Mochmann, Ekkehard

Working Paper

e-Science Infrastructure for the Social Sciences

RatSWD Working Paper, No. 115

Provided in Cooperation with:
German Data Forum (RatSWD)

Suggested Citation: Mochmann, Ekkehard (2009) : e-Science Infrastructure for the Social Sciences, RatSWD Working Paper, No. 115, Rat für Sozial- und Wirtschaftsdaten (RatSWD), Berlin

This Version is available at:
https://hdl.handle.net/10419/186221

German Council for Social
and Economic Data (RatSWD)

www.ratswd.de

RatSWD
Working Paper Series

Working Paper No. 115

e-Science Infrastructure for the Social Sciences

Ekkehard Mochmann

July 2009

Federal Ministry
of Education
and Research

Working Paper Series of the Council for Social and Economic Data (RatSWD)

The *RatSWD Working Papers* series was launched at the end of 2007. Since 2009, the series has been publishing exclusively conceptual and historical works dealing with the organization of the German statistical infrastructure and research infrastructure in the social, behavioral, and economic sciences. Papers that have appeared in the series deal primarily with the organization of Germany's official statistical system, government agency research, and academic research infrastructure, as well as directly with the work of the RatSWD. Papers addressing the aforementioned topics in other countries as well as supranational aspects are particularly welcome.

RatSWD Working Papers are non-exclusive, which means that there is nothing to prevent you from publishing your work in another venue as well: all papers can and should also appear in professionally, institutionally, and locally specialized journals. The *RatSWD Working Papers* are not available in bookstores but can be ordered online through the RatSWD.

In order to make the series more accessible to readers not fluent in German, the English section of the *RatSWD Working Papers* website presents only those papers published in English, while the the German section lists the complete contents of all issues in the series in chronological order.

Starting in 2009, some of the empirical research papers that originally appeared in the *RatSWD Working Papers* series will be published in the series *RatSWD Research Notes*.

The views expressed in the *RatSWD Working Papers* are exclusively the opinions of their authors and not those of the RatSWD.

e-Science Infrastructure for the Social Sciences

Ekkehard Mochmann

(E. Mochmann[at]web.de)

Abstract

When the term „e-Science“ became popular, it frequently was referred to as “enhanced science” or “electronic science”. More telling is the definition ‘e-Science is about global collaboration in key areas of science and the next generation of infrastructure that will enable it’ (Taylor, 2001). The question arises to what extent can the social sciences profit from recent developments in e- Science infrastructure?

While computing, storage and network capacities so far were sufficient to accommodate and access social science data bases, new capacities and technologies support new types of research, e.g. linking and analysing transactional or audio-visual data. Increasingly collaborative working by researchers in distributed networks is efficiently supported and new resources are available for e-learning. Whether these new developments become transformative or just helpful will very much depend on whether their full potential is recognized and creatively integrated into new research designs by theoretically innovative scientists.

Progress in e-Science was very much linked to the vision of the Grid as “a software infrastructure that enables flexible, secure, coordinated resource sharing among dynamic collections of individuals, institutions and resources’ and virtually unlimited computing capacities (Foster et al. 2000). In the Social Sciences there has been considerable progress in using modern IT- technologies for multilingual access to virtual distributed research databases across Europe and beyond (e.g. NESSTAR, CESSDA – Portal), data portals for access to statistical offices and for linking access to data, literature, project, expert and other data bases (e.g. Digital Libraries, VASCODA/SOWIPORT). Whether future developments will need GRID enabling of social science databases or can be further developed using WEB 2.0 support is currently an open question. The challenges here are seamless integration and interoperability of data bases, a requirement that is also stipulated

by internationalisation and trans-disciplinary research. This goes along with the need for standards and harmonisation of data and metadata.

Progress powered by e- infrastructure is, among others, dependent on regulatory frameworks and human capital well trained in both, data science and research methods. It is also dependent on sufficient critical mass of the institutional infrastructure to efficiently support a dynamic research community that wants to "take the lead without catching up".

