compiling and harmonising aggregate context variables for survey analyses
- experimenting with alternative (technical and traditional) methods of translation to improve equivalence
- investigating and mitigating longstanding problems in the collection and classification of occupation and education
- improving the capacity to pilot and pre-test new questions on emerging issues of public concern
- experimenting on a multinational basis with methods of improving response rates

All this work will be in addition to designing and coordinating the biennial ESS and to the conduct of fieldwork, coding and keying in some 30 European nations.

1.3 SURVEY OF HEALTH, AGEING AND RETIREMENT IN EUROPE - SHARE

SHARE was founded in 1999 in order to create an equivalent to the US-American Health and Retirement Survey. Preliminary data collection started in 2002, and in 2004 a first wave of data on the economic, health and family conditions of about 27,000 respondents aged 50 and over were collected in 11 European countries.

The major strength of these data is the ex-ante harmonized cross-national dimension that allows comparing the effects of different welfare systems (e.g. pension and health care systems) on the lives of mid-

dle-aged and older European Citizens. The second wave of data collection is currently going on and includes two new EU member states, Poland and the Czech Republic, and Ireland (EU-funded by an Integrated Infrastructure Initiative and a STREP). A third wave of data collection specializes on the life histories of the SHARE respondents (EU-funded as Integrated Project) in 2007.

SHARE will add more waves to monitor the ageing process in Europe and to foster multidisciplinary cross-national research on the social and economic implications of ageing.

1.4 COUNCIL OF EUROPEAN SOCIAL SCIENCE DATA ARCHIVES - CESSDA

CESSDA has been a mature network of European data archives for over twenty years. It is to-day covering each of the European countries. The effectiveness and range of activities need however to be enhanced.

- Development of common standards, tools, instruments and services
- Development, maintenance and improvement of already existing tools and instruments
- Establishment of special programme for the smaller, less-developed and less-resourced CESSDA member organizations in order to enable them contribute fully and on an equal basis to the CESSDA-based programme of activities
- Establishment of a seed-money programme in order to extend the existing CESSDA network and foster the development of national data archiving initiatives in those countries that are not currently part of CESSDA.

1.5 COMMON LANGUAGE RESOURCES AND TECHNOLOGY INFRASTRUCTURE – CLARIN

The project is a large-scale pan-European collaborative effort to coordinate and make language resources and technology available and useful to scholars of all disciplines, in particular the humanities and social sciences. It will overcome the present fragmented situation by harmonizing structural and terminological differences based on a Grid-type of infrastructure and by using Semantic Web technology. In creating an integrated and interoperable domain of language resources and technology it will establish the basics of eHumanities. The envisaged infrastructure CLARIN will be stable, persistent, highly accessible and open for extensions to meet the grand challenges of the research disciplines, in particular in the humanities. CLARIN will draw on the accumulated professional

experience and it will integrate the experiences from infrastructure and knowledge dissemination projects and associations.

1.6 EUROPEAN RESEARCH OBSERVATORY FOR THE HUMANITIES AND THE SOCIAL SCIENCES - EROHS

EROHS will be designed as a new European-wide infrastructure (RI) addressing the major impediments of existing research infrastructures, whether it be lack of coordination and cooperation, insufficient training, outdated policies, inadequate access, unsatisfactory tools or incomplete resources. In some areas it will be based on mature existing resources and infrastructures, supporting, promoting and extending their scope – and will in other areas facilitate the development of new infrastructures based on the need of the humanities and the social sciences by bringing together national initiatives, organisations and individuals in order to provide upgraded and enhanced European wide actions.

EROHS will be organised to promote and ensure cooperation and integration of data, technologies and policies. This calls for setting up a central and coordinating unit, an enabling infrastructure, explicitly and specially addressing the cultural, economic, legal, and institutional constraints to the realisation of Europe as a natural laboratory for the social sciences and humanities. EROHS will operate both as a central and distributed facility with a strong physical hub working in close conjunction with a number of spokes across Europe, harnessing European expertise through a coordinated yet decentralised network.

1.7 DIGITAL RESEARCH INFRASTRUCTURE FOR THE ARTS AND HUMANITIES - DARIAH

The aim of DARIAH is to provide an infrastructure that as an end result could support access to all surviving humanities and cultural heritage information for Europe. It would do this by focusing on four key elements:

- Promote and bring together the best efforts of national initiatives, organisations and individuals in order to develop truly pan-European actions, initiatives and services.
- Develop national services and digitisation programmes aimed particularly at those European countries without such services and programme
- Bring together the different sectors involved in cultural heritage and humanities information management and access – education, memory and cultural heritage institutions and organisations, and the commercial sector – in order that they might work together for the benefit of both themselves and the research communities across Europe
- Enhance and promote digital scholarship in the humanities and arts across Europe

The scientific case for DARIAH is driven by three key factors:

- The changing nature of research practice, knowledge creation, and information sharing
- The highly distributed and dispersed nature of much cultural heritage and humanities information
- Increasingly pervasive broadband connectivity and a growing range of technologies and applications

1.8 RECOMMENDATIONS ABOUT THE ROADMAP PROCESS IN THE FUTURE

The two Expert Groups received a large number of proposals for assessments. The SSH RWG is however fully aware of the narrow selection of European research infrastructure for the social sciences and humanities the **recommended** proposals constitute and will recommend strongly that ESFRI in later rounds will try to cover a broader area of the Humanities and Social Sciences. Furthermore the SSH RWG recommends that no proposals reviewed should be rejected from future considerations by ESFRI.

2 SCIENTIFIC LANDSCAPE OF SOCIAL SCIENCES AND HUMANITIES

2.1 STATUS

The practice of the social sciences and humanities has slowly but profoundly been transformed along with the emergence of new information technologies. Digital resources, computer networks, and software tools to a great extent, influence the sense of the human record, the way it is understood and the way those understandings are communicated.

Although much of the technology is developed for other scientific purposes and domains, the social sciences and humanities are different from the physical sciences in the variety, complexity, incomprehensibility, and intractability of the entities that are studied. However, diversity and size of infrastructure needs in the social sciences and humanities are often very similar to those needs in life science.

Digitizing the products of human culture and society poses intrinsic problems of complexity and scale. The complexity of the record of human cultures—a record that is multilingual, historically specific, geographically dispersed, and often highly ambiguous in meaning—makes digitization difficult and expensive.

The present major challenge is therefore to create pan-European infrastructural systems that are needed by the social sciences and humanities to utilize the vast amount of data and information that already exist or should be generated in Europe. Today the social sciences and humanities are however hampered by the fragmentation of the scientific information space. Data, information and knowledge are scattered in space and divided by language, cultural, economic, legal, and institutional barriers.

A similar situation is found for comparative social science research. As a consequence too much of the research is based on data from a single nation, carried out by a single-nation team of researchers and communicated to a single-nation audience. This state of affairs is preventing the development of a comparative and cumulative research process integrating the European Research Area.

Information for research purposes is however not a scarce resource in Europe neither for the social sciences nor for the humanities. Well-developed official statistical systems combined with a variety of academically driven data gathering programs and activities are producing a wealth of data and information about various aspects of the European societies. This also includes simulation systems and collections of multimedia content—images, text, moving images, and audio.

However, the majority of these resources are country or nation specific. They are produced to meet national requirements and collected by means of nation and language specific instruments based on local methodologies and classifications. They are normally documented in local languages only and rarely published for general use outside the country of origin. On top of this nation specific access restrictions will often prevent information to travel abroad.

Yesterday's answers to these challenges would probably have been formulated in terms of centralization and establishment of large-scale European-wide institutions. Today's answers should rather focus on standardization, the power of emerging information and communication technologies, harmonization of data access restrictions and strengthening of and collaboration among already established groups and organizations engaged in the development of the European Research Area. Concerted efforts on a European scale are needed to bring about the necessary changes.

2.2 INFRASTRUCTURE NEEDS

To encourage and nurture research within the social sciences and humanities in Europe research infrastructures offering a series of crucial functions and resources must be established. Many of these are extensions of functions and resources already existing at a national level in some European countries. In some cases the model service might even be present at a European level, but would need to be extended in scope, strengthened or multiplied to other disciplines.

2.2.1 DATA COLLECTION

Within the social sciences as well as for the humanities collection of data is essential and could be referred to as the fuel for the research process. Although the nature of the information needed might differ between the two domains, the nature of the needed infrastructure will be the same.

For the social sciences high quality cross-national data are essential to comparative research. Long established data gathering projects like the Eurobarometers, ISSP and the World Value Studies have for decades been the only available data sources designed to produce comparative empirical evidence. The newly established European Social Survey (ESS) has recently set new standards for this type of operations and demonstrated what can be achieved when

data are collected according to the most stringent scientific methods. ESS can in this respect serve as a model and an example of best practice for other European wide data gathering projects within other social science disciplines. To facilitate this type of cross-national data gathering operations concerted effort are needed to standardize survey instruments, develop generic and multilingual survey modules to be used in a variety of projects, streamline cross-national sampling methodologies etc.

Our modern digital societies are also producing a wealth of new data that up to recently hardly has been the object of scientific research. Examples of potentially new data resources might be transactional data produced by loyalty cards or Internet behaviour evidence deriving from web-traffic data. Digitization of artifacts, manuscripts, music, etc offer us new ways of describing and analyzing art, sculptures, human languages, etc. Concerted effort is needed to develop the methodologies to collect and standardize this type of data at a European level.

2.2.2 DIGITALIZATION

Within the humanities the new technological possibilities lead to a public as well as scientific demand for open and online access to the full range of primary source materials housed in repositories such as museums, historical societies, local libraries and research libraries, special collections, archives, and privately held collections. This includes books and journals, newspapers and magazines, government documents, manuscripts, maps, photographs, satellite images, census data, recorded sound, film, and broadcast television.

The new information technology offers ways to reunite disseminated collections, to compare exemplars, or to bring together the works of single creators. The prospects are inconceivable. Today the Web give access to billions of pages with information and the annual digital output totaling many times what any library in the World holds. Nevertheless the potential of networked cultural heritage information is far from implemented.

2.2.3 INTEROPERABILITY OF DATA

One major objective is to provide seamless access to data across repositories, nations and research purposes. Data generated for one purpose should be open for use in many ways. They should consequently be created, described, and preserved in ways that facilitate use for a variety of purposes. The use of standards is crucial to the sharing of data. The culture of standards has shown to be weak in the social sciences and humanities.

To encourage data sharing and interoperability across communities and software systems, standardization

of metadata and data are needed and should be encouraged. For social sciences data important initiatives and efforts are however underway on an international level, most notably the Data Documentation Initiative (DDI) carried out by an international alliance of data producers and archives, and Statistical Data and Metadata Exchange (SDMX) initiated by data producers at the international level, like International Monetary Fund (IMF), Eurostat and European Central Bank (ECB).

In the vast majority of the repositories resources are however put into online information resources that cannot be search or analyzed horizontal between data collections. The advantage of the new technical possibilities is not utilized. In many cases one replicates the problem found in the analogue world of being unable to readily assemble physically dispersed information.

Within the social sciences the majority of existing data from Europe are not comparative by design. However, data collected at a national level might still be of value for comparative research if properly documented and harmonized. An example of this type of post hoc documentation and harmonization of data is the work carried out by Luxembourg Income Studies or the efforts undertaken by ICORE¹ regarding election studies.

Documentation and harmonization of data for comparative research is however a time-consuming and expensive activity that seldom will be funded by individual research projects. It should be addressed at European-level and funded as an infrastructure activity.

To make sure that important requirements are met, it is important that the European research community plays an active part in the development of these standards. Even more important is it to encourage and facilitate the use of these standards to document and publish the existing inventories of research data sitting in national data archives and research institutes across Europe. This is a major undertaking that only can happen if the relevant incentives, tools, training and support networks are in place.

2.2.4 INTEROPERABILITY BETWEEN SCIENTIFIC FIELDS AND LANGUAGES

One of the major challenges in the building of new European research infrastructures lies in the interoperability between different domains of research. Currently, research infrastructures are not only separated across national borders, but also by disciplinary fields

¹ ICORE stands for International Committee for Research into Elections and Representative Democracy, and was established at an ECPR Research Session on Electoral Studies held in Rimini in 1989.

and, above all, by a long tradition of isolating the humanities from the social sciences and vice versa.

Likewise, a further invisible border can be identified which results from the diversity of different languages within the European Research Area. Thus, many of the nationally available digitalized objects or researchers outside a specific language community cannot access national data. In this sense, overcoming the language barriers becomes an essential need both for the social sciences and, even more so, for the humanities.

It must become one of the main tasks of the new European research infrastructure to leave these old distinctions and divisions behind and to create research tools and instruments that can be readily accessed and used both by the social sciences and humanities, irrespective of language differences.

2.2.5 CENTRAL DATA ACCESS/LOCATION SERVICES

In the majority of European countries data are preserved and made available for the research communities by national repositories. It is of crucial importance that these data resources can be found and even accessed through central web-based cataloguing and location services. This does not imply that data as such should be centralized in a “European Data Archive” – only that the knowledge of similarly structured data and metadata are made available through a virtual and central access point.

Among the functions that should be supported by central location and access service are: a) a registration service allowing existing and new resource providers to register their data with the system, b) a resource location service, allowing researchers and other end users to find data easily, c) functionality that facilitates easy identification of comparable data across datasets and sites, d) functionality that provides multi-lingual resource-location services.

Technologies meeting a subset of these requirements have been developed by EU-funded projects, like the MADIERA project. The Council of European Social Science Data Archives (CESSDA) has also made a decision to encourage the development of a European data location and access service building on the technologies coming out of these projects. However, there is still a long way to go before a service including the majority of countries and existing information resources becomes a reality.

2.2.6 HARMONIZATION OF DATA ACCESS POLICIES

One major obstacle to access to empirical data in Europe is produced by the multitude of data access policies and regulations implemented by the national

governments. To make it easier to bring together data from various countries mapping and even harmonization of regulations are needed.

In the ideal world we could envisage a European-wide agreement for data providing a central registration and authentication service for scientific data users based on state-of-the-art PKI-technology. A service like this could also function as a centre of knowledge about national access regulation and work to promote the ideals of open access and data sharing.

2.3 LONG-TERM STRATEGIC GOALS FOR RESEARCH INFRASTRUCTURES

Having identified major current and future needs for research infrastructures the next task at hand lies in the explication of major long-term goals and strategies that should become the targets or attractors for the process of European construction of research infrastructures in the area of the social sciences and humanities. In essence, three major goals can be identified that address the six different infrastructure needs. These goals can be formulated as:

- European Comparative Data and Modelling
- Data Integration and Language Tools
- Coordination and Enabling

The first strategic goal is linked to the need for European-wide data, the second goal encompasses the interoperability of data and languages and the third goal is directed towards the harmonization of data access policies, the standardization of digitalization processes as well as the interoperability between the social sciences and the humanities in general.

2.3.1 EUROPEAN COMPARATIVE DATA AND MODELING

More concretely, the first strategic goal lies in the specification of best practice standards for European data collections both for survey and for panel research. Here, the European Social Survey has already established itself as a European data infrastructure for surveys which has been explicitly founded on the premises of being the European standard in terms of data reliability, data quality and data comparability. A similar best practice standard has also to be created for panel data where currently SHARE offers the prospect of becoming the European model for highly reliable and comparative panel data. Finally, a substantial number of current comparative data programs like ISSP or Election Studies should be upgraded to the model of the ESS or to SHARE so that European data collection proceeds, as a long-term strategic goal, in a similar high quality fashion with a growing stock of best practice tools and instruments necessary to conduct European surveys or panels.

2.3.2 DATA INTEGRATION AND LANGUAGE TOOLS

Turning to the second strategic goal one can see two major tasks that are largely independent from each other. The first task lies in the field of data-integration where the long-term vision is to grant access to researchers from the social sciences and the humanities across Europe to all relevant nationally collected data as well as to all comparative data collections. The current model for this vision can be seen in the dissemination policies of the ESS which becomes available free of charge to the entire European research community as soon as the necessary data integration and data cleaning processes are completed.

Likewise, the second long-term task lies in the proliferation of tools and instruments that help to reduce the language barriers which inhibit currently the access to many national collections or to understand the available texts or documents linked.

This is a task that involves concerted efforts of research communities in language technology in all European countries over several years, which will yield tangible results in an incremental manner.

2.3.3 COORDINATION AND ENABLING

The third long-term target is probably the most difficult to accomplish and it lies in the enabling and in the overall coordination of, very broadly speaking, the interoperability of research infrastructures across national borders, across scientific fields as well as across language barriers in Europe.

Difficult as it may sound, a substantial number of steps must be undertaken within the next years towards the third long-term goal especially because the new information and communication technologies, if implemented in an intelligent pan-European manner, would allow for a giant leap towards harmonization, standardization and accessibility both of data and document collections across Europe

2.3.4 RECOMMENDED PROPOSALS AND STRATEGIC GOALS

The choice of recommended infrastructure projects reflects both the major infrastructure needs and the three strategic goals, outlined above. In essence, each of the three strategic goals has been provided with two new or upgraded infrastructure projects.

With respect to the first goal, the ESS and SHARE have been identified as necessary first steps since they are currently seen as best practice examples for comparative survey or panel studies. In relation to the second strategic goal, CESSDA and CLARIN have been regarded as contributing to data integration (CESSDA) and to overcoming the language barriers (CLARIN). Finally, the third and probably the most difficult strategic goal has been given two infrastructure projects (EROHS, DARIAH) which strive, on the one hand, towards common standards, harmonization processes and pan-European policies in the field of data (EROHS) and of digitalization (DARIAH) as well as, on the other hand, to enable the interoperability between the social sciences and the humanities irrespective of long-standing national or disciplinary traditions.

3 SOCIAL SCIENCES AND HUMANITIES ROADMAP WORKING GROUP - SSH RWG

3.1 MEETINGS HELD BY THE SSH RWG

1	January 7, 2005	Brussels
2	February 3, 2005	Oslo
3	April 29, 2005	Bergen
4	September 22, 2005	London
5	January 20, 2006	Brussels
6	April 19, 2006	Brussels
7	May 16, 2006	Brussels

3.2 MEMBERS OF SSH RWG

Member			Comments
Henrichsen	Bjørn	Chair, Norway	
Müller	Karl	Austria	
van Doninck	Bogdan	Belgium	
Krejci	Jindrich	Czech Republic	
Ploug	Niels	Denmark	
Sors	Andrew	EC representative	To November 2005
Corpakis	Dimitri	EC representative	From April 2006
Schmoltzer	Andrea	EC representative	To April 2006
Mustajoki	Arto	Finland	
Henriot	Christian	France	
Friedrich	Reinhold	Germany	To March 2005
Wagner	Gert	Germany	From April 2005
Vokos	Gerasimos	Greece	To March 2005
Kallas	John	Greece	From April 2005
Kenesei	Istvan	Hungary	
Bric	Maurice	Ireland	
Padoa Schioppa	Antonio	Italy	
Beinarovica	Baiba	Latvia	To December 2005
Kocere	Venta	Latvia	From January 2006
Petuchovaite	Ramune	Lithuania	
Schaber	Gaston	Luxembourg	
Ferring	Dieter	Luxembourg	Substitute for Schaber
Mikkelsen	Egil	Norway	
Marek	Tadeusz	Poland	
Manuel Mandes	Jose	Portugal	
Kranjc	Andrej	Slovenia	To December 2005
Vodusek-Staric	Jerca	Slovenia	From January 2006
Ackum	Susanne	Sweden	To February 2005
Åberg	Rune	Sweden	From January 2006
Calvo-Armengol	Antonio	Spain	From January 2006
Joye	Dominique	Switzerland	
Koppen	Jan Karel	The Netherlands	
Barkcin	Savas	Turkey	
Schurer	Kevin	United Kingdom	
Kiberg	Dag	Secretary	

3.3 LIST OF OPPORTUNITIES

The SSH RWG received the following proposals to the List of Opportunities:

- Visualizing Data for Education, Science and Society - VIDESS
- European Data Engine - EDEN
- European Language Acquisition Database - ELAD
- CESSDA Catalogue - C_Cat
- Representative behavioural experiments on economic behaviour and individual preferences - REEBIP
- Survey of Health, Ageing and Retirement in Europe - SHARE
- International Agency for Violence Research - IAVR
- Old age Provision Observatory for Migrants in Europe - OPOME
- A Europe-wide interactive, dynamic database and microsimulation model for empirical research on material well-being with special emphasis on the effects of public policies - COM-MONtools
- A European Virtual Centre for Comparative Survey Methodology - ECCM
- European Longitudinal Cohort Studies: From Birth to Employment - ELCS
- European Research Observatory for Cultural Change - EROCC
- European Network of Economic and Social Science Research Centers - EUNESS
- European Social Orders - ESO
- Corpus of Roman Findings in the European "Barbaricum" - CRFB
- International Database for European Archaeology – Literature - IDEA-L
- German Micro Data Service Centre - GEMS
- Institutional Reforms - European Observatory - IREO
- The European Social Survey - ESS
- Data Bank of the European Written Heritage – BewriH
- Interdisciplinary Research Library - Beic Project
- An International Legal and Judicial Repository
- EROHS European Research Observatory for the Humanities and Social Sciences
- Archives in the Digital Age - ADA
- Single Access to Digital Data and Documents in the Human and Social Sciences - ADONIS
- Central European Opinion Research Group Foundation - CEORG
- Challenges of the present: ethnographic and comparative studies of the modern cultural complexities

- European Heritage in Spatial Culture - EHSC
- European Digital Museum Project - EuDiMuP
- Generations and Gender Survey - GGS
- Humanities Access Grid Node - HAGNODE
- Centre for GIS Applications in the Humanities - HUMANGIS
- Research on economic, social and cultural sources of inequalities and determinants of life-success: common European strategy for participation in large scale international longitudinal studies
- Medieval European Libraries Network - MEDLIB
- Répertoire International des Sources Musicales - RISM
- Standardization of European Digital Text Resources - SEDiTeR

The SSH RWG decided to bring forward two projects to the ESFRI Joint Proposal:

- European Social Survey -ESS
- European Research Observatory for the Humanities and Social Science – EROHS
-
- Both projects were included in the final ESFRI proposal.

3.4 PRESENTATIONS OF RESEARCH INFRASTRUCTURE PROJECTS

The following research infrastructure projects have been presented in SSH RWG meetings:

- "The ADONIS project", Andrea Iacovella, France
- "Council of European Social Science Data Archives", Kevin Schurer, United Kingdom
- "The DANS project", Peter Doorn, the Netherlands
- "Trans-European Language Resources Infrastructure", Tamás Váradi, Hungary
- "The projects BewriH and Beic", Antonio Padoa-Schioppa, Italy
- "VIDESS", Karl Müller, Austria

Later on all projects under final consideration for the Roadmap was presented by a rapporteur.

3.5 EXPERT GROUPS

The SSH RWG proposed two Expert Groups (EG):

- European Cultural Heritage Expert Group – ECH EG, chair: Maurice Bric, Ireland
- European Research Observatory for the Humanities and Social Sciences (EROHS) Expert Group – EROHS EG, chair: Niels Ploug, Denmark

European Cultural Heritage Expert Group		
Bric, Maurice	School of History, University College Dublin	Ireland; Chaire
Menu, Michel	Palais du Louvre-Porte des Lions	France
Renn, Jürgen	Max Planck Institute for the History of Science	Germany
Fotakis, Costas	Institute of Electronic Structure and Laser, FORTH	Greece
Varádi, Tamas	Institute for Linguistics Research, Hungarian Academy of Sciences	Hungary
Granelli, Andrea	Telecom Italia Net	Italy
Doorn, Peter	DANS	The Netherlands
Jubb, Michael	British Library	UK
Douka, Maria		EC; observer
Conlon, Tim		
EROHS Expert Group		
Ploug, Niels	The Danish National Institute of Social Research	Denmark; Chair
Cermak, Frantisek	Faculty of Philosophy, Charles University	Czech
Boerch-Supan, Axel	Mannheim Research Institute for the Economics of Aging (MEA), University of Mannheim	Germany
Canny, Nicholas	Centre for the Study of Human Settlement and Historical Change, National University of Ireland	Ireland
Stefanizzi, Sonia	Faculty of Sociology, University of Milan-Bicocca	Italy
Kvalheim, Vigdis	Norwegian Social Sciences Data Services	Norway
Erjavec, Tomaz	Jozef Stefan Institute	Slovenia
Van der Knaap, G.A.	Erasmus Universiteit Rotterdam	The Netherlands
Özcan, Yusuf Ziya	Department of Sociology, Middle East Technical University	Turkey
Theofilation, Maria		EC; observer
Christensen, Lars	Secretary	

3.6 MAPPING OF POTENTIAL NEW (OR MAJOR UPGRADE) PAN-EUROPEAN RESEARCH INFRASTRUCTURES FOR ESFRI CONSIDERATION

The SSH RWG conducted a mapping of potential new (or major upgrade) pan-European Research Infrastructures for ESFRI consideration. Deadline for returning the questionnaire was set to November 10.

The motivation for the mapping was to have a broader input from then was available via the “List of Opportunities” (LoO). It was agreed in September 2005 by the SSH RWG and accepted by the ESFRI Executive, to send out a questionnaire to all members of the SSH RWG for a mapping of proposals for new or up-graded European RI within the SSH domain. The completed questionnaires together with the input to the LoO were to be reviewed by the two EGs.

The total numbers of received proposals were 97² including both proposals from LoO and the mapping. The proposals were divided between the EGs according to relevant fields. ECH EG reviewed the proposals for research infrastructures (RI) for the Humanities while EROHS EG reviewed the RIs for the Social Sciences.

² The actually number of received proposals were 109. Some of the proposals were however submitted both to the LoO and via the mapping.

4 METHODOLOGY USED BY EXPERT GROUPS

4.1 EUROPEAN CULTURAL HERITAGE EXPERT GROUP – ECH EG

4.1.1 INTRODUCTION

The European Cultural Heritage Expert Group (ECH EG) was established by ESFRI to identify and develop research infrastructures (RI) for the area. It is chaired by Dr. Maurice J. Bric (IE), MRIA, Chairperson of the Irish Research Council for Humanities and Social Sciences and a member of the Humanities and Social Sciences Roadmap Working Group (SSH RWG). The other members of the ECH EG were after nomination from ESFRI delegations. They were invited not to represent their national and private interest and agreed to stated protocols on potential conflicts of interest. Their names are listed in appendix 1.

The ECH EG had its first substantive meeting on 23 November 2005. It also met on 12 January, 8 March and 3 April 2006.

4.1.2 DEFINITIONS

In general terms, “infrastructures” was taken from the working language of the European Commission (2005) as referring to facilities, resources or services that are needed by the research community to conduct research in all scientific and technological fields. This definition covers:

- major equipment, or sets of instruments, used for research purposes
- knowledge-based sources such as collections, archives, structured information or systems relating to data management, used in scientific research
- enabling information or communication technology-based infrastructures, such as GRID, computing, software and communications, and,
- and other entity of a unique nature that is used for scientific research³.

4.1.3 CRITERIA

The ECH EG applied the criteria as set by ESFRI that proposals should:

- have a scientific case
- provide a clear business case

- be timely, viable and sustainable
- reflect a real need for the development of the humanities in Europe
- be supported by the appropriate scientific community at the European level
- be of pan-European interest
- offer possibilities for European partnership, including commitment from major stakeholders
- be multi-user facilities offering an open access, physical or virtual, for scientists from all over Europe
- be relevant at international level.

The ECH EG was further asked to distinguish between proposals which were

- mature
- embryonic
- immature

4.1.4 PROCESS

The ECH EG considered two sets of questionnaires. The first of these followed a call that had been made for the Mapping exercise that was made during the early months of 2006 (listed by acronym in 5.1). The second had been assembled earlier for the List of Opportunities of 2004. The two sets of proposals were referred to the ECH EG by the Secretariat of the WPHSS. The actual questionnaires are on file with the Secretariat. Only those questionnaires being recommended by the ECH EG are appended to this report.

Following a preliminary discussion on 23 November 2005, and as agreed at its meeting of 12 January 2006, the ECH EG assessed the questionnaires according to the following rubric:

Mature

- “A” an application which displays clear maturity and scientific excellence
- Embryonic
- “B” an application which suggests early stage, that it is recently established and not yet mature. Such proposals were expected to grow over time and could serve as propositions for future development or support

New

- “C” an application which is relatively new or immature, or reactive, in the sense that the proposal was written in response to the call for proposals, and not much else.

³ As from papers for WGHSS, 2 September 2005.

- “D” an application which was new but could benefit by association with other and similar proposals.

Not Relevant

- “E” an application which was not considered relevant to the remit of our exercise, or provided information which, in the opinion of the ECH EG, was insufficient or incomplete.

4.1.5 SUBMITTED QUESTIONNAIRES

35 proposals were included for the *Mapping Exercise* and were discussed according to the rubric described in 4.2 above. These are here listed by their acronyms or where none such has been provided, by their proposers. Those proposals, which are recommended for further consideration, are highlighted in bold type. The letters C, D, or E indicates those, which were rejected. The following were discussed:

Visualizing Data for Education, Science and Society: VIDESS	C
Croatian Language On-Line Repository	D
Glagolic Script	D
Hyper Learning	E
European Research Observatory for Cultural Change: EROCC	Referred to EROHS Expert Group
Research Infrastructures for Cultural Heritage: RICH	A
Node for Secondary Processing: NSP	Referred to EROHS Expert Group
Humanities Bibliographic Database: HBD	C
European Database for Musicology and Choreology	C/D
Irish Virtual Research Library and Archive: IURLA	C
(O Cathain)	D
Medieval European Libraries Network : MEDLIB	B
(Douglas)	C
(Fennell)	C
Wetland Heritage and Environment Network: WHEN	C
Irish Humanities and Social Sciences Research Database	C
European Network for Communicating Cultural Heritage Values	C
Digitized Data Base	A/B
Interdisciplinary Research Library: BEIC	D/E
Hanseatic Historical Archives Network	B
Wittgenstein Archives: WAB	B
Medieval Nordic Text Archive: MENOTA	B
Hamsun in Context	E
Language Data Corpus and Translation and Language Training Centre	C
Trans-European Language Resources Infrastructure: TELRI	A
Swiss Theatre Collection	D
Institute of Dialect and Folk Life Studies	D
Herausgabe des Historischen Lexicon der Schweiz: HLS	D
NCCR Mediality. Historical Perspectives	E
The European E-Reference Collection for Cultural Heritage: ARTeFACT	B?
LangWeb	A
(Mandemakers)	Referred to EROHS Expert Group
(Van Dijk)	Referred to EROHS Expert Group
Data Infrastructure for the Humanities and Social Sciences: DISH	A
Europaen Arhive for Language Resources: EARL	A
European Minority Language Library: EMILL	C
A Website Address for a Digital Research Library for European Culture	D

37 proposals were included from the *List of Opportunities*. Some of these had been re-submitted from the Mapping Exercise; these are indicated by asterisk (*). Those relating to the social sciences and as such,

deemed to be outside the remit of the ECH EG, or more appropriate to the EROHS Expert Group, are indicated by a double asterisk (**). Those which were considered by the ECH EG are listed in bold type:

Visualizing Data for Education, Science and Society: VIDESS**	
European Data Engine: EDEN**	
European Language Acquisition Database :ELAD	C
Council of European Social Science Data Archives: CESSDA/ C-CAT**	
Representative Behavioural Experiments on Economic Behaviour and Individual Preferences: REEBIP**	
Survey of Health, Ageing and Retirement in Europe: SHARE**	
International Agency for Violence Research: IAVR**	
Old Age Provision Observatory for Migrants in Europe: OPOME**	
COMMONtools**	
A European Virtual Centre for Comparative Survey Methodology: ECCM**	
European Longitudinal Cohort Studies: ELCS**	
European Research Observatory for Cultural Change: EROCC*	C
European Network of Economic and Social Sciences Research Centers: EUNESS**	

European Social Orders: ESO**	
Corpus of Roman Findings in the European “Barbaricum” : CRFB	C
International Database for European Archaeology - Literature	C
German Micro Data Service Centre: GEMS**	
Institutional Reforms – European Observatory: IREO**	
European Social Survey: ESS**	
Data Bank of the European Written Heritage: BEWRIH	C ?
Interdisciplinary Research Library: BEIC*	C/D
An International Legal and Judicial Repository: SCIL**	
European Research Observatory for the Humanities and Social Sciences: EROHS**	
Archives in the Digital Age: ADA	C
Single Access to Digital Data and Documents in the Human and Social Sciences: ADONIS**	C
Central European Opinion Research Group Foundation: CEORG**	
Challenges of the Present: Ethnographic and Comparative Studies of Modern Cultural Complexities	C
European Heritage in a Spatial Culture: EHSC	C
European Digital Museum Project: EDiMuP	C
Humanities Access Grid Node: HADNODE	C
Centre for GIS Applications in the Humanities: HUMANGIS	C
Research on Economic, Social and Cultural Sources of Inequalities and Determinants of Life Sciences: Common European Strategy for Participation in Large Scale International Longitudinal Studies **	
Medieval European Libraries Network: MEDLIB	B
Répertoire International des Sources Musicales: RISM	C
Standardization of European Digital Text Resources	
Digital Text Resources: SEDiTeR	C

4.1.6 RECOMMENDATIONS: MATURE PROJECTS

4.1.6.1 First Step

Of the listed proposals, 6 were deemed to be appropriate for preferential treatment in the sense that, in the opinion of the ECH EG, and on the basis of the information supplied, they met the criteria as listed for mature projects.

4.1.6.2 Second Step

In order to ensure that this would be the case, as well as to further explore the detail of the proposals, the ECH EG invited the *rapporteurs* of each of the 6 projects to give individualised presentations at a special meeting on 8 March 2006. As a result of this process, the ECH EG proposed, and the relevant *rapporteurs* agreed, that the best interests of promoting research infrastructures in Cultural Heritage would be served by addressing overlap in some of the proposals and as a result, associating these proposals as follows:

- CLARIN (as the developed association of TELRI, LangWeb, EARL and the Digitized Data Base)
- DARIAH (as clarified from DISH)
- EURICA (as developed from RICH)

4.1.6.3 Third Step

Following detailed discussions and the incorporation of further detailed suggestions, especially in relation to business plans, financial flows, and the clarification of matters in relation to pan-European value and maturity, the ECH EG concluded as follows:

4.1.6.4 RECOMMENDATION ONE: Mature Projects

The European Cultural Heritage Expert Group recommends that the following three mature projects

be included in the Roadmap. Each of these represents distributed facilities. They are either based on existing European infrastructures which require major up-grades, and new functional nodes, or on national facilities which should be joined to enhance pan-European distributed facilities. These mature infrastructure projects are:

- CLARIN
- DARIAH
- EURICA

The requested template for each proposal is enclosed to this report.

4.1.7 RECOMMENDATIONS: EMBRYONIC PROJECTS

4.1.7.1 RECOMMENDATION TWO: Embryonic Projects I

Having regard to the ESFRI criteria, the ECH EG considered that the following were of great potential importance but were immature at this stage. However, it strongly felt that they should be recommended as embryonic proposals, provided that in each case, they should provide the type of detail as in those projects which have been recommended above as mature, including the provision of a clear business plan, statement of pan-European value, and scientific value. The ECH EG observed that in some cases, the information submitted from the List of Opportunities was vague.

The relevant projects are:

- CRFB

- MEDLIB on the further understanding that the relevant *rappporteur* should associate his/her project with IDEA-L

4.1.7.2 RECOMMENDATION THREE: *Embryonic Projects II*

The ECH EG also recommends that the following should be considered as embryonic proposals but because of seeming matters of scale, in a lesser light than in the case of Recommendation Two.

- The Hanseatic Historical Archives Network, in association with the Wittgenstein Archives (WAB) and MENOTA
- ARTeFACT

4.1.7.2.1 General Observations

The ECH EG is aware of the new and special importance which in recent months, European policy makers and funders have been attaching to the Humanities and Cultural Heritage. It welcomes this attention. However, it also recognises that while some humanities infrastructures are being, or have been, developed in recent years, the area does not have the same co-ordination as other areas of research. Accordingly, while the ECH EG took the task which was entrusted to it by ESFRI, this task would have been facilitated if more time had been given to mobilise relevant and interested parties. It also observed that some areas which were part of its remit, were not reflected in the questionnaires. Nonetheless, the ECH EG is satisfied that even within these and other limits, it can recommend three strong projects which can promote research infrastructures for projects in cultural heritage.

It also observed that many of the other proposals had within them the basis for further elaboration and refinement and would encourage them to pursue their proposals with a view to engaging with the ESFRI process at a future stage.

4.1.7.3 RECOMMENDATION FOUR

The ECH EG strongly recommends that the ESFRI process accord the Humanities real as well as rhetorical support in the current phase of the Mapping Exercise.

4.1.7.4 RECOMMENDATION FIVE

The ECH EG strongly recommends that an Expert Committee on Cultural Heritage be retained as part of the ESFRI process.

4.1.7.5 RECOMMENDATION SIX

The ECH EG strongly recommends that a more efficient system be devised to publicise the calls for questionnaires and to involve the scientific communities in its business.

4.2 EUROPEAN RESEARCH OBSERVATORY FOR THE HUMANITIES AND SOCIAL SCIENCES (EROHS) EXPERT GROUP – EROHS EG

4.2.1 INTRODUCTION

The Expert Group on the European Research Observatory for the Humanities and Social Sciences (EROHS) was established in the autumn 2005. It was chaired by Niels Ploug, the The Danish National Institute of Social Research, a member of the SSH RWG. The Group was set up with two interrelated tasks:

- Given its name to specify and substantiate the proposal for the setting up of EROHS
- To evaluate the proposals for new and upgraded pan-European research infrastructures in the social sciences

The remit of the Expert Group was thus to cover both humanities and social sciences in its deliberations. Consequently it was not set up as an exclusively social science enterprise. This acknowledgement is important when evaluating the outcome of the Group's deliberations.

The responsibility of the Expert Group is also obvious when considering the composition of the Group as it consisted of members with expertise both within the humanities and social sciences.

The Expert Group has held five meetings in the period since its inception. An introductory meeting 2 September 2005 and four substantive ones - 3 November 2005, 11 January 2006, 7 March 2006 and 28 March 2006 and what follows is a short report presenting the outcome of the Expert Group's deliberations and work.