Are advances in socio-economic research driven by data or technology? Claims and inspired deliberations pondering on this alternative are not new. As Norman Nie asserted without reservation “that all science is fundamentally data driven” (Nie 1989, 2) others argue “that progress in science rather depends on formal modelling” (Rockwell 1999, 157): More recently “methodological and substantive rigour” (http://www.europeansocialsurvey.org/) are emphasized as necessary preconditions to create reliable sources of knowledge about social change. Both, information technology and the social science data base, have developed remarkably over past decades, from poverty of data to a rapidly expanding production of all kinds of empirical evidence beyond survey and statistical microdata – now including e.g. electronic texts, event data bases, videos, geo-information and new kinds of data, e.g. transaction data (Lane 2009; Engel 2009) or biomarkers (Schnell 2009; Gampe 2009). Access to comprehensive databases and advanced data analysis increasingly allow modelling of complex social processes.

To efficiently support future empirical research 'The present major task is ... to create pan-European infrastructural systems that are needed by the social sciences ... to utilise the vast amount of data and information that already exist or should be generated in Europe. Today the social sciences ... are 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' (ESFRI Report 2006).

1. e-Science, e-Social Science, the Grid and Web 2.0

Though progress fuelled by new kinds of measurement, expanding data bases and technological support has already been observed over past decades, there are revolutionary new systematic approaches to analyse research challenges. Based on the results of these analyses they implement comprehensive technological infrastructures to facilitate innovative research. These “e-Science” approaches were initially referred to as enhanced science or electronic science. More telling is the definition “e-Science is about global collaboration in key areas of science and the next generation of infrastructure that will enable it” (Taylor 1999). Basically, e-Social Science is following these ideas, with emphasis on providing advanced IT services to “enable” social research. The National Centre for e-Social Science at Manchester (NCeSS) states: “e-Social Science is a term which encompasses technological developments and approaches within Social Science. We are working with Social Scientists and Computer Scientists on tools and research which Social Scientists can take and use to

help their research. These tools might either allow a Social Science researcher to conduct new research, or else, conduct research more quickly. These tools can be used across a variety of Social Science domains. “Within NCeSS, we refer to the 'e' in eSocial Science as 'enabling'.” (http://www.ncess.ac.uk/about_eSS/)

Progress in e-Science was very much linked to the vision of the Grid as “a software infrastructure that enables flexible, secure, coordinated resource sharing among dynamic collections of individuals, institutions and resources and virtually unlimited computing capacities” (Foster et al. 2000). As such, it was based on multi gigabit broad band width fibre cables connecting distributed and loosely coupled computing resources, using open standards in the Grid. In co-ordination with the National Research and Educational Networks (NRENs), they would provide a globe spanning net with virtually unlimited computing capacity, intelligent middleware to support interoperability of network services and control of access and authentication. To support information handling and support for knowledge processing within the e-scientific process, the future developments point towards the Semantic GRID (De Roure et al.9.

The Enabling Grids for E-SciencE (EGEE) project is a prominent globe spanning example to build a secure, reliable and robust Grid infrastructure with a light-weight middleware solution intended to be used by many different scientific disciplines. It is built on the EU Research Network GÉANT and exploits Grid expertise generated by many EU, national and international Grid projects, including the EU Data Grid (http://eu-datagrid.web.cern.ch/eu-datagrid/.) Just to show the new dimensions: At present, it consists of approximately 300 sites in 50 countries and gives its 10,000 user's access to 80,000 CPU cores 24 hours a day, 7 days a week. This project came to the conclusion that the state of computer and networking technology today facilitates extensive computing grids that integrate geographically distributed computer clusters, instruments, scientific communities and large data storage facilities. The resulting benefits include a large increase in the peak capacity, the total computing available and data management power for various scientific projects, in a secure environment“ (http://www.eu-egee.org). Critics, however point to the fact that these new developments can not be used outside high energy physics, so far.

The Grid idea followed the computer scientists´ blueprint for a perfectly designed distributed infrastructure. Lessons learnt from early developments emphasize that it is very important to have application scientists collaborate closely with computer scientists. “Successful projects were mostly application and user driven, with a focus on the development of standard and commodity components, open source, and results easy to

understand and to use" (Gentzsch 2007).

It is remarkable that the German Grid initiative (http://www.d-grid.de/), which started in 2005 with six science projects, now also includes Text Grid (http://www.textgrid.de/) from the Humanities and none from the Social Sciences. Over the past few years more than 10 new projects from the sciences were added. In this field the Social sciences belong certainly not to the early adopters. This pattern can also be observed in most other countries, apart from UK and USA, where the social science communities made particular efforts to boost their e-infrastructure. In this context it may be noteworthy that the first attempt to support retrieval of data by machine was actually conceived by a social science project described in 1964 already (Scheuch, Stone, Harvard 1964) and ideas for the researchers dialogue with interactive data analysis and retrieval systems date to 1972 (Scheuch and Mochmann 1972, 154f). With respect to trans-national data infrastructure the Council of European Social Science Data Archives (CESSDA) is studying the feasibility of Grid enabling. This activity investigates current developments and applications in Grid technologies in order to find efficient and sustainable ways for the implementation of a cyberinfrastructure for the Social Sciences and Humanities and to identify the issues for implementing Grid technology.

Instead of an enthusiastic uptake of Grid technologies, a number of initiatives followed a bottom up approach in collaborative systems development, e. g. for access to virtually distributed data bases using the World Wide Web in a more sophisticated way. These new trends in the use of WWW technology to enhance collaboration as well as information and data sharing are referred to as Web 2.0 technologies. They are still based on the so far known World Wide Web specifications. Results of these developments are possibly less perfect than those designed for GRID applications, but they are facilitated by cooperative approaches within the science community and they take usually much less time to implement.

2. Social Research Infrastructure, e-Infrastructure, Cyberinfrastructure

As Social Sciences had a long record of infrastructure development in terms of service institutions, databases, data laboratories and researcher networks in the field of international comparative research (Scheuch 2003). Thus it was no surprise that the social scientists pointed to the need to distinguish their already existing infrastructure from the emerging IT based infrastructure (Serenate Report 2003). The e-Infrastructure concept was then proposed in 2003 to coin a term for the development of the next generation of trans-national ICT research infrastructure in Europe: "e-Infrastructure refers to this new research environment in

which all researchers – whether working in the context of their home institutions or in national or multinational scientific initiatives – have shared access to unique or distributed scientific facilities (including data, instruments, computing and communications), regardless of their type and location in the world." (Building the e-Infrastructure. Computer and network infrastructures for research and education in Europe).

At the same time, the National Science Foundation Blue-Ribbon Advisory Panel identified similar objectives for what they called "Cyberinfrastructure" (Atkins et al. 2003, 12): "We envision an environment in which raw data and recent results are easily shared, not just within a research group or institution but also between scientific disciplines and locations. There is an exciting opportunity to share insights, software, and knowledge, to reduce wasteful re-creation and repetition. Key applications and software that are used to analyze and simulate phenomena in one field can be utilized broadly. This will only take place if all share standards and underlying technical infrastructures." Cyberinfrastructure is defined in relation to known infrastructures so far: "Although good infrastructure is often taken for granted and noticed only when it stops functioning, it is among the most complex and expensive things that society creates. The newer term cyber-infrastructure refers to infrastructure based upon distributed computer, information and communication technology. If infrastructure is required for an industrial economy, then we could say that cyberinfrastructure is required for a knowledge economy" (Atkins et al. 2003, 5).

In Europe, the provision of network services to research and education is organised at three levels: the Local Area Network to which the end-user is connected, the national infrastructure

provided by the National Research and Education Network (NREN) and the pan-European level provided by GÉANT.

GÉANT currently interconnects the national research and education networks (NRENs) from all over Europe, including Russia. In terms of geographical coverage, technology used and services made available, GÉANT considers itself the number one research network in the world, which attracts requests for interconnection from all over the world. Under the GÉANT2 project it has grown to include more than 100 partners already. This is much more than the Social Sciences need so far, but it gains importance when we think about International Data Federations to support continuous global comparative and trans-disciplinary research. While the technical back bone network is in place many application tools, standards and content with rich metadata have to be developed in order to make full use of these technologies.

3. Data infrastructure needs of the Social Sciences

(Major results of the SERENATE project and the AVROSS study)

Exciting visions of the future potentials of new technologies like to travel with appealing descriptions of actual implementations in working environments. Closer examinations frequent-ly show that services, which are actually needed by end-users on a continuous basis are often far from satisfactory. Economic potential to implement new technologies, the level of expertise in societies to support these technologies and to adjust them to the specific needs of their user communities, as well as data management and methodological skills vary from country to country.