4.2.2 SUBSTANTIATION OF THE EROHS-PROPOSAL (TASK 1)

4.2.2.1 *Background*

In May 2004 a working group handed in a report to the European Strategy Forum for Research Infrastructures (ESFRI) with a proposal to establish a "European Research Observatory for the Humanities and Social Sciences (EROHS)".

The report argues that the humanities and social sciences are hampered by a raft of problems caused by the current state of research infrastructures at European level. This has severe consequences for the possibilities of conducting comparative European empirical research. Against this background a European strategy is proposed that addresses the current problems.

The first report was a blueprint, setting the stage for the further specification of EROHS. The Expert Group takes over where the first report on EROHS ended. Whereas the first report primarily focused on '*science case*' for EROHS, this work of the EROHS Expert Groups concentrated on the '*business case*' of EROHS.

As a consequence, the enclosed report has taken as its backdrop the general recommendations from the first report for granted and given, and will concentrate on the aspects of implementing EROHS – the organisational principles and platform, governance system, financial structure and work programme.

4.2.2.2 Methodology

The method applied was very much to study the success and features of other infrastructural enterprises – and first of all the flagships within the confines of the humanities and social sciences. The recent years' success of e.g. Council of European Social Science Data Archives (CESSDA), European Social Survey (ESS), and many more serves as ready inspiration. As does many national efforts within the composite disciplines. The current experiences and collaborations clearly illustrate both the readiness and determination of the scientific communities to cooperate and the ability to do so. Indeed, an unprecedented opportunity exists today to take advantage of these enterprises by taking research infrastructure collaboration to a level that could not have been foreseen just a few years ago.

Facilities within other sciences also inspires, however is also clear from the outset that EROHS cannot be straightforward replica of the large scale single-sited facilities so successful in the natural sciences. However also within other sciences exist the kind of distributed facilities that reminds very much of EROHS. To mention just two examples the Global Biodiversity Infrastructure Facility (GBIF) can be put forward as can the International Neuroinformatics Coordinating Facility (INCF) both born out of OECD. Both have great resemblance to EROHS as they origin from the scientific communities arguing for a common integrated, yet distributed facility.

4.2.3 THE EVALUATION OF PROPOSALS FOR NEW AND UPGRADED PAN-EUROPEAN RESEARCH INFRASTRUCTURES IN THE SOCIAL SCIENCES (TASK 2)

4.2.3.1 The proposals

The EROHS Expert Group was presented with two sets of questionnaires for new and updated research infrastructures in the humanities and social sciences. The first set of questionnaires was collected by ESFRI for the compilation of a first List of Opportunities in the autumn of 2004. The second set was collected on a basis of a call for new and upgraded re-

search infrastructures by the SSH RWG, autumn 2005. There proved to be some overlaps in the proposals received from the two exercises.

All together the EROHS Expert Group evaluated 54 proposals over several meetings and agreed and coordinated with the other RWGHSS Expert Group – Expert Group on Cultural Heritage (EGCH) – to leave 34 proposals for their consideration only.

4.2.4 GUIDELINES FOR THE EVALUATION

Every proposal has been evaluated according to the two general criteria as they are presented in the procedural guidelines for the Expert Groups:

- The Scientific Case (=Pan-European)

The proposed new RI should correspond to future needs of the scientific communities in Europe, demonstrate impacts on scientific developments, support new ways of doing science in Europe and participate to the enhancement of the European Research Area. It should be supported by the appropriate scientific community at European level, should demonstrate its pan-European value, setting the scene for the infrastructure in a European and an international context, as well as its relevance and quality.

- The Concept/Business Case (=Maturity)

The proposed new research infrastructure should be technologically and financially feasible and meet the necessary degree of maturity which is defined as (a) the existence of a technical concept for the realisation of the project, and of feasibility studies, including identification of technical challenges and risks, (b) the existence of a projection about construction, operating and decommissioning costs, including a clear timetable.

Additionally the Expert Group were asked to evaluate the proposals according to the following classification:

- Mature proposals
- Embryonic proposals
- Immature proposals

Each member of the Expert Group evaluated the proposals based on the above-mentioned criteria. The result of this process was discussed at one of the last EG meetings and resulted in a categorisation of proposals according to the three criteria mentioned above.

Lastly but not least it is important to notice that the Expert Group did focus its discussions on what it found to be mature proposals. The division below between embryonic and immature proposals is therefore less validated and the Expert Group recommends that none of the many very interesting proposals that was received during the mapping procedure and that are not listed as mature in this short

report at this stage should be excluded from further consideration in future work of establishing research infrastructure for the humanities and social sciences.

The major conclusion of the EG is thus that the mature proposals mentioned below should be considered for the first roadmap for the humanities and social sciences.

Below are each proposal listed by its acronym or alternatively by its full name and are categorised into mature, embryonic and immature proposals and finally the proposals left for the EGCH to evaluate.

4.2.5 MATURE PROPOSALS

The Expert Group has identified the following four mature proposals according to the above criteria:

- EROHS. European Research Observatory for the Humanities and Social Sciences
- ESS. The European Social Survey
- SHARE. Survey of Health, Ageing and Retirement in Europe
- CESSDA / C_CAT. Council of European Social Science Data Archives

As mentioned above given the name of the Expert Group, EROHS has been given special attention by the Group and was consequently both addressed under task 1 and 2. It is worth noticing that in the contest with the received proposals under the mapping-exercises, the EROHS-proposals is ranked as mature.

The requested template for each proposal is enclosed to this report.

4.2.6 EMBRYONIC PROPOSALS

The Expert Group found that the proposals listed below falls in the category embryonic proposals. Currently they are not sufficiently mature and/or pan-European, but they have the clear potential to become so in the longer run. Compared to the mature proposals they in general fall below on parameters as European-wide access to data and quality insurance. This judgement has in some cases and areas been difficult in comparisons with some of the mature proposals. The EG will there support that the embryonic proposals are taken seriously into consideration in the continued work of the SSH RWG in the coming years.

The proposals identified as embryonic are:

- ISSP. The International Social Survey Programme
- GGP. Generation and Gender Programme
- CSES. The Comparative Study of Electoral Systems

4.2.7 IMMATURE PROPOSALS

The following list contains the proposals that the Expert Group decided not to put in contention as the were not considered sufficiently pan-European or mature:

- ECCM. A European Virtual Centre for Comparative Survey Methodology
- ELCS. European Longitudinal Cohort Studies
- EROCC. European Research Observatory for Cultural Change
- ESO. European Social Orders
- GEMS. German Micro Data Service Centre
- OPOME. Old Age Provision Observatory for Migrants in Europe
- REEBIP. Representative Behavioural Experiments on Economic Behaviour and Individual Preferences
- EUNESS. European Network of Economic and Social Sciences Research Centers
- COMMONtools
- IDEA-L. International Database for European Archaeology - Literature
- IREO. Institutional Reforms – European Observatory
- EDEN. European Data Engine
- An International Legal and Judicial Repository
- ADA. Archives in the Digital Age
- ADONIS. Single Access to Data and Documents in the Human and Social Sciences
- CEORG. Central European Opinion Research Group Foundation
- Challenges of the Present: Ethnographic and Comparative Studies of the Modern Cultural Complexities
- EHSC. European Heritage in Spatial Culture
- ELAD. European Language Acquisition Database
- EuDiMuP. European Digital Museum Project
- HAGNODE. Humanities Access Grid Node
- HUMANGIS. Centre for GIS Applications in the Humanities
- Research on Economic, Social and Cultural Sources of Inequalities and Determinants of Life-Success: Common European Strategy for Participation in Large Scale International Longitudinal Studies
- RISM. Répertoire International des Sources Musicales
- SEDiTeR. Standardization of European Digital Text Resources
- IAVR. International Agency for Violence Research
- Electroencephalogram (EEG) and Transcranial Magnetic Stimulation (TMS/Single Pulse)
- IZA Data Service Center
- FDZ-BA. The Research Data Centre of the Federal Agency of Labour

- FDZ-Bund. The Research Data Centre of the Federal Statistical Office
- FDZ-Länder. The Research Data Centre of the Statistical Offices of the Länder
- ZA. Central Archive for Empirical Social Research, University Cologne
- FDZ-RV. The Research Data Centre of the German Statutory Pension Insurance
- ZEW. The Centre for European Economic Research
- IFO. Institute for Economic Research at the University of Munich
- IZ. Social Science Information Centre
- GESIS Service Agency Eastern Europe
- ZUMA. Centre for Survey Research and Methodology
- SOEP. The German Socio-Economic Panel Study
- Research Encompasses All Aspects of Past and Present Expressions of Folklore and Folk-life
- SIDOS. Data Archive for the Social Sciences in Switzerland
- HSE. The Historical Sample of Europe
- CCWW. Crosscultural Women's Writing
- The European Facility Life Course
- EUROMESS. European Infrastructure for Measurement and Experimentation in the Social Sciences.
- RIEPA. Research Infrastructure for European Panel Analysis

- European Centres for Analysis in the Social Sciences
- ETUS. European Time Use Study
- Methods for Research Synthesis Centre

4.2.8 RECOMMENDATIONS

On the basis of its deliberations the EROHS Expert group would like to put forward the following deliberations:

That the following four proposals identified as mature are included in the roadmap. The four proposals were:

- EROHS. European Research Observatory for the Humanities and Social Sciences
- ESS. The European Social Survey
- SHARE. Survey of Health, Ageing and Retirement in Europe
- CESSDA / C_CAT. Council of European Social Science Data Archives

That at this stage and on the basis of the current information coming from the mapping-exercises a too fixed distinction between embryonic and immature proposals is dangerous. The Expert group recommends that none of the proposals ranked as embryonic or immature should be rejected from future considerations.

5 RECOMMENDED PROPOSALS FOR NEW AND UP-GRADED RESEARCH INFRASTRUCTURES

5.1 RECOMMENDATIONS

Based on the assessments from the two Expert Groups, and discussions in the Group the Social Sciences and Humanities Roadmap Working Group (SSH RWG) recommend the following proposals for new (or major upgrade) pan-European Research Infrastructures for ESFRI consideration.

Template for each proposal follows in appendices.

5.1.1 MATURE PROPOSALS

The proposals recommended as mature cover different infrastructural objectives and functions. To reflect this and to give a more perceptive picture of these objectives, the recommended proposals are grouped under three headings (see chapter 2).

5.1.1.1 *European comparative data and modelling*

- ESS
- SHARE

5.1.1.2 *Tools for data access, sharing and integration*

- CESSDA
- CLARIN

5.1.1.3 *Data preservation, enabling and cooperation*

- EROHS
- DARIAH

5.1.2 EMBRYONIC PROPOSALS

- ISSP
- GGP
- CSES
- CRFB
- MEDLIB
- The Hanseatic Historical Archives Network, in association with the Wittgenstein Archives (WAB) and MENOTA
- ARTeFACT
- EURICA

The ECH EG proposed the EURICA project as a mature project. After a presentation of the project to the SSH RWG and thorough discussions in the Group it was decided to include the project as an embryonic proposal to ESFRI.

The two Expert Groups received a large number of proposals for assessments. The SSH RWG is however fully aware of the narrow selection of Euro-

pean research infrastructure for the social sciences and humanities these proposals constitute and will recommend strongly that ESFRI in later rounds will try to cover a broader area of the Humanities and Social Sciences.

5.1.3 CONFLICT OF INTERESTS

As respectively President and Vice-President of CESSDA, Kevin Schurer (UK) and Dominique Joye (Switzerland) were not present and did not take any part when decision was made on CESSDA.

As a member of the Central Coordination Team for ESS, Bjørn Henrichsen (Chair SSH RWG, Norway) was not present and did not take any part when decision was made on ESS.

5.2 EUROPEAN SOCIAL SURVEY – ESS

The European Social Survey (ESS)

The facility: The ESS was set up in 2001 to monitor long-term changes in social values throughout Europe and produce data relevant to academic debate, policy analysis and better governance. It now covers 27 European countries. A long-term pan European instrument such as the ESS requires long-term funding commitments. A major upgrade is now sought to fill debilitating gaps in the present programme.



Background: The ESS has a complex network of management and advisory committees, representing national teams and funders on the one hand, and academic specialists on the other. It covers the whole of the EU (apart from Latvia, Lithuania and Malta), and includes both associated countries and a number of accession and candidate countries. It was built as a

multi-funded enterprise. Its costs have been shared between the EC, the ESF, and 27 national academic funding bodies. Two-thirds of the ESS is now provided by the nations and one third from the Commission. ESS data are made publicly available on the web as soon as they are available – with no prior "privileged" access. This makes the publication of each dataset a major event in the European social science calendar.

What's new? Which impacts? The purpose of the proposed major upgrade is to unify, regularize and secure the funding for the RI as a whole over an extended period, though naturally with periodic reviews. A large and complex time series such as the ESS requires such continuity of funding, which is a pre-requisite of appropriate planning. But a major upgrade would also help to fill debilitating gaps in the pre-sent programme of work – allowing much-needed new programmes of work on:

- compiling and harmonising aggregate context variables for survey analyses
- experimenting with alternative (technical and traditional) methods of translation to improve equivalence
- investigating and mitigating longstanding problems in the collection and classification of occupation and education
- improving the capacity to pilot and pre-test new questions on emerging issues of public concern
- experimenting on a multinational basis with methods of improving response rates

All this work would be in addition to designing and coordinating the biennial ESS and to the conduct of fieldwork, coding and keying in some 30 European nations.

Timeline for construction and first operation with related estimated costs: The total annualised cost of the ESS infrastructure at present, combining all sources of finance amounts to around €6m per year. The major upgrade proposed would bring the total annual costs to around €9m per year (a 50% overall increase). If the ESS were to assure funding for three further rounds, including its infrastructure responsibilities, the overall commitment would be around €54m over 6 years.

Leading consortium:

ESS has adopted a careful balance between 'top down' and 'bottom-up' elements in its organisational structure. Thus, a structure has been adopted that combines tight central co-ordination with strong participation from all participating countries, and independent advice from teams of independent academics.

5.2.1 INTRODUCTION

The **European Social Survey (ESS)** has now been in existence for five years. It has three main aims:

- To monitor long term change in social, political and cultural values within and between European nations, especially in relation to Europe's changing political and social institutions
- To improve the accuracy of comparative social measurement in Europe and beyond as an aid both to more informed analysis of social change and ultimately to better governance
- To provide the material and methods via which to develop robust social indicators that will measure and compare improvements or otherwise in the quality of life of European citizens

The ESS thus combines substantive, methodological and policy objectives, each of which it has demonstrably been fulfilling since its inception (under Framework 5) in 2001. During that period it has, of course, undergone a variety of peer-reviewed assessments and has on every occasion been successful. These include:

- referees' assessments in respects of each of four *separate* applications over the five years to EC Framework programmes for successive rounds of the ESS time series (Rounds 1-4)
- a separate peer-reviewed application in 2005/2006 for Framework 6 Infrastructure support (i3), one of the first social science projects ever to win such support
- regular support over all three rounds to date (together with substantial pump-priming funds) from the European Science Foundation
- the first ever successful bid by a social science project for the coveted Descartes Prize, Europe's top science prize awarded annually "for excellence in collaborative scientific research"
- separate independent applications in 27 separate nations to their respective national academic funding bodies for short or longer term support for the ESS in that country, resulting in an extraordinarily high degree of continuity from round to round

So it is by no means just our own claims about the long-term value of the ESS that persuades us to seek ESFRI support. On the contrary, the project has consistently cleared every funding and critical hurdle placed in its path since its inception, overcoming on the one hand the fiercely competitive environments

of national academic funding councils in rich and poor countries alike and, on the other, the longstanding doubts about the value of the social sciences as a whole among some of the natural and physical scientists who made up the Descartes Grand Jury.

It would be comforting to believe that this state of affairs can continue. But it cannot. The future of the ESS is almost certainly unsustainable in the long term under its present funding regime. A time series of this scale and diversity cannot realistically rely on a series of nearly 30 independent and largely discrete funding decisions at each biennial round. And even though the ESS's round by round funding regime has recently been supplemented (for the next five years) by some Infrastructure support from the Commission, this support explicitly excludes the core business of the ESS – which is the design, coordination and data collection associated with each biennial round. The present infrastructure support is solely for outreach activities and major methodological refinements and advances. So for the moment the continuity of the ESS proper still has to depend wholly on the success or otherwise of its bids under STREP (or equivalent) instruments in successive EU Framework Programmes, as well as on the vagaries of countless individual national funding decisions.

So, however popular or successful the ESS may be, its future as a time series depends critically on being included as an enhanced infrastructure on ESFRI's Roadmap. Its omission from the Roadmap would, in effect, almost certainly create severe problems for what is a mature and demonstrably thriving project, whose data are in great demand and whose methods are widely admired, not only in Europe. (Even the traditionally stand-offish US National Science Foundation has now approached the European Commission to request the presence of US scholars at upcoming ESS workshops, seminars and conferences.) So the omission of a flagship project such as the ESS from the ESFRI Roadmap would cause a substantial stir within the wider social science community, whose members – whether academics or policy makers and almost regardless of discipline - have for once given near universal endorsement to a specific high-profile project. In conjunction with their natural science counterparts, they have backed the project's blend of good science with good methodology, recognising that it contributes too to better governance and a better understanding of how Europeans view their world and themselves.

Now, the very appearance of ESFRI on the horizon has created high expectations that the fragility of the present funding arrangements for the ESS might soon be remedied and that its status as a research infrastructure for the future will be secured and enhanced. It would be a bitter blow if the ESS's very success to date turns out to be partly responsible for its omission from the Roadmap on the false

grounds that its continuation is somehow magically assured.

The ESFRI document (Procedures for the Expert Groups on Preparing the European Roadmap) identifies two criteria for evaluating proposals for major upgrades or new infrastructures - the ‘scientific case’ and the ‘concept case’. Within each of these two categories are a number of specified sub-categories. So we refer to each of these main and sub- categories below in presenting the ESS’s case for a major upgrade.

5.2.2 ‘THE SCIENTIFIC CASE’ FOR THE ESS

The ESFRI document refers under this heading to ‘the needs of the potential user scientific community (ies) within the next 10-20 years’. It provides six factors against which these criteria should be evaluated, which we summarise as follows:

- the infrastructure must be relevant to Europe’s future scientific needs
- it must have demonstrable influence on scientific developments
- it must support new ways of conducting scientific research in Europe
- it must contribute to the enhancement of the European Research Area
- it must produce pan-European added value

We are pleased to report that the ESS is able to tick all these boxes.

As noted, the principal substantive aim of the ESS is to monitor long term change in social, political and cultural values within and between European nations, especially in relation to Europe’s changing political and social institutions. The European Commission and European Parliament are together responsible for a more diverse mix of nations, religions, cultures and languages than is any other democratic institution in the world. And the expansion of democracy in Europe, followed by the expansion of the EU itself, has been rapid and widespread.

Thanks largely to the efforts of Eurostat and the major National Statistical Institutes, we now know more than ever about the *characteristics* of the EU’s population, such as the ethnic composition of the population in different countries, or the make up of their respective labour forces, and so on. But we still know surprisingly little about the *character* of the EU’s population, such as the importance that attaches to democratic or humanitarian values in different countries, or whether, where, to what extent and with what consequences religiosity is rising or falling. Similarly, while we can chart falling turnout in elections, our knowledge of reductions in political trust is at best patchy. And while we know about trends in crime, we still know much less about the growth or

fall in fear of crime and the consequences this has for the quality of life in different areas.

The case for more and better data on these aspects of society is overwhelming. As Mohler and Wagner have said,⁴ cross-section surveys such as the ESS are needed by academics for a range of purposes, such as “checking distributions of characteristics and screening for untypical or counterfactual patterns on an aggregate level”. They go on to argue for an enlarged ESS not only because the present ESS is “not yet sustainably financed”, but also because it does not have enough space for a range of demands upon it. The explosion of use of the ESS among the wider social science community is testimony to this argument. There is now an expanding base of over 8,000 registered users of the ESS data website, around forty percent of whom have downloaded the data for statistical analysis. Meanwhile, seven books based on the ESS are either in print or in press, and countless articles and papers have already been published or are in preparation.

So the scientific relevance and impact of the project is amply demonstrable. In addition, the impact of the ESS’s methodology on comparative survey practice has already been profound – with several nations having not only arrested but reversed a precipitous downward trend in survey response rates. Despite the best endeavours of the ESS and other major projects, however, serious weaknesses still exist in comparative social research that the European Commission above all needs to see remedied. More informed European governance depends upon it. As we will argue, an enhanced ESS infrastructure will enable us to make further methodological inroads into these very difficult areas of social measurement.

5.2.3 ‘THE CONCEPT CASE’ FOR THE ESS

The ESFRI document refers under this heading to the ‘technological and financial feasibility’ of the infrastructure and its ‘degree of maturity’. It provides six factors against which these criteria should be evaluated, which we summarise as follows:

- it must demonstrate that it can overcome its technical challenges
- it must have a clear plan and timetable of action
- it must have demonstrable peer group support
- it must be capable of sharing risks and costs across nations
- it must provide access to others
- it must ensure that the human investment in the infrastructure produces high dividends

⁴ Mohler P and Wagner G, September 2004, Paper to the RISSH Group, “Foundations for Continuous Measurement on a European Level Using Survey Technology”.

Once more, we are able to report that the ESS is able to tick all of these boxes.

As noted, the ESS is a mature project that has successfully tackled its technical, organisational and managerial challenges. It has enjoyed unrivalled peer support not only from the academic social science community, but also from the policy community and, unusually, from the natural science community too. Its costs to date have been shared between the EC, the ESF and 27 national academic funding bodies. It covers all but three EU countries (only Latvia, Lithuania and Malta are missing for the moment) and includes associated countries (Norway, Switzerland, Iceland, Israel) and a number of accession and candidate countries (Bulgaria, Romania, Turkey). However, the ESS is and always has been open to *all* European countries that are capable of meeting its high technical requirements. So, for instance, Russia and Ukraine are already participating and several others are attempting to overcome the present financial barriers of joining – trying to raise sufficient finance to cover their national costs.

The dividends of the ESS are already proving to be substantial. As noted, its outputs (substantive and methodological) are far-reaching and are influencing worldwide audiences. Policy analysts, journalists, academics and students are all among the 9,000 users to date and we are glad to report that the ESS is being used increasingly in universities both for teaching and research purposes. Students in particular are growing as a proportion of all users.

The wider international interest in and impact of the ESS is already apparent. A foundation in the US has funded a team at Georgetown University to design and carry out a nationwide US replica of part of the ESS to facilitate intercontinental comparisons of values. And a similar initiative at the Australian National University is now underway.

5.2.4 WHAT WILL AN ENHANCED ESS INFRASTRUCTURE COMPRISE?

Despite its successes, the ESS remains a fragile and hybrid entity, funded from 30 independent sources and presently costing a total of around €6m per year over the next five years (ie while it has i3 support in addition to its core support). The bulk of this cost (around €4m per year) presently comes from national sources.

We hope that the ESFRI Roadmap will now recommend that the ESS becomes a more secure, better-funded, more unified and more centralised infrastructure than hitherto. Only then can the ESS incorporate important new features which the present uncertain and largely episodic funding regime does not permit.

One key element of this enhancement would be to enable the biennial multinational survey that is at the heart of the ESS to become part of the same structural arrangements as an expanded experimental, developmental and outreach programme that – in a reduced form – is presently at the heart of the i3 for the next five years. Such a development would represent a great deal more than a mere re-organisation. It would not only make the ESS infrastructure secure – which is a prerequisite for a multinational time series of this magnitude – but it would also make it more inclusive, comprehensive and influential.

We cannot in a short document of this nature provide the sort of detail that a full peer-reviewed research proposal would contain. So we provide below a summary of what we have in mind for the enhancement.

5.2.4.1 Fusion and central funding

The ideal model for an enhanced ESS infrastructure would involve central funding of all its activities over a period of, say, six years in the first instance. Then, instead of over 30 potential funding sources, each with uncertain long-term commitments (as at present), there would be one central source of funds over a sustained period with regular reviews. This does not by any means preclude national funding, which needs to continue in one form or another so that participating nations continue to feel ownership of the project. But national funding could be pooled with central funding for the ESS infrastructure, enabling national fieldwork and coordination costs to be covered from the centre, but with the possibility that some countries are larger proportional contributors than others.

The precise arrangements will of course be the object of discussion and fine-tuning between the interested parties – notably the Commission, the ESS Funders' Forum and the European Science Foundation – in consultation with the ESS Central Coordinating Team. But the principle would be that a substantially greater proportion of funding than now would come from the centre, while maintaining the model of the European Research Area which involves, among other things, joint funding from the centre and the periphery.

We should stress that these proposed changes are by no means just for administrative convenience or ease of operation. On the contrary they are based on the premise that long-term research infrastructures such as the ESS need to be planned as such, with appropriate funding arrangements, organisational structures and continuity. Unless these arrangements are properly in place, a great deal of time is inevitably spent on issues of process as opposed to performance. Having experienced an uncomfortable set of funding arrangements for the last five years, we are convinced that the ESS – as with other infrastructures

– requires substantial improvements in this respect in order to thrive long-term.

5.2.4.2 *Additional substantive work*

A major omission in the present ESS corpus of work is the provision to data analysts of harmonised aggregated data at a European level that would inform and explain national and regional differences and similarities.

Despite the recent growth of analyses based on multi-country data throughout Europe, surprisingly little account is usually taken of the possible effect of each country’s social and political context on the reliability and validity of the comparisons. On the contrary, sophisticated analyses are often carried out as if the individual respondents were all living in the same society (or perhaps in different societies with identical histories, geographies, social relationships and institutional structures). In one of the ESS’s earliest documents - its ‘Blueprint’⁵ – reference is made to the necessity of multi-level analysis to take account of the impact on survey responses of the “social and institutional environment in which individuals are embedded”.

In short, in order to understand micro data in a particular population, it is helpful if not essential to consider them in the context of their country’s aggregate characteristics and circumstances. But academically-driven regional and contextual databases are hard to find. While some excellent aggregate databases have been compiled by Eurostat, they are not of course designed to meet academic research objectives. Their content is tailored to the needs of official statistics, and in any case their availability is often limited. Occasional social science databases can be found, such as the one compiled by electoral political scientists of elections and electoral systems, but they are in very short supply. But an excellent compilation of aggregate information alongside survey data, which also employs a sophisticated inference methodology and innovative cartographic displays, has been provided by the PartCom Multilevel project (HPSE-CT-1999-00029). And although its data are focused closely on issues to do with political participation, we believe we can learn a lot from it as an exemplar of what we wish to do.

Our aim is to build a Europe-wide aggregate database designed not only for the benefit of ESS users but for users of a variety of multinational social surveys (eg Eurobarometer, EVS, ISSP, WVS, etc). We will collect, harmonise and integrate academically-relevant information from diverse sources throughout Europe, in the process generating added value to these sources. But the added value to the ESS and its

users will be particularly high because we will then be able to harmonise and integrate individual-level data with aggregate-level data and distribute it on-line alongside the main ESS datasets. The aggregated data we produce will also be linked to a system of coordinates at European regional level, enabling mouse-over data browsing and cartographical and graphical presentations showing territorial and spatial units - such as municipalities, communes, census tracts, voting districts, counties, nations, and so on. Once compiled, it will at last enable researchers to link territorial data to descriptive statistics about demography, health, the economy, welfare, elections and so on, as well as to survey data - enriching the possibilities for multi-level analysis.

This will, of course, be a massive undertaking for the ESS infrastructure and will involve extensive work by specialists from at least three of its institutions – NSD, ZUMA and SCP (Netherlands) – each of which has directly relevant experience and expertise to contribute. It will, we believe, have an important and overdue impact on the European Research Area, which has long struggled against the constraints of existing aggregate databases, which are often prohibitively expensive, and of limited analytical value. Our database, like the rest of ESS data, will in contrast be available to all on-line and at no charge.

The facility will adapt to the new metadata standard for the social sciences, the Data Documentation Initiative (<http://www.icpsr.umich.edu/DDI/>) - a joint undertaking by data archives which paves the way for integrated portals for both survey-type data and aggregate data. The integration of these formerly separate types of data will help researchers to convert mere facts and figures into more usable, policy-relevant information, adding substance and context to the figures. It will also open up possibilities for much more sophisticated search operations than are available now.

This major new initiative is only one of many similar substantive enhancements that the ESS infrastructure needs to address. A related but separate need is a programme of work to improve the equivalence of occupational and educational classifications across Europe. Useful as the ESS has proved to be, and despite the fact that it employs ISCO and ISCED ‘standard’ classifications of occupation and educational attainment respectively, substantive specialists in the field of class and class mobility remain frustrated by unaccountable differences in these classifications between countries. Although the ESS data on these variables seem to be a great deal more equivalent than some past data, there are still some serious flaws. We are already in detailed discussions with specialists in occupational and educational classifications to examine in detail how the building blocks of such classifications are designed and implemented cross-nationally. But we face the problem that com-

⁵ “The European Social Survey (ESS) - a research instrument for the social sciences in Europe” (ESF 1999)

parative academic quantitative research of this scale and standard is still in its relative infancy.

So the ESS infrastructure aims to become a forum for debate and improvement in the measurement of these critical analytical tools, building on our existing contacts within Eurostat, National Statistical Institutes, universities and commercial survey houses. And since the main classifications we employ are international ones (not just European), we also need to keep in touch more closely than hitherto with world class American centres in the field – at the Universities of Michigan, Maryland and Chicago. And we need to form similar links with specialists in other countries.

Another weakness of the ESS is its relatively poor provision for the piloting and pre-testing new items for inclusion in the questionnaire (including the 100 or so new questions in each round that form two rotating modules). The budgetary limits of STREP vehicles simply took its toll on piloting more than on other aspects of the project. There is, of course, some provision for piloting but – with survey costs rising inexorably – the once relatively generous provision has become inadequate. In any case, we have never made provision for serious question development of new items on emerging issues. Although the ESS does not deal with topical issues of the sort that opinion polls measure, it does need from time to time to introduce new topics in response to changes on the ground in member states.

An example of an emerging issue is the conflict these days between libertarian values on the one hand and security concerns on the other. To what extent and in what circumstances do citizens of different countries suspend their libertarian values in favour of greater protection of citizens against terrorism? Devising questions on such topics is always tricky and extensive testing on analysable sample sizes should be *de rigueur*. A serious study such as the ESS should have an annual budget devoted to the piloting of new items such as these rather than having to rely – as at present – on fragile pieces of evidence that they ‘work’.

5.2.4.3 *Additional methodological and experimental work*

Above all other priorities for methodological work within the European Research Area is the need for serious improvements in the protocols and practices of translation. Equivalence of survey measurements depends first and foremost on equivalence of language. And although the ESS has made great strides in updating and enhancing translation methods with a view to increased levels of equivalence, there is still much to do. It remains the case that several differences in the national distribution of ESS responses seem to owe more to differences in translation than to real differences in attitudes, values or behaviour patterns between nations. So, just as comparative research in Europe requires aggregate data to make

important strides forward, so does it require improved systems of making and checking translations. In particular it needs a process that better captures connotations rather than just denotations, concentrating even more than now on equivalence of meaning rather than on direct translation of words and phrases.

We propose to concentrate our attention on two primary aspects of translation. First, recent advances in translation technology have meant that translators in other fields are increasingly working with Computer Aided Translation (CAT). Programs such as TRADOS, MultiTrans, Déjàvu, Transit and Wordfast are all used to aid translation, as well as to check and manage the process. These tools might turn out to have considerable application for translations in comparative surveys. But none of the present systems is designed for a multi-lingual, multi-national survey. Therefore we wish to investigate and evaluate the particular strengths and weaknesses of each package to establish whether or not it might nonetheless be fit for use in complex comparative surveys – perhaps with adjustment and adaptation. A thorough review of existing CAT systems is envisaged followed by a detailed report on problems and prospects. To the extent that the results of such a review seem promising, we would then discuss possible adaptations with the program developers to investigate the prospects of equipping them for survey translations.

The second focus of our attention on translation is at once less glamorous and more urgent, since it can be put into practice experimentally without delay. Our experience with three rounds of the ESS has demonstrated that the problem of faulty translations lies less in the protocols themselves than in their uneven application at a national level. The ESS protocol involves national teams of four people – two to undertake the initial translation, one to review the process, and one to help resolve difficult decisions. All these actors have to be competent in the two languages involved and are expected to have a working knowledge of survey practice with its sometimes arcane vocabulary and protocols.

Yet, despite the mostly conscientious application of these rules, serious errors still occasionally slip through, rendering potentially useful cross-national comparisons null and void. So we wish to experiment at the first available opportunity with a system of independent centralised checking of all translations in all languages (including different versions of the same language, eg French in France, Belgium and Switzerland).

This would be an expensive and time-consuming procedure if applied throughout a 30-nation survey. So we want in particular to experiment with economical methods of achieving this objective, such as by training social science students with appropriate mother tongues who happen to be at one or more

of our institutions' universities to undertake the initial task with appropriate supervision. In order to test the methodology, we will need a series of sizeable experiments. Our strong belief is that independent centralised checking of translations will have a sizeable positive impact on the equivalence of translations. And if this does prove to be the case, a sharp reduction in the incidence of faulty item translations will probably more than offset the cost of getting it wrong. In any case, if the aim is to improve the precision and comparability of cross-cultural research, there needs to be a 'gold standard' available for research projects to adopt or reject after due deliberation. As always, the work itself will of course be described together with its outputs in a series of reports that will be widely disseminated.

The second item on a long agenda of items for methodological enhancement of comparative studies is a series of experiments in improving survey response rates in countries where it is lagging behind. Alongside falls in turnout in elections, survey response rates have been diminishing over the years in most countries, threatening the credibility of surveys and the representativeness of their samples. As noted, one of the ESS's most impressive achievements to date has been to arrest and in some cases reverse this trend. But a great deal more work is required over the years, especially in conjunction with our existing experimental work on survey modes - which might eventually permit different modes of data collection in the ESS, such as telephone interviewing and the internet. The greatest danger of these new methods is that currently low response rates could become derisory response rates in the absence of the persuasive powers of the survey interviewer.

During the first two rounds of the ESS, a few mini-experiments on response rate enhancement were conducted, thanks to the help of countries such as Switzerland, Norway and The Netherlands. But a full-blown programme of such experiments in a range of different countries, each using different modes of data collection, would perform an invaluable role in advancing comparative survey methods worldwide. In particular, the fast-growing availability of the internet is heralding a possible new era of data collection which some organisations have rushed into rather recklessly. A mixed mode future for multinational surveys such as the ESS remains an attractive future option, but not before we are a great deal more certain than we are now that it will not bring in its wake serious problems of representativeness and artefactual variations in findings.

5.2.5 WHY THE ESS NEEDS TO REMAIN INDEPENDENT BUT LINKED

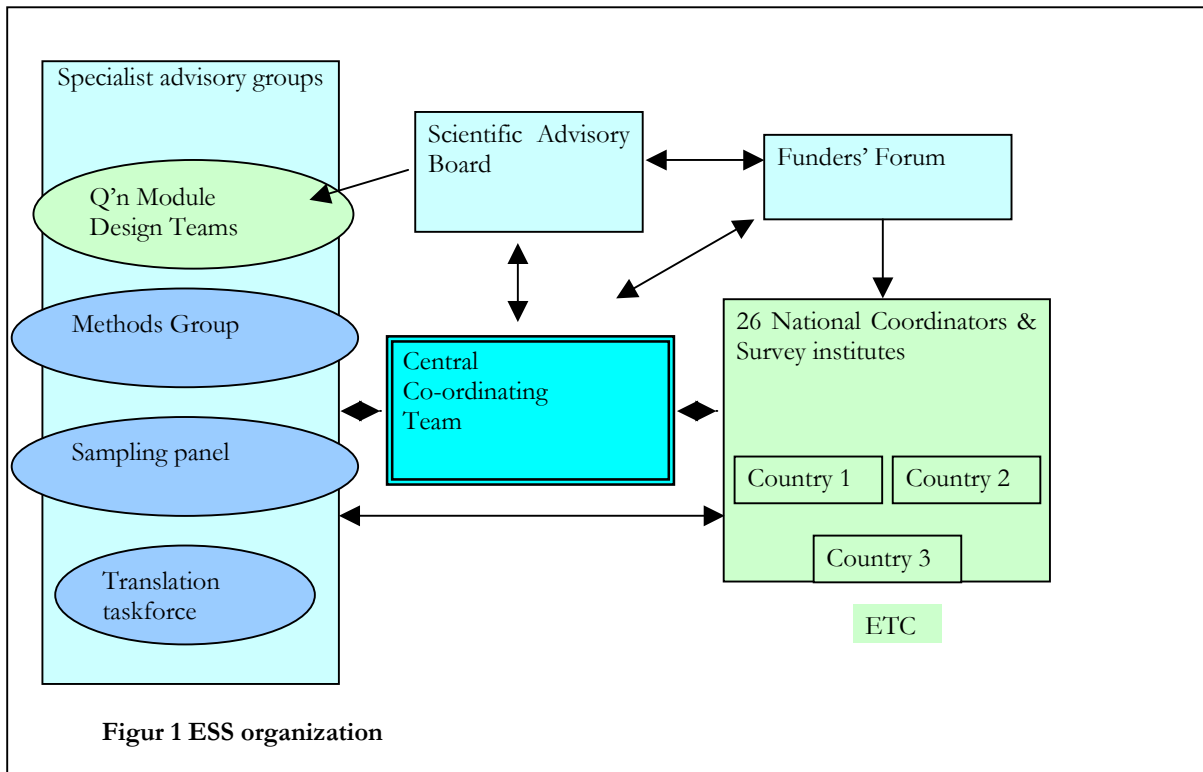
We welcome the possibility of other infrastructures in the social sciences being created as a result of the ESFRI initiative. The options we have had some

contact with in the past are EROHS and SHARE, but we have also heard wind of an application for a language initiative within the humanities. We single out these three possibilities simply to illustrate the sorts of links that independent European humanities and social science infrastructures may be able to forge with one another.

ESS's work is of such obvious relevance to SHARE's, and *vice versa*, that the possible mutual benefit of links needs no elaboration. Whether or not either becomes an ESFRI-led infrastructure, these two projects need to keep in close touch, exchanging experience and learning from one another whenever possible, on occasions holding joint meetings and possibly publishing joint papers.