Needs, challenges and obstacles in relation to these new technologies have been analysed by the **S**tudy into **E**uropean **R**esearch and **E**ducation **N**etworking **a**s **T**argeted by ***e***Europe (Serenate). Security features were highlighted by a large number of the respondents dealing with sensitive data or even medical images. Another requirement is for mobile access to the network services including both home access for researchers, particularly for non-laboratory-based research such as humanities and social sciences, and also access when on visits to other countries. As a consequence of these usage patterns, the deployment of "Authentication, Authorization and Accounting" (AAA) services across the various networks was stipulated to give the necessary controls on access. The report from the final workshop also noted that access is possible to a rich variety of data from many sources and identified the potential for software to support collaborative working, sharing of data bases and data integration at many levels. Finally, the networks offer the means to include the "future generation of scientists in

schools" (Serenate Report 2003, 14).

In the spirit of e-Science approaches to systematically examine options and challenges for enhancing scientific research, Serenate includes some tough observations on contextual requirements into its findings: “We have learned that many people – national and European politicians, ministries and agencies in the national governments, the European Commission, telecoms vendors, equipment vendors, various service suppliers, local and regional authorities, universities and user communities all have to be mobilised, and to move in the same direction, if we are to make progress. If we do not make plans to maintain and even improve the situation over the next 5-10 years, then the sustained pace of technical, organisational and political change will inevitably lead to rapid decay” (Serenate Report 2003).

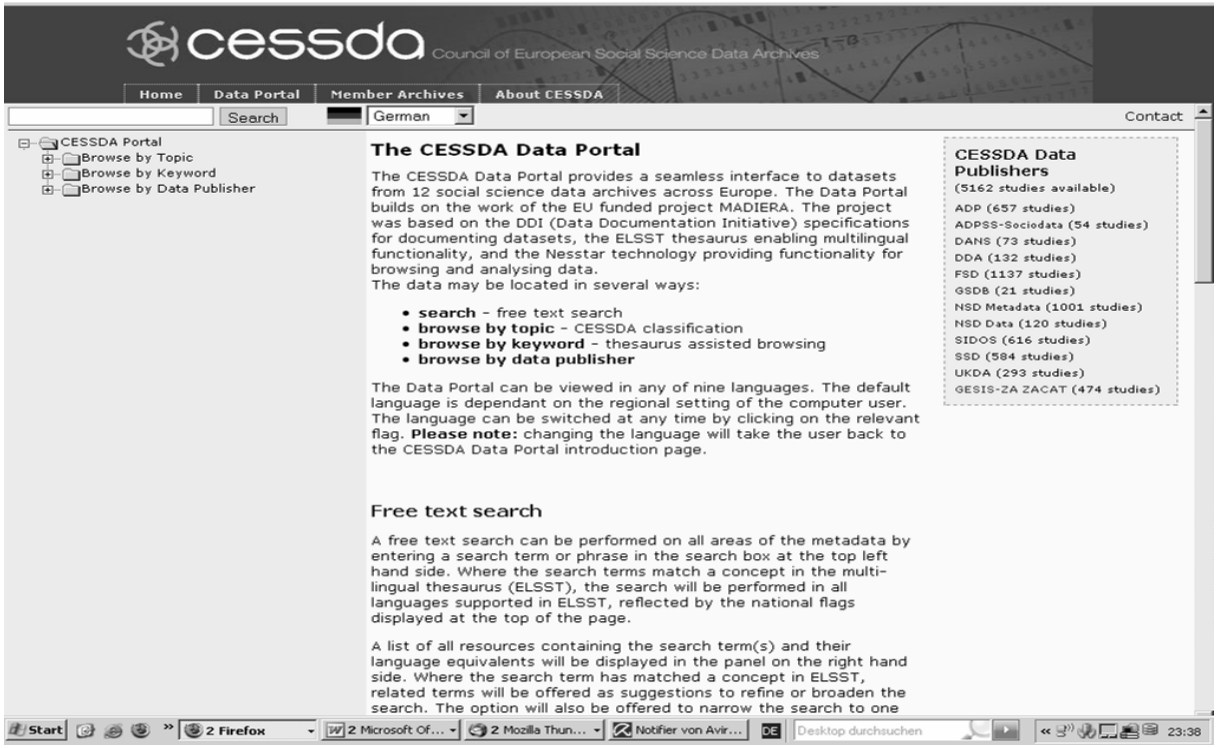
Analyses based on the AVROSS Study concluded that US and the UK’s efforts appear as an exception as no other European country has adopted an initiative that promotes e-Infra-structure uptake by the social science or humanities. At the same time, the European Strategy Forum on Research Infrastructures, ESFRI has recognized the importance of including these domains of science in the ESFRI Roadmap report. This foundational report identified three long-term strategic goals for SSH research infrastructures: comparative data and modelling, data integration and language tools, and coordination (European Strategy Forum on Research Infrastructures 2006). These aims create a potential for researchers in SSH who want to develop or use e-Infrastructure.