It may come as more of a surprise that the ESS also needs to forge close links with a possible humanities infrastructure. In the first place (see above) one of the ESS's chief concerns continues to be the uncertain equivalence of language in multinational surveys. We are already in discussions with language and translation experts about these issues, but would welcome the opportunity to engage with academic specialists in the broader field. Language equivalence, even in survey questionnaires, is not of course a problem that will be solved quickly or completely. But more and wider attention to the issue (from a variety of angles) is a clear priority within the EU generally and for the ESS in particular. In addition, humanities specialists are as eligible as their counterparts in the social sciences to enter the competition for rotating modules that forms part of each ESS round. One such application - on cultural participation in different countries - has already been mooted and we hope that others will follow. A successful link of this kind would provide an excellent opportunity for multidisciplinary work. In any case, it should in our view be a requirement of any body included in the ESFRI Roadmap to investigate areas of mutual interest with other such bodies.

The relationship between a possible ESS infrastructure and a possible EROHS infrastructure is more promising still, though apparently a little more complicated. Questions have been raised as to whether ESS should be a spoke in the EROHS wheel or an independent infrastructure in its own right. And our answer remains that - although we would welcome close and cordial links between the two possible infrastructures - we can see real risks in changing the existing successful structure of the ESS in order to do so. As a mature investment whose initial architecture goes back ten years under the guidance of a European Science Foundation Expert Group, the ESS has grown organically and operates within a complex, tailor-made organisational structure (see below) that in many ways mirrors the proposed EROHS structure.



Each of the above component groups is active, meets regularly, and is peopled by leading experts in their respective fields. The organisational structure represented above comprises about 100 people from all parts of Europe, a number that rises as participation in the ESS continues to expand. Most important of all, the organisational structure works. Thanks partly to the overwhelming levels of help and guidance that we receive from all these specialists, the ESS has so

far always met and sometimes exceeded its delivery and timetable objectives. The large group of distinguished specialists who help to manage and influence the ESS tend to take their 'ownership' of the project very seriously - to the study's great benefit. We can see no possible reason for disturbing the structure or independence of what is a mature and thriving enterprise which encompasses a large and complex organisational network across 31 countries (see below).

5.2.6 ESS PARTICIPATING NATIONS RS 1 AND 2

Denmark	Austria	Czech Rep
Finland	Belgium	Estonia**
Iceland**	France	Hungary
Norway	Germany	Poland
Sweden	Ireland	Slovak Rep**
Luxembourg	Slovenia	Israel*
Greece	Netherlands	Ukraine**
Italy	Switzerland	Portugal
UK	Turkey**	Spain

* = Round 1 only to date

** = Round 2 additional countries

Note: R3 will also include Bulgaria, Cyprus, Romania, Russia

Within each country, there is a still greater variety of specialists, some within the survey houses, who help to keep the show on the road. Though closely coordinated from the centre, a massive amount of the vital work of the ESS takes place at a national level to ensure that around 50,000 interviews per round get carried out to the exact specification and within the exact timetable, and that they are then keyed and delivered to the data archive for prompt and timely release. In any event, for one reason or another the system works remarkably well, maintains high standards and delivers its promises. Indeed, the system is now being adopted by similar enterprises elsewhere.

So when we argue for a major upgrade of the ESS infrastructure, we strongly advise that its present fragile but highly successful structural arrangements should nonetheless not be unnecessarily disturbed. We would, however, be absolutely committed to ensuring that our links with a possible EROHS infrastructure were both firm and frequent. After all, the very existence of EROHS as a powerful centre of social science endeavour at the heart of Europe would be a boon not only for the ESS but for a range of other international projects too. But the close links we establish ought, we believe, to be between two large infrastructures of different sizes and ages and with different foci and structures.

5.3 SURVEY OF HEALTH, AGEING AND RETIREMENT IN EUROPE – SHARE

Survey of Health, Ageing and Retirement in Europe (SHARE)



Background: Preliminary data collection started in 2002, and in 2004 a first wave of data on the economic, health and family conditions of about 27,000 respondents aged 50 and over were collected in 11 European countries. The participating countries covered all EU15 regions. The data is harmonized cross-national. The second wave of data collection is currently going on and includes Poland, the Czech Republic, and Ireland. A third wave of data collection specializes on the life histories of the SHARE.

What's new? Which impacts? The first wave of SHARE data was collected in 2004 and the second wave is currently fielded in 2006, further waves are envisaged from 2008 onwards bi-

annually. In the years in-between these waves, experimental modules will be tested, such as the collection of life-histories in 2007. The 24 months between the end of wave t and the end of wave t+1 can roughly be divided into 12 months of preparation and 12 months of data collection (including experimental modules). The SHARE data infrastructure is accessible free of charge through an archive operating as internet platform.

Timeline for construction and first operation with related estimated costs: Costs are roughly proportional to the number of participating countries and the number of waves. Preparatory costs amount to approximately 250k€ per country and wave (= 3.75m€ for 15 countries). Construction costs (i.e. data collection) amount to about 400k€ per country/per wave (= 6m€ for 15 countries). Annual operating costs for the entire SHARE infrastructure (data distribution and documentation) amount to some 300k€ per year. Hence, one wave of the bi-annual data collection in 15 countries thus costs about 10.5m€. Three waves in all 25 member states cost approx. 51m€. There are no decommissioning costs. Current funding is from EU, NIA and national sources, and such cost sharing is also expected in the future.

Leading consortium: Originally, 8 countries participated in the 2004 SHARE baseline study (Sweden, Denmark, Germany, Netherlands, France, Italy, Spain, and Greece). In some of these countries, co-funding has been obtained from member states to increase sample size. Belgium, Austria, Switzerland joined the 2004 wave fully financed by member-state funds. Israel joined in 2005, funded by the German-Israeli Foundation. Poland, the Czech Republic and Ireland joined SHARE in 2006, the first two funded by the EU, Ireland fully by own funds. Three further countries (Finland, Portugal, and Slovenia) have expressed interest in joining SHARE in 2007 providing their own funding. In addition to the central EU funding, about 5m€ in funding has been obtained from the US National Institute on Aging.

The science community connected to SHARE is truly world-wide: SHARE has been developed departing from the English Longitudinal Study on Ageing (ELSA) and the US Health and Retirement Study (HRS) and is in close and ongoing co-operation with these studies, and the SHARE concept is now being copied by teams in South Korea and Japan.

5.3.1 ABSTRACT

This project proposed for ESFRI will generate a longitudinal research infrastructure in order to monitor the ageing process in Europe and to foster multidisciplinary cross-national research on the social and economic implications of ageing in Europe. Core of the activity is the collection of an ex ante harmonized panel data set on health, ageing, labor force participation, retirement, economic, family and social conditions of individuals aged 50 and older.

5.3.2 THE GENESIS OF SHARE

SHARE was founded in 1999 in order to create an equivalent to the US-American Health and Retirement Survey. Preliminary data collection started in 2002, and in 2004 a first wave of data on the economic, health and family conditions of about 27,000 respondents aged 50 and over was collected in 11 European countries. The participating countries covered all EU15 regions: Nordic countries (Sweden, Denmark), Western European countries (Netherlands, Belgium, France, Germany, Austria, Switzerland), and the Mediterranean (Spain, Italy, Greece). The major strength of these data is the ex-ante harmonized cross-national dimension that allows comparing the effects of different welfare systems (e.g. pension and health care systems) on the lives of middle-aged and older European Citizens.

First analyses of the data were also financed by the EU in the AMANDA project (Advanced Multi-Disciplinary Analysis of New Data). They showed the vast potential of these data to answer a number of crucial academic and policy questions [refs. 3,4,5,10,16,18,20,33, 36,37,39,44,46,47,53,60,61,62, 63,64,82,88,89,98].

The second wave of data collection is currently going on and includes two new EU member states, Poland and the Czech Republic, and Ireland (EU-funded by an Integrated Infrastructure Initiative and a STREP). A third wave of data collection specializes on the life histories of the SHARE respondents (EU-funded as Integrated Project) in 2007.

5.3.3 THE SCIENTIFIC CASE

The European Commission has identified population ageing and its social and economic challenges to growth and prosperity to be among the most pressing challenges of the 21st century in Europe [ref. 23]. Responding to the March 2000 Special European Council in Lisbon, a communication by the European Commission to the Council and the European Parliament calls to “*examine the possibility of establishing, in co-operation with Member States, a European Longitudinal Ageing Survey.*” [ref. 26, p.61]. The 5th framework

project “Survey of Health, Ageing and Retirement in Europe (SHARE)” has responded to this call and has established a baseline for such a longitudinal ageing survey in Continental Europe. This I3 project will integrate existing UK data to this infrastructure, make the SHARE infrastructure longitudinal by adding a second wave of data as requested by the call, include two accession countries to the infrastructure, and provide free and user-friendly access for the entire research and public policy community.

Important knowledge and information gaps with respect to healthy and successful ageing exist in Europe. In addition to exploring the molecular and cellular determinants of healthy ageing [ref. 29], the implications of the ageing process for the well-being of the population and the societal costs of improving public health and maintaining social insurance schemes will only be understood once the interactions between individual traits and behaviour on the one hand, and the socio-economic and public policy environment on the other hand, is taken into account [22].

Research on these interactions requires an infrastructure of easily accessible micro-data on the relevant dimensions of ageing, i.e. data on the health, work, economic and social conditions of individuals as they age and the resulting quality of life and well-being. Much as physicists need an infrastructure like CERN to understand particle physics, and astronomers need an infrastructure of telescopes, social scientists need an infrastructure of survey data to base their research on quantifiable and falsifiable hypotheses. Unlike the United States with its successful Health and Retirement Study (HRS), Europe does not have such an infrastructure [ref. 26, 71]. It is, however, all the more urgent given a more pronounced ageing process in Europe, and a much more precarious financial state of its pension, health and long-term care systems, threatening long-run economic growth and sustainable prosperity in Europe.

In order to provide that much needed infrastructure, we propose to strengthen and extend the access to an ongoing European data collection effort on ageing, the “Survey of Health, Ageing and Retirement in Europe (SHARE)” funded under the 5th framework programme. The existing informational infrastructure of SHARE carries data on health, work and economic status, and family and social networks in 11 countries in the autumn of 2004. A second wave is being collected during 2006, now encompassing 15 countries funded under the 6th framework programme, and a life-history module will be added in 2007, also funded by the 6th framework programme. First results have been analysed and the potential of the data has been demonstrated by the “Advanced Multidisciplinary Analyses of New Data on Ageing (AMANDA)” project also funded under the 5th framework programme.

Ageing, however, is a dynamic process that can only be observed longitudinally. The environment in which people age also changes over time, e.g., by way of pension and health care reforms. Without adding a substantial time dimension, the current data remain a torso because the process of ageing cannot be observed, reactions to the ongoing changes in the institutional environment cannot be traced, and related behavioral hypotheses cannot be tested. The proposed ESFRI infrastructure will therefore give the existing SHARE infrastructure a genuinely longitudinal dimension by collecting a third, fourth and fifth wave of data in 2008, 2010 and 2012 and by linking these two waves to understand the changes over time.

A genuine time dimension is essential because ageing is a process, not a state. Comparing two individuals of different ages at a point in time is no substitute for observing the same person over time since the two persons have been born at different dates and thus belong to two different generations. Moreover, the time dimension is essential for policy analysis because institutions such as pension and health care systems dramatically change in the current reform process and the citizens react to these changes over time. Finally, time is essential to break purely statistical correlations and to establish causality between interventions and effects. Adding the time dimension will therefore raise the performance of the existing multi-disciplinary and cross-national data collected under SHARE by the quantum leap that is needed to close the research disadvantage in ageing research vis-à-vis the United States. Due to the cross-nationality of SHARE, the new infrastructure has the great potential to even create a research advantage vis-à-vis the HRS.

A second extension to the existing data concerns the countries covered. The current 15 countries cover Continental West European countries from Scandinavia to the Mediterranean, two Central European accession countries and Ireland, and, indirectly through data collected in the English Longitudinal Study of Ageing (ELSA), also parts of Great Britain. This covers much of the European Union, but many policy analysis at the EU level require comparisons among all member states. SHARE is already being used as a policy evaluation tool by the Directorate General of Employment and Social Affairs as well as the Directorate General of Economic and Financial Affairs. Most significantly, the new European Commission's projections of pension and health care expenses could be based on a scientifically state-of-the-art methodology thanks to the SHARE data.

The proposed ESFRI project will create a research infrastructure that will enable researchers from a very broad set of disciplines to analyse the Europe-wide ageing process; to better understand adaptive behaviour e.g. in retirement, saving and health service utilisation in response to population ageing and its in-

duced policy changes; and to develop cross-nationally comparable indicators for key concepts relevant to EU policy, e.g. pension replacement rates, retirement incentives, savings adequacy, disability rates, health status, prevalence of health conditions, and well-being of the elderly.

The European Commission has identified in its Lisbon agenda a number of policy problems that need to be resolved in order to be better prepared for the future, including a low labour force participation of older individuals, unsustainable pension and health care systems, and vastly diverging health expenditures across member states without discernible impacts on health outcomes [refs. 23,27,85,86,89]. These policy problems are unresolved also because of our insufficient understanding of the implications of ageing on individuals' behaviour and the complex interactions between health, social, and economic issues. By the way of example:

(1) Planned reforms will reduce the generosity of public pension systems [13,23,74,75] but we do not have data on savings behaviour in order to assess whether people will save enough on their own to make up the resulting pension income gap [28,30,36,98]. We do not know how much people will save for retirement if this saving is voluntary [10,25] and we do not know how much other kinds of savings are displaced when more saving is devoted to retirement [17,24,44].

(2) A certain extent of labour force participation may be an important ingredient for active and healthy ageing [8,19,22,33,37,56,88]. However, we have insufficient longitudinal data about labour force participation of the elderly and the process of retirement [14,35]. While we have some information from household and labour surveys, we do not have health and social inclusion data going with it, supposedly major determinants of the retirement process [12,14,16,19]. We do not understand how partial retirement is related to job conditions, in particular chronic work stress [84,87,88], and how retirement is co-ordinated in the increasing number of two-worker households. We partially understand how pension policies affect the actual retirement age, but we know less about the choice between alternative pathways to retirement such as claiming disability or unemployment benefits [11,14,69,84].

(3) Family and social support is a major resource in old age [4,5,6,53,68]. However, we do not have sufficient data to assess the impact of social policy (such as long-term care insurance) on potential family support. We know very little about time and money transfers between generations (including taking-in parents), and to which extent this is displaced or encouraged by state support [4,5,6, 34,55,73,80]. We do not understand how bequests will react to changes in the public intergenerational redistribution through pensions, health and long-term care insur-

ance [5,47,54,90]. We only have case studies on the interaction between the economic status of the elderly and the social, psychological and health effects of social exclusion [7,18,60,79].

(4) We do not have sufficient data to study the complex relation between health, economic and socio-psychological status [9,39,42,63]. We lack longitudinal data to understand how economic status, including job characteristics, during the active life affects health during retirement [40,65], or to understand the observed correlation between wealth and health [91]. Since data-linking mortality to wealth is largely lacking, we do not know if pensions redistribute from the poor who live shorter to the wealthier who live longer [67,72,83].

(5) Disability insurance has turned out to be a costly social programme. Uptake rates vary dramatically across EU member states [16,27]. While some relate this to the generosity of disability insurance, we did not have the medical data to measure disability rates comparably and reliably to substantiate this [46,61,62,82]. We lack longitudinal data to investigate the side effects of changes in pension and unemployment insurance benefits on disability claims [1,12].

Unlike the United States, Europe is ill equipped for research on these issues because it lacks an adequate longitudinal data infrastructure to investigate the multifaceted implications of the ageing process and its interactions with public policy [71]. Until today, most surveys focus on only one discipline such as health or economics; they are mostly national; and almost all lack a longitudinal dimension essential to understand the ageing process. These deficiencies of the European research infrastructures effectively prevent the conduct of high-quality research similar to that in the US. Moreover, Europe with its huge cultural, historical and policy diversity is currently left unexploited as a “natural laboratory” to understand the effect of public policy on the behaviour and well-being of its citizens. The provision of integrated *multidisciplinary*, *cross-national*, and *longitudinal* data on a European level, freely available to the research community, is therefore an essential infrastructural prerequisite for a better understanding of population ageing and its consequences. *Multi-disciplinary data* is required because health, economic and social conditions are closely interrelated and these interactions are under-researched. *Cross-nationality* is essential in order to understand the impact of different policies on the health, economic and social status of the citizens in Europe, and *longitudinality* is needed to observe the individual and societal processes over time, identify policy reactions and permit the detection of causal mechanisms.

5.3.4 THE BUSINESS CASE

SHARE has developed to maturity in a six year process starting from its US-American mother survey, the US Health and Retirement Study. The longitudinal core-questionnaire of SHARE has undergone a four-year development process and is well-established, via the first large cross-sectional 2004 wave, conducted in 11 countries and covering 27,000 individuals, and now collecting the first longitudinal data as well as additional samples in the Czech Republic, Poland, Ireland and Israel during 2006. Further waves are envisaged from 2008 onwards bi-annually. In the years in-between these waves, experimental modules will be tested, such as the collection of life-histories in 2007.

The SHARE project family has been peer-reviewed several times with great success in competitive applications in Europe and the US:

- five times as EU DG Research projects in the 5th and 6th framework programme,
- twice as competitive interagency grants funded by the US National Institute of Aging,
- several national competitive grants
- a competitive German-Israeli-Foundation grant.

Funding acquired so far has exceeded 26m€.

The research team currently involves about 150 researchers in 18 countries. The core participants in SHARE have worked together for a long time; their collaborations have been in place through various funding mechanisms (SPES, TMR, RTN, etc.) for more than 15 years. The multi-disciplinary nature of the past projects (economics, public health, psychology, sociology) has proven to be sustainable as shown by the joint research papers and a set of joint RTN and RTD projects. In addition to the core senior researchers, the group has succeeded in attracting a large number of junior researchers who write their academic work along with building up SHARE. This experienced team is set to continue for a bi-annual longitudinal extension beyond 2006. The assembly of this multidisciplinary and experienced team provides an exceptionally large value added in leveraging previously EU-funded research.

In order to manage this large number of researchers efficiently, SHARE has developed a management structure which is a combination of top-down and bottom-up. A matrix structure combines country teams in all involved countries with working groups of all involved subject matters as follows:

Core management, i.e. the co-ordinator with the core management group, sets constraints within which the participating teams operate. For instance, the working groups are responsible for the development, integration and analysis of questionnaire modules. They

work decentrally, but the amount of time for each module will be centrally determined, as are the interfaces to other modules. The country teams decentrally negotiate and communicate with the fieldwork agencies, but the structure and essential scientific details of the survey are set centrally to achieve genuine ex ante harmonization.

SHARE currently has 16 country teams representing Austria, Belgium, the Czech Republic, Denmark, France, Germany, Greece, Ireland, Israel, Italy, the Netherlands, Poland, Spain, Sweden, Switzerland, and the United Kingdom. SHARE is open access – we welcome new participant states and aim at covering all 25 member states of the EU. Large parts of SHARE have been copied by similar surveys in South Korea and Japan. Since SHARE is fully comparable with (and actually has been developed departing from) the English Longitudinal Study on Ageing (ELSA) and the US Health and Retirement Study (HRS), the science community connected to SHARE is truly world-wide.

Since 2001, SHARE has created 16 working groups which are devoted to the analysis of physical health, mental health, social participation and well-being, assets, income, consumption, employment and pensions, expectations, housing and family networks, intergenerational transfers, health service utilisation, poverty and social exclusion, reaching the oldest old, sampling, data base management, and response behaviour.

A core feature of the SHARE business case is the close link between researchers and professionals. Unlike many other data sets which are collected for administrative reasons by national statistical agencies, the set up of the questionnaire is solely research-driven. We stress our belief that the best infrastructure is being constructed by those who then also analyse the data. This makes sure that the SHARE infrastructure is a true research instrument and targets at relevant scientific work. The methods of data collection are developed in close cooperation between professional data collection agencies and the researchers to combine high data quality with immediate research input.

SHARE has therefore built up a network of small and medium enterprises (SMEs) that assist the SHARE scientists with highly specialized professional tasks. Examples are programming tasks (subcontracted to a Dutch SME), translation tasks (subcontracted to special survey translators), and data collection tasks (subcontracted to professional survey agencies). Innovative questionnaire elements and methods used in SHARE require particular training efforts for survey agencies and interviewers. SHARE has developed a train-the-trainer (ITT) program to facilitate such decentralised training in the participating small and medium enterprises (SME). This is expected to have a significant quality impact on the survey industry.

The SHARE data is being used widely and currently has about 370 registered users from across Europe and outside Europe. It has been used by DG Employment and DG EcFin for projections and analyses. Free and easy access is a crucial feature of SHARE also in the future.

5.3.5 NECESSARY FUNDING

The 24 months between the end of wave t and the end of wave $t+1$ can roughly be divided into 12 months of preparation and 12 months of data collection (including experimental modules).

Preparation costs in the first 12 months amount to approximately 250k€ per country and wave. These costs mainly include salaries for the involved junior scientists in the working groups and country teams (one full time junior scientist in each team; senior scientists are paid by their universities), costs of management and training activities, plus costs for regular meetings and conferences about 4 times a year.

Construction costs (mainly data collection) in the second 12 months amount to about 400k€ per country/per wave at a country sample of 2500 individuals. Construction costs are dominated by survey costs (about 120€ per computed-aided personal interview, including physical measurements for health status and panel maintenance), plus salaries for 1 full-time country operator, costs of management and training activities, plus costs for regular meetings and conferences about every two months during data collection.

Preparation and construction costs are roughly proportional to the number of participating countries and the number of waves. At the current number of 15 participating countries, preparation costs amount to 3.75m€ and construction costs to 6m€ per wave, totaling 9.75m€ per wave. For the envisaged participation of all 25 EU member countries, preparation costs amount to 6.25m€ and construction costs to 10m€ per wave, totaling 16.25m€ per wave.

Annual operating costs for the entire SHARE infrastructure (data distribution, documentation and user support services) amount to some 300k€ per year. There are no decommissioning costs.

Upgrading the current SHARE infrastructure to a 5-year panel for all 25 EU member countries in the three waves 2008-2010-2012 would therefore require funds of about 51m€.

Current funding is from EU, NIA and national sources, and such cost sharing is also expected in the future. In particular, cost sharing is expected between participating member states and central sources (EU, potentially NIA). The fraction of national funding

should not exceed 50% to maintain central control in order to ascertain ex ante harmonization of the data which is crucial for the project.

Costs do not include the analysis of the data by the researchers involved in their construction, and they

also exclude the costs of experimental modules between the bi-annual waves. Such costs are expected to be borne by additional grants tailored to the specific research questions involved in the analyses and modules.

5.3.5.1 References

- 1 Aarts, L.J.M., R.V. Burkhauser, Ph.R. de Jong (eds.) (1996), *Curing the Dutch Disease, An International Perspective on Disability Policy Reform*, Aldershot: Avebury.
- 2 Alcer, K. and G. Benson (2005), The SHARE-SRC train-the trainer programme, in: A. Börsch-Supan et al. (eds), *Health, Ageing and Retirement in Europe: First results of SHARE*, University of Mannheim.
- 3 Andersen-Ranberg, K., I. Petersen, J.-M. Robine and K Christensen (2005), Who are the oldest old, in: A. Börsch-Supan et al. (eds), *Health, Ageing and Retirement in Europe: First results of SHARE*, University of Mannheim.
- 4 Attias-Donfut, C., J. Ogg and F.-C. Wolff (2005), Family Support, in: A. Börsch-Supan et al. (eds), *Health, Ageing and Retirement in Europe: First results of SHARE*, University of Mannheim.
- 5 Attias-Donfut, C., J. Ogg and F.-C. Wolff (2005), Financial Transfers, in: A. Börsch-Supan et al. (eds), *Health, Ageing and Retirement in Europe: First results of SHARE*, University of Mannheim.
- 6 Attias-Donfut, C. and M. Segalen (eds.) (2001), *Le siècle des Grands-parents*, Paris, Autrement.
- 7 Baltes, P., and K.-U. Meyer (1999), *The Berlin Aging Study*, Cambridge University Press.
- 8 Banks, J. and M. Casanova (2003), Work and retirement, in Marmot et al. (eds), *Health, wealth and lifestyle of the Older Population in England: The 2002 English Longitudinal Study of Ageing*, The Institute for Fiscal Studies, London.
- 9 Banks, J., S. Karlsen and Z. Oldfield (2003), Socio-economic position, in Marmot et al. (eds), *Health, wealth and lifestyle of the Older Population in England: The 2002 English Longitudinal Study of Ageing*, The Institute for Fiscal Studies, London.
- 10 Bernheim, B.D. (1999), Taxation and Saving, *Handbook of Public Economics*, forthcoming, North-Holland Publishing Company, Amsterdam.
- 11 Blanchet, D., A. Brugiavini, and R. Rainato (2005), Pathways to Retirement, in: A. Börsch-Supan et al. (eds), *Health, Ageing and Retirement in Europe: First results of SHARE*, University of Mannheim.
- 12 Blöndal, S. and S. Scarpetta (1998), *The retirement decision in OECD countries*, OECD Economics Department Working Paper No. 202, Paris.
- 13 Börsch-Supan, A. (2000), A Model under Siege: A Case Study of the German Retirement Insurance System, *The Economic Journal*, Vol. 110 No. 461, F24-45.
- 14 Börsch-Supan, A. (2001), Incentive Effects of Social Security under an Uncertain Disability Option, in: D. Wise (ed.), *Themes in the Economics of Aging*, University of Chicago Press.
- 15 Börsch-Supan, A. (2005), The SHARE Development Process, in: A. Börsch-Supan et al. (eds), *Health, Ageing and Retirement in Europe: First results of SHARE*, University of Mannheim.
- 16 Börsch-Supan, A. (2005), Work Disability and Health, in: A. Börsch-Supan et al. (eds), *Health, Ageing and Retirement in Europe: First results of SHARE*, University of Mannheim.
- 17 Börsch-Supan, A. (ed.) (2001), *International Comparisons of Household Saving*, New York: Academic Press.
- 18 Browning, M. and E. Madsen (2005), Consumption, in: A. Börsch-Supan et al. (eds), *Health, Ageing and Retirement in Europe: First results of SHARE*, University of Mannheim.
- 19 Brugiavini, A., E Croda and F. Mariuzzo (2005), Labour Force Participation of the Elderly: Unused Capacity?, in: A. Börsch-Supan et al. (eds), *Health, Ageing and Retirement in Europe: First results of SHARE*, University of Mannheim.
- 20 Brunner, H., C. Riess, and R. Winter-Ebmer, (2005), Public and private pension claims, in: A. Börsch-Supan et al. (eds), *Health, Ageing and Retirement in Europe: First results of SHARE*, University of Mannheim.
- 21 Carton, A., and G. Loosveldt (1999), *How the initial contact can be determined from the final response rate in face to face surveys*, International Conference on Survey nonresponse, Portland, Oregon.
- 22 Economic and Social Committee of the European Parliament (2000), *Older Workers*, Opinion SOC/039, October 2000.
- 23 Economic Policy Committee (2000), Progress Report of the Ecofin Council on the Impact of ageing populations on public pension systems, Economic Policy Committee, EPC/ECFIN/581/00-EN-Rev.1, Brussels.
- 24 Engen, E., W.G. Gale, and J.K. Scholz (1996), The Illusory Effects of Saving Incentives on Saving, *Journal of Economic Perspectives* 10, 113-138.
- 25 European Commission (1999), *Towards a single market for supplementary pensions*, Commission Communication COM(99) 134 final/2.
- 26 European Commission (2000), *The Contribution of Public Finances to Growth and Employment: Improving Quality and Sustainability*, Communication from the European Commission to the Council and the European Parliament.
- 27 Eurostat (1999), Key Data on Health 2000, in: *Theme 3: Population and Social Conditions*, Brussels: European Commission.
- 28 Feldstein, M. (1974), Social Security, Induced Retirement, and Aggregate Capital Accumulation, *Journal of Political Economy* 82, 905-925.
- 29 Finch, C.E., J.W. Vaupel and K. Kinsella (2001), *Cells and Surveys: Should Biological Measurements Be Included in Social Science Research?* Washington, D.C.: National Academy Press.

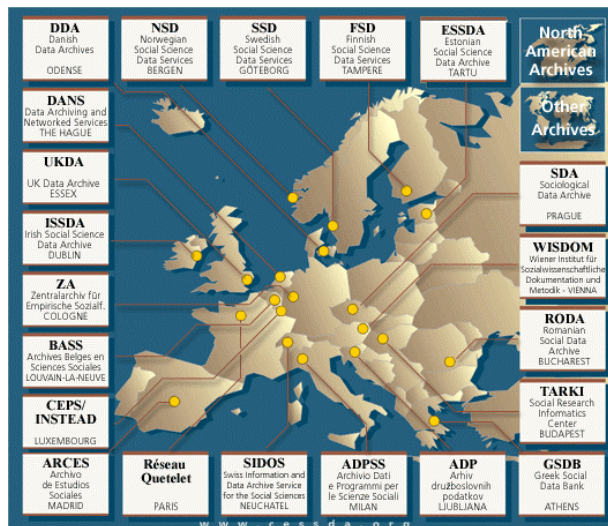
- 30 Gale, W. (1998), The Effects of Pensions on Household Wealth: A Reevaluation of Theory and Evidence, *Journal of Political Economy* 106, 706-723.
- 31 Garber, A. (1990), Long-term care, wealth and health of the disabled elderly living in the community, in: D. Wise (ed.), *The economics of aging*, University of Chicago for NBER.
- 32 Gjonka, E. and L. Calderwood (2003), Socio-demographic characteristics, in Marmot et al. (eds), *Health, wealth and lifestyle of the Older Population in England: The 2002 English Longitudinal Study of Ageing*, The Institute for Fiscal Studies, London.
- 33 Gonzales, J. and E. Croda (2005), How Do European Older Adults Use Their Time, in: A. Börsch-Supan et al. (eds), *Health, Ageing and Retirement in Europe: First results of SHARE*, University of Mannheim.
- 34 Greene, V., et al. (1993), Do community-based long-term care services reduce nursing homes use?, *Journal of Human Resources* 28 (2), 297-317.
- 35 Gruber, J., and D.A. Wise (1999), *Social Security and Retirement around the World*, Chicago University Press.
- 36 Guiso, J., A. Tiseno and J. Winter (2005), Expectations, in: A. Börsch-Supan et al. (eds), *Health, Ageing and Retirement in Europe: First results of SHARE*, University of Mannheim.
- 37 Hank, K. and M. Erlinghagen (2005), Volunteer work, in: A. Börsch-Supan et al. (eds), *Health, Ageing and Retirement in Europe: First results of SHARE*, University of Mannheim.
- 38 Harkness, J. (Ed.) (1998), Cross-Cultural Survey Eqivalence (ZUMA Nachrichten Spezial No 3.), Mannheim, ZUMA.
- 39 Holly, A., K. Lamiraud, H. Chevrou-Severac, and Tarik Yalcin (2005), Out of pocket payments for Health Care Expenditures, in: A. Börsch-Supan et al. (eds), *Health, Ageing and Retirement in Europe: First results of SHARE*, University of Mannheim.
- 40 House, J.S., J.M. Lepkowski, A.M. Kinney, R.P. Mero, R.C. Kessler, and A.R. Herzog (1994), The Social Stratification of Aging and Health, *Journal of Health and Social Behavior* 35, 213-234.
- 41 Hyde, M. and M. Janevic (2003), Social activity, in Marmot et al. (eds), *Health, wealth and lifestyle of the Older Population in England: The 2002 English Longitudinal Study of Ageing*, The Institute for Fiscal Studies, London.
- 42 Idler, E., and Y. Benyamini (1997), Self-Rated Health and Mortality: A Review of Twenty-Seven Community Studies, *Journal of Health and Social Behavior* 38, 21-37.
- 43 Janevic, M., E. Gjonka and M. Hyde (2003), Physical and social environment, in Marmot et al. (eds), *Health, wealth and lifestyle of the Older Population in England: The 2002 English Longitudinal Study of Ageing*, The Institute for Fiscal Studies, London.
- 44 Jappelli, T., D. Christelis, and M. Padula, (2005), Wealth and portfolio composition, in: A. Börsch-Supan et al. (eds), *Health, Ageing and Retirement in Europe: First results of SHARE*, University of Mannheim.
- 45 Jowell, R. (1986), The codification of statistical ethics, *Journal of Official Statistics* 2 (3), 217-253.
- 46 Jürges, H. (2005), Cross-country differences in general health, in: A. Börsch-Supan et al. (eds), *Health, Ageing and Retirement in Europe: First results of SHARE*, University of Mannheim.
- 47 Jürges, H. (2005), Gifts, inheritances and bequest expectations, in: A. Börsch-Supan et al. (eds), *Health, Ageing and Retirement in Europe: First results of SHARE*, University of Mannheim.
- 48 Jürges, H. (2005), Interviewer effects in SHARE. Mimeo, MEA, University of Mannheim.
- 49 Juster, F.T., and J.P. Smith (1997), Improving the Quality of Economic Data: Lessons from the HRS and AHEAD, *Journal of the American Statistical Association* 92.
- 50 Juster, F.T., and R. Suzman (1995), An Overview of the Health and Retirement Study, *Journal of Human Resources* 30, S7-S57.
- 51 Kish, L. (1965), *Survey Sampling*, New York: Wiley.
- 52 Klevmarken, A. (2005), Sample design, in: A. Börsch-Supan et al. (eds), *Health, Ageing and Retirement in Europe: First results of SHARE*, University of Mannheim.
- 53 Kohli, M., H. Kuehnemund, and T. Zähle (2005), Housing and Living Arrangements, in: A. Börsch-Supan et al. (eds), *Health, Ageing and Retirement in Europe: First results of SHARE*, University of Mannheim.
- 54 Kotlikoff, L.J., and L.H. Summers (1981), The Role of Intergenerational Transfers in Aggregate Capital Accumulation, *Journal of Political Economy* 89, 706-732.
- 55 Künemund, H., and M. Rein (1999), There is more to receiving than needing: Theoretical arguments and empirical explorations of crowding in and crowding out, *Ageing and Society* 19, 93-121.
- 56 Lehr, U., (1996), *Psychologie des Alterns*, UTB 55, Quelle und Meyer.
- 57 Lipps, O. and G. de Luca (2005), Fieldwork and Sample Management, in: A. Börsch-Supan et al. (eds), *Health, Ageing and Retirement in Europe: First results of SHARE*, University of Mannheim.
- 58 Loosveldt, G. (1995), The profile of the difficult-to-interview respondent, *Bulletin de Méthodologie sociologique* 48, 68-81.
- 59 Loosveldt, G., J. Pickery, and J. Billiet (1999), Item nonresponse as a predictor of unit nonresponse, Paper presented at the International Conference on Survey nonresponse, Portland, Oregon.
- 60 Lyberaki, A. and P. Tinios (2005), Poverty and Social Exclusion: A New Approach to an Old Issue, in: A. Börsch-Supan et al. (eds), *Health, Ageing and Retirement in Europe: First results of SHARE*, University of Mannheim.
- 61 Mackenbach, J., M. Avenado, K. Andersen-Ranberg and A.R. Aro (2005), Physical health, in: A. Börsch-Supan et al. (eds), *Health, Ageing and Retirement in Europe: First results of SHARE*, University of Mannheim.
- 62 Mackenbach, J., A. Aro, M. Avendano (2005), Health behaviour, in: A. Börsch-Supan et al. (eds), *Health, Ageing and Retirement in Europe: First results of SHARE*, University of Mannheim.

- 63 Mackenbach, J. , A. Aro, M. Avendano (2005), Socio-economic Disparities in Physical Health in 10 European Countries, in: A. Börsch-Supan et al. (eds), *Health, Ageing and Retirement in Europe: First results of SHARE*, University of Mannheim.
- 64 Mariuzzo, F. and A. Börsch-Supan (2005), Validation with Other Data Sets, in: A. Börsch-Supan et al. (eds), *Health, Ageing and Retirement in Europe: First results of SHARE*, University of Mannheim.
- 65 Marmot, M., and M. Shipley (1996), Do socio-Economic Differences in Mortality Persist After Retirement? 25 years Follow up of Civil Servants from the First Whitehall Study, *British Medical Journal* 313, 1177-1180.
- 66 McMunn, A., M. Hyde, M. Janevic and M. Kumari (2003), Health, in Marmot et al. (eds), *Health, wealth and lifestyle of the Older Population in England: The 2002 English Longitudinal Study of Ageing*, The Institute for Fiscal Studies, London.
- 67 Menchik, P.L. (1993), Economic Status as a Determinant of Mortality Among Black and White Older Men: Does Poverty Kill? *Population Studies* 47, 427-436.
- 68 Mira, P. and M. Martinez-Granado (2005), The Number of Living Children, in: A. Börsch-Supan et al. (eds), *Health, Ageing and Retirement in Europe: First results of SHARE*, University of Mannheim.
- 69 Mitchell, O.S., and J.W.R. Phillips (2000), *Retirement Responses to Early Social Security Benefit Reductions*, Pension Research Council Working Paper, Wharton School.
- 70 Müller, W., U. Blien, P. Knoche, and H. Wirth (1991), Die faktische Anonymität von Mikrodaten, Bd. 19 der Schriftenreihe *Forum der Bundesstatistik*. Statistisches Bundesamt (Hrsg.). Stuttgart: Metzler-Poeschel.
- 71 National Academy of Sciences (USA) (2001), *Preparing for an Aging World: The Case for Cross-National Research*. Washington, D.C.: National Academy Press..
- 72 Nelissen, J.H.M. (1999), Mortality Differences Related to Socioeconomic Status and Progressivity of Old-age Pensions and Health Insurance: The Netherlands, *European Journal of Population* 15, 77-97.
- 73 Norton, E. (2000), Long-term care, in: *Handbook of Health Economics*, J. Newhouse and A. Culyer (eds.), North-Holland.
- 74 OECD (1988), *Ageing Populations: The Social Policy Implications*, Paris.
- 75 OECD (1998), *Maintaining Prosperity in an Ageing Society*, Paris.
- 76 OECD (2000a), *Reforms for an Ageing Society*, Paris.
- 77 Peracchi, F. (2002), The European Community Household Panel: A Review. *Empirical Economics* 27, 63-90 .
- 78 Peracchi, F. and G. de Luca (2005), Survey response, in: A. Börsch-Supan et al. (eds), *Health, Ageing and Retirement in Europe: First results of SHARE*, University of Mannheim.
- 79 Perelman, S., Bonsang, E. and K. van den Bosch (2005), Income, Wealth and Consumption inequality, in: A. Börsch-Supan et al. (eds), *Health, Ageing and Retirement in Europe: First results of SHARE*, University of Mannheim.
- 80 Philipson, T., and D. Lakdawalla (1999), *Ageing and the growth of long-term care*, NBER w.p. nr. 6980.
- 81 Pickery, J., and G. Loosveldt (1999), *The respondent, the interviewer and the questions as sources of item nonresponse*, Paper presented at the International Conference on Survey nonresponse, Portland, Oregon.
- 82 Prince, M. and M. Dewey (2005), Mental Health, in: A. Börsch-Supan et al. (eds), *Health, Ageing and Retirement in Europe: First results of SHARE*, University of Mannheim.
- 83 Reil Held, A. (2001), *Einkommen und Sterblichkeit in Deutschland: Leben Reiche länger?*, Universität Mannheim, Sonderforschungsbereich 504.
- 84 Rust, J. (1999), A Structural Model of the Disability Claiming Process, mimeo, Yale University.
- 85 Santos-Eggimann, B., J. Junod, and Sarah Cornaz (2005), Health Services Utilisation in Older Europeans, in: A. Börsch-Supan et al. (eds), *Health, Ageing and Retirement in Europe: First results of SHARE*, University of Mannheim.
- 86 Santos-Eggimann, B. , J. Junod, and Sarah Cornaz (2005), Quality of Care delivered to Older Europeans, in: A. Börsch-Supan et al. (eds), *Health, Ageing and Retirement in Europe: First results of SHARE*, University of Mannheim.
- 87 Schnall, P., K. Belkic, P. Landsbergis, and D. Baker (eds.) (2000), *The Workplace and Cardiovascular Disease, Occupational medicine*, State of the Art Reviews 15.
- 88 Siegrist, J., O. v.d. Knesebeck, and M. Wahrendorf (2005), Quality of Employment and Well-Being, in: A. Börsch-Supan et al. (eds), *Health, Ageing and Retirement in Europe: First results of SHARE*, University of Mannheim.
- 89 Siegrist, J., O. v.d. Knesebeck, M. Hyde, P. Higgs and A. Kupfer (2005), Quality of Life and Well-Being, in: A. Börsch-Supan et al. (eds), *Health, Ageing and Retirement in Europe: First results of SHARE*, University of Mannheim.
- 90 Sloan, F., and E. Norton (1997), Adverse selection, bequests, crowding out, and private demand for insurance: evidence from the long-term care insurance market, *Journal of risk and uncertainty* 15 (3), 201-19.
- 91 Smith, J.P. (1999), Healthy bodies and thick wallets: The dual relationship between health and economic status, *Journal of Economic Perspectives* 13, 145-166.
- 92 Statistisches Bundesamt (2000), *Handbuch für das Europäische Haushaltspanel*, available under: www.statistikbund.de/download/micro/absch_1.doc.
- 93 Steel, N., F. Huppert, B. McWilliams and D. Melzer (2003), Physical and cognitive functioning, in Marmot et al. (eds), *Health, wealth and lifestyle of the Older Population in England: The 2002 English Longitudinal Study of Ageing*, The Institute for Fiscal Studies, London.
- 94 Stuck, S. (2005), Country specific coding in SHARE. Mimeo, MEA, University of Mannheim.
- 95 Suzman, R.M., D.P. Willis, and K.G. Manton (eds.) (1992), *The Oldest Old*, Oxford: Oxford University Press.
- 96 Taylor, R., L. Conway, L. Calderwood and C. Lessof (2003), Methodology, in Marmot et al. (eds), *Health, wealth and lifestyle of the Older Population in England: The 2002 English Longitudinal Study of Ageing*, The Institute for Fiscal Studies, London.

- | | |
|----|---|
| 97 | van Soest, A. and A. Kalwij (2005), Item Response, in: A. Börsch-Supan et al. (eds), <i>Health, Ageing and Retirement in Europe: First results of SHARE</i> , University of Mannheim. |
| 98 | Weber, G. and O. Paccagnella (2005), Household Income, in: A. Börsch-Supan et al. (eds), <i>Health, Ageing and Retirement in Europe: First results of SHARE</i> , University of Mannheim. |

5.4 COUNCIL OF EUROPEAN SOCIAL SCIENCE DATA ARCHIVES – CESSDA

Council of European Social Science Data Archives (CESSDA)



The facility: CESSDA is a distributed RI that provides and facilitates access of researchers to high quality data and supporting the use of them. The CESSDA network of organizations currently extends across 21 countries in Europe. It holds some 15,000 data collections and provides access to over 20,000 researchers. CESSDA already operates within a global data environment. Data access arrangements and agreements are already in place with other data holding organizations worldwide. The international dimension and function is an important element of the proposed major upgrade.

Background: The CESSDA data archives have a long-standing record for the acquisition, support and supply of data range across official government censuses and surveys, election studies, longitudinal and cohort studies, opinion polls, and surveys addressing most issues relating to society and human activity. The CESSDA organisations provide access gateways to important European (and International) data materials and EU investments such as the European Social Survey, the Eurobarometers, the International Social Survey Programme and the European Values Surveys.

cial government censuses and surveys, election studies, longitudinal and cohort studies, opinion polls, and surveys addressing most issues relating to society and human activity. The CESSDA organisations provide access gateways to important European (and International) data materials and EU investments such as the European Social Survey, the Eurobarometers, the International Social Survey Programme and the European Values Surveys.

What's new? Which impacts? CESSDA suffers from a number of weaknesses. In effect CESSDA needs to create a European "passport", which enables researchers and data alike to move virtually across national and organisational boundaries. A major upgrade is proposed to develop CESSDA from the current situation in which the member organisations work with limited national resources to create a common platform, sharing a common mission, to a fully integrated data archive infrastructure for the SSH. This major upgrade will discover datasets and data-related materials in a cross-national environment; understand in detail the origin, methodology and structure of the underlying data collections; compare and link data from different locations; connect to other experts and share analyses, experiences and knowledge; enforce confidentiality and intellectual property rights whilst maintaining accuracy, security and open access to data sources; preserve and maintain data collections over time; design new research instruments. Although initially looking to serve the CESSDA community, the technologies, tools, resources and standards used are such that they may be exploited in order to be used by other agencies.

Timeline for construction and first operation with related estimated costs: CESSDA will continue to operate in a distributed way. Construction costs will be zero, given that the purpose is to upgrade rather than construct CESSDA. Decommissioning costs will be zero as it is expected that CESSDA will continue in operation for the foreseeable future. The upgrade costs are 30 M€ covering the upgrading of the existing technical RI (common standards, tools, instruments and services through the creation of middleware); capacity building (a hub for strategic development, maintenance and coordination); supporting less-developed and less-resourced organisations; and extending and deepening the CESSDA network to new and associated CESSDA-Members

Leading consortium: The CESSDA network has been in existence since the mid 1970s. The member organisations are funded entirely by national contributions, usually via ministries and national research councils as well as public funding from the education/university sector. Across Europe some 200 staff are employed within these CESSDA member organisations.

5.4.1 UPGRADING THE CESSDA RI – THE SCIENTIFIC CASE

5.4.1.1 Background

‘The ability of Europe’s research teams to remain at the forefront of all fields of science and technology depends on their being supported by state-of-the-art infrastructures.’ (EU document providing general description of the action in support of research infrastructures in the specific programme in *Structuring the European Research Area*).

In 2004 the report of the European Strategy Forum for Research Infrastructure (ESFRI) working group in the Humanities and Social Sciences (the EROHS report) recommended the establishment of a European Research Observatory which will build upon existing resources and both actively and systematically promote synergy and coherency. Subsequently, this proposed EROHS Observatory has been recognised by ESFRI as one of only two social science infrastructure initiatives to be placed on the formal ‘List of Opportunities’ presented to the Commissioner.

This proposal for a major RI upgrade is entirely in keeping with the general principles and vision set out in the EROHS report. To quote directly from the report, such an enhanced RI would ‘strengthen interdisciplinary and cross-border collaboration and comparative research on a European dimension. Further, it will enhance the building of research infrastructure capacity in the less resourced European countries of today. Finally, it will increase the opportunity to improve knowledge on social processes and thus holds great potential in terms of advising European and national policy-makers on how to manage the challenges currently faced by the societies of Europe.’

Building on the existing operational structure of the Council of European Social Science Data Archives (CESSDA) network of data archives and its associated partners, this proposal for a major RI upgrade will seek to address the major gaps and deficiencies identified by the EROHS report on Research Infrastructures 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 and the **generation of new and genuinely European data** for the comparative researcher.

5.4.1.2 Rationale

The over-riding objective of upgrading the existing CESSDA RI project is to develop a social science of the highest quality in order to ensure that European researchers may have access to the data resources they require to conduct comparative cross-national research, irrespective of their location in the infrastructure.

Data are the single most important component necessary for a science-based understanding of society and to promote and facilitate access to these data is to promote research. Although there have been significant advances in recent years in making data available for scientific use, these advances have not been European-wide. There still remains a large difference in both the actual availability of data as well as in the value that is attached to accessing data across Europe.

In the present environment comparative research across countries in Europe is mainly restricted to the analysis of data from a single dataset, collected specifically for that task. The resources involved in identifying and harmonising potentially comparable data from multiple and diverse datasets across geography or time, make such an activity economically unviable. The main objective of this proposal hopes to rectify this situation; its success will be measured in the future increase in the amount of comparative research using data from distributed sources.

This main objective will be achieved through providing support for the operation, enhancement, extension and exploitation of existing synergies of the CESSDA network and its associated partners. It will focus on four key scientific objectives.

- 1) **Support data-based research and decision making** by a) fostering interoperability across heterogeneous technology domains, which allows the seamless linking of social science data for large-scale horizontal (cross dataset) and parallel (simultaneous) analysis, and thereby reinforce user community engagement, promote collaboration amongst researchers in Europe and enable leading-edge research; b) providing quality associated knowledge-products and resources, thus creating bridges between texts in electronic journals, question banks and the underlying data.
- 2) **Improve mechanisms for the data resource discovery of virtual distributed data collections** located both within and outside of the existing CESSDA network through the further development of underlying semantic Web technologies (standard metadata, thesauri and ontologies).
- 3) **Increase resource sharing and widen access to archived social science data** by breaking down both 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.
- 4) **Ensure the security of data objects** through the

development of management tools addressing authentication, authorisation and statistical confidentiality.

Traditional comparative research will be greatly enhanced should these scientific objectives be fulfilled, reducing the time required to locate, gain access to and understand comparative datasets as well as making the task of designing and implementing new comparative surveys, such as the European Social Survey, that much easier. Thus the success of this project will be measured in the general increase in comparative social science research as a whole.

In enhancing and further facilitating data sharing within the social sciences, one of the key aims of this upgrading of the RI is not only to make access to data easier, especially across national boundaries, but in so doing increase data usage and, thereby, increase the volume of high quality cross national and cross comparative social science research.

In upgrading the existing CESSDA RI it is essential to address the needs of the social scientific community to:

- **Discover** datasets and data-related materials in a cross-national environment;
- **Understand** in detail the origin, methodology and structure of the underlying data collections (social science datasets are generally modest in size but big in complexity);
- **Compare** and link data from different locations;
- **Connect** to other experts **and share** analyses, experience and knowledge;
- **Enforce confidentiality** and intellectual property rights whilst still maintaining accuracy, security **and open access** to data sources;
- **Preserve** and maintain data collections over time;
- Design new research instruments.

The following table summarises some of the detailed scientific and technological objectives of this proposal and how they address the above needs of the social scientific community.

Discover	Understand	Compare	Connect and Share	Enforce confidentiality and open access.	Preserve	Design new research.
via a European data portal	via enhanced metadata incorporating controlled vocabularies for origin, methodology and concepts used in component structures	via the inclusion of contextual metadata	via the additional knowledge products available from the European data portal	via the development and adoption of management tools addressing statistical confidentiality	via the inclusion of fixidity and version control metadata	via the creation of a European question bank, registry and related tools
via an extended European language multi-lingual thesaurus	via translation of metadata using a European language multi-lingual thesaurus	via mappings between European standard category codes for individual countries	via the creation of a 'Wiki' to enable comments from the European research community to be added to the data and the multi-lingual thesaurus	via the development and adoption of management tools addressing authentication and authorisation	via training for a new generation of data archivists	via grid-enabling data and associated resources
via a unique reference id for major components such as data set, questionnaire, variable	via component structures of data sets and questionnaires, and the variables and code lists they use	via the creation of a European question bank registry	via new opportunities for knowledge transfer between data archives	via a single registration agreement to enable researchers and data to move virtually across national and organisational boundaries		
	via the creation of a European question bank registry linking published data to the original source					

Table: Addressing the needs of the social scientific research community

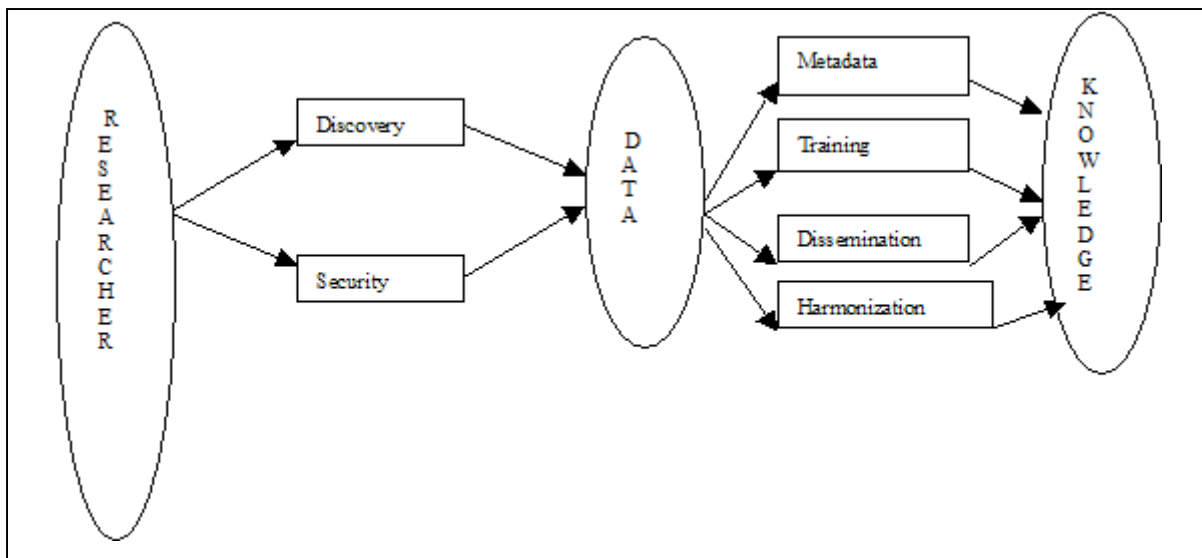
5.4.1.3 Identifying needs

This proposal aims to integrate and strengthen the European Social Research Area by furthering the availability of and access to European data. It will achieve this objective by strengthening and extending the existing CEESDA RI in key areas:

- metadata standards;

- tools for resource discovery;
- tools for data dissemination;
- tools for data security;
- tools for data harmonization;
- training resources and human capacity building

These key objectives can be seen to inter-relate between researchers and the development of knowledge in the following way.



The above diagram shows how a researcher's access to the data is made through resource discovery tools and security systems, and that any knowledge gained from those data is obtained through the metadata, training products or harmonization tools that are delivered within the dissemination system.

In addition to upgrading the underlying technical infrastructure of the CESSDA network, building the middleware that underpins the creation of unified data documentation, integrated data storage, data preservation and universal data access systems, and developing the training and human capacity building, as mentioned above, effort centrally is also required to both strengthen and deepen CESSDA involvement within the European Research Area.

First, a special programme is needed for the smaller, less-developed and less-resourced CESSDA member organizations in order to enable them contribute fully and on an equal basis to the CESSDA-based programme of activities. This not only involves direct help in implementing centrally developed middleware tools, but may also extend to special localised projects to support the acquisition of important national data resources, or the translation of metadata into English in order to facilitate wider accessibility.

Second, a special seed-money programme is needed in order to extend the existing CESSDA network and foster the development of national data archiving initiatives in those countries which are not currently part of CESSDA. The obvious purpose of this activity is to spread the CESSDA-network to each EU-member state and to create and maintain a 'complete' pan-European SSH network, including representation from emerging and candidate countries. Equally, some countries with organizations within CESSDA have specialised research-led project-based teams whose data currently fall outside of the CESSDA network, and mechanisms likewise need to be put in

place to create better comprehensiveness and coherency.

5.4.1.4 Potential impact

CESSDA has provided a Research Infrastructure for the social sciences for the past 30 years through the acquisition, support and supply of data for the European social science research community. Over this period it has grown both in geographical and substantive terms. The CESSDA network now extends to 21 countries across Europe and also has associate partners in many of the emerging member states (formerly referred to as the East European Data Archives Network (EDAN)). Collectively the constituent CESSDA organisations serve some 10,000 social science researchers across Europe each year, providing access to and delivering some 40,000 data collections per annum and acquiring a further new 1,000 data collections per year for research purposes. In addition, the CESSDA organisations provide access gateways to important social science data materials and EU investments such as the European Social Survey, the Eurobarometers, the International Social Survey Programme and the European Values Surveys.

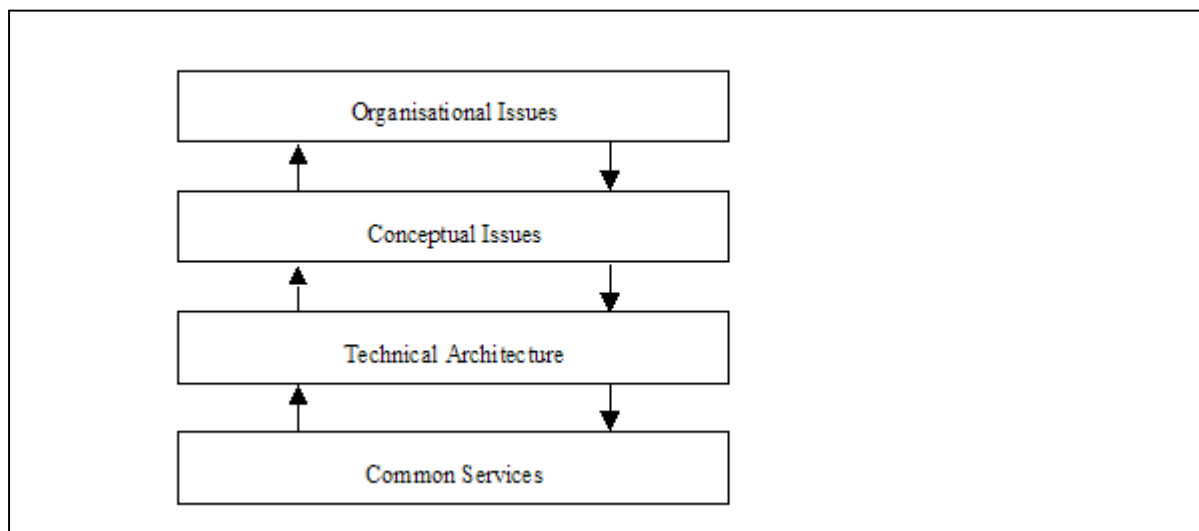
In recent years CESSDA has developed, largely through a series of EU-funded projects, a number of complex tools to facilitate cross-national resource discovery (LIMBER - Language Independent Metadata Browsing of European Resources; MADIERA - Multilingual Access to Data Infrastructures of the European Research Area; MetaDater - Metadata Management and Production System for Surveys in Empirical Socio-economic Research) and data management and access (NESSTAR - Networked Social Science Tools and Resources; FASTER - Flexible Access to Statistics, Tables and Electronic Resources). Each of these has incrementally built upon one another to produce a cross-national infrastructure for social science research.

Despite the important developments which collec-

tively form the basis of the current CESSDA Infrastructure, much work remains to be done in order to transform the existing set of arrangements into a fully-functioning RI in which data resources can be easily and seamlessly located and accessed; that enables participants to share computing and information resources across organisational boundaries in a secure and highly efficient manner; that enables re-

search-oriented organisations to solve problems in ways that were previously not feasible, due to existing computing, data access or data integration constraints.

The proposed RI upgrade is composed of four distinct, yet inter-related, layers:



These four layers are explained below.

5.4.1.5 Organisational Issues

The creation of a fully functional RI for social science will only be possible if the underlying contractual requirements for the virtual movement of researchers or data across national boundaries are in place. Regardless of the technical solutions in place, the problem fundamentally rests on the solving of a number of key organisational issues. In effect CESSDA needs to create a European ‘passport’ which enables researchers and data alike to move virtually across national and organisational boundaries.

A further concern that needs to be addressed is the fact that while the current CESSDA network is extensive, including 21 countries, it is not totally comprehensive. Equally, the CESSDA network is currently rather heterogeneous, with some country members being younger and less-developed. These imbalances need to be addressed. Obviously certain social science data collections currently reside in organisational repositories outside of the existing CESSDA network, and the proposed upgrade will need to consider how these can be incorporated into the CESSDA RI, and thereby become widely accessible and available to European researchers. Regardless of the technical infrastructure at hand, organisational issues need to be resolved before such collections can be plugged into the upgraded RI or a wider integrated European Data Grid. Action is needed on two fronts: to *strengthen* and develop capacity in the smaller existing CESSDA organization; and to *widen* participation

in the CESSDA RI both directly by fostering membership in new countries and indirectly by deepening involvement and extending the Data Grid to agencies and organizations who which to remain outside of CESSDA yet continue to host important data collections.

It is obvious that the work proposed within this submission can only be carried out at a European level.

5.4.1.6 Conceptual Issues

If sharing is the most important single keyword characterising a working RI or Data Grid, the key to realising the benefits of such technologies is standardisation through agreed metadata standards. Standardisation facilitates the development of middleware so that the diverse resources that make up the RI or Data Grid 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. There is a pressing need to develop a life-cycle model for social science data and to identify and formalise the descriptive and functional metadata requirements of this process.

In order to facilitate the free movement of data and solve identification and version control issues, the concept of a unique dataset reference id, somewhat akin to the ISBN of the published book world, needs to be addressed.

5.4.1.7 *Technical Architecture*

Social science data often consist of information streams relating to individuals and organisations. Due to increased concerns regarding privacy and confidentiality within social science research there is increasingly a trade off between data content in terms of the depth and detail at which data are made available and free open data access. This not only limits the potential of the original dataset for academic research but also is inefficient as it requires an additional process between data generation and data release.

To address this problem middleware is required that matches a potential user's access rights against differential confidentiality risk levels and automatically produces a view or version of the underlying data that satisfies both sides of the equation and, in so doing, addresses appropriate statistical disclosure requirements. Linked to this problem, middleware is also necessary which allows full and proper authentication of registered users, thus ensuring security of the system and the data contained within it.

5.4.1.8 *Common Services*

Building on these above layers, this upgrade will turn the existing CESSDA RI into a fully operational, one-stop shop portal through which researchers across Europe can register, search for and gain access to data collections and associated knowledge products housed in a fully distributed environment. In essence, this is the primary goal of this submission.

5.4.1.9 *Summary*

In summary, therefore, the main strategic impact of this application to upgrade the existing CESSDA RI and in so doing build the foundations for an integrated European social science Data Grid will be to address the major gaps and deficiencies identified by the EROHS report on Research Infrastructures. In will achieve this 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 **generation of new and genuinely European data** for the comparative researcher.

Although initially looking to serve the CESSDA community, its associated partners and, through these, researchers across Europe, the technologies, tools, resources and standards used are such that adoption by other organisations would be relatively straightforward, and thus, interoperability with other data resources held in alternative systems and metadata formats would be made that much easier. This also means that these technologies, tools, resources and standards may be exploited in order to be used by these other agencies, especially National Statistical Offices and other digital data repositories.

5.4.2 UPGRADING THE CESSDA RI – A BUSINESS CASE

5.4.2.1 *Outline*

In order to achieve the objectives summarised in the Scientific Case, this proposal initially suggests a programme of work of consisting of four main functions and twelve initial sub-tasks as outlined in the summary table below and expanded in the sections that follow:

5.4.2.2 *Function 1 – Technical Development*

Upgrading the existing technical RI developing common standards, tools, instruments and services through the creation of middleware.

- Task 1 – Development of Access and Publishing Tools for the C-CAT portal
- Task 2 – Enhancing Metadata Standards
- Task 3 – Extending the Multi-lingual Thesaurus
- Task 4 – Developing Middleware for Authentication and Authorisation
- Task 5 – Developing Middleware for Statistical Disclosure

This list of tasks represents only a preliminary list. New tasks would evolve over time and be identified by the Expert Steering Group and in collaboration with such bodies as EROHS. Obvious candidates would include the development of a European Question Bank Repository, Grid-Enabling of key resources; creation of harmonised data collections; mapping of coding schemes and classifications.

5.4.2.3 *Function 2 – Capacity Building*

Providing strategic development, co-ordination, management and training.

Task 1 – A Hub for Strategic Development, Maintenance and Coordination

Task 2 – Training and Exchange Programmes

5.4.2.4 *Function 3 – Strengthening*

Supporting less-developed and less-resourced organisations in order to create greater homogeneity within the European Research Area.

Task 1 – Direct Help

Task 2 – Language Harmonization

Task 3 – National Upgrades

5.4.2.5 *Function 4 – Widening*

Extending and deepening the CESSDA in order to create and maintain a 'complete' pan-European SSH network.

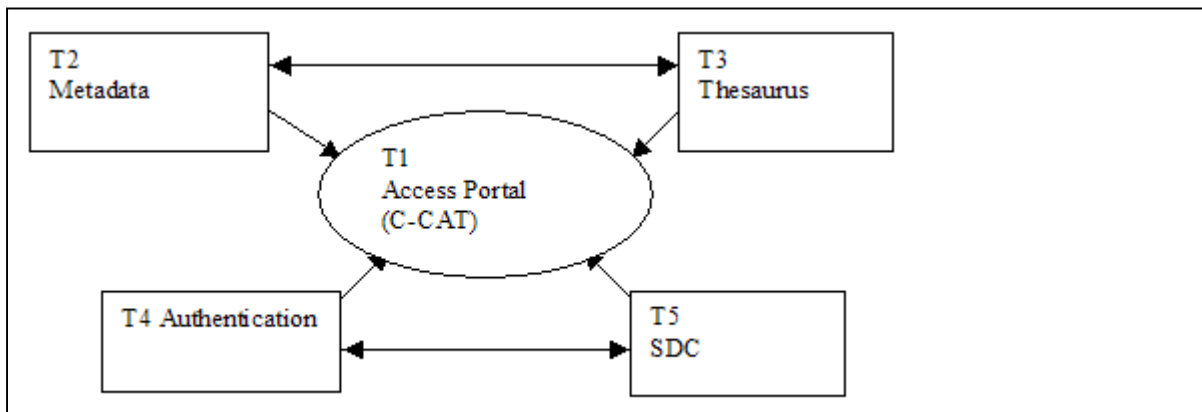
Task 1 – Integrating New National CESSDA-Members

Task 2 – Integrating Associated CESSDA-Members

All of the five technical tasks are inter-related and will benefit from cross-working and synergies. However,

only Task 1 (the Access Portal) is mutually dependent on the other four technical tasks. The inter-

relationship between these tasks can be summarised in the following diagram:



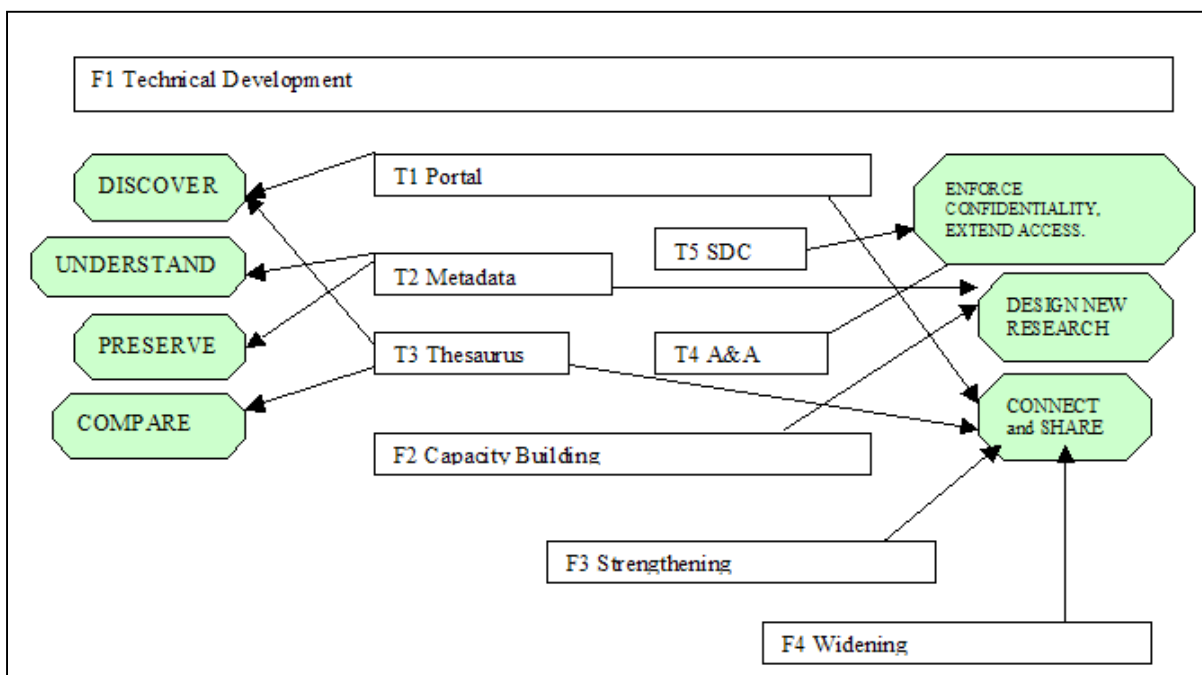
As the above diagram shows, Task 1 with its emphasis on the development of the C-CAT Access Portal is central and dependent on each of the other technical tasks. In this sense Task 1 will implement the middleware tools to be developed for authentication and authorisation under Task 4 and for statistical disclosure control under Task 5, and will also integrate the thesaurus system developed under Task 3 and the metadata standard improvements made under Task 2.

standard under Task 2 will feed into aspects of the thesaurus system development and the question bank system development, while the SDC middleware modules developed under Task 5 will be required to integrate fully with the registration and authentication middleware tools produced by Task 6.

Sitting across all these Technical tasks, of course, will be the Training, Strengthening and Widening Functions which will not only require the implementation of the outputs of the Technical development tasks.

Strong integration and synergies will also occur between the metadata Task 2 and the thesaurus Task 3 and similarly between the authentication and authorisation Task 6 and the statistical disclosure Task 5. In particular the work on extending the DDI metadata

The relationship of these Functions and Tasks with the identified objectives of addressing the needs of the social scientific community (see table 1) can be represented as follows:



Taking these identified Functions and Tasks as a single ensemble, the resulting upgraded CESSDA RI will have a critical impact on the social science re-

search community and will become the major research infrastructure of its kind. The availability of this upgraded RI will provide an immense opportunity for social scientists. Such a facility will

nity for social scientists. Such a facility will enable researchers, not only between disciplines but also between countries, to work together, developing leading-edge research methods and efficiently analysing the large datasets which constitute the major existing infrastructure in the social sciences. In essence, integration through a Data Grid platform could make it possible for social scientists to sit at their computer, locate, access, merge and analyse data from a number of different sources. Based on Open-Source principals and agreed metadata standards the RI will also built the foundations for interoperability and cross-overs with other domains, facilitating the potential for increased cross disciplinary research.

5.4.3 FUNCTION 1 – TECHNICAL DEVELOPMENT

5.4.3.1 Task 1 – Development of Access and Publishing Tools for the C-CAT portal

5.4.3.1.1 Key Objectives

To establish a one-stop shop for data location, access, analysis and delivery across the social science community of Europe;

- To increase the type of data available through such a portal by refinements in existing software for data publishing, data linkage, comparison and harmonisation, and to include data held outside the CESSDA and its associated partners;
- To create a more dynamic knowledge management oriented Web where knowledge-products are fed back into the metadata supporting the data, thus creating bridges between text in electronic journals and the underlying data;
- Expand on present metadata through collecting and disseminating community-produced metadata;
- To ensure quality data and services through the implementation of best practice resources.
-

5.4.3.1.2 Description of work

The present CESSDA infrastructure is based on those national data archives that have implemented a Nesstar server and published their data holdings with accompanying metadata marked up to the DDI XML standard. The Madiera portal brings these distributed resources together through a single Web entry point that utilises browsing techniques based on a multi-lingual thesaurus. The tools available through this portal for resource discovery, authentication and access, data analysis and delivery will be enhanced to cover the needs of researchers wanting to investigate cross national phenomena from various distributed data resources. This workpackage will aim to further develop the metadata concept and data location, analysis and visualisation tools, including translation, data harmonisation, resource integration, thesaurus,

and graphical and cartographical tools. The current CESSDA gateway will be developed into an integrated and effective distributed social science portal giving researchers and decision-makers access to a range of data for instant analysis.

It is essential that the data available to a European social science data infrastructure is not restricted and that tools exist to make such a variety of data readily available and useable. The publishing tools will be refined to ensure that the metadata being published is of a sufficient quality and richness to enable comparison and harmonisation between datasets. Publication software will be enhanced to ensure compatibility with other standards besides DDI; to improve methods for semi-automated indexing from multi-lingual thesauri; to deal with metadata for data complexities and spatial coverage and to incorporate instrument metadata capture. It is also an explicit aim to expand on the metadata concept, to develop the concept of secondary metadata to include knowledge and experiences stemming from substantive use.

The present CESSDA infrastructure is also inconsistent in data coverage across countries. This activity will investigate two methods that could enhance European data content. Firstly interoperability with other existing data resources, and secondly harmonised acquisition policies for CESSDA to address the imbalance of data coverage.

Given that the challenge of bringing substantial amounts of social science data into a Data Grid will be overcome, the next natural step is to create bridges between these data and the knowledge products generated from them. These bridges will vary in functionality and purpose. They can be linked references from a scientific text to the data used in the research or links directly embedded in tables or graphs that reproduce the underlying statistical analyses on the data or, indeed, links from the metadata of a study to on-line scientific texts that are produced on the basis of the study. The concept of secondary metadata will be actively developed and specified.

On top of the software tools the portal should also provide instructional guides. These on-line guides will form the basis of a best practices Web resource, including templates for data complexities, controlled vocabulary listings, indexing guidelines and OAI recommendations. Best practice guides provide rules needed by organisations and individuals to be effective participants in producing, providing and using data on a European level. In particular they will address a range of data providing functions and expand consensus on the elements and processes for data description and data access. Similar to the criteria of the OAI (Open Archival Information System) a consistent standard will be developed that is based on a number of unifying concepts to overcome the barriers to reliable comparative research using existing

data. The OAIS Reference Model provides a high-level definition of the environment of an archival information system, defines the function of a digital archive, and presents a logical model for the information stored in an archive.

5.4.3.2 Task 2 – Enhancing Metadata Standards

5.4.3.2.1 Key Objectives

- To expand the existing data models to cover complex dataset types, like the repeated cross-national datasets, the repeated cross sections, the harmonization studies, panels, hierarchical datasets, time-budget studies.
- To contribute to the activities of the DDI Alliance in its efforts to expand the DDI standard and take benefit of the experience of the other members of the Alliance
- To create appropriate documentation of the developed data models
- To collect information about international standard or harmonized classifications and the conversion keys used to recode from the local classification schemes
- To create a database for managing the information collected and publishing it on line
- To create the web application, which gives access to the information collected
- To draft a plan about how the information can be periodically updated

The CESSDA community is planning to build a RI whose scope is to give world wide access to the data stored by its members. The main challenge lies in the dimensionality of data in the social sciences, i.e. the multiple dimensions on which the data transferred must be defined. Social science data bear meanings which have to be interpreted. The interpretation is grounded on a detailed knowledge of the conditions under which the data were collected: question structures and formulations, categories from which the respondent has to choose from, data collection method, scope of the data collection project, time and space co-ordinates, etc.

The data model currently in use is appropriate for simple cross-section studies and needs to be extended to cover more complex data designs in a well defined manner. This work will build on the activities of the MetaDater project, which started with the modeling of the repeated cross-national study and on the work done by the DDI Alliance – of which several partners to this activity are members.

A well documented model covering several types of complexities will be developed and made available to the CESSDA community as well as to the DDI. This development will make it possible to continue the development of applications like Nesstar. The services and applications using the model will be inter-

operable, a condition for the development of an efficient *e*-social science community.

5.4.3.2.2 Description of work

1) *Selection and analysis of complex dataset types* – The dataset types to be handled will need to be selected and resulting conceptual work undertaken.

2) *Modeling complex dataset types* – It will be possible for the modelers to retro-act on the conceptual work and ask for refinements where necessary. The resulting UML model will be accompanied by a discussion of the interfaces to and differences with other models already in use in the CESSDA community, namely the current DDI, the MetaDater model and the Nesstar metadata model.

3) *Documentation of the models* – Data models are usually not self-sufficient and have to be accompanied by a documentation giving the keys to the core characteristics. This documentation will be done by one of the partners, normally not the one responsible for the final presentation, under the control of all others, acting as a potential public. Whatever the techniques used for documentation, there will be a hyperlinked interface to it, to warrant an easy and navigable access to it.

4) *Cooperation with DDI* – The Data Documentation Initiative data model acts presently as an important international standard for social sciences metadata. Several CESSDA organizations are involved in the Expert group, which elaborates extension to the version of the model presently in use. The models elaborated by this task will be fed into the Expert Group's work as a well prepared CESSDA input. Reports will be elaborated to account for the actual collaboration and the progress made.

The last four objectives address another level of complexity in metadata management for the social sciences: the variety of classifications used in various countries or cultural contexts for rendering key variables, which are at the core of any cross-national or cross-cultural comparison. It has a development aspect in that appropriate information structures must be defined and implemented to show relevant information; it has also an informational aspect, since information will be collected to start populating the information system.