4. Status quo and best practice examples from Social Sciences

Predominantly Social Scientists do not see a particular need to use the Grid technology for e-Social Science developments, as most of their data and computation needs could be handled by the existing Internet capacities. Numerous Internet solutions for access to specific collections, even with local AAA procedures were employed. While many of them provide sufficient user support for their constituency, interoperability of databases and metadata (see Metadata chapter in this book) as well as world wide networked access are rarely possible. There are, however, a few remarkable examples for trans-national data access in virtually distributed data bases.

Building on extensive experience in international data transfer, the Council of European Social Data Archives worked towards networked solutions that ideally would allow interested researchers to access the holdings of member archives from any point in the world. This is

operational now as the CESSDA Portal, providing seamless access to datasets from currently 12 social science data archives across Europe (http://www.cessda.org/). Among others, it includes prominent reference studies from international comparative research, like European Social Survey, Eurobarometers, International Social Survey Programme and the European Values Studies (http://www.cessda.org/accessing/catalogue/). The Data Portal builds on the work of the EU funded MADIERA project (http://www.madiera.net/). All content is based on the DDI (Data Documentation Initiative) specifications for documenting datasets including relevant metadata (http://www.ddialliance.org/org/). Multilingual functionality is supported by the ELSST thesaurus and the Nesstar technology provides functionality to the user for browsing and analysing data (http://www.nesstar.com/). The software consists of tools which enables data providers to disseminate their data on the Web. Nesstar handles survey data and multidimensional tables as well as text resources.



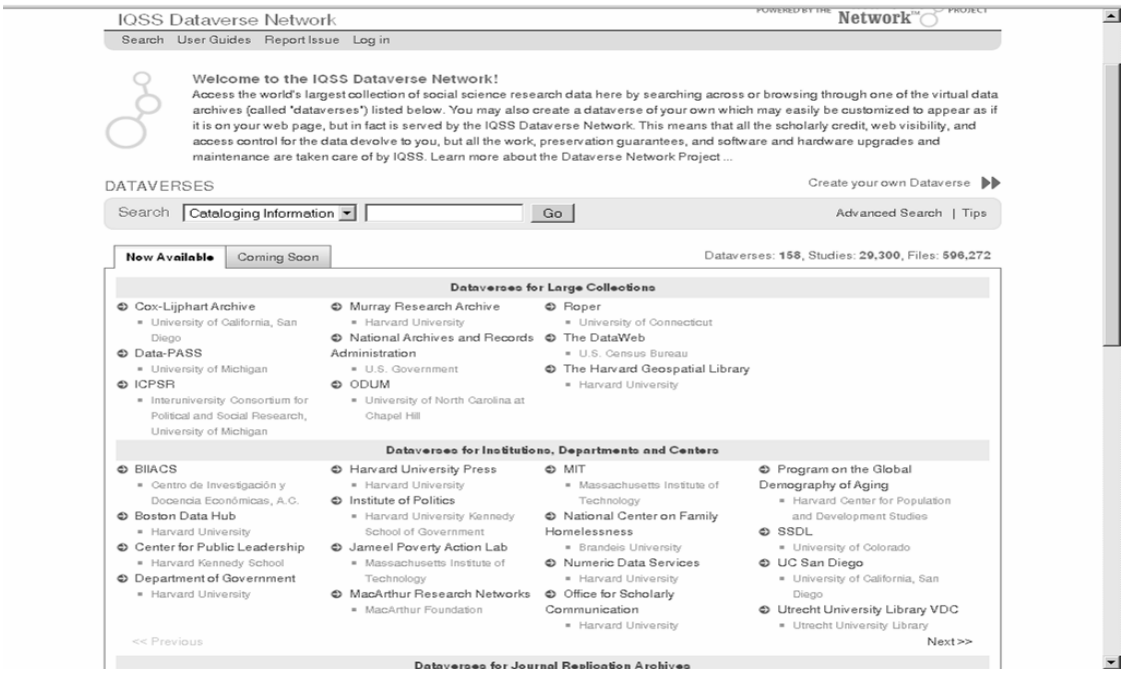
As a recent user survey conducted by the Institute for Social Research (ISR, Michigan Ann Arbor) in co-operation with GESIS under the auspices of the European Science Foundation with more than 2000 users show, there is a high level of satisfaction with these new technologies. These are efficiently supporting simultaneous data access to thousands of studies in a virtual distributed network, frequently including the option to check the measurement instrument, methodological and technical background information and then to proceed to data analysis in the same session. As a precondition to take advantage of this