To support more efficiently comparative analysis between datasets stored in the European archives, this activity will collect and redistribute information about existing standard or harmonized classifications, locally used classification and conversion keys for a limited number of key variables. If available, discussions on conversion difficulties or critical remarks will be integrated into the information system. The information system will be structured in such a way that makes it available to integrated changes in classifi-

cations and multiple conversion keys for the same pairs of classifications. Links to publications will help users of the system going deeper into the issues involved.

The information will be made available on line, either as a part of the C-CAT portal or as a distinct portal linked to the C-CAT portal. In the long run, this work will make it possible to create in the application used in the C-CAT data portal a function which computes the standard (and comparable variable) from a classification for which a conversion key to the standard exists.

5) Identifying key variables for cross-national comparison; defining the dimensions of the information to be collected and distributed – This will identify selected key variables and the dimension of information to be collected. The work will be undertaken on two strands, one dedicated to the development of the information base and its publication, the other dedicated to the collection of relevant information in each of the partner's countries.

6) Create a data model for handling the collected information; create an application for handling the collected information; create a web-application giving access to the collected information – A data model must be elaborated to handle the information collected: current classifications, standard classifications in various versions, conversion keys from current classifications to standard classification, documentation of problems encountered in conversion. An interface to the most used data models in the CESSDA community will be proposed. A prototype database will implement the data model, so that the information collected by the partners can be stored and managed in an appropriate way. The design of that database will take into account the requirement of publishing the collected information on the web.

7) *Publish the information available on-line and elaborate plans for the update of the information available* – The on line publication of the collected and edited information will make it broadly accessible. The ELSST vocabulary will be extended to mark up in the data holdings of the archives the classifications for which conversion keys are available. A strategy will be proposed to update the information available as versions become available for classifications and conversion keys in the database. The information system would also be available for adding new key variables to the set initially considered.

5.4.3.3 Task 3 – Extending the Multi-lingual ThEsaurus

5.4.3.3.1 Key Objectives

- To create a centralised resource for the European social science research area.
- To create mechanisms for the maintenance and management of a multi-lingual thesaurus for use by European-wide organisations.

- To extend the existing languages of German, Spanish and Italian, to be in line with the other 7 European Languages already in the multi-lingual thesaurus (ELSST).
- To expand ELSST to include further European languages, including those from eastern countries.
- To enhance comparative research and harmonisation of datasets by the construction of controlled
- vocabularies for metadata.
- To make the maintenance of the thesaurus and related resources self-financing.

5.4.3.3.2 Description of work

1) *Translation of outstanding concepts into Spanish and German.* Including the distribution of hierarchies and the clarification of concepts through the existing comments interface and database with the addition of scope notes to the thesaurus as necessary. After the actual translation of the concepts, consistency checks will be performed and a peer review undertaken culminating in the release of ELSST version 2.3.

2) *Translation of all ELSST concepts into additional European languages.* This will involve the same processes as outlined above and culminate in the release of ELSST version 3.

3) *Maintenance and management tools for ELSST.* The development of software to deal with candidate term proposals, new term additions, tracking of translation status, version control and history, term deletion and local extensions. This will involve software design and specifications, the development of an initial prototype and its evaluation, the software then being refined and resulting in a final version of tools.

4) *Centralised resources for ELSST.* The development of Web resources to provide on-line tutorials on navigating and using the thesaurus both for indexing and as a finding aid, a shopping basket collection of index terms from the thesaurus and assignment to resources by downloading in certain metadata formats or directly to databases. This will involve design and specifications, the development of an initial prototype and its evaluation, the Web resource then being refined and resulting in a final version.

5) *Develop self-financing mechanisms and products.* This will include the establishment of a subscription scheme for organisation outside of CESSDA and the production of national bilingual thesauri. It would also need to establish intellectual property rights and copyright frameworks.

6) *Construction of controlled vocabularies to aid comparative research.* After the identification of DDI metadata elements needing controlled vocabularies, terms from the ELSST thesaurus will be selected to construct

controlled vocabularies.

5.4.3.4 Task 4 – Developing Middleware for Authentication and Authorisation

5.4.3.4.1 Key Objectives

- To review the Trans Border Access Agreement and CESSDA members' access arrangement(s) and requirements.
- To establish a one-stop shop for user registration, allowing users to register once for access to services from each member.
- To apply a common user interchange architecture to the cross-border system, enabling standardised user authentication and authorisation.

The CESSDA community of archives holds a vast quantity of highly valuable data. Each archive maintains agreements with its depositors to make these data available to some or all members of its own country, and sometimes to other nations. Should a user from one country within the CESSDA community wish to make use of the data from another country's archive, a mediated, archive-to-archive transaction takes place. The practical transfer of data is currently undertaken primarily via communication between officers in the requesting archive and the supplying one. The work in this task will speed up this process by providing a mechanism which will take the user directly to the data. It will create a system of direct user-to-archive communication by establishing a one-stop shop registration and access service.

The system envisaged will allow users throughout Europe to register for access to any CESSDA archive locally, log into the archive of their choice and download those data that are available to them. This activity will produce a registration and authentication module which can be plugged into member archives' systems. It will fast test and demonstrate this functionality between the UK and Norway, integrating it with the C-CAT portal (Task 1), thus providing users with a streamlined, user-friendly and consistent access path into data resources. The service will also be developed in order to integrate with the SDC module of Task 5 in order to enable user risk analysis to take place based on users' registration characteristics and authentication attributes.

The system will allow local control over access rights as well as local registration within a dispersed environment. Local registration will retain the visibility of smaller archives and allow each archive to apply the restrictions required by its own, unique data depositors, but will also mean that any user from any CESSDA archive will be able to gain access to data held elsewhere, by virtue of the properties associated with their usernames being carried to any archive's resource. The solution envisaged will thus be sophis-

ticated, allowing for local flexibilities, but also simple to apply. Its modularity will mean that new non-CESSDA, data-holding organisations, whether national or sub-national, will also be able to contribute to the portal by joining the access system.

5.4.3.4.2 Description of work

1) *Review of TBAA/local access arrangements.* This will require Desk research, a fact-finding survey and intensive discussion and regulation via workshops and focus groups, in order to review the existing TBAA and assess local archives' access arrangements and requirements. The results of the workshop will be analysed and an agreed list of registration requirements drawn up. Recommendations for a future TBAA will also be made.

2) *Creation of a one-stop registration service.* To speed up each user's route to data held in other archives and to provide greater flexibility of access, this activity will create a dispersed and modular registration and access system, building on the work of the EU-funded FASTER project and the past experience of the UK Data Archive in developing a one-stop shop for census, economic and social data services. The specification drawn up will be informed by the results of the TBAA and local access requirements review. Following the first release version, beta testing and further refinement will take place. After this, the module will be implemented in other candidate archives and its transferability tested. The connection between the last archives will be established and will serve as a pilot for implementing the cross-border registration and access system across the whole CESSDA RI.

3) *Application of standardised authentication protocols.* Standardised, international authentication (and authorisation) protocols will be evaluated, such as Shibboleth or the Grid Security Infrastructure from the GLOBUS toolkit. Following this, one protocol will be adopted and implemented in the UK and Norway.

5.4.3.5 Task 5 – Developing Middleware for Statistical Disclosure

5.4.3.5.1 Key Objectives

To design an interactive SDC model that integrates data dissemination and preservation procedures within an e-Social Science environment.

To write the software for and implement a prototype of the resulting model within Nesstar.

To install the final version of the resulting model within the C-CAT portal.

The inclusion of a disclosure control procedure in the process of data dissemination has several significant benefits. First, it minimises the loss of information imposed on current users by providing them with tailor-made subsets that better fit their research purposes than the standard PUF datasets, while omitting sensitive data. Second, it provides the data pro-

ducer and the data archive with the option of modifying parameters of statistical confidentiality over time. Under the suggested model, data producers can change levels of security or variable tagging whenever disclosure control standards are changed. Third, this approach maximises the amount of information available for future generations by allowing data archives to preserve further-detailed data. It also simplifies the data preservation task by reducing the number of data versions that have to be documented, maintained and preserved. Last but not least, it upgrades the role of European data archive to that of a 'senior partner' in the mission of encouraging and facilitating public use of data.

5.4.3.5.2 Description of work

1) *Reviewing the current technologies.* The current SDC methods, software and requirements will be reviewed, as will Web-based data dissemination systems.

2) *Designing the SDC module.* This Task proposes a model that integrates data dissemination and preservation procedures with Statistical Disclosure Control (SDC) in an e-Social Science environment. By moving part of the SDC operation from the phase of preparing a dataset to the point of accessing it, this model is expected to provide users with the data most appropriate for their research needs, while leaving richer data for preservation and consequently for future social, economic and historic research. The user will be able to define a subset and let a confidentiality clearance Web-engine check the defined subset against a given level of security. If a subset is cleared for a given user, it will be immediately available for download or for on-line analysis. Otherwise, the 'high-risk' variables will be displayed and the user requested either to remove them from the request or to let the system take care of their disclosure risk. In the process of modifying the requested subset, users who drop key or sensitive variables may try to add other "neutral" variables to 'reimburse' their research plan. Once the user replies, the simulator runs again and displays updated SDC results. This simulation process may go on for several iterations until both the user and the confidentiality clearance engine are satisfied or the user withdraws her application. This Activity would work closely with A3 and A6 in order to feed its results into the C-CAT portal and to ensure the authentication system was suitable for the SDC module's needs.

3) *Testing the module.* The module will be tested after design and after the prototype version has been installed in Nesstar. Additionally, the project model will be tested.

4) *Installing the modules.* The prototype will be installed in Nesstar and then, following testing and further refinement, will be introduced to the Data Grid via the C-CAT portal.

5.4.4 FUNCTION 2 – CAPACITY BUILDING

Providing strategic development, co-ordination, management and training.

This second function is focused on the construction of a co-ordinating hub or unit capable, on the one hand, of strategically developing, maintaining and improving the already existing tools and instruments (Task 1) and, on the other hand, of conducting a number of academic exchange programmes in the field of data documentation, data storage and data use, including training programmes for technicians and academic researchers. (Task 2)

5.4.4.1 Task 1 – A Hub for Strategic Development, Maintenance and Coordination

5.4.4.1.1 Key Objectives

- To coordinate the overall CESSDA upgrade
- To develop long-term strategies and policies for data availability and data comparability within the European Research Area
- To maintain and to continually improve the operability of the existing CESSDA-tools and instruments.

The key objectives for Task 1 follow directly from the basic function of capacity building in general and from the construction of a hub or a co-ordinating unit in particular. The hub will be responsible for the upgrade-process of CESSDA, will co-ordinate the different programmes and will work on the maintenance and the gradual improvement of the current infrastructure tools and instruments. Thus, the hub will have to fulfil a large number of organizational and administrative tasks, but also a substantial amount of technical maintenance work.

It should be added though, that the hub will assume an important role as a strategy centre for European data infrastructures and for European data policies and will act as an active adviser, evaluator and propagator for European data infrastructures both at the European level and at the national levels.

5.4.4.2 Task 2 – Training and Exchange Programmes

5.4.4.2.1 Key Objectives

- To develop training and exchange programmes for CESSDA-personnel.
- To introduce a special training programme for social scientists or technicians from new CESSDA-member countries.
- To initiate a number of international and national workshops and seminars on the topics of data documentation and data comparability for SSH-researchers and for SSH-students.
- To organize an annual Summer School for European students on the topics of data comparability and comparative research designs.

The second key task for the hub lies in the organization of a substantial amount of training and exchange programmes both for CESSDA staff members, European SSH-researchers and European students. Four different types of training programmes will be developed which will serve different producer and user groups.

The first key objective addresses the issue that the overall CESSDA upgrade will only function effectively if younger staff who currently provide day-to-day support for the network of data archives are encouraged to develop and exchange their knowledge and skills and so become the next generation of managers. The CESSDA community has a history of networking its junior staff members via short expert seminars for groups of young archive staff engaged in common activities such as data preparation, data distribution, end user and data depositor support. This enables staff across Europe who are 'at the coal face' to exchange views and ideas related to different working practices across Europe to feed these back to their managers by formal means. It is usually this same network of staff that need the skills and knowledge to implement new software and systems developed with project funding. The problem for most archives is that, although the costs of such meetings are kept as low as possible and the duration of the seminar is short, CESSDA as an organisation does not have a training budget and relies on its own individual archives having the resources to fund attendance at the seminars. Consequently, the hub will develop new formats for more extensive seminars and will provide support for some smaller organisations, especially those in emerging European countries. Additionally, an easier access of CESSDA staff members to international conferences, such as IAS-SIST will be made possible by financing the costs of attending these meetings. Thus the hub will include an annual amount for attendance at such events, to be competitive and justified on the training benefits expected.

Furthermore, another strand of intra-CESSDA training activities will be achieved by the establishment of staff exchanges between national data archives. The goal of these exchanges will be to encourage knowledge transfer between CESSDA partners. There is a notable imbalance of experience and funding between CESSDA and its associated archives. This is largely due to the simple fact that the longer established archives have had time to develop skills and expertise, to establish a sound funding base for their work and, critically, large user bases. It is a fact that, until archives can offer high quality resources and efficient services, their user bases are unlikely to expand. This activity will provide one means of overcoming some of the disadvantages that this causes by offering organisations, at whatever stage of their development, the opportunity to fill specific gaps in knowledge or skills by allowing sufficient time for practical, detailed

and in-depth knowledge transfer. There will be multiple benefits to such exchanges: the filling of skills gaps in the host organisation; career development for the exchange staff; long-term and sustainable improvements in the preservation of and access to cultural and scientific resources.

The second approach to training will focus on new CESSDA-members who are setting up a data archive for their national SSH-environments. Here one cannot assume a high familiarity and a technical acquaintance with available infrastructure tools and standards. Thus, the training programme will focus on a series of well-defined modules of one month to three months in the areas of data documentation standards, the utilization of available tools like NESSTAR, MADIERA, MetaDater or FASTER, the problems of data-comparability or key issues for comparative research designs. Some of these modules will be organized by the hub directly, some modules in collaboration with national data archives so that new CESSDA-members will become familiar with the day to day operations in other data archives and will be exposed to a continuous process of learning by doing.

The third objective ensures that the new developments arising from Function 1 are distributed and disseminated widely not only across the CESSDA community of archives and associated archives such as those of EDAN (Eastern European Data Archives Network), but to the European SSH community at large. Thus, the hub will also conduct workshops, seminars and conferences, partly in combination with national data archives, partly as hub-activities alone, in which new tools and instruments will be presented and discussed with SSH researchers across Europe. Moreover, the hub will take a strong initiative in promoting seminars and conferences on the topic of data comparability and comparative research designs which should be of special relevance for SSH researchers already engaged or potentially engaged in comparative research projects. In this way, a direct and permanent link can be established between the continuous flow of innovations of new data infrastructure tools and instruments for comparative research and a rapid knowledge transfer to the SSH community in order to upgrade and to improve their comparative research designs, their data organization and their data documentation practices.

The fourth objective has European students in the SSH-domain as its main clientele. Here the major new element lies in the organization of an annual Summer School which is the main responsibility of the hub itself. It is of special importance that students in the SSH-domain become familiar with new tools, instruments, standards and new comparative research designs very early in their career in order to use and implement this type of knowledge in their theses or dissertations. Thus, the main goal of the Summer

School will be a theoretical and practical guide to use best-practice standards in the selection of comparative data and in the design of comparative research.

In addition, each of the four main training activities will be required to prepare manuals, guidelines and training material as appropriate to their activity. The content and delivery of documents will be the responsibility of the hub but funds will be allocated to ensure consistency of quality and presentation and to co-ordinate effective distribution across the European Research Area.

5.4.5 FUNCTION 3 – STRENGTHENING

Supporting less-developed and less-resourced organizations in order to create greater homogeneity within the European Research Area.

Special programmes are needed for the smaller, less-developed and less-resourced CESSDA member organizations in order to enable them contribute fully and on an equal basis to the CESSDA-based programme of activities. This not only involves direct help in implementing centrally developed middleware tools (Task 1), but may also extend to the translation of metadata into English in order to facilitate wider accessibility (Task 2) or to special localised projects to support the acquisition of important national data resources. (Task 3).

5.4.5.1 Task 1 – Direct Help

5.4.5.1.1 Key Objectives

- To implement new middleware tools in CESSDA-archives.
- To organize a direct knowledge-transfer between the producers of new middleware tools and national operators.
- To use quality control standards throughout the CESSDA-RI.

The first task has a well-developed agenda with three main objectives which need little further specification or elaboration. The overall goal for Task 1 lies in the direct technical assistance and in knowledge transfers from the hub to individual CESSDA-members. In doing so, the hub will gradually be able to develop common quality and best practice standards and to operate within a far more homogeneous European data technology.

5.4.5.2 Task 2 – Language Harmonization

5.4.5.2.1 Key Objectives

- To facilitate and to finance the translation of available metadata descriptions into English.
- To organize knowledge-transfers between the available language-instruments like MADIERA and national data archives.
- To co-ordinate the process of language harmonization between CESSDA-members.

Likewise, the second task has a straightforward distribution of objectives and aims to overcome the existing language barriers which in many cases of comparative research impede the use of data from other national data archives. It is the clear mission that within a period of a few years the current data fragmentation due to language barriers will be successfully overcome.

5.4.5.3 Task 3 – National Upgrades

5.4.5.3.1 Key Objectives

- To implement a programme for small national data archives to upgrade their facilities both technically and in manpower.
- To act as a European strategic player for mobilizing additional national resources for small data archives.
- To become an evaluation centre with respect to the success of national upgrades.

Task 3 will be of vital importance for a rapid homogenization of the distributed CESSDA-RI. The main goal lies in the development of national work plans from small and very small national data archives which have to meet several criteria simultaneously. These criteria lie partly on the side of available data technologies and the implementation of new tools and instruments, partly on the side of increasing the current volume of national data holdings substantially and partly on the diffusion side of offering better and easier data access to the respective national SSH-community. Based on these criteria, the hub will support and evaluate a number of national work plans and will make certain that the implementation of national upgrades leads to a massive reduction in the current imbalances within CESSDA with respect to data holdings or data-transfers.

5.4.6 FUNCTION 4 – WIDENING

Extending and deepening the CESSDA in order to create and maintain a ‘complete’ pan-European SSH network.

A programme of special seed-money is needed in order to extend the existing CESSDA network and foster the development of national data archiving initiatives in those countries which are not currently part of CESSDA. The obvious purpose of this activity is to spread the CESSDA-network to each EU-member state and to create and maintain a ‘complete’ pan-European SSH network, including representation from emerging and candidate countries (Task 1). Equally, some countries with organizations within CESSDA have specialised research-led project-based teams whose data currently fall outside of the CESSDA network, and mechanisms likewise need to be put in place to create better comprehensiveness and coherency (Task 2)

5.4.6.1 Task 1 – Integrating New National CESSDA-Members

5.4.6.1.1 Key Objectives

- To implement a seed-programme for the implementation of new national data archives
- To act as a European strategic player for mobilizing national resources for emerging data archives.
- To become an evaluation and accreditation centre for new national data archives.

Task 1 has one single overall mission and that is to bring in new national data-archives into the CESSDA-network and to close the existing gaps of CESSDA-membership across Europe. Thus, the purpose of the seed programme is to ensure that the new CESSDA-members are highly qualified to carry out the day to day operations of a national data archive and to fulfil its international functions within the CESSDA-umbrella. Consequently, the seed programme will be based on national applications for setting up a national archive and the hub will also use an accompanying evaluation in order to ensure the success of the national seed-programmes.

5.4.6.2 Task 2 – Integrating Associated CESSDA-Members

5.4.6.2.1 Key Objectives

- To offer an Associated CESSDA-membership for international project-consortia with special data-holdings, well-established European SSH-networks with specific data-collections or research organizations with specialized large data-bases.
- To enable technical assistance, knowledge transfers and financial support for upgrading the data holdings of new Associated CESSDA-members
- To become an evaluation and accreditation centre for Associate CESSDA-members.

Finally, Task 2 is directed toward another form of widening the existing data-holdings within the CESSDA-RI. Here, the primary goal lies in a new type of CESSDA-membership which has been labelled as ‘Associated’. Due to the rapidly growing number of data stakeholders within the European Research Area Task 2 will become of increasing importance. Under the auspices of Task 2, the CESSDA-RI will attempt to integrate new types of data-holdings from international project-consortia, European SSH-networks or research organizations and to provide active assistance in the implementation of best practice standards with respect to resource discovery and access tools, instruments and data-documentations.

5.4.6.3 Management

Having operated for some 30 years the existing CESSDA network obviously has an existing man-

agement structure and governance. However, following a major upgrade, this would need to be modified to take account of its extended range of activities and functions.

The objectives stated in the current CESSDA constitution are as follows:

- To promote the acquisition, archiving and distribution of data throughout Europe.
- To promote projects and procedures for enhancing exchange of data and technologies among data organizations.
- To stimulate the development and the use of these procedures throughout Europe.
- To encourage new data organizations to further these objectives.
- To promote the integration of the European database.
- To associate and cooperate with other international organizations sharing similar objectives.

Membership is limited to those institutions and organizations within Europe having the capacity and expertise to further these stated objectives. Membership of CESSDA requires organizations to undertake a number of obligations, including: the support of inter-archival data transfers, adhering to the CESSDA Transborder Data Access Agreement; to cooperate in CESSDA generated projects; to support training and information exchange in methods and techniques of data service and secondary analysis.

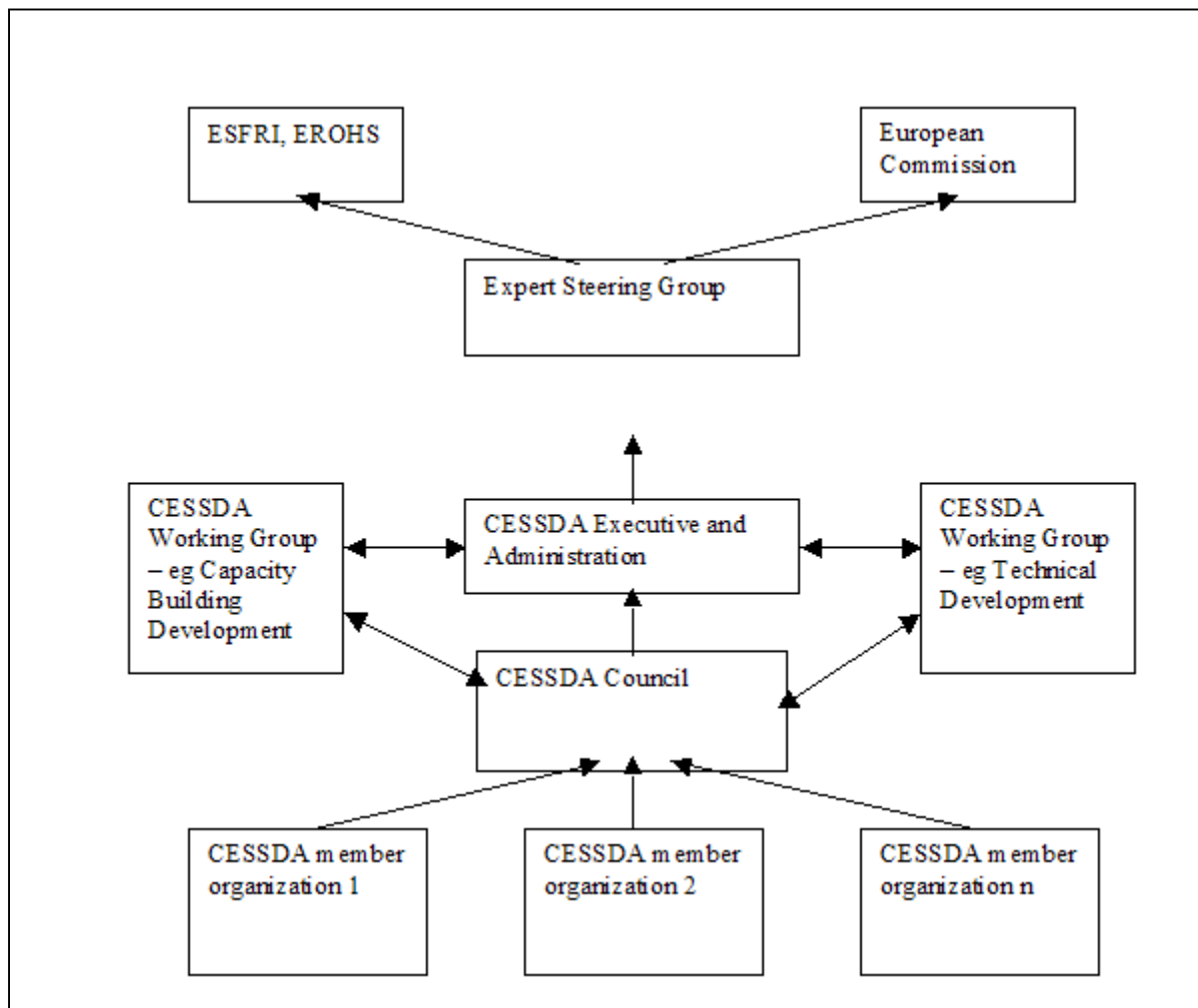
The CESSDA Council, consisting of one member per organization, elects an Executive Committee, consisting of a President, Vice-President, Secretary and up to two other Executive Officers.

Whilst this management structure is generally robust, in an upgraded infrastructure it will need to be strengthened and enhanced. Currently, one of the weaknesses of CESSDA is that it has no real central co-ordinating and administrative hub. All the Officers and essentially employed in national organizations and are primarily responsible for national agendas. In an upgraded RI a small co-ordinating and administrative hub will need to be formed in order to directly service pan-European interests, and to implement the objectives under Functions 2, 3 and 4. Likewise, the administrative co-ordinating hub will need to be informed by focused Working Groups drawing representatives from the member organizations, over-seeing work on Technical Development, Capacity Building, etc. Likewise each of the technical development tasks identified will need to have Task Co-ordinator and Project Team.

Equally, the accountability of CESSDA will also need reconstituting. It is initially proposed that an Expert Steering Committee is established to which the Executive will report. This will have broad international

representation and be composed of nominated experts and stakeholders, external to the CESSDA member organisations. This body could also link to the Commission and other appropriate bodies such as EROHS.

The proposed management structure could be illustrated as follows:



5.4.7 DRAFT COSTINGS AND TIMETABLE

At this stage it is only possible to produce broad indicative costs and associated time-scales. In order to be fully effective, it is proposed that the upgrade should be funded for an initial period of five years. A time horizon significantly shorter than this would run the risk of not fulfilling its stated objects, particularly in terms of technical infrastructural developments, their implementation and the widening and deepening of the CESSDA network.

A summary of the expected costs and timescale is provided for illustrative purposes in the table below. It is expected that costs will highest for the Technical Development function in years 1 to 3, placing major efforts into developing urgently needed infrastructural tools. Once these are ready for testing and implementation that more effort will be switched to the Strengthening and Widening functions.

	Costs M€					Total
	Year 1	Year 2	Year 3	Year 4	Year 5	
Function 1 Technical Development						
<i>Task 1</i> –Tools for C- CAT portal	0.3	0.3	0.2	0.1	0.1	1.0
<i>Task 2</i> – Metadata Standards	0.4	0.4	0.4	0.4	0.4	2.0
<i>Task 3</i> – Multi-lingual Thesaurus	0.4	0.4	0.4	0.1	0.1	1.4
<i>Task 4</i> – Authentication and Authorisation	0.4	0.4	0.4	0.1	0.1	1.4
<i>Task 5</i> – Statistical Disclosure	0.5	0.4	0.4	0.1	0.1	1.5
Future tasks	0.0	0.4	0.6	0.6	0.6	2.2
Function 2 Capacity Building (including central Administration and Coordination)	0.5	0.75	0.75	1.0	1.0	4.0
Function 3 Strengthening	0.5	1.5	1.5	2.0	2.0	7.5
Function 4 Widening	0.5	1.5	2.0	3.0	2.0	9.0
Total	3.5	6.05	6.65	7.4	6.4	30

5.5 COMMON LANGUAGE RESOURCES AND TECHNOLOGY INFRASTRUCTURE – CLARIN

Common Language Resources and Technology Infrastructure (CLARIN)



The facility: CLARIN is a large-scale pan-European coordinated infrastructure effort to make language resources and technology available and useful to scholars of all disciplines, in particular the humanities and social sciences. It will overcome the present fragmented situation by harmonizing structural and terminological differences, based on a Grid-type of infrastructure and by using Semantic Web technology.

Background: The volume of written texts (either as continuous discourse or, for example, descriptions of objects of cultural heritage) and (more recently) recorded spoken texts is enormous, and it is growing exponentially. The sheer

size of this material makes the use of computer-aided methods indispensable for many scholars in the humanities and in neighbouring areas who are concerned with language material.

What's new? Which impacts? The CLARIN Infrastructure aims to provide a comprehensive and easily accessible archive of language resources and technology, covering not only the languages of all member states, but also languages and language issues related to migration.

The tools and resources will be interoperable across languages and domains. They will contribute towards addressing the issue of preserving and supporting multilingual and multicultural European heritage. An operational open infrastructure of web services will introduce a new paradigm of distributed collaborative development and will allow many contributors to add new services ensuring reusability and allowing scaling up to suit individual needs. CLARIN will provide preferably off-the-shelf tools and solutions and the necessary training and advising to customize the resources in order to suit the particular needs of humanities researchers. It will strengthen the European position in standardization efforts, function as a pivotal and exemplary case for international initiatives and it will help Europe to train young researchers in not only using the benefits of an infrastructure enabling eHumanities, but, more importantly, to contribute to it.

Timeline for construction and first operation with related estimated costs: Due to the range of experience of the partners and the high degree of synergy between them, a preparatory phase of 12 months will be sufficient. It will allow starting in parallel with first design and implementation work in 2007. In 2008 CLARIN will start with the main construction work, which will go on for five years to meet the needs of the various disciplines and to stability. In parallel to the construction work first services will be offered to allow early tests and feedback strategies. The CLARIN infrastructure is scheduled to reach its full functionality based on fully operational resource and service centres in 2012. The total costs for the five year period will amount to 146 Mio € based on an estimate of about 20 distributed resource and service centres in Europe.

Leading consortium: Now 47 institutions from 30 European countries are registered as CLARIN members. The list includes universities, academies of sciences and institutes funded by regional, national or private funding bodies. In most countries exploratory discussions have taken place with the various bodies and agencies that are already funding resources infrastructures at the local level, and on the whole they seem to show a keen interest in further exploration of the CLARIN concept. In order to obtain any formal commitment to a collaborative effort involving over 30 funding bodies a clear legal international framework is required.

5.5.1 INTRODUCTION

The proposed Language Resources Infrastructure intends to serve all those who are active in the field of the humanities in a broad sense, who are concerned with language (spoken, written, multimodal, as carrier of information, object of study, or otherwise), and are dependent on or could benefit from using computer support for their language processing needs. The volume of written texts (either as continuous discourse or, for example, descriptions of objects of cultural heritage) and (more recently) recorded spoken texts is enormous, and it is growing exponentially. The sheer size of this material makes the use of computer-aided methods indispensable for many scholars in the humanities and in neighbouring areas who are concerned with language material. Language processing is costly, both in terms of the creation of digitised material (e.g. large collections of written text or recorded speech) and in terms of the skills required to provide the necessary tools to manipulate the data. The cost of collecting, digitising and annotating large text or speech corpora, dictionaries or language descriptions is huge in terms of time and money, and the creation of tools to manipulate these language data is very demanding in terms of skills and expertise, especially if one wants to make them accessible to professionals who are not experts in linguistics or language technology.

Therefore, all areas of the humanities that deal with texts in any form can benefit from an enabling research infrastructure that would optimally ensure the provision and coordinated creation of a vast amount of resources and technologies and that includes the necessary training and advising. We propose to create a European Resources Infrastructure that would be based on an open European Network of strong service centres and repositories that would jointly provide the whole European humanities community with

- knowledge about the existence of language resources,
- coordinated creation of and access to such resources,
- access to services that would allow for the adaptation of resources to specific needs or purposes,
- bundling of and access to expertise related to specific language processing problems

5.5.2 CONTENT

We encounter language, in all its modalities and varieties, basically as *text*, whether written, spoken or containing multimodal interaction. Hence, the primary language resources are *corpora*, i.e. collections of texts assembled, analysed and annotated according

to unified principles. For certain projects, a corpus may contain an exhaustive collection of the texts to be studied. However, for the purposes of building a research infrastructure, a corpus that is designed to be a general reference corpus serving to represent the totality of language use for a given language community would be more typical and important. The compilation and maintenance of such a general reference corpus for any language clearly go beyond the means of any individual research institute and must be provided for every European language as part of the research infrastructure.

In order to invest computers with the linguistic knowledge we humans use when analysing texts, corpora have to be linguistically analysed and annotated in line with standard principles. This requires a set of extra language resources, chiefly **lexical and terminological databases** that contain among other things detailed morphosyntactic descriptions of the behaviour of lexical units. To extend the analysis of language beyond its formal structure to its semantic level, particularly in response to the Semantic Web initiative, semantic lexicons **Wordnets** and **ontologies** are required.

Europe is inherently multilingual, and to address the issues of multilingualism, large scale **parallel corpora** (aligned texts that are translations of each other) and multilingual lexical databases have to be developed for at least some language pairs for any language. For a few languages, there exists an abundance of language resources, for others, particularly the languages of members that have recently joined, there is an acute scarcity even of basic resources. And it is certainly true for all languages that whatever resources do exist, they are fragmented and not interoperable.

We propose to create a minimum level of uniform language resources and tools for all the languages supported by the infrastructure. The fact that they will be interoperable will in itself contribute a great deal towards addressing the issue of preserving and supporting multilingual and multicultural European heritage. They will also be compiled in full recognition of the increasing importance of multimodal communication, which is rapidly gaining popularity with the young generation of today and which, we anticipate, is likely to become dominant in the time-scale we are considering.

5.5.3 TOOLS

To produce and exploit the language resources mentioned above, one needs to resort to software tools and technologies in the form of standards of annotation, data structure and processing. Software tools include tokenisers, morphological analysers, taggers, chunkers, named entity recognisers, semantic taggers, and aligners as well as tools for the stochastic

modelling of language. These tools are applied in a chain to build up a standard representation of language structure, which will be adequate for serving a number of particular applications. Advanced Semantic Web technologies which will allow researchers to participate in “knowledge weaving”, i.e., to create a semantically rich domain of language resources, will complete the scope of tools.

The use of standards at all levels of analysis and with respect to all language resources is essential in making language resources reusable and interoperable. The relevant technologies include the Unicode standard for character encoding, various XML technologies, and initiatives like the TEI, EAGLES, ISLE, ISO TC37/SC4 which address various aspects of encoding language.

5.5.4 COMMUNITIES

5.5.4.1 *The infrastructure providers*

The envisaged language resources and language technology infrastructure will cater for every European language (national and minority languages alike) and will rely on a network of research centres (universities, public, private or commercial research institutes) dedicated to at least one language. Present members of the CLARIN community represent the key players in the Language Resources and Technologies field and can look back on a solid track record with a high amount of synergy between different networks (ELRA, ELSNET, TELRI etc.) now set to join forces in a truly pan-European initiative.

On top of the national grid, there will be a number of competence centres established for the purpose of pooling resources and providing increased efficiency in serving the humanities user community. Each centre will conduct the development of resources in its own field and at the same time it will support users through its portal and advise them through electronic contact on a case by case basis. A dedicated competence centre will serve the following areas:

- Monolingual synchronic corpus development
- Historical corpora
- Parallel corpora
- Multilingual services portal
- Ontology development
- Archiving and sharing
- Grid and Web technologies
- Language technology integration and adaptation

5.5.4.2 *The user communities*

Computer aided language processing is already used by a wide variety of communities in different sub-disciplines of the humanities and the social sciences,

such as history, arts, literature, cultural studies, law, anthropology, sociology, philosophy, theology, and, of course, within linguistics (the study of language) it is used in different sub-disciplines such as dialectology, discourse studies, language acquisition, language documentation, language and speech technology, lexicography, phonetics, typology, etc.

All these communities address one or more of the multiple roles language plays in the study of the humanities, such as: (i) a carrier of cultural content and knowledge, both synchronically and diachronically,

- an instrument for inter-human communication within and across languages,
- one of the central components of the identity of a person, a group, a culture or a nation and
- an object of study or preservation.

At the same time it should be noted that many of the data collections and tools produced for one purpose for one community can be re-used for other purposes by other communities, provided that the other communities are aware of the existence of these materials, that they have access to these materials, and that they have access to the specific expertise needed to put the materials to use in their own (research) context or for their own needs.

Creating the envisaged infrastructure inevitably requires a high amount of linguistic and technological knowledge and data to be made amenable to computers. However, it should be emphasized that all this is just the required means to the end of serving the needs in humanities research, which may not be concerned with language for its own sake. The overall objective of the infrastructure is to bring computational analysis of texts and semantic annotation, for whatever purpose, within the means of researchers not necessarily interested in language per se (such as archaeologists, historians, sociologists, etc.), and certainly lacking the required skills in language technology to even consider the development of the necessary infrastructure for computational methods.

5.5.5 CONSERVATION

The issue of conservation is addressed at the synchronic and the diachronic levels. In order to fully capture the linguistic variety that is an inherent part of the European identity, all languages spoken within Europe including minority languages must be recorded in the widest possible variety of settings. While this has been partially achieved for a handful of languages, the majority of EU languages are without adequate language resources. This is particularly true for most of the recent members and the candidate countries.

Language resources also play a crucial role in conser-

vation by preserving different chronological stages in language, an entity subject to constant imperceptible change. The diachronic aspect is especially important for the wide range of disciplines in the humanities that study cultural heritage. Finally, there is also the technical issue of preservation of recorded language data, where compliance with annotation standards and appropriate technological migration can safeguard long-term conservation and accessibility of language resources.

5.5.6 DIGITISATION

Digitisation is fundamental to both the conservation and the dissemination of language resources.

Only digital representations that allow copying without loss of quality will finally allow inexpensive preservation – not of the physical object – but of its immaterial content and memory. Digitisation also changes the rules of giving access to objects: While physical objects have to be protected since every access means a degradation of their physical substance, digital representations can be accessed as often as is wished.

Language resources and, in the broader sense, cultural heritage objects become accessible for the public and not only for a select group of experts. Digitisation therefore means democratisation of our cultural heritage.

In particular, the digital availability of spoken language and multimodal communication will boost speech and multimodality research. Both will become increasingly necessary since new generations seem to turn away from written language.

5.5.7 INTERPRETATION

The envisaged research infrastructure should contain language resources that are processed in a standard manner and have the same richness and granularity of annotation across the languages where possible.

The analysis should involve three levels: formal aspects of language structure, semantic annotation and stochastic modelling. The guiding principles include accessibility and extensibility, scalability, interoperability and user-friendliness.

To ensure maximal interoperability across languages and resources, a minimal level of analysis carried out uniformly and following set standards has to be imposed on all the language resources in the infrastructure. Accordingly, corpora should be equipped with a consistent and rich set of metadata that describe their source and record every aspect of their processing so that relevant research queries will help in locating useful resources. The texts should be morphosyntactically analysed and disambiguated. A common set of

proper names of persons, places and institutions should also be identified and annotated.

Humans have an immense facility for processing and analysing text without reflection or perceived effort. In order to simulate such a processing feat, computerized text processing must equip texts with the result of a vast amount of analysis that is annotated in a standard manner. Recording the formal aspects of language structure may be relevant to varying degrees across the different humanities disciplines. Interpretation crucially involves the semantic and pragmatic aspects of language use. Accessing these higher levels of language structure depend on technologies like semantic tagging, language understanding, discourse representation, creation of typed relations and related areas. Such research is driven by the concept of the Semantic Web and leads to the development of ontologies, i.e. words arranged in hierarchical conceptual schemes. These technologies are highly experimental at present but they are bound to become mature technologies within the timescale of the envisaged infrastructure.

For some limited domains, the envisaged infrastructure will also contain a rich set of hypertext annotations providing links to relevant additional digital content that very often exists in other media such as pictures and sounds.

5.5.8 DISSEMINATION

We are fully aware that the accessibility of Language Resources and Technology alone is not sufficient. Rather than relying on merely serving current practice within the humanities, the field of Language Technologies must look ahead and be catalytic by actively promoting the use of Language Resources and Language Technologies within the humanities. The linguistic community is not only a consumer of its own resources and its technology; it has to especially understand that it has a mission as a service provider for other disciplines. Since it is certainly true that only a small fraction of the group of scholars in the humanities is fully aware of the opportunities of, for example, Grid computing, distributed PetaByte databases and advanced Semantic Web frameworks, pure dissemination has to be accompanied by training, education and awareness-raising programs.

Every aspect of the development of Language Resources and Technologies is implemented in strict compliance with standards to ensure accessibility and interoperability of tools and resources. However, widespread use of these resources by the target user community is only realistically attainable if the resources and technologies are maximally user friendly and can be customised with minimal effort and expertise. To this end, continuous training and support is an integral part of the dissemination policy of the infrastructure.

These services can best be established by a research infrastructure that is based on a European network of strong service centres and repositories that interact closely with emerging centres for the humanities such as AHDS and DANS.

Given the variety of services and the need to overcome the fragmentation of effort and resources, these services have to be built upon widely agreed standards and registered web services and web-applications that allow both programmed access by whatever application and human access. By widely spreading them, these repositories will also have the strength to take care of the stability and persistence of the services and, in particular, of the long-term preservation of the resources.

5.5.9 COSTS

The envisaged research infrastructure can only be realized when it is based on strong national commitments and formation processes. The following figures are indications for a 5 years construction and a five years operation period:

- resource repository formation process 20 (10 sites/5 years)
- knowledge centre formation process 18 (10s/5y)
- setting up a pan-European service infrastructure 6 (5y)
- gathering, converting and encapsulating resources 20 (5y)
- gathering, adapting and encapsulating technology 20 (5y)
- developing sample applications and services 10 (5y)
- comprehensive training and education programs 16 (10y)
- operating the infrastructure (repositories, services) 30 (5y)
- management 6 (10y)
- total investments 146 (10 years)

Roughly speaking one can say that these expenses cover all member states which means in average 0.6 Mio € per year per country. Part of this sum is already covered by the running efforts of the member states in maintaining a number of such centers and repositories.

5.5.10 CLARIN BUSINESS PLAN

5.5.10.1 Introduction

In this document we give a brief overview of the CLARIN business plan. You will find the following sections, some of which refer to separate documents:

- Mission: this is distributed as a separate document (CLARIN Mission Statement)
- Characteristics
- Competition
- Feasibility
- Languages and opportunities
- Strategy (phases and cost per phase)
- Funding
- Management: this is distributed as a separate document (CLARIN Governance and Management)
- Budget: a detailed cost breakdown is distributed as a separate document (CLARIN Budget Justification)

5.5.10.2 Mission

See separate mission statement document entitled “CLARIN Mission Statement”

5.5.10.3 Characteristics

The infrastructure we envisage is not a physical installation, located in one place, and to be used by the research community on a time sharing basis, as would be the case with e.g. a particle accelerator or a specialized laboratory.

The infrastructure is mainly electronic in nature and aims at providing access to language resources and technologies, and to services connected to these. The infrastructure is distributed in nature, and its participants constitute an open ‘federation’ of organisations who are willing to share their resources, technologies and expertise, and who are willing to provide additional services based on these resources and technologies.

The infrastructure is special in that it will (at least initially) largely be based on the language technology and linguistics communities in Europe, but that it intends to serve the European humanities community at large, more specifically all those who have an interest in language in one of its many roles.

5.5.10.4 Competition

At this moment we see a number of projects and initiatives that are already carrying out tasks similar to the ones that we envisage for the infrastructure. They vary in size, scope and ambition, and there is very little systematic coordination between them.

The CLARIN infrastructure aims at creating a framework in which these activities can be united and their efficiency and their effectiveness and impact can be increased. No competition with others is foreseen. As it is open there are no obstacles for others to join if they are doing similar things. As it is coordinated duplication of work can be avoided.

5.5.10.5 Feasibility

The feasibility of an enterprise of this type depends on a number of factors. Apart from the obvious financial factors we would like to mention three factors that are crucial for its success:

- Critical mass within the community. At the moment of writing 44 institutions in 29 countries have already expressed their willingness to actively participate in the creation and operation of the CLARIN infrastructure (see <http://www.mpi.nl/clarin>), and more are expected to sign up in the coming weeks.
- Critical mass in terms of available resources. A brief investigation conducted amongst 42 European institutions in 2005 (including most of the CLARIN members referred to above) led to the conservative estimation that the total investment represented by the resources owned by these institutions represented a value of over one billion euro.
- Maturity of the technology to be deployed. The core of the infrastructure will be based on existing and broadly supported standards, and existing technologies that have already demonstrated their value and their feasibility in infrastructure projects initiated by some of the key players in CLARIN, such as the EC funded DAM-LR project.

5.5.10.6 Languages and opportunities

The infrastructure is initially intended to cover all language communities in EU, associated and applicant countries, but in principle there are no obstacles for others to join. Some specific opportunities for cooperation and synergies are worth mentioning here:

- Smaller (or technologically less mature) languages can benefit from the experience and expertise gained for the bigger languages, and rely on best practice and (emerging) standards
- A common definition of a minimal reference set of resources needed for a language to do any language technology at all (in research, development and training) will facilitate the organization of coordinated actions to bring the language resources coverage for smaller languages up to the standard
- The infrastructure offers cooperation possibilities for languages shared by communities in different countries
- The fact that a number of EU languages are also spoken in countries outside the EU (e.g. English, French, Spanish, Portuguese) offers a good starting point for sharing of resources and cooperation at the international level.

5.5.10.7 Strategy

- phase 0, networking:
 - activity: gather critical mass of resources and service providers so that the infrastructure can be populated; start reaching out to the humanities community at large
 - aims:
 - creating a network covering at least all EU and associated states and all major resources creation and distribution projects & organisations
 - setting up alliances with organisations and infrastructures that are well-embedded in the humanities
 - ensuring support from funding agencies
 - agreeing on functionality and shape of phase 1 infrastructure
 - duration: 2 years, 2006-2007
 - cost: 1.0 M€ (travel and networking)
 - funders: participants
- phase 1, setting up the initial infrastructure
 - activity: creating an initial federative service infrastructure on the basis of existing infrastructures
 - aims:
 - setting up management structure
 - selecting a small set (ca 5) of resources centers, knowledge centers and service centers to act as the first pilot infrastructure, with a balanced spread in size (smaller and larger centers) and location (old and new countries)
 - setting up mechanisms for others to join
 - setting up mechanisms to promote the infrastructure to the humanities community and to elicit their feedback
 - setting up a framework for coordination of activities
 - setting up working groups
 - evaluate and prepare for phase 2
 - duration: 2 years, 2008-2009
 - cost: 22 M€
 - funders: participants, national and regional funding agencies, EC
- phase 2, building and promoting the full infrastructure
 - activity: expanding the infrastructure with a view to covering all EU and associated states and their languages
 - aims:
 - inviting other knowledge, resources and service centers to join in order to quantitatively and qualitatively expand the infrastructure in a coordinated fashion
 - gathering resources and technolo-

- gies to populate the infrastructure
 - defining the concept of standard reference corpora and creating them in case they are not available
 - building sample applications
 - establishing links with humanities community at large
 - organising dissemination and training actions for the humanities
 - establishing and promoting standards
 - establishing and promoting best practice recommendations for IPR issues
 - evaluate and prepare for phase 3
- duration: 3 years, 2010-2012
- cost: 85 M€
- funders: participants, national and regional funding agencies, EC
- phase 3, first fully operational phase
 - activity: operating and gradually enhancing and expanding the infrastructure
 - aims:
 - maintaining and operating a sustainable infrastructure
 - ensuring that services and resources keep pace with user needs and technological developments
 - maintaining and promoting standards
 - ensuring continuous dissemination and training
 - periodical evaluation of and reflection on the role and functioning of the infrastructure
 - deciding about whether and if so how to continue beyond this phase
 - duration: 5 years, 2012-2016
 - cost: 38 M€
 - funders: participants, national and regional agencies, EC

5.5.10.8 Funding

We see four main classes of potential funders for this infrastructure (which are not always completely disjoint):

- EC programmes
- National and regional funding agencies (e.g. National Research Councils, National or Regional Academies of Sciences, etc)
- Private funding agencies (e.g. MPG, Volkswagen Foundation)
- Participants (e.g. universities and research institutes)
- Beneficiaries (e.g. humanities scholars or students who make use of the resources and services provided by the infrastructure)

All five parties listed above have already invested and are still investing significant amounts in the creation of language resources and infrastructures to make them accessible. The creation and operation of the CLARIN infrastructure will require additional investments for coordination, dissemination and training, creation and delivery of services, etc.

The expected benefits are manifold (and described in more detail in the questionnaire and the other supporting documents): coordination will help exploiting synergies and avoiding duplication; dissemination and training will lead to better exploitation of what is available and to innovative, cross-discipline approaches in the humanities; services will enable computationally naïve users to get tailor-made solutions. How the costs of the creation and the operation of the CLARIN infrastructure should be shared between the various parties is beyond the scope of this document.

5.5.10.9 Management

- see separate document on management entitled CLARIN Governance and Management Structure

5.5.10.10 Budget

Phase	Activity	Period	Total	Grid
0	“Networking”	2006-07	1 M€	0 M€
1	“Setting up the initial infrastructure”	2008-09	22 M€	6 M€
2	“Building and promoting the full infrastructure”	2010-12	85 M€	6 M€
3	“First fully operational phase”	2013-16	38 M€	4 M€
Total		2006-16	146 M€	16 M€

The total amount per phase is given in the column “Total”. In the column “Grid” we indicate how much of the total is budgeted for the Grid.

A more detailed breakdown of the costs is presented in a separate document, entitled CLARIN Budget Justification

5.5.11 CLARIN MISSION STATEMENT COMMON LANGUAGE RESOURCES AND TECHNOLOGIES INFRASTRUCTURE

The CLARIN project is a large-scale pan-European collaborative effort to create, coordinate and make language resources and technology available and readily useable to scholars of all disciplines, in particular the humanities and social sciences.

It intends to rise to the challenge that language (in spoken, written, multimodal form, as carrier of information, object of study, or otherwise) presents in our age when the volume of texts (either as continuous discourse or, for example, descriptions of objects of cultural heritage) and (more recently) recorded spoken texts is enormous, and it is growing exponentially. The sheer size of this material makes the use of computer-aided methods indispensable for many scholars in the humanities and in neighbouring areas who are concerned with language material.

The CLARIN infrastructure is based on the firm belief that the days of pencil-and-paper research are numbered even in the Humanities. Computer aided language processing is already used by a wide variety of sub-disciplines in the humanities and social sciences, addressing one or more of the multiple roles language plays (i.e. carrier of cultural content and knowledge, instrument of communication, component of identity and object of study). There is a high degree of commonality in the methods and objectives of current practice, and it is also evident that to reach the higher levels of analysis of texts which non-linguist scholars are typically interested in, such as their semantic and pragmatic dimensions, requires an effort of a scale that no single scholar could, or indeed, should afford.

The cost of collecting, digitising and annotating large text or speech corpora, dictionaries or language descriptions is huge in terms of time and money, and the creation of tools to manipulate these language data is very demanding in terms of skills and expertise, especially if one wants to make them accessible to professionals who are not experts in linguistics or language technology.

Hence the benefits of computer enhanced language processing become available only when a critical mass of coordinated effort is invested in building an enabling infrastructure, which can then provide services in the form of provision of tools and resources as well as training and counselling across a wide span of domains. This is the mission of the CLARIN infrastructure initiative.

To realize the above objectives, CLARIN will create a comprehensive and free to use archive of language resources and technologies covering not only the languages of all member states, but also minority

languages and language phenomena addressing the issue of migration. The infrastructure will be based on a number of resource, service and expertise centres and will commit itself to collaborating with education organizations and to providing training programs with the aim of enabling the widest possible range of users to exploit the benefits in their own field. Through the fact that the tools and resources will be interoperable across languages and domains will in itself contribute a great deal towards addressing the issue of preserving and supporting multilingual and multicultural European heritage. An operational open infrastructure of web services will introduce a new paradigm of distributed collaborative development. It will allow many contributors to add all kinds of new services based on existing ones thus ensuring reusability and allowing scaling up to suit individual needs.

The CLARIN community unites all the leading institutions in the Language Resources and Technologies field across Europe representing decades of experience in tools and resources development, standardisation and infrastructure initiatives. The existing networks (ELRA, ELSNET, TELRI etc.) are now set to join forces in this truly pan-European initiative. The governance and management plans for the CLARIN infrastructure are detailed in a separate document.

Regarding the potential users of the CLARIN infrastructure, we should emphasize that the overall objective of the infrastructure is to bring computational analysis of texts and semantic annotation within the means of humanities and social sciences, be it archaeology, history, psychology and sociology to name but a few relevant fields. Researchers in these fields are not necessarily interested in language, per se, and certainly lack the required skills in language technology to develop themselves the infrastructure required for computational methods to be even considered. Every effort will be made not only to make resources available but to provide preferably off-the-shelf tools and solutions and the necessary training and advising to customize the resources in order to suit the particular needs of humanities researchers.

5.5.12 GOVERNANCE AND MANAGEMENT STRUCTURE

5.5.12.1 Preface

The CLARIN infrastructure has a two-tier organisational structure. The pan-European infrastructure is built on top of national infrastructures. In this respect CLARIN has an underlying structure similar to that of GEANT and DANTE, for example. Participation and representation are a result of national formation processes. CLARIN is operating at the European level, bundling and coordinating all efforts, defining generally agreed standards and is responsible for the European layer of services.

5.5.12.2 *Membership*

We distinguish two types of members of the CLARIN infrastructure:

Members (M): Private and public organisations who have been identified by the national bodies, who have committed themselves to the creation and operation of the CLARIN infrastructure and who have the financial and organisational capacity to jointly ensure its medium to long term sustainability. The members are the shareholders of CLARIN. They meet at the regular LREC conferences, elect the supervisory board and discuss all issues that are put on the agenda by the members.

Associate members (AM): Private and public organisations who have committed themselves to actively contribute to the creation and operation of the CLARIN infrastructure. Such contributions may include providing access to their resources, providing technology, and providing expertise needed to perform specific tasks for the network, possibly with financial support from EC or other funding bodies, and possibly for a delimited period of time.

5.5.12.3 *Management*

The Supervisory Board (SB) (ca. 15 members) is responsible for taking strategic decisions, policy making, high-level coordination and in particular for overall budget decisions. The SB meets at least once per year; its members have a term of office of two years and will be elected at the bi-annual LREC conferences. The SB will appoint an Executive Board to carry out the work in full responsibility. The chairperson of the Supervisory Board keeps in close contact with the activities of the Executive Board, can request special reports and can arrange extraordinary meetings of the Supervisory Board.

The Executive Board (at least 5 members) will take care of all project management aspects, lead and represent the CLARIN work within the strategic decisions of the Management Board. The EB will have at least five members: Chief Executive Officer (CEO), Chief Scientific Officer (CSO), Chief Technological Officer (CTO), Chief Financial Officer (CFO) and Chief IPR Officer (CIPRO). The Executive Board has to report to the Supervisory Board twice a year, in particular, present the state of the work and the perspectives at the annual meeting of the SB. Members of the Executive Board are responsible for the progress in the Working Groups. Reports of the Working Groups are submitted to the Executive Board which will include the results in its annual reports to the Supervisory Board. The EB has to offer an appropriate mix of expertise.

5.5.12.4 *Working Groups*

In consultation with the Supervisory Board the Executive Board will set up permanent or temporary Working Groups (WG) for specific tasks and prob-

lem areas. The Working Groups work out solutions to problems and tasks that have been identified as key issues. The Working Groups are essential parts of the management work of CLARIN. Members of the WGs are drawn from member organisations, but can also include well-known experts for the topics to be discussed. They are suggested by the Executive Board and are finally appointed by the SB.

The need for seven WGs was recognized. Others may be created as the need arises.

- The Standards Working Group (SWG), responsible for the adoption and implementation of relevant standards in all aspects of the infrastructure.
- The Grid and Services Working Group (GSWG), responsible for the definition and implementation of the Grid and services infrastructure.
- The Language Resources Working Group (LRWG), responsible for the creation and selection of language resources to be integrated and for matters of integration and interoperability.
- The Language Technology Working Group (LTWG), responsible for the creation and selection of language technology components to be integrated and for matters of integration and interoperability.
- The IPR Working Group (IWG), responsible for the discussion of all legal and ethical aspects and for working out guidelines, declarations etc.
- The Education and Dissemination Working Group (EWG), responsible for training, education and dissemination activities.
- The User Consultation Working Group (UCWG), responsible for the collection of feed-back from users and potential users.

The Supervisory Board may set up task forces for specific tasks.

5.5.12.5 *Advisory Councils*

Three councils advise the management boards with respect to their strategic role. All councils will get access to internal reports.

- The International Advisory Council (IAC) is an advisory body to the MB consisting of independent internationally recognized experts.
- The National Representatives Council (NRC) is an advisory body that consists of representatives of the National Science Organizations that support the CLARIN infrastructure and take care of the national contributions. In particular, in strategic budget decisions the reports of this council are of extreme relevance.

- The Associate Member Council (AMC) is an advisory body that consists of representatives of all associate members that can give advice about all CLARIN matters.

5.5.12.6 Physical Center

The CLARIN infrastructure is highly distributed in nature. With respect to the management it is advisable that the Executive Board is physically located at one institution, i.e., office spaces and meeting facilities have to be available for all matters of the Executive Board. It is expected that the CEO, CSO and CTO will share a core time at the anchoring institution to take care of the necessary synchronization tasks. This solution does not require a building, however, one of the institutions involved should offer the necessary space.

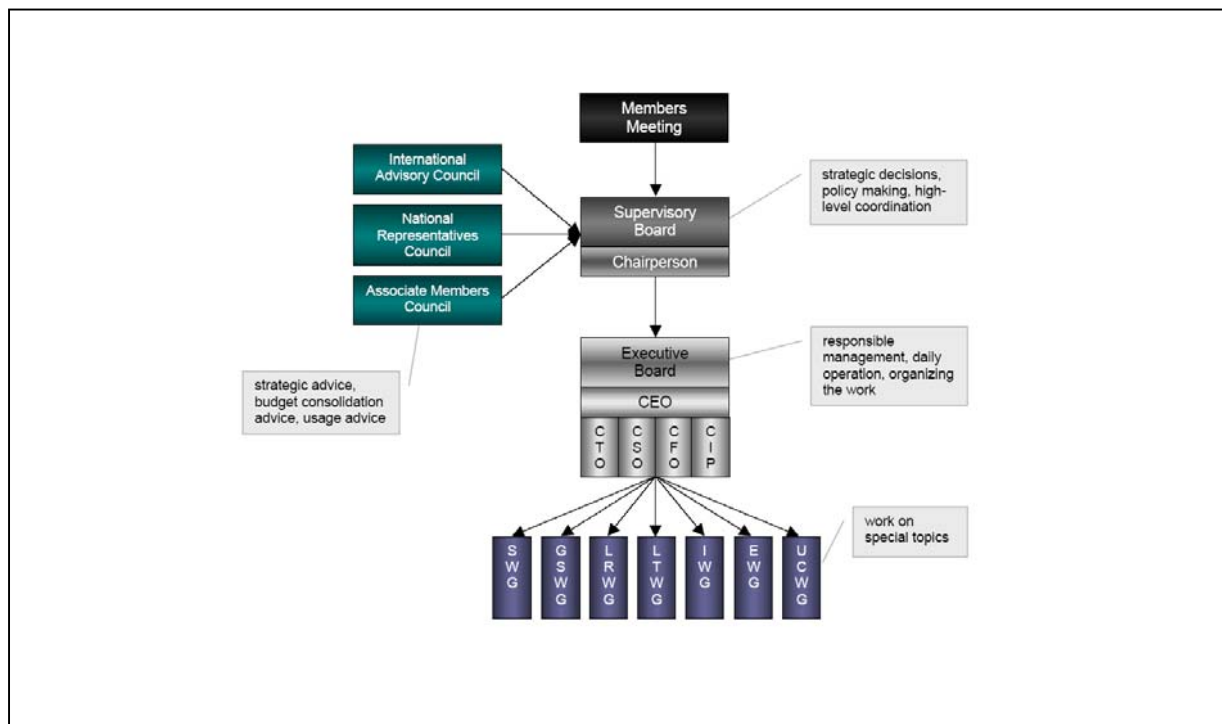
5.5.12.7 Phases

The CLARIN members understand that there is a difference between the building and start-up phase on the one hand and the operational phase on the other hand. With respect to the basic management structure there is no need to change this. However, a completely different support structure will be required. The Working Groups will be replaced partly by Service Groups. Yet it is too early to make detailed statements. It is expected that the building and start-up phase will take five years. From the third year on plans have to be made for the operational phase.

5.5.12.8 Graphical Overview

The management structure is kept very simple and in analogy to typical company organizations. It combines efficient decision taking at operational level in full responsibility with strategic decision taking at

control level. Further, it includes working groups that are actually carrying out the detail work on certain content topics as well as overall advisory councils to help defining the strategic goals.



5.5.13 CLARIN BUDGET JUSTIFICATION

This note is intended to explain and justify the CLARIN budget. It also makes an attempt for a realistic funding scheme.

5.5.13.1 General Statements

The CLARIN budget was created to present the total cost summed up for the period of 10 years. It was made without taking into account that existing centers and infrastructures, for example, could take over important tasks without requiring additional costs of the amount that were specified in the budget. Part of the funds are already covered by existing funding schemes. In this document we will make an attempt to include such schemes, although these can only be

rough estimates, since accurate specifications can only be made after a process of negotiations.

5.5.13.2 Justification of the CLARIN Budget

The described costs are calculated on a five years period.

5.5.13.2.1 Resource Repository Formation¹

The repositories have to run the servers and services of the CLARIN infrastructures. The services are widespread and range from archiving services and Grid services to managing stable resource access services. The “Live Archives” flyer gives an impression of the tasks of modern digital language resource archives.

new server, storage and network technology all 5 years in average	400 k€
a system manager to run the computer network and basic services	300 k€
two archive managers to take care of a consistent and accessible repository	600 k€

a software developer to develop and maintain archiving and access software	300 k€
assistants for various tasks	300 k€
expenses for long-term archiving	100 k€
Subtotal:	2.0 M€

It was estimated that in Europe we will have at least 10 centers of this sort. Again this is very much dependent on the formation processes in the countries. Some countries may decide to have more than one of these centers at national level, others may decided to join with other countries to share the costs. During the formation process the countries have to make selections about their national strategy, the EC has to make statements about their contributions and the participating institutions at national level have to contribute. A stepwise procedure as taken in the DAM-LR project could be chosen as well, i.e. a pilot project of two years with a limited number of centers will be carried out to establish even more solid budget and load numbers. It is obvious that these budget numbers are only suitable when one can build

upon already available expertise in some of these centers. It is also obvious that a larger number of smaller “national or regional centers” will exist. These have to be included in the Linguistic Data Grid stepwise.

5.5.13.2.2 Knowledge Center Formation

The knowledge centers have the linguistic and engineering knowledge to give linguistic advice with respect to languages, to work out language technology and develop appropriate services. These services can either be offered via own facilities or preferable by making use of the facilities of the resource repositories. In the latter case the costs for local infrastructure can be neglected here.

seven language and language technology experts (part time contracts)	900 k€
--	--------

three language technology and service developers	900 k€
--	--------

Subtotal	1.8 M€
-----------------	---------------

Again, it was estimated that in Europe we will have at least 10 centers of this sort. The same arguments as made above will hold, i.e., the national formation process will finally define how many of these centers will be established. Also in this respect it would make sense to first start with a small-scale pilot project and to build on the expertise and infrastructure that is already in place. Even more important here is the inclusion of existing smaller “national and regional centers” to make use of the existing knowledge.

5.5.13.2.3 Pan-European Service Infrastructure
This is the part of CLARIN that will develop the necessary pan-European middleware that will enable the CLARIN infrastructure. It is scheduled to be developed to maturity within a five years time period so that it can afterwards be maintained with less effort. It will be built upon the experiences made with establishing Grid middleware such as in the DAM-LR project and different standardization and web services initiatives. The service infrastructure will be realized at the resource centers, i.e., they can take profit from the existing facilities and knowledge. The activities will have five aspects:

- documenting the developed and installed software
- training other developers and system managers
- setting the software up at all resource centers that will participate

The range of tasks can be described by the following keywords²: a PKI system for trusted servers and services, a system of unique resource and tool identifiers including redundant resolution, an integrated metadata and registry system for all kind of services, browsing and searching facilities for this registry, standards for web services, encapsulation prototypes, distributed authorization mechanisms, accounting system, workflow system, conversion services for a number of standard formats, content search services, ontology manipulation framework, collaboration tools, semantic annotation and weaving frameworks, etc. For all this work we estimate the following effort over 5 years:

- defining standards and selecting existing components where possible
- development and testing of a middleware layer

project management	700 k€
documentation and training	600 k€
installation, training, helpdesk, ebugging at many sites	1.2 m€
sw development at all layers	3.0 M€
localization of the sw in all major EU languages	500 k€
Subtotal	6.0 M€

The work will be spread maximally across 5 institutions with expert knowledge to ensure feasibility.

5.5.13.2.4 Integrating existing Resources and Technology

In the European member states a gigantic value has been established during the last decades and even beyond. A huge amount of language resources and language technology components has been created³. It is now time to make these resources accessible to the scholars within the humanities. This means that as many as possible language resources and tools will be integrated into the emerging domain to create a critical mass establishing the eHumanities. Negotiations with other disciplines and large institutions such as libraries will take place to make sure that resources of many disciplines will be integrated. The integration can only be done by having a core group of experts that will do acquisition and evangelization and establish temporary contracts with the owners or archivists to carry out the organization, adaptation, metadata descriptions, encapsulation and integration of the resources and tools. For the integration small teams are made consisting of a computer or corpus linguistics expert and two developers each. For the contracts with resource and tool providers lump sums are estimated that cover all costs at the providers side (licenses, help, expertise, etc).

Resource Integration

project management	700 k€
10 acquisition and integration teams	9.0 M€
Assistants	500 k€
contracts with resource providers	10 M€
Subtotal	20 M€

Tool Integration

project management	700 k€
10 acquisition and integration teams	9.0 M€
Assistants	500 k€
contracts with resource providers	10 M€
Subtotal	20 M€

5.5.13.2.5 Sample Applications

It is of crucial relevance to develop new sample applications for different disciplines in the humanities that can show the great potential of the new eHumanities infrastructure for the researchers, that can be used for training courses and that can demonstrate in particular to the coming generations of researchers and developers how to use the infrastructure in an advanced way. Finally, the new infrastructure is intended to motivate many persons, from students to scholars, to contribute in various forms to a living eHumanities landscape. Again here it is intended to form temporary teams covering language resource and technology experts, discipline experts, assistants and developers to tackle interesting tasks. The size of the teams will depend on the task. It is intended to write calls for tenders so that interested groups can submit proposals. The most interesting proposals in terms of their scientific potential and their usefulness for educational programs will be selected.

The work will start at the third year after the first infrastructure pillars will be ready. In average we estimate teams with a 3 years contract existing of the following persons:

LRT expert	180 k€
2 discipline experts	360 k€
2 developers	360 k€
assistants	50 k€
Subtotal	0.95 M€

This amounts to the following overall costs:

10 teams	9.5 M€
project management	700 k€
Subtotal	10 M€

The teams will be associated with the applying institutions, however, close relations with the CLARIN management will be maintained.

5.5.13.2.6 Education and Training Program

The education and training program has to operate

along five dimensions will it help to fulfil the CLARIN goals:

- First, PR material such as flyers will be created to attract the interest of many researchers in the humanities.
- It has to operate in all member countries and even address the issues in a number of languages
- It has to be attractive for the persons involved in the research and education programs, i.e., from students to scholars, to convince them about the emerging possibilities and about actively contributing and enriching the infrastructure. For students whole courses have to be worked out and offered in a number of languages.
- It has to carry out programs devoted to a number of disciplines in the humanities
- It has to address the developers and computer linguists to contribute with new tools.

Such a multifaceted training and education program is only possible with considerable efforts in designing a layered program and a large amount of organizational and management effort. With different foci it will cover the whole period of 10 years. In average it is suggested to spend 1.6 M€ for these programs covering the following activities: organizing and managing the production of material, attracting the researchers, designing courses of various types from short management seminars (2 days), researcher and developers courses (typically a week) up to developing material for whole student courses, translating the course material (some to all EC languages to address the students) and to carry out the courses. It would be

new server, storage and network technology all 5 years in average	400 k€
system manager	300 k€
archive manager	300 k€
2 developers	600 k€
assistants	200 k€
Subtotal	1.8 M€

For 10 centers this would amount to 18 M€

At the knowledge centers personal will be necessary to take care of typical tasks such as maintaining the

Knowledge Centers

2 computer linguists	600 k€
2 developers	600 k€
Subtotal	1.2 M€

too early to make detailed statements, however, we can state that besides having a central management and coordination team the work will be spread thematically and with respect to languages to maximally 20 teams associated with research institutions.

education and training management	1.2 M€
assistants and contracts	800 k€
distributed teams	14 M€

5.5.13.2.7 Operating the Infrastructure

After the five years startup and development phase the CLARIN infrastructure will be turned into an operating infrastructure, i.e., the costs will mainly be spent for the typical tasks such as help desk, debugging, adapting, operating etc. Further developments should they be necessary have to be funded by other programs. We assume about 20 resource and knowledge centers having received an official task at the European level and an integration of these centers in well-functioning institutions where the experts can take profit of the already existing expertise and infrastructure.

At the resource centers personal will be necessary to take care of the typical operational functions such as maintaining the infrastructure, developing and adding new functionality, integrating new resources, installing the software at additional centers, upgrading the computer networks, access management at system level, helpdesk, etc. In addition the computer network has to be upgraded completely according to the five years rule.

integrated tools, developing new tools and adding them into the infrastructure, giving advice, helpdesk etc

For 10 centers this would amount to 12 M€

5.5.13.2.8 Overall Management and other costs

Overall management of the infrastructure over 10 years: 8 M€

As this overview focuses on the cost of the infrastructure per se, some cost categories have not been included here, as they are relatively small compared to the cost of labour and computer equipment (e.g. legal advice on IPR issues), or are part of what normally counts as indirect costs (housing, clerical assistance, etc) for which different funding agencies apply different schemes.

Notes:

- The experience of the MPI for Psycholinguistics, that archives currently about 20 TeraByte of data and offers a wide range of access and archiving services, is taken as an example here.
- It is not the purpose of this document to describe the necessary services in detail. They cover typical services known from the Grid and Semantic Web layers and typical eScience applications.
- Also in this respect it is not the task of this paper to make detailed statements.

5.6 EUROPEAN RESEARCH OBSERVATORY FOR THE HUMANITIES AND THE SOCIAL SCIENCES – EROHS

The European Research Observatory for the Humanities and the Social Sciences (EROHS)



The facility: EROHS will operate both as a central and distributed facility with a strong physical hub working in close conjunction with a number of spokes across Europe, harnessing European expertise through a co-ordinated yet decentralised network. It will be organised to promote and ensure cooperation and integration of data, technologies and policies.

Background: EROHS is an instrument for planned cooperation and integration. EROHS will in some areas be based on mature existing resources and infrastructures, supporting, promoting and extending their scope – and will in other areas facilitate the development of new infrastructures based on the need of the humanities and the social sciences by bringing together national initiatives, organisations and individuals in

order to provide upgraded and enhanced European wide actions.

What's new? Which impacts? The proposed research infrastructure – EROHS – is aiming at organising the communication, coordination, documentation and sharing of information in ways that will set standards for research infrastructures worldwide and make Europe the world leader in this area. It will provide Europe with the tool to realise the vision of open, distributed, well coordinated, well documented, on-line access to data for the humanities and the social sciences. Through EROHS more researchers will be introduced to the possibilities created by existing infrastructures. Training and seminar activities organised by EROHS will also be to the advantages of both existing and emerging research infrastructures. Emerging infrastructures will be important stakeholders of EROHS, be they national or pan-European infrastructures. EROHS will also strengthen the national infrastructure. By using the same model in different nations, national hubs will be linked both to each other as well as to the international hub.

Timeline for construction and first operation with related estimated costs: EROHS will gradually become active in more and more areas. The budget will therefore increase year by year during the first five years of EROHS life time, from Euro 5.000.000 for the first year to 12.000.000 for the fifth year. The personnel costs will in the initiating phase cover approx 10 employees increasing to approx 15 working full time in the EROHS-hub and the outreaching activities cover travel and meetings as a part of EROHS' broker function and mediator function. Decommissioning costs will be zero as it is expected that EROHS will continue in operation for the foreseeable future.

Leading consortium: EROHS must operate as an autonomous body, to secure the necessary legitimacy of the research communities. EROHS will also be accountable to its funders and the funds received in terms of expenditure and efficiency. To undertake and manage its operations and to ensure that the work is carried out according to the planned intentions EROHS will have a separate legal status.

There are a number of stakeholders in EROHS. Beside nation states these will for example be content owners and suppliers, standard communities, digital libraries and archives – all of which are either by national or European in the scope of their work. In its initial phase EROHS will be dependent of the experiences of existing national or pan-European infrastructures. A number of countries have informed that they are interested in bringing it forward (for instance the Nordic countries through a Nordic collaboration).

5.6.1 THE CHALLENGE

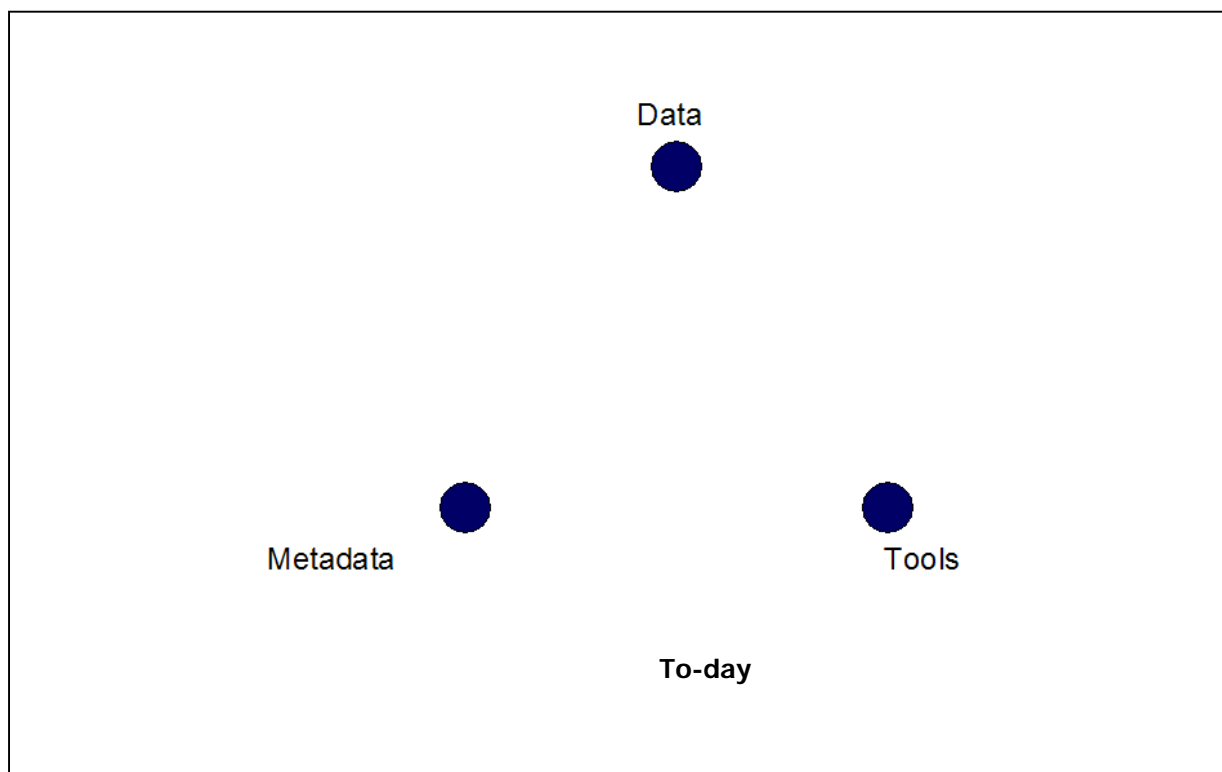
The humanities and social sciences are hampered by a fragmentation of the scientific information space. Data and its derivatives, sources in cultural heritage institutions, information and knowledge, are geographically dispersed and divided by language and institutional barriers. Consequently, European research is predominantly based on data from single nations or single data sources. Against this background the establishment of a European Research Observatory for the Humanities and the Social Sciences (EROHS) is proposed to address the current problems.