functionality on the output side there are non trivial investments on the input side. To close the knowledge gap between principle investigators, who designed the study and followed the steps through fieldwork and data management up to the analysis ready files a lot of methodological and technical details covering the research process so far have to be communicated to enable further informed analysis (MetaDater project http://www.metadater.org/9.

A frequently stipulated development is the integration of data, literature, project documentation and expert data bases. One development in this direction is SOWIPORT, which includes among others references to social science literature and data resources offered by different providers (http://www.sowiport.de). The Dutch Data Archiving and Networked Services (DANS) have started to store data for long term preservation and access in the Grid (http://www.dans.knaw.nl/en).

There are several other technological developments that are successfully applied to social science data service for larger international user communities, e.g. the Data Service for the European Social Survey (EES http://www.europeansocialsurvey.de/), the ZACAT-Data Portal of GESIS providing access to most of the continuous international survey programmes (http://zacat.gesis.org), the JDSystems Survey Explorer (http://www.jdcomunicacion.com/ISSPSpain.asp) or SDA: Survey Data and Analysis, a set of programs for the documentation and Web-based analysis of survey data (http://sda.berkeley.edu/) which includes e.g. the General Social Survey (GSS) and the American National Election Study (ANES).

A more recent development, is the IQSS Dataverse Network supported by the Institute for Quantitative Social Science of Harvard University (http://dvn.iq.harvard.edu/dvn/). “The Dataverse projected aimed to solve some of the political and sociological problems of data sharing via technological means, with the result intended to benefit both the scientific community and the sometimes apparently contradictory goals of individual researchers” (King 2007, 1). Dataverse provides open source software to host Dataverse networks at larger institutes or to create individual “dataverses” as archives of individual owners that may be just for long term archiving and analysis, or for access by other user over the Internet. This way individual created data bases and trusted archives can be networked as the Networks homepage depicts (http://dvn.iq.harvard.edu/dvn/):



As software is only part of the solution, IQSS also provides citation standards for the content to be stored. The digital library services of each dataverse include data archiving, preservation formatting, cataloguing, data citation, searching, conversion, subsetting, online statistical analysis, and dissemination.

5. Conclusions

As we can observe already today: A comprehensive infrastructure based on advanced data communications, computing and information systems are extremely supportive for conducting high-quality research. They are indispensable for progress, which so far has been unlikely to be achieved in many fields of research. Outstanding examples are the mapping of the human genome and the discovery of new elementary particles, which were facilitated by advanced computational, data storage and network technologies. Being in touch with widely dispersed research communities, collaborative working and data access in globe-spanning comparative social survey programmes including 40+ countries are already receiving strong support from these new technologies. The rapidly growing social science data base, including methodologically controlled data bases and new kinds of data with related metadata increasingly leans to data linkages across topical domains. Thus modelling of complex social processes that may need collaboration in dispersed researcher networks in need of large scale data access and computation resources can be supported more efficiently than ever before.

One example for creating that kind of research environment is PIREDEU (http://www.piredeu.eu/), a Design Study - Providing an Infrastructure for Research on Electoral Democracy in the European Union), bringing together all kinds of empirical evidence from survey data, aggregate statistics to party manifestoes on European level, while the Comparative Study of Electoral Systems is taking a global approach (http://www.umich.edu/~cses/).

The technical backbone and the e-infrastructure for advanced Grid applications are in place and actually used by many international and national science communities. In principle and in practice there are technological solutions to provide researchers with computational resources on demand, the capability to share complex, heterogeneous and widely distributed data repositories, and the means to enable researchers to collaborate