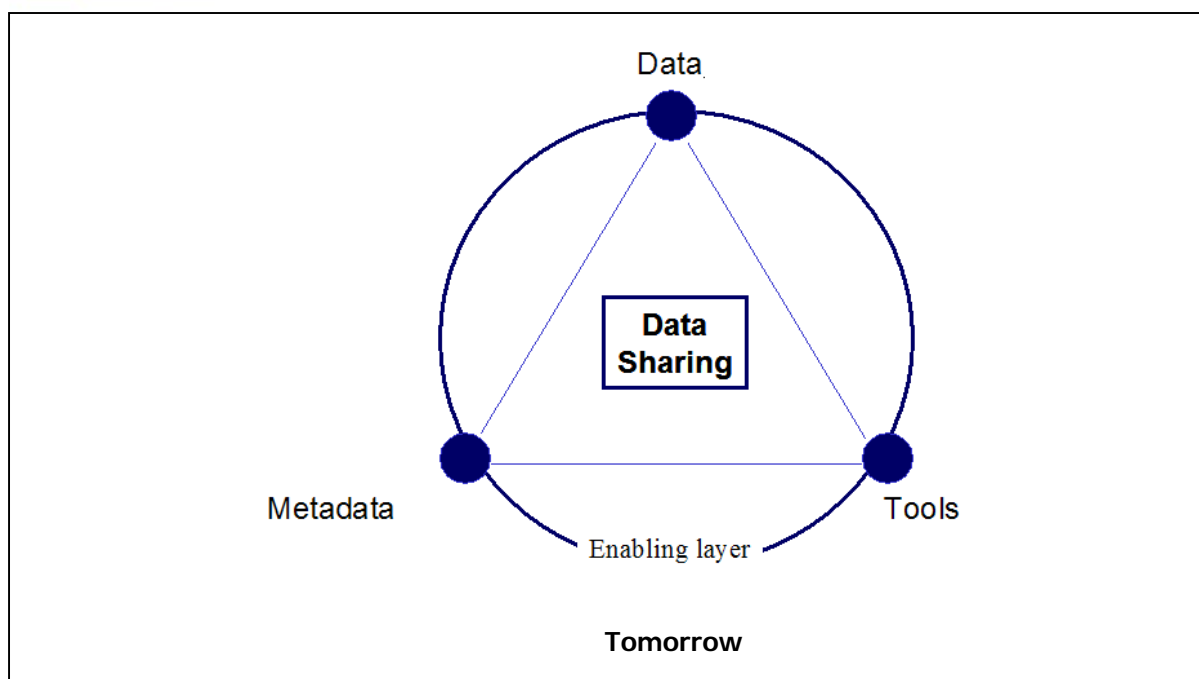
EROHS will be designed as a new European-wide infrastructure (RI) addressing the major impediments of existing research infrastructures, whether it be lack of coordination and cooperation, insufficient training, outdated policies, inadequate access, unsatisfactory tools or incomplete resources.

EROHS will in some areas be based on mature existing resources and infrastructures, supporting, promoting and extending their scope – and will in other areas facilitate the development of new infrastructures based on the need of the humanities and the social sciences by bringing together national initiatives, organisations and individuals in order to provide upgraded and enhanced European wide actions.

EROHS will be organised to promote and ensure cooperation and integration of data, technologies and policies. This calls for setting up a central and coordinating unit, an enabling infrastructure, explicitly and specially addressing the cultural, economic, legal, and institutional constraints to the realisation of Europe as a natural laboratory for the social sciences and humanities.

EROHS will operate both as a central and distributed facility with a strong physical hub working in close conjunction with a number of spokes across Europe, harnessing European expertise through a coordinated yet decentralised network.





The most important single keywords and principles of EROHS are “sharing and coordination”: sharing and coordination of data and other sources, of the best efforts nationally and at the European level for everybody to be able to join in, and sharing and coordination of metadata and tools.

The challenge is to add value to existing data and sources within the humanities and social sciences, by bringing together, supporting, upgrading, promoting and extending their current scope with the goal of ensuring genuine comparative research.

The national and European benefit of data and sources already funded, collected and stored elsewhere will thus be maximised. From this it follows logically that EROHS from the outset will base itself on the current crop of work and data, source material, metadata and tools funded and collected by other – national as well as European – organisations

The vision of EROHS is to facilitate and promote availability, access, quality and comparability of data and sources at European level in order to increase scientific quality and comparative research at European level. EROHS will create synergies and opportunities on the basis of the existing base of data and source material thereby fostering and supporting a culture of collaboration and sharing among European researchers.

Its mission will be to fulfil the four following purposes:

- The facilitation of access to and sharing of existing European and national data and source material, thereby more efficiently and effectively linking data resources already in existence

- The promotion and development of the highest standards and documentation relating to European and national data and source material in order to enhance the scientific quality of data and sources and their potential for interoperability
- The facilitation of new and genuinely European data. This will involve the stimulation of research-driven data collection and the digitisation of currently non-computerised research materials
- The provision of research training programmes for the next generation of data and source material providers and users

The initial task of EROHS is thus to ensure accessibility, comparability and quality of current data and sources and facilitate the generation of new truly European data.

In the longer run EROHS will also play a role in facilitating the augmentation of the current stock of data and source material in the humanities and social sciences through initiatives to coordinate and enhance the collection and generation of new data and sources in relation to e.g. national initiatives or initiatives in relation to European Framework programmes.

EROHS will not only be important on the international scene. The model can also be used to strengthen the national infrastructure. By using the same model in different nations, national hubs can be linked both to each other as well as to the international hub. In that way we will secure a clear coordination between the national infrastructures and the international infrastructures.

5.6.2 THE HUB AND SPOKE SYSTEM

EROHS will be established as a three-layers distributed infrastructure based on:

- *The enabling layer* (a central hub) addressing institutional policies and the broader legal and, political context and social norms – e.g. the intellectual property and privacy rights regimes – the development of standards and the creation of cooperation and synergy between European researchers and research communities.
- *The technical layer* providing the middleware, applications, exchange protocols etc.
- *The data layer* providing the actual content and serving as the base for the two previous layers.

These layers will be organized as a distributed system - a hub and spoke system- where a central and enabling infrastructure, a hub, is set up to ensure coordination and integration of data, sources, metadata, tools and policies across institutions and borders.

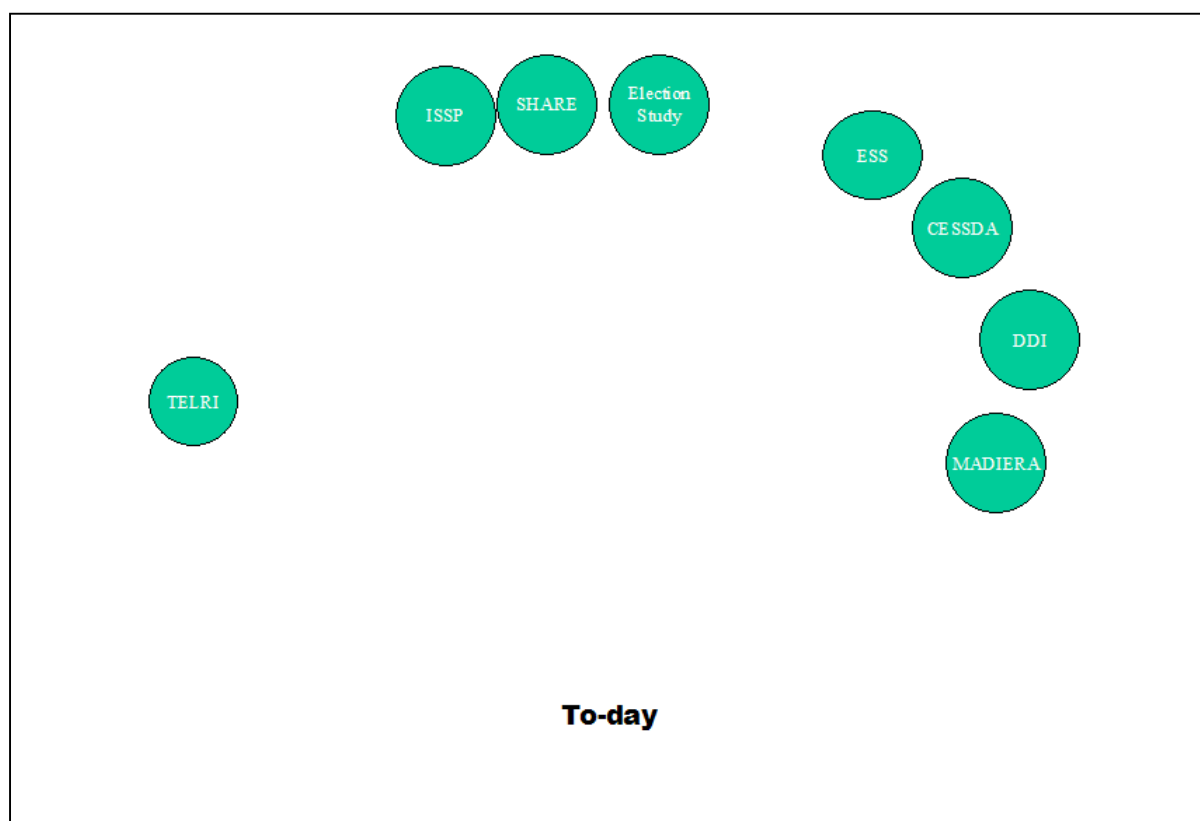
The system will be based on ‘communicative integration’ with the hub ensuring the inclusion of the spokes that wants to join EROHS and are able to fulfil the EROHS vision of open access and high quality as it is described above. The hub and spoke system is based on cooperation – it is no hierarchy with the hub deciding the work of the spokes. The

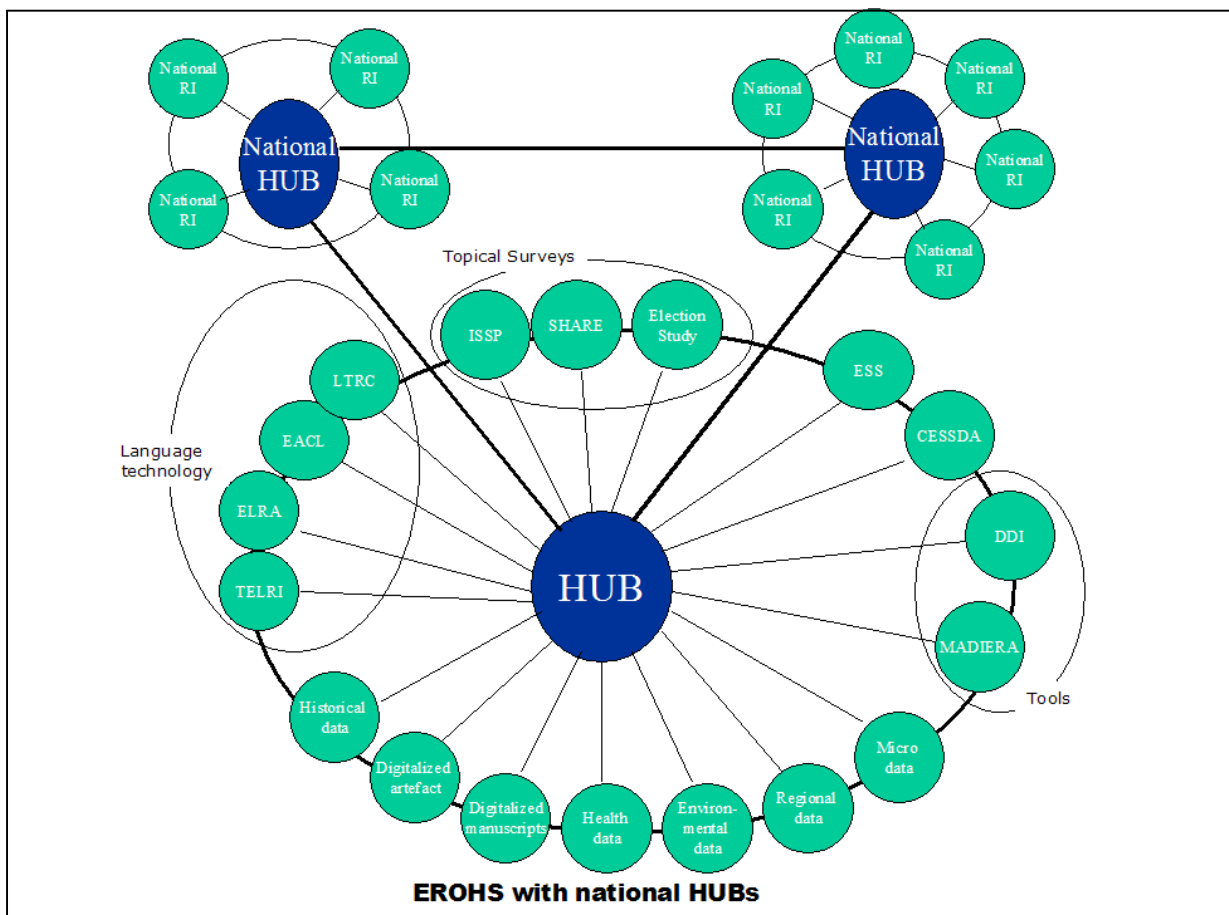
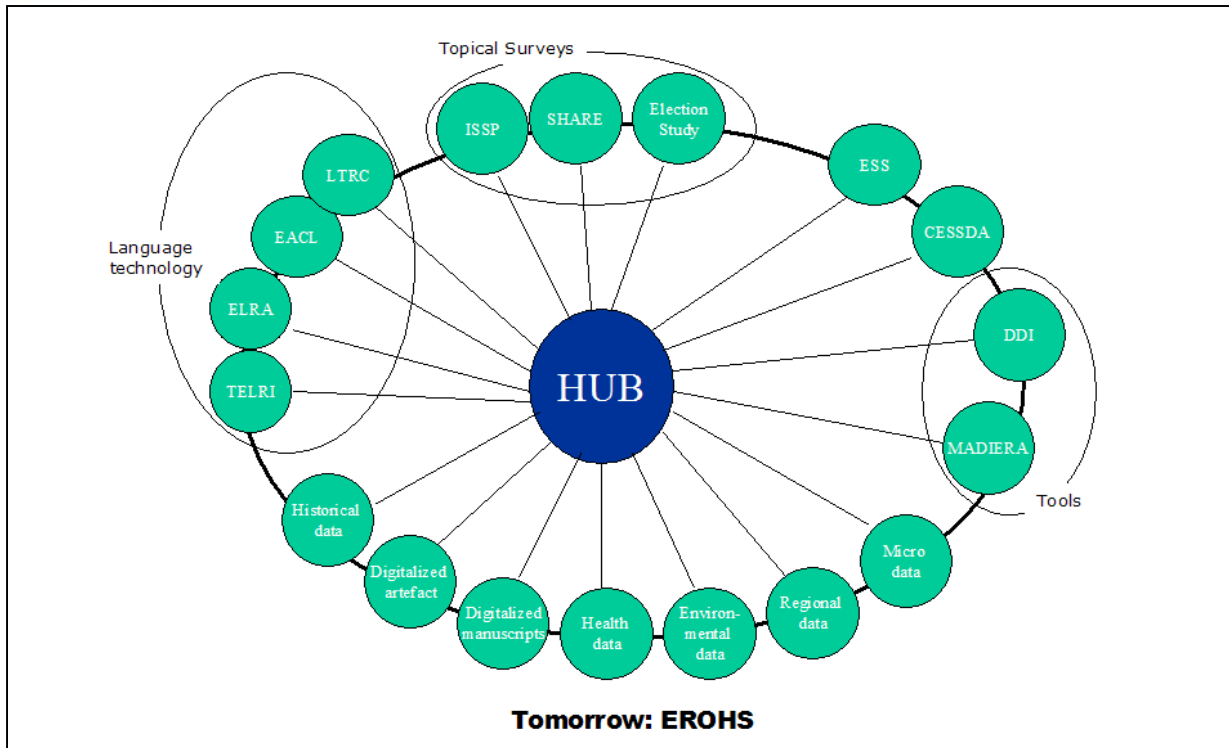
spokes will however; need to live up to the policies and principles guiding EROHS.

In order to have the greatest impact on research in the humanities and social sciences, EROHS must be based on strong national commitment and support. The success of EROHS relies on whether or not existing infrastructures join EROHS. The benefits for countries, institutions, researchers, and data, source and tool providers are at least four:

- Increased use of what they offer and by extension visibility
- Increased access to data, source material and metadata
- Increased access to tools
- Possibility to get funding to add quality and functionality to data, source material, metadata and tools

EROHS as a collaborative system aims to ensure data quality and documentation of the highest scientific standards. In order to be included in EROHS, the spokes have to live up to the high standards and principles set by EROHS. EROHS is more than ensuring democratic access to data. It is about sharing in the widest meaning of the word, which makes EROHS a system based on the highest scientific ambitions for the humanities and the social sciences in Europe.





In general terms, the large scale-public goods that EROHS promises to deliver will never come about, without collective and centralized action. This owes to the inherent complex character of its operations

and is thus the fundamental argument for the establishment of a strong, physical enabling hub.

European comparative research is dependent on sophisticated systems for the creation, curation and preservation of information. It is also dependent on a policy, economic, legal environment that encourages unimpeded access to digital information and digital tools. The most difficult barriers to overcome are not primarily technological, but economic, legal and institutional. The role of the hub is a key factor in this regard. Through its enabling functions the hub will ensure cooperation and integration between researchers, technologies and information across institutions and countries, provide support and tools and other shared capacities such as an authentication and authorisation portal and negotiate standardised access agreements across Europe.

While coherency and coordination is critical for the purposes of EROHS, it is equally clear that a single unit will not be able to deliver promised results in the form of more high quality cross national comparative research based on a wide range of high quality data and sources from distributed data sources. Indeed, EROHS is in general intended to build on an existing base of infrastructures and efforts; the data infrastructures and technical infrastructures already established throughout Europe. EROHS will in effect add an enabling layer on top of existing endeavours widespread across Europe, as an incremental but very important addition to the current data and technical infrastructural base.

In its totality EROHS will thus operate both as a central and distributed facility with a strong physical hub working in close conjunction with a number of spokes across Europe, harnessing European expertise through a coordinated yet decentralised network.

The spokes are crucial for EROHS. Their primary role is to provide access to data resources and the accompanying domain knowledge that normally follows from this role. Some spokes might also serve as providers of specific functions, mainly because of their expertise in areas important to the overall network. This could be specific domain knowledge or more generic methodological expertise.

The spokes will be spread across Europe; however work within them will not be performed in isolation from one another. Here the role of the hub is of critical importance in ensuring that the various activities are tied together, serving the overall vision of EROHS.

It is at this stage premature to put together an inclusive and exhaustive list of spokes. Indeed, the concrete spokes will have to be identified at a later stage in the lifetime of EROHS. Suffice to mention is that the sheer complexity of the matter underline the importance of the enterprise being distributed – or rather, why one centre cannot command the competence on all relevant issues.

- The spokes can represent different scientific fields – these may be organised along traditional disciplinary lines, or may be cross-disciplinary and organised according to subject, theme or approach. As quite disparate scientific cultures will be working together, there must be a system that brings these scientific communities together
- The spokes may represent one or several competences, for example methods, sampling, data archiving, standards, accessibility, as this expertise is distributed across Europe
- The spokes can reflect the rich diversity of national differences. All data gathered are traditionally anchored in the context of national communities. The detailed knowledge of national rules, procedures and legalities are both respected and brought into play at European level

However, a few examples of infrastructures that it is important to work closely with:

ESS comes immediately to ones mind. Even though ESS is a very mature research infrastructure with its own governance systems it could be of mutual benefit for ESS and EROHS to join forces at least in some areas. EROHS will need the experience of ESS staff and experts as input in its work with core activities like data documentation and dissemination. On the other hand EROHS could be useful to ESS – and other major cross national surveys like SHARE, ISSP, Election Studies – in organising scientific workshops on the development of quality in survey research and training activities for both users and creators of large cross country survey data.

CESSDA is a corporation between data archives in Europe. CESSDA promotes the acquisition, archiving and distribution of electronic data for social science teaching and research in Europe. It encourages the exchange of data and technology and fosters the development of new organisations in sympathy with its aims. It associates and cooperates with other international organisations sharing similar objectives. CESSDA contains knowledge and expertise that is highly relevant for EROHS. But EROHS is also of interest for CESSDA as this initiative is without an actual hub. EROHS could then act as the hub for CESSDA.

For art and the humanities there are enormous potential in working not only on bringing data sources together at a European level but also to organise national data sources in ways so that they can be used both nationally and comparatively. In many cases data does not even exist at the national level in ways that makes it possible to join forces at the European level. It will be an important challenge for EROHS to take on the task of working on how in the best ways

possible to solve the problems of data collection, storage and sharing within the arts and humanities. The important work of the expert group on the humanities showed that also in this field both national activities and the willingness to work together at the European level exist. So even though a lot of work needs to be done to solve the problems and release the potentials for data for art and the humanities this endeavour will be able to build on national activities and to be enhanced by the help and support of EROHS.

5.6.3 THE SCIENCE CASE

For the humanities and social sciences data and data sources are the fuel of empirical research – just as they are for the other sciences.

All scientists need scientifically driven data of the highest quality based on common standards if they are to understand, improve and critically test and validate empirical knowledge about a given subject. Meaning, interpretation, criticism and re-interpretation of sometimes fuzzy data are the core of empirical research within both the humanities and the social sciences.

The data which these scientific fields are using consist of a variety of human artefacts such as catalogues and bibliographies, national or pan-European datasets with macro and micro data information on the structure of society, the output and outcome of policies and the actions, and the behaviour and attitudes of individuals and detailed source material held as text, video, audio, images, cultural 3D objects, maps and so on. Both disciplines thrive on the massive wealth of data on demographic, societal, constitutional, institutional, political, legal, economic, cultural, linguistic, religious, and historical variations in Europe.

The data infrastructure is well in place in many European countries. The mapping exercise of the ESFRI roadmap working group for the social sciences and humanities (RWG SSH) conducted in late 2005 showed a massive wealth of activities in creating, organising, disseminating and managing data in a great many member countries of the European Union. Most of these important activities had a national scope. Only few were pan-European in nature like the collaboration between national data archives CESSDA, the creation and running of a number of surveys within the social sciences like ESS, ISSP, Election Studies and SHARE, and the cooperation between researchers within language technology like TELRI.

Data exist in many but not all cases at the national level, and in some cases at a multi-national level, organisations to manage and disseminate these national based data both within the humanities and the social sciences is in many cases in place and develop-

ing further every year – and there are good examples both within the humanities and the social sciences of best practices on how to try to deal with this at the European level following the highest scientific standards

The technical infrastructure is an area that have witnessed incredible developments in recent decades, pointing to a situation where a decentralised, distributed facility is not just possible but also the only sound approach to developing future research infrastructures. Grid technology and web services make possible the sharing, using and managing of massive volumes of data and information. These technologies, by their very existence also highlight the need for streamlining connectivity, communication and collaboration. A decentralised data providing system would otherwise easily turn into a nightmare for anyone who wants access to data from multiple providers.

The technologies to facilitate a decentralised, distributed facility exist and are used and being developed. The European project MADIERA is engaged in the development of tools to facilitate data sharing. In relation to ESS immediate online access to data and data documentation for everybody at the time of the release of the data has shown to be possible.

The enabling infrastructure is the one that needs installation and development - and the hub is the enabling layer that the humanities and the social sciences need. And the creation of the hub does not start on bare ground though.

The conclusions from the first EROHS report – which were delivered to ESFRI in June 2004 – highlighted the need for action and further development in three areas:

- Coherency and funding

Data and sources within the humanities and social sciences are primarily collected and stored within the confines of the nation-state and often through specific isolated research projects with little or no follow-up. Or data and sources are gathered by the efforts of international administrative bodies with little focus on the needs of researchers. The absence of coordination at a European level leads to sub-optimality and even duplication of efforts and incommensurable local solutions.

The funding of research infrastructures is often short term and often secured nationally or regionally. But research infrastructures at a European level are beyond the capabilities of single countries. Indeed the EU increasingly finances research infrastructures – and especially access - but funding is allocated on the level of specific projects, consequently only adding to

a patchwork of research infrastructures. The lack of coherency at a European level is evident.

- Accessibility

Access is restricted due to many reasons: juridical, privacy, confidentiality, ownership rights, linguistic, financial, pricing systems, institutional impediments, lack of online availability, variety of storage formats, and so on. Data and research infrastructures are in general a public good. However, whilst data are not a scarce resource in Europe, they are not as translationally available for secondary analysis as they could be.

- Standardisation and quality

Research in the social sciences and the humanities in Europe is currently often carried out within national contexts and based on nationally generated data and sources with a large variation in quality. This fragmentation and compartmentalisation has severe implications for the quality of the data in a European perspective as crossing borders within the humanities and social sciences comes at a cost. European research based at a European level often falls below the standards applied at the national level as the available data are not immediately comparable due to differences in standards and documentation – sampling, collection, variables, size, formats. Conducting European research one has to frequently rely on post harmonisation of national data at the level the lowest common denominator. As a result, quality and detail are both compromised.

These conclusions have been presented and discussed widely within the European research community in the humanities and the social sciences both at a national and European level since the launch of the first EROHS report.

They have been discussed and adapted by a number of national funding organisations – and they have been given the highest priority on the list of opportunities delivered by ESFRI to the Commissioner in March 2005. There has in general been created hopes and anticipation among important stakeholders that EROHS will materialise within the coming years.

The need for EROHS is demonstrated by the results the mapping of research infrastructures in the humanities and the social sciences conducted by the ESFRI SSH RWG.

This mapping exercise shows major research infrastructure activities nationally which underlines the need for coordination in order to hinder duplication of efforts and ensure cooperation, synergy and the move towards common European standards for distributed facilities in Europe.

The need is there now when the national based efforts are mainly emerging – not in ten years time because of the danger that an array of different practices without coordination efforts will develop in the member states.

EROHS is the way to fulfil this need. Already in its initial phase it will be able to draw on the experiences of existing infrastructures like the ESS and coordination initiatives like CESSDA and to communicate their experiences to emerging infrastructure initiatives at the national and European level. This will be the first important phase in the endeavour to ensure coordination at the European level. But more steps will be taken.

With the technological evolution Europe has the potential of becoming a natural laboratory for the humanities and social sciences. With the provision of data on European culture and societies with the impressive and important cultural heritage of Europe and the members states almost optimal combination of diversity in social models and homogeneity in social goals Europe can constitute the 'best case' research object for historical and between nations and social models comparisons for scientists working within the humanities and social sciences.

The recent US Cyber infrastructure for the Arts and Humanities report proposed as its grand vision 'access to all surviving humanities and cultural heritage information across all time and space'. The enabling research infrastructure – EROHS – proposed in this paper is aiming at just that. It will have the potency of organising the communication, coordination, documentation and sharing of information in ways that will set standards for research infrastructures worldwide and make Europe the world leader in this area. It will provide Europe with the tool to realise the vision of open, distributed, well coordinated, well documented, on-line access to data for the humanities and the social sciences.

EROHS is an instrument for planned cooperation and integration. It will be based on existing resources and infrastructures, supporting, promoting and extending their scope – and facilitate the development of new infrastructures based in the need of the humanities and the social sciences.

The proposal of adding value to existing investments makes intuitive sense when taking into account first that e-science research infrastructures in the humanities and social sciences in its essential quality are public goods, being non-exclusive and non-rival. Secondly, public funded data as a matter of principle, and certainly in the interest of the funders, should be accessible for secondary analysis and replication.

The vision of EROHS is to facilitate and promote the availability, high quality and comparability of data

in order to reach the highest level of scientific quality and comparative research at European level. EROHS will create openness, synergies and opportunities on the basis of the existing base of data, thereby fostering and supporting a culture of collaboration and sharing among European researchers.

5.6.4 WORK PROGRAM

The EROHS work programme encompasses in the outset 11 activities containing 20 promises on areas where EROHS will work to make a difference for stakeholders in the humanities and the social sciences all over Europe. EROHS will promote progressive work by listing to the wishes and visions of others and by bringing people and organisations into the EROHS network to work in designated areas.

The work program are described in generic terms as it comprise all data types within the humanities and social sciences, safeguarding the plethora of theories, methods, data, subjects and fields subsumed under the disciplines of humanities and social sciences. There is no intention of pre-empting the scope of EROHS' operations or prioritising any discipline, data-type etc. over another and it is envisaged that the work programme will develop based on the interplay between EROHS and the research communities in the Community as EROHS matures.

- Accessibility and coordination

EROHS will provide information, support and guidance on common regulations relevant for the access to data and sources throughout Europe and advocating a uniform access infrastructure across Europe and a commonly agreed upon framework of operational principles, good practices and tools ensuring researchers and projects easy access to rich data and sources.

EROHS will promote policies that foster openness and access and various procedures will be trialled with a view to establish the best practice examples for optimal access, e.g. data licensing. This requires professional knowledge on common regulations on data accessibility throughout Europe and the ability to act as a broker between data providers and users and forming partnerships and negotiating with data providers and holders on European as well as national level.

- Standardisation, documentation and digitisation

EROHS will promote the European use of metadata and data standards with a view to interoperability and reusability and policing the compliance.

EROHS will offer advice to and work to upgrade datasets according to the highest standards.

EROHS will develop spaces for collaboration among toolmakers and standardbearers, as well as scholarly validation of these activities

- Preservation and maintenance of data

EROHS will promote policies and negotiate standard requirements for depositing publicly funded data in (national) data archives. The facilitation of future-proof preservation of data and sources involves validation, normalisation, storage, migration and delivery to users that have been authenticated and authorised to receive the data.

- Researcher-driven data collection

EROHS will promote researcher-driven collection of data and sources. Many data especially in the social sciences are collected for administrative purposes. Only as a side effect is data used for scientific purposes. Data and instruments need to be tailored to the scientific task. Researcher-driven data is an essential ingredient for modern empirical research in the humanities and social sciences.

EROHS will investigate possible benefits creating an European scientific survey agency to fulfil the needs of researcher driven data to be collected using the highest scientific standards for data collection

- Seal of approval

EROHS will issue certificates for trademark datasets that fulfils a set of requirements in terms of quality, sustainability, documentation, accessibility and usability, all consistent with accepted international standards.

- Harmonisation and data linkage

EROHS will explicitly promote new data collection that facilitates cross-national comparisons on a pan-European scale. While data may best be collected at the national level, coordination and harmonization is necessary at the European level. Generating a consensus on standards for such coordination and harmonization is an essential task.

EROHS will work to promote the active harmonisation of existing datasets with a view to enable comparability. It will facilitate data linkage between different sources of data and information, enabling the enrichment and harmonization of research data in new areas and in areas where no datasets currently exist.

EROHS will work to develop a central variable/classification database, standard modules of

background variables and a European question bank.

- Central data access / portal

EROHS will promote the creation of a central data registry, providing a clearinghouse for European data in the humanities and social sciences. The central function will be the provision of registration service allowing resource providers to register their data with the system, a multi-lingual resource location service allowing researchers and other end users to locate data easily and the development of functionalities that facilitates easy identification of comparable data across datasets and sites.

- Exploratory instruments and methods

EROHS will promote the employment of recognised standards, but it should also nurture a culture of experimentation and risk-taking. Activities can include calibration, quasi-experimental settings and new modes of data collection and measurements.

- Long-term data collection

EROHS will promote the long-term continuity of data assembly and data collection efforts that may span the duration of e.g. two or more EU framework programmes or several national funding cycles. Currently, the European Research Area does not provide a home for such infrastructures. However, many infrastructures in the humanities and social sciences gain substantially value by increasing their time dimension.

EROHS will work as a mediator towards the funding regimes to promote the long-term viability.

- Training programmes

EROHS will provide training programmes for both data brokers and users with a view to develop the skills, knowledge and abilities of data producers, data archive and dissemination staff and users with regard to the methods and possibilities of data use.

EROHS will also provide training for researchers to introduce them to the methods and possibilities of data use. Specific activities will be workshops, summer schools but also the production of curricula to promote wider and informed use of data.

- Technological development

EROHS will at its centre of its attention *not* have technical infrastructure. However, the implementation of the above operations will require accompanying and development of the technical architectures that is demanded for the implementation of its other operations. This will involve the development and

maintenance of technological interfaces, protocols, and middleware software tools. Although EROHS will not function as a research council, it will need funding to initiate specific actions in accordance with the work program.

5.6.5 STAKEHOLDERS

In accordance with the ESFRI rules and procedures the countries participating in this collaboration will be the primary stakeholders in EROHS. It goes without saying that EROHS will need the support of a sufficient number of ESFRI member states to become a reality.

For member states - joining EROHS ensures influence on the development of this infrastructure cf. the description of governance below. It also ensures that national data, which live up to the EROHS criteria of sharing and high quality, will be actively used by researchers across Europe. This will promote the scientific interest for the specific countries data and problems. By joining EROHS member states will also facilitate access to the European scene and to European collaboration for their national research communities in the humanities and the social sciences.

Therefore EROHS will be an advantage for the participating countries – but EROHS also need authority mandated by the countries. EROHS need a hub with a strong mandate from the participating countries and the EU. Unambiguous political commitment and even an official designated authorisation is required for EROHS to engage as a broker between parties and to perform its monitoring function of for instance data access, policies and practices.

There will be a number of stakeholders in EROHS beside nation states. Among others these will be content owners and suppliers, researchers and professional societies, funding agencies, standard communities, digital libraries and archives – all of which can either by national or European in the scope of their work.

EROHS will in its initial phase be dependent of the experiences of existing national or pan-European infrastructures like e.g. ESS and CESSDA. But existing infrastructures will also have an interest in joining the collaboration ensured by EROHS. Through EROHS more researchers will be introduced to the possibilities created by these infrastructures leading to increased research activities and use of best practices in the areas covered by the specific structures. This will increase the interest for these structures, it will enhance quality as more users will give input to its development – and even though EROHS will not be in a position to fund specific structures joining EROHS can become a decisive argument when decisions on funding are taken by national and pan-European funders in the humanities and the social sciences.

Training and seminar activities organised by EROHS will also be to the advantages of both existing and emerging research infrastructures bringing together researchers from all over Europe and all over the world who are experts in a specific area. The methods used by e.g. ESS have given rise to a number of journal articles on methodology in cross-country surveys. It could be the role of EROHS to bring together in an open forum all who takes an interest in this issue for discussion and development.

Emerging infrastructures will be important stakeholders in EROHS. This goes both for national and pan-European research infrastructures. The concept 'emerging' infrastructures covers a variety of activities. It can be the very many initiatives in relation to e.g. digitalisation, which based on the mapping exercise of the SSH RWG seem to be emerging in a number of member states. It can also be established data bases like e.g. the ISSP wanting to use the high standard tools for language check of questionnaires used in many countries and for data documentation developed by the ESS. In both cases EROHS has a role to play, first as facilitator of the process leading to the establishment of a spoke for digitalisation and secondly enabling an important upgrading of the quality of an existing database.

5.6.6 GOVERNANCE

EROHS must operate as an autonomous, organisational body, drawing its expertise from the scientific communities. If EROHS is to secure the necessary legitimacy of the research communities within the humanities and social sciences at large it needs to be self-governed according to excellence alone.

EROHS must also be accountable to its funders and the funds received in terms of expenditure and efficiency. This implies that EROHS needs an effective governance structure that ensures the backing and trust of the funders.

Additionally a key requirement will be to obtain a separate legal status enabling EROHS to undertake and manage its operations and to ensure that the work is carried out according to the planned intentions. Different models are at play here: three options will be a non-profit organisation and intergovernmental organisation or a European Economic Interest Grouping (EEIG), a legal entity based on Community Laws to facilitate and encourage cross-border European cooperation. Turning to the governance system, the hub will take the lead in managing EROHS and it is envisaged that the co-ordinating hub will have the following main governing bodies:

- The Council

The Council will decide on the overall strategic plans and priorities. It will ultimately be responsible for all important decisions; it will set out EROHS' policies in scientific, technical and administrative matters, approve the programme of activities, adopt budgets and review expenditure. It is suggested that alongside representation from the European Union, each member state should have two official delegates, one representing his or her government's administration, the other the national scientific interests. The Governing Council elects a Chairperson and Deputy Chairperson, whose terms are three years. The Chairperson and Deputy Chairperson are to govern the Council. The Council can set up advisory ad hoc-committees within chosen focus areas where special competence is required.

- The Director

The Director and the hub-secretariat will be the executive body responsible for the day-to-day management of EROHS, being accountable to the Council. Together they will be the drivers of EROHS. The Director must be an internationally highly respected scientific figure, supported by a Directorate-team of equal standing. The Director of EROHS is to serve as the Executive Secretary to the Council.

- The Directorate

The Directorate, consisting of the Director and the heads of the spokes will be the forum where the coordination and implementation of the EROHS-activities and division of labour between centre and nodes will be planned and agreed.

5.6.7 THE FINANCIAL CASE

The financial support necessary for the creation of EROHS will have to come from public funds – and pooled public funds as the whole endeavour is truly European by nature and clearly exceeds the national competences available. EROHS – or other research infrastructures for that matter – cannot be run like a private business-enterprise due to its character as a public good.

In comparison to research infrastructures within other sciences EROHS requires a modest investment. And certainly in the year of the commencement as EROHS does not demand huge investments in the year for the construction of the facility.

The funding will very much be top-off funding, based on the national and EU investments – as 'in-kind' funding – already made to data and infrastructure and hence capitalising on those investments. This especially goes for the decentralised costs for the spokes they will entirely be funded outside the EROHS-

budget, but will altogether represent massive investments reaching millions of euro. In that sense, the potential benefits for both the humanities and social sciences and Europe clearly exceeds the modest needed funding for EROHS.

Another perspective on the funding aspect is that EROHS also promises to reduce the duplications and inefficiency that currently beset the existing investments in research infrastructure. The funding demanded for EROHS might then very well correspond to the expected savings.

Being a new construction without actual precedents, the budget for EROHS is arrived at by extrapolating from alike national efforts in the humanities and social sciences on the one hand and comparable international endeavours within other sciences on the other hand.

A tentative budget for the first five years of EROHS' existence is as follows (€):

	Year 1	Year 2	Year 3	Year 4	Year 5
Personnel, outreaching and operating expenses in the hub	3.000.000	3.250.000	3.500.000	3.750.000	4.000.000
Work Programme	2.000.000	3.500.000	5.000.000	6.500.000	8.000.000
Total	5.000.000	6.750.000	8.500.000	10.250.000	12.000.000

A few further remarks need to be made to the budget. The personnel costs will in the initiating phase cover approx 10 employees increasing to approx 15 working full time in the EROHS-hub and the outreaching activities cover travel and meetings as a part of EROHS' broker function and mediator function.

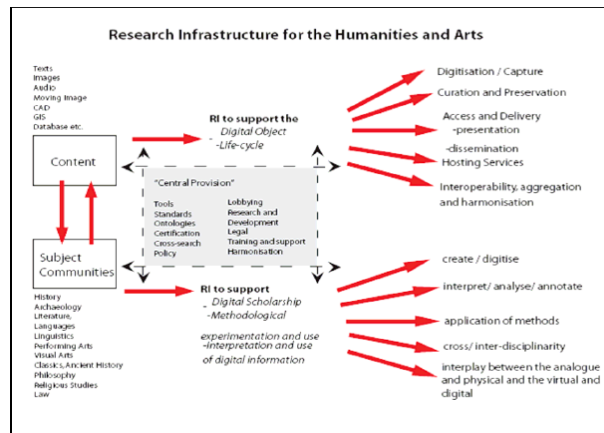
The costs for the work programme include all the eleven operations' gradual realisation, including for instance the training activities. As EROHS will gradually become active in more and more areas of the work program the budget for the work programme is set to increase year by year during the first five years of EROHS life time.

After the initial five years, the funding agreements – simultaneously with the general review of EROHS – are to be reviewed, with the option of bringing in other funding sources and modes of funding.

By proposing EROHS we are putting forward an organisational frame for optimizing access to data. This is necessary to make the European Research Area come thru for the Humanities and the Social Sciences. We are proposing a distributed system where it will be possible for all European countries to take part and benefit. The system will also give improved standards and better coordination of the research infrastructure for the Humanities and the Social Sciences at a national level for each European country.

5.7 DIGITAL RESEARCH INFRASTRUCTURE FOR THE ARTS AND HUMANITIES – DARIAH

Digital Research Infrastructure for the Arts and Humanities (DARIAH)



The facility: DARIAH will be based upon an existing network of Data Centres and Services based in Germany (Max Planck Society), France (CNRS), the Netherlands (DANS) and the United Kingdom (AHDS). The model is however open and will be able to embrace new fields. It will also profit from European Cultural Heritage Online (ECHO), an Open Access Infrastructure to bring essential cultural heritage online, of which the MPG is the host.

Background: Just like astronomers require a virtual observatory to study the stars and other distant objects in the galaxy, researchers in the

humanities need a digital infrastructure to digitise, get access to and study the sources that are until now hidden and often locked away in cultural heritage institutions. Much of the source material so vital to humanities is scattered across libraries, archives, museums and galleries, and as yet only a fraction is available in a digital form.

What's new? Which impacts DARIAH aims at providing an infrastructure for the entire field of the art and humanities and access to cultural heritage of Europe. It contributes therefore to creating a common understanding of the cultural diversity and its history in Europe. It will help to cross cultural boundaries and to create a new European coherence based on mutual understanding and on true integration of the uniquely rich European traditions. This aspect is particularly relevant in education.

Research and development will find rich primary data for scholarly interpretation which will also allow not only for comparative research over time periods, cultures, languages, or regions, but trigger the possibility of novel research questions, that with traditional access to cultural heritage sources dispersed over a multitude of different sites and institutions could up to now not be approached. Likewise, due to its distributed and modular architecture the proposed infrastructure can build on and integrate existing endeavours and will also help to enhance national infrastructures.

Timeline for construction and first operation with related estimated costs: DARIAH will be based on partner organisations that have national infrastructures in place and are already collaborating. A core group of national institutions will directly contribute to DARIAH by offering their existing infrastructures (thus ensuring a quick take-up within three years at most). Preparatory cost is estimated at €2M and Construction cost at €8M. Operational cost is estimated at €4M per year.

Leading consortium: AHDS, CNRS, DANS and MPG are ready to share their national resource centres in order to allow a quick implementation of this joint infrastructure. Define a coordinated policy and management structure that allows them to speak with one voice at an EU and international level; Implement a series of concrete actions that will lead to the implementation of a joint European infrastructure.

5.7.1 INTRODUCTION

The recent US Cyber Infrastructure for the Arts and Humanities report proposed as its grand vision ‘access to all surviving humanities and cultural heritage information across all of time and space’. Since the publication of that report, the National Science Foundation in the US has produced a draft strategy aimed at implementing many of the recommendations contained in the report. The challenge for Europe is to ensure that the development of a research infrastructure for cultural heritage and the humanities that can match that of the US.

The Research Infrastructure (RI) proposed in this paper would seek to take some small steps towards achieving this grand vision for European humanities and cultural heritage information, aiming to provide an infrastructure that could support access to all surviving humanities and cultural heritage information for Europe. It would do this by focusing on four key elements:

- Providing a coordinated infrastructure across Europe that would act as a catalyst to bring together the best efforts of national initiatives, organisations and individuals in order to provide upgraded and enhanced European wide actions, initiatives and services that could not be provided at local or national level.
- Providing a coordinated infrastructure that would act as both a catalyst and support for the development of national services and digitisation programmes aimed particularly at those European countries without such services and programmes.
- Providing a coordinated infrastructure that would act as a catalyst to bring together the different sectors involved in cultural heritage and humanities information management and access – education, memory and cultural heritage institutions and organisations, and the commercial sector – in order that they might work together for the benefit of both themselves and the research communities across Europe
- Providing a coordinated infrastructure that would act as a catalyst for the enhancement and promotion of digital scholarship in the humanities and arts across Europe, including facilitating cross-disciplinary research and the sharing of content, tools and methods across communities of practice and discipline domains. This would also ensure that the arts and humanities did not work in isolation but took note of developments across the social, physical and medical sciences.

5.7.2 THE SCIENTIFIC CASE

The case for the Research Infrastructure described here is driven by three key factors: first, the changing nature of research practice, knowledge creation, and information sharing; second, the highly distributed and dispersed nature of much cultural heritage and humanities information; third, increasingly pervasive broadband connectivity and a growing range of technologies and applications that offer opportunities for creating and sharing digital information and knowledge across a highly distributed environment. These ‘grand challenges’ require a new way of thinking about the kinds of infrastructure and support services that are provided for research, teaching and learning across Europe.

Research practice in the arts and humanities is about criticism and meaning, interpretation and re-interpretation, and about extracting meaning from often incomplete and fuzzy data. It requires researchers to seek out and track down a wide range of primary and secondary sources, to organise and structure these, and to analyse and interpret them, and to publish the results. In the era of pervasive broadband connectivity the way in which these processes are undertaken is changing, and in some cases, the processes themselves are changing. Increasingly, research practitioners are using the power of the web, new tools, and the range of digital information that is available to them, to create their own personal network spaces, to publish on-line highly interactive themed collections of research information and knowledge, and to visualise and reconceptualise their interpretations and analysis. New forms of collaboration are also emerging as the tools available encourage and enable ‘web-working’ across the globe. Meeting the challenges created by these changing research practices requires a new kind of Research Infrastructure that can respond easily and seamlessly.

The nature of the source material on which research is based ranges from simple catalogues and bibliographies, through to detailed source materials (both analogue and digital) held as text, video, audio, images, cultural 3D objects, maps, and so on. This material is highly distributed, dispersed across a wide range of different organizations including libraries, archives, museums, galleries, and archaeological and cultural heritage sites. An increasing amount is now digitised, but much, much more remains hidden in small archives or libraries with minimal cataloguing, where access may be extremely difficult and expensive. A significant digitisation programme is required to make these information sources more widely available and accessible. For those resources that are digitised, they are likely to be described differently using different metadata standards, may well have used different technical standards in their creation, and will be presented and published through a large number of disparate sites, some simple, and some highly complex. Moreover, some of these source

materials will be protected by copyright or will be available only as commercial offerings. Making sense of this plethora of information and ensuring that all that is available can be found and accessed is a huge task.

We are also witnessing rapid changes in technology and in the digital environment in which we conduct research. The emergence of Web Services and RSS, Wikis, and Blogs to streamline connectivity, communication and collaboration; the rise of the Grid for sharing, managing and using massive volumes of data and information; the phenomenal growth of Google, Yahoo and Amazon, and the rise of their related services such as Google Scholar, Google digitisation, Flickr, Wikipedia and the like, are providing significant new opportunities for the way in which the research communities can create, share, and use information, and communicate and collaborate. The challenge is surely to engage with these new technologies and seek to use them to create a European Research Infrastructure that is second to none.

The impact of these three drivers is a new kind of collaboration out of which arises new scholarship. Research scholars and research technologists are coming together as equals to grapple with the problems and questions that arise in arts and humanities research, and through the use of digital scholarship are creating innovative solutions and new knowledge. Research practitioners working in the arena of humanities computing, or digital scholarship as it is sometimes known, are formulating new theories of the interaction between content (the source materials on which research is based), analytical and interpretative tools and technologies and data structures, methodological approaches, and disciplinary kinships. Much of this has been expressed by Harold Short and Willard McCarty in their work on Methodological Commons which seeks to outline an intellectual and disciplinary map of humanities computing that incorporates the complexity of digital scholarship.

The Methodological Commons approach argues that the digital life-cycle is part of an iterative process, where digitisation processes, the structuring of digital content, the curation and preservation of digital content, and the results of research are binding together, each informing and feeding into the practices of the other. We are witnessing this through a massive increase in the on-line availability of dynamic and growing digital information resources with multiple interpretive layers and presentations targeted at different user communities, and where research results are increasingly used as navigational devices, linking research data with publications, and with other relevant sources.

However, much of this work still takes place within discipline and institutional silos, and the ability to engage meaningfully across domains and discipline

areas remains problematic. The challenge is to bring these different elements together in a pan-European coordinated framework where different professions (research, technologists, archival etc.) can work together, where cross-disciplinarity is encouraged and facilitated, where shared understandings and use of standards can flourish, where research practitioners can learn from and teach others about different methods and practices, and where not only the digital content is preserved, but also the interpretive and presentational layers are recorded and preserved, and made available for others to add their own interpretations and annotations. This new paradigm requires an integrated Research Infrastructure at the European level that acts as an umbrella to draw together the disparate parts.

The primary aim for European Cultural Heritage and Humanities Research outlined in this paper is to establish a coherent and coordinated RI that can take forward this agenda, that can take advantage of all these new technologies have to offer, take account of the new ways of working – and encourage new ones to develop, and that can facilitate the creation and digitisation of new information, and hence new knowledge about our histories, our cultures, our knowledge and understanding, and our role in the society of today.

5.7.3 INFRASTRUCTURE OPPORTUNITIES

Significant impact could be made by developing a Research Infrastructure that addressed the following three interlocking functional areas:

- Digitise – Curate – Preserve
- Discover – Access – Deliver
- Connect – Collaborate – Use
-

These three functions would be guided by needs and requirements of the research communities in the arts and humanities, and be based around the iterative digital life-cycle as discussed above. The RI, through addressing these three functions, would seek to bring together the research communities with the information managers and providers, underpinned by a strong and reliable technical framework. We envisage that such a comprehensive model would support a European wide digitisation programme, enable researchers to engage fully with the widest possible range of information and knowledge, and ensure best use of the available (and forthcoming) technologies. Crucial to this would be a focus on open access to the digital information on which research relies, and open access to the results of research. If Europe is to maintain its world class research profile then an emphasis on open access is essential.

The RI should support **digitisation** by providing advice and guidance on legal issues and copyright

clearance; advice and guidance on digitisation methods and standards, including metadata and technical standards; and by providing digitisation services at dedicated digitisation centres with expertise in different information types e.g. video, images etc. These Centres would also offer mobile facilities for use in Archives and Libraries and other places where content may not or cannot be moved. The RI should support multiple **content** types, including texts, databases, images, moving images, sound, spatial data.

The RI should support **curation and preservation** by providing advice and guidance on preservation and sustainability issues, and by providing dedicated curation and preservation services that could be used by those unable to undertake these complex tasks themselves. The RI should also undertake R&D into sustainability of deep websites, and the preservation and long-term access of complex data, and the interpretive layers embedded within much multi-media digital content.

The RI would comprise:

- Legal Services that would provide
 - Advice and guidance on licencing and access issues
 - Model agreements and licences, including a European Creative Commons
 - Copyright negotiation and resolution of IPR issues
- Digitisation Services that would provide
 - Support and advice on the creation and digitisation of cultural heritage information and objects, including metadata and related documentation
 - Facilitate the take-up of standards and best practice, including those for the long-term curation and preservation of digitised information, including the provision of model solutions and case studies
- Digitisation services specialising in the digitisation of particular information types e.g. video, sound, museum artefacts, bibliographic and catalogues etc.
- Preservation Services providing
- Advice and guidance on preservation and sustainability issues;
- Preservation services on behalf of cultural heritage and memory organisations;
- Research and development into the preservation sustainability of cultural heritage information and object
- Registries and technology watch services

The RI should support **discovery, access and delivery** of digital information by using advanced methods of data and text mining to support enriched and enhanced aggregation services, and ensure that the enriched material is exposed to Google type ser-

vices to ensure the fullest possible exposure. This should include the development and implementation of ontologies and multi-lingual searching. The RI should also support visual searching (non text based methods) for image, moving image, and audio materials, and should work alongside Google and others to develop new non-textual methods for searching and locating relevant sources of non-textual information.

The RI should provide grid-enabled hosting services for large volumes of digital information, for example, satellite imagery or large quantities of video content. The RI should provide hosting services for cultural heritage and memory organisations that are unable to provide these services. Included in these hosting services should be facilities for authorisation and authentication. The RI should provide advice and guidance on the creation of sustainable websites and services, and the use of appropriate technologies.

The RI would comprise:

- Aggregation and Discovery Services providing
 - Aggregation of related resources by, for example, subject, theme, content type (including digitised information and objects, publications, catalogue records and bibliographies) from across Europe to create a growing universal working space, that adds value through harmonisation, clustering and adding additional information using advanced data and text mining techniques
 - The application of ontologies for discovery and use at a deep, rich level of granularity across multiple collections
 - Applying new methods for visual searching and browsing
 - Service layers interacting with Google and other such Services to ensure full exposure of content
- Hosting Services providing
 - Grid-enabled platforms for the access and delivery of resources for those organisations and individuals unable to support the development of their own sustainable services
 - Authorisation, authentication and payment services for access to protected information (for example for access to restricted materials, or those requiring permissions to use or requiring payment to use)
 - Advice and guidance for presentation and on-line publication, website development, interface development and related HCI issues, use of web technologies and web services

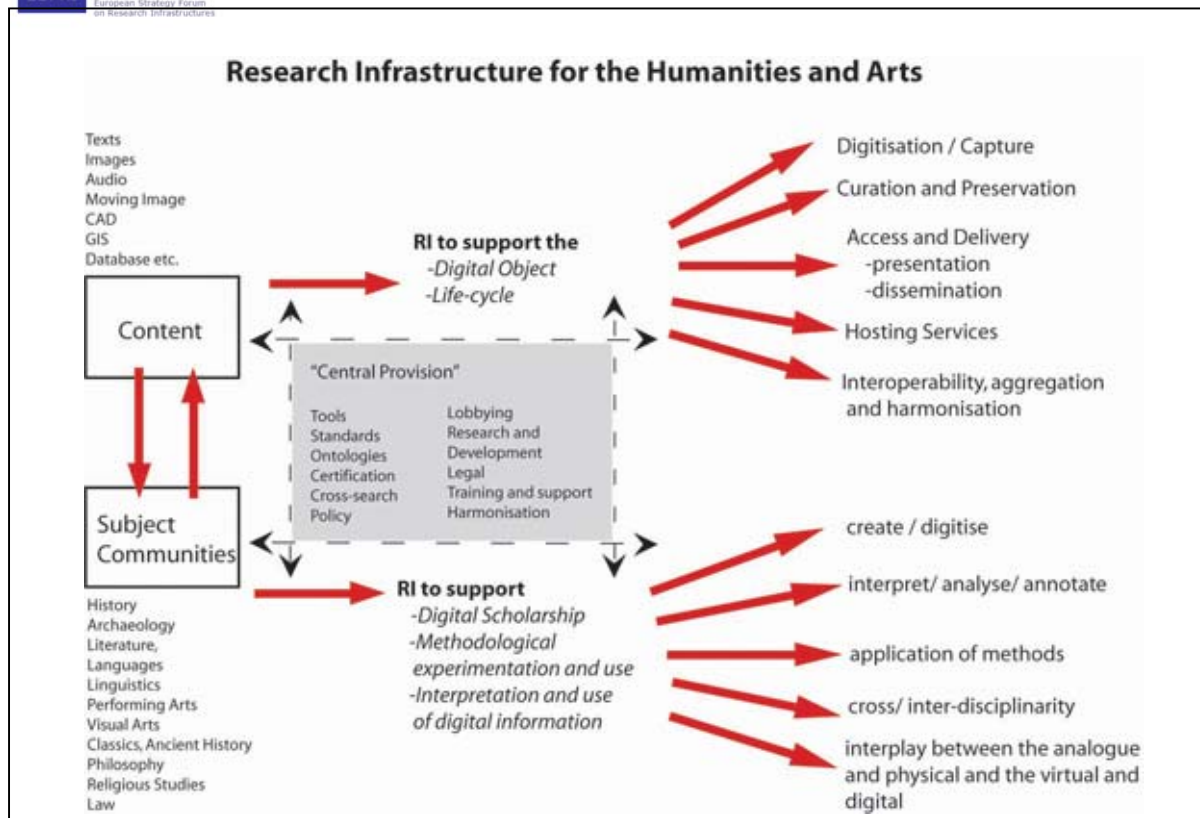
The RI should support **connectivity, collaboration, and use** of digital information by using Access

Grid and VRE technologies to support collaborations, and to develop shared collaborative working spaces with groups of research practitioners. The RI should provide a training and awareness programme to promote the European-wide use of standards and best practice. It should also take the lead in developing user and usage analysis services that would be used to further enhance and upgrade services. The RI should also engage with the linguistics and information science communities to develop multi-lingual ontologies to further enhance the text and data mining initiatives. Most importantly, it should support the emerging practice of digital scholarship that brings together research practitioners and research technologists, that encourages and facilitates cross-disciplinary work not only between arts and humanities scholars but also across other disciplines in the social, physical and medical sciences, that supports communities of practices formed around methods, disciplines, content, and theory, and that supports the development and sharing of tools and technologies.

The RI would comprise:

- Provision of universal working and shared collaborative spaces
 - A technical infrastructure based around Grid technologies and Virtual Research Environment technologies to support collaborative working and creative endeavour
 - Provision of physical meeting and research space dedicated to the advancement of digital scholarship
 - Administrative and research support for Communities of Practice who would undertake short to medium term activities in particular fields or subject areas of benefit to the Research Infrastructure, or existing networks of experts that wish to be affiliated to the RI. These activities would also seek to identify areas, tools, methods etc. for cross-fertilisation across different disciplines
 - A programme of expert seminars to share methods, theories and practice
- A programme of workshops to promote cross-disciplinary exchange of tools, methods and expertise
- Development of best practice guidelines, model solutions and case studies
- Research papers and conference presentations
- European wide training and awareness programme
 - Training programmes to promote the use of standards and best practice, particularly focused on the digitisation of cultural heritage information and objects, and creating sustainable resources and websites
 - Training programmes in digital scholarship
 - Training and awareness raising programme to encourage sharing expertise, methods, content and tools
- Tools development and registry services
 - Assessment of tools requirements and a related programme for the development of open source tools
 - A tools registry providing information, advice and access (to open source tools) to tools for digitisation, organisation, interpretation and analysis
- Development of multi-lingual thematic ontologies that would support and enhance research
- Usage analysis services that would use data mining techniques to analyse patterns of usage which services and organisation would use to improve and prioritise services
- Development of new business and cost models appropriate for a European wide environment of shared services and shared content

The diagram below provides a conceptual model of how the RI would operate and integrate the various component parts. It would bring together content, digitisation, curation and preservation, and dissemination and publication with communities of practice, tools, and interpretation and analysis embedded within research practice.



5.7.4 DARIAH IMPLEMENTATION AND COST

5.7.4.1 Goals

The 'grand vision' for the DARIAH Research Infrastructure is to facilitate long-term access and use to all European humanities and cultural heritage information for the purposes of enhancing and expanding research, thereby increasing our knowledge and understanding of our histories, heritage, languages and cultures.

To achieve this vision we have the following goals:

- To build capacity for digitising European humanities and cultural heritage information to quality standards and following best practice
- To build capacity for the expert curation and preservation of European humanities and cultural heritage information
- To build capacity for the dissemination, presentation and publication of European humanities and cultural heritage information and, wherever possible, the research outputs and knowledge based upon the interpretation and analysis of that knowledge
- To facilitate the use, interpretation and analysis of digital humanities and cultural heritage information
- To increase e-humanities capacity by developing, promoting and supporting digital scholarship and the application of advanced ICT methods

- To support the cross-fertilisation of ideas, methods and expertise, transfer of competence from one domain to another, and use of best practice and standards that will ensure interoperability across collections of information
- To leverage national, regional, and institutional infrastructure investment across Europe for the benefit of humanities and cultural heritage research
- To facilitate the development and sharing of expertise, tools, and ICT methods for the creation, curation, preservation, access and dissemination, and use of humanities and cultural heritage digital information
- To promote synergies across Europe, coordinate activities, encourage mutual support, and lead the development of common policies and technology standards

5.7.4.2 Partners

The Infrastructure proposed here will be based upon an existing network of Data Centres and Services based in Germany, France, the Netherlands and the United Kingdom. Each of the participating partners has a strong track record within their own country and is mature in terms of their organisational structure and level of activity. In addition, the partners have worked to develop a framework for collaboration across the Centres on a pan-European level and it is with this in mind that an initial infrastructure based around these four organisations is proposed.

The Arts and Humanities Data Service (AHDS) is a UK national service funded by the Arts and Humani-

ties Research Council and the Joint Information Systems Committee to collect, preserve and promote the use of digital information resources which result from or support research and teaching in the arts and humanities. By preserving this digital information the AHDS encourages widespread research and educational use of this material both nationally and internationally. The AHDS is also active in identifying and promoting the use of shared standards and developing and promoting an integrated approach to resource creation, discovery, access and use using relevant technologies. The AHDS seeks the widest possible collaboration across its range of activities and acts to establish fruitful partnerships nationally, across Europe and internationally. Within the UK it works in partnership with the AHRC ICT Methods Network which provides a national forum for the exchange and dissemination of expertise in the use of Information and Communication Technologies (ICT) for arts and humanities research.

The Centre National de la Recherche Scientifique (CNRS) is a publicly-funded research organisation that defines its mission as producing knowledge and making it available to society. CNRS service and research units are spread throughout the country and cover all fields of research. CNRS is active in all major scientific fields including the humanities. CNRS strives to develop collaboration between specialists from different fields of expertise. These interdisciplinary programs and actions offer a gateway into new domains of scientific investigation and enable CNRS to address the needs of society and industry. CNRS has established digital curation activities to manage and curate complex digital information.

DANS is the national organisation in the Netherlands responsible for storing and providing permanent access to research data from the humanities and social sciences. To this end DANS collaborates with researchers and encourages them to work in partnership with one another. DANS operates as a network, with a centre responsible for organising the data infrastructure. DANS works closely with partner organisations to share digital information, expertise and new technological developments.

The Max Planck Society for the Advancement of Science (MPG) is an independent non-profit organisation that supports and promotes research across a number of institutes. It encourages and supports cutting-edge research involving collaboration and cooperation across disciplines and nations, and has a strong record for facilitating pan-European research activity. The Max Planck Digital Library has recently been approved by the Senate of the MPG with a significant investment in the management of digital research information over the coming years.

5.7.5 COMMUNITIES

The RI will work with or seek to include in DARIAH a number of stake-holders across Europe including:

- National and regional data services
- Content owners and suppliers
- Digitisation services
- Research practitioners, subject communities, professional societies
- Standards communities
- Digital libraries, archives, and support services
- Funding agencies

5.7.5.1 Organisation and Structure

The present situation of research infrastructures in the humanities is such, that in a number of European countries national (or sometimes thematic international) organisations exist that have an important role in the organisation of the national or thematic digital resource infrastructure. The AHDS in the UK, CNRS in France, DANS in the Netherlands and MPG in Germany are strong actors in this area. Each organisation is responsible for a number of subject centres, digital resource centres, data archives or institutes that play a vital role in providing long term access to research data in a particular field. Between them, they have enormous experience and expertise in content creation, standards and best practice, curation and preservation, access and use of a wide range of digital content.

The RI aims to empower and design new facilities for pioneering research of an international and interdisciplinary nature. Its organising principle is that of a decentralised network with a strong core. Initially, the core will be formed from bringing together four existing expert centres - AHDS, CNRS, DANS, MPG - that already provide advanced services at the national level - to leverage their services and expertise across the EU. These four will be supplemented by the establishment of a coordinating body located at one or more of the existing centres. This core network will bear responsibility for organising and supporting the network, for the basic infrastructure and initial set of services, and for the method and means of communication. In addition, the core network will develop partnerships with digitisation centres, and other providers of services, content, and advice to add to its portfolio.

The decentralised network will bring in specific thematic or disciplinary expertise under the heading of thematic groups. Thematic Group network members will be prominent institutes, organisations and research networks with a leading role within the European context. The model is an open one and will be able to embrace new, promising fields that are as yet unable to play such a leading role in Europe.

The consortium proposes a three-tiered structure as its organisational model, which is to be flexible and responsive to future developments, with a particular remit to expand to include more partners at the national and thematic levels. The model is based on reciprocal relationships and subsidiarity. The elements of DARIAH are:

- Core Coordinating Body
- Core Network Partners
- Thematic Groups (subject or topic related)

5.7.6 FUNCTIONS

Core Network Coordinating Body: This is a new body that would best be located at one or more of the consortium partner institutions. It will be established in the preparatory phase.

The functions of the coordinating body include:

- Management of DARIAH
- Enabling, coordinating, funding
- Collating and promoting best practice and standards
- Coordinating legal advice and model licences and contracts
- Coordinating and developing training and awareness programmes
- Harvesting, harmonisation and combination of digital resources
- Overseeing development of EU-wide technical architecture
- Ontology and metadata development
- Tools registry
- Monitoring and certification

Core Network built upon National Services and Organisations: Initially the network will comprise the four consortium partners - AHDS, CNRS, DANS, and MPG. Over time we expect the network to expand to include at least an additional 10 partners from across Europe. Additional partners will be required to have a firm foundation at a national, regional or organisation level, with an appropriate level of funding, and to be able to make a significant contribution to the RI.

The functions of the core network include:

- Working with the Coordinating body - enabling, coordinating, and stimulating best practices and standards
- Advice and guidance to Thematic Groups
- Services for Thematic Groups
- Digitisation services
- Hosting services
- Curation and preservation
- Access and delivery (publication and presentation)

- Interoperability, aggregation and harmonisation
- Grid-enabling selected collections
- User Portals

Thematic (domain) Groups: Initially we expect to work with 3 already established Thematic Groups in the domains of History, Languages, and Practice-based Art. In the operational phase we will encourage other thematic groups to join the RI. Thematic groups will be expected to collaborate with the RI and to adhere to standards and best practice agreed with the RI.

The functions of the Thematic Groups include:

- Building communities of practice and other subject coalitions
- Research and digitization projects
- Analysis and interpretation of digital information
- Publication and presentation of digital information
- User portals
- Sharing methods, expertise and tools
- Building research capacity and digital scholarship across Europe

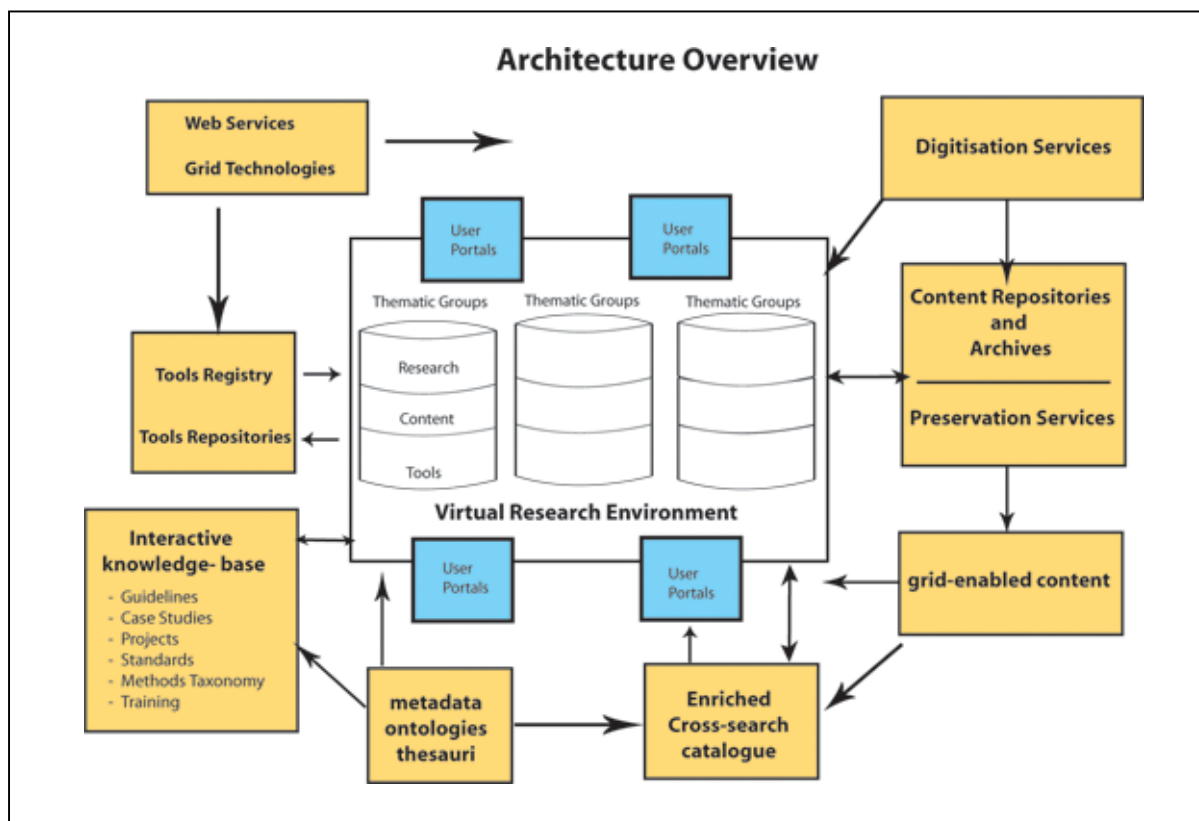
Once operational, the RI will expand, grow, and consolidate into a full-fledged research infrastructure for the humanities, with a Central Coordinating Institute at the centre, surrounded by a network of domain and national centres, and which collaborate and contribute to standard and subject groups. We also envisage that other local groups, centres and projects may participate for a limited period with the RI. The DARIAH research infrastructure also has a key role to play collaborating and sharing expertise with similar structures that may exist for the social sciences and the science.

5.7.7 ARCHITECTURE OVERVIEW

The diagram below provides an overview of how the RI will integrate the component parts and where primary responsibility lies for the different elements:

- Yellow elements will be the primary responsibility of the coordinating centre and the core network and their partners
- White elements will be the primary responsibility of the Thematic Groups and their partners
- Blue will be jointly developed between the network partners and thematic groups

Management of the whole and ensuring interaction and sharing of content, tools, methods, expertise etc. will be the primary responsibility of the coordinating centre.



Metadata, ontologies and thesauri will enable and support the creation, discovery and use of digital information and tools.

Grid-enabled content will ensure access and use of large volumes of complex materials, and aid the harmonisation and aggregation of dispersed content

The Tools registry will provide access to **tools repositories** that contain tools for use by researchers, curators of digital content, and digitisers of digital content.

User Portals will provide integrated access to harmonised collections and tools, together with authentication mechanisms, thus allowing further interpretation and analysis of RI content. It is envisaged that users of the portal environments will be able to contribute their own content, tools, annotations etc.

Virtual Research Environments (such as Sakai) will help support distance researchers and enable sharing of research processes, research resources, tools, methods, and presentation and publication of results. It is envisaged the Thematic Groups will also work within such environments.

The Interactive Knowledge Base will provide an on-line source of rich information on standards, guidelines, projects, methods, case studies, training opportunities, and tools. Users will be able to rank training and tools, comment on methods and, via

moderation, add their own suggestions. A version of the Knowledge Base already exists for the UK, developed and managed by the AHDS. We propose that the core network extend this for other EU countries.

5.7.8 IMPLEMENTATION AND FINANCE

A key aim of DARIAH is to assist, through the provision of training and advice and guidance, the development of humanities research infrastructure capacity across Europe, and the organisation and implementation of DARIAH reflects this aim. We envisage the Coordinating Centre remaining throughout the life of DARIAH to manage and coordinate the activities of the Network. Alongside this, we envisage both the Core Network Partners, and the Thematic Groups expanding significantly over the lifetime of DARIAH. It is this mix of strong national, regional and thematic centres of excellence working within a pan-European collaborative framework that will support excellence in research across the humanities and arts in Europe.

5.7.8.1 Preparatory Phase: Year 1

Cost: 2m Euro

The preparatory stage is intended to set up the physical and human elements of the Research Infrastructure, to establish the Coordinating Centre, and to ensure it is on a firm legal and financial footing.

- Establish Coordinating Centre
- Secure physical space at Core Network sites

- Recruit staff at centre and core sites
- Finalise contract between partners
- Project planning for the Construction Phase
- Establish web presence and Research Infrastructure design
- Purchase initial equipment
- Establish communication methods between partners
- Fund travel between partners
- Identify and open negotiations with initial set of Thematic Groups

5.7.8.2 Construction Phase: Years 2 - 3 Cost: 8m Euro

The construction stage will build **additional pan-European capacity** at the core network sites and **initiate** work with the Thematic Groups and new National Partners. Slovenia has expressed a wish to collaborate with DARIAH with the aim of becoming a core network partner, and we will seek an additional partner during the construction phase. DARIAH has also had discussion with the following Thematic Groups - Languages (CLARIN), History (DISH), Practice Arts (MARCEL), and Archaeology to include them in the DARIAH network. CLARIN has a vital role to play in working with DARIAH on multi-lingual issues, and the development of ontologies. Both MARCEL and the Archaeology Network have vital role to play in the development of technologies to support use of GRID technologies (in particular ACCESS Grid), and the archiving of complex digital content.

The implementation process will be relatively quick as DARIAH will take advantage of the maturity of the partner sites to extend and enhance their existing services and expertise to operate at a pan-European level. The Coordinating Centre site will take a management and coordinating role and will also take responsibility for key activities that are better done at a Central level. Wherever possible services and activities will take advantage of existing activities, and seek to develop collaborative partnerships. A key function of the Central site will be to identify existing services and activities (for example, legal advisory services) and to develop partnership agreements. This phase will also include a series of agenda setting workshops with the purpose of engaging research communities across Europe to identify their needs and requirements from the RI. These workshops will ensure that the RI is driven by and meets user needs.

Core Network (CN) Sites:

- Build Access-Grid-enabled i-Lab research facilities and technical infrastructure for research collaborations and universal working
- Grid-enable selected content
- Build content hosting service infrastructure (layered onto existing service)

- Build preservation service infrastructure (layered onto existing service)
- Contribute to standards and best practice guidelines (revise existing, create new)
- Contribute to advisory and guidance service
- Digitisation
- IPR and copyright
- Preservation
- Standards
- Extend knowledge base (exists for the UK at AHDS)
- Contribute to legal advisory service
- With Centre, establish legal agreements, licences, and contracts for hosting and preservation services
- Contribute to work with thematic groups
- Build web presence
- Start limited operational services
- Identify R&D programme for curation and preservation of digital humanities content

Coordinating Site (CC):

- Detailed project planning for Operational Phase
- With core sites, organise and hold agenda setting workshops
- Policy development
- Build web presence
- Develop multi-lingual approach
- Partnership development and agreements
- Establish and coordinate legal advisory service
- Establish tools registry
- Establish and coordinate advisory service
- With core sites and others develop training programmes
- Establish certification service for hosting, preservation and digitisation services
- Invite proposals from digitisation services for certification by RI
- Develop pan-European cross-search catalogue
- Collate existing standards and best practice guidelines and identify new requirements. Commission new guides and case studies.
- Establish communication, promotion and marketing strategy and implementation plan
- Invite proposals and coordinate activities with initial set of Thematic Groups - in the first instance these are likely to subject-based groups for History, Languages, and Archaeology, and standards groups such as the TEI consortium
- Identify additional thematic groups for subjects, standards and content
- With core sites and others identify and promote cross-disciplinary sharing of expertise, methods and tools, and set up workshop, seminar and training programmes

Thematic Groups:

- Detailed planning
- Fund-raising
- Identify key sources and tools
- Digitisation
- Preparation and contribution of content
- Contribute to guidelines, standards work etc.
- Contribute to development of first set of user portals

5.7.8.3 Operation Phase: Years 4 - 15 Cost: 40m Euro

The operations stage of the RI will seek to roll-out the services and activities outlined above, and to start to facilitate the building of additional capacity in other European countries at national, regional and organisation levels by providing training and advice and guidance. A key goal of the RI is to develop capacity across Europe and to facilitate sharing of expertise, methods, ideas and approaches, tools and technology for the creation, curation, preservation, access and use of humanities and cultural heritage digital content. Detailed planning for the operational stage of the RI will be undertaken during the con-

struction phase and will respond to the results of the agenda setting workshops.

- Full roll-out of services and activities at core sites
- Develop enhanced services through:
 - Multi-lingual search services
 - Ontology development and building
 - Aggregation services
 - Enrichment of metadata through data and text mining
 - Visual searching and browsing
 - Advanced training programmes implemented
 - Implement R&D programme
 - Other activities as identified during project planning
- Open call for Services to join the Core RI to build capacity across all EU countries
- Open call for thematic groups to join the RI

CN: Matching funding would be expected from national, regional, or other sources.

TG: Significant other funding would be expected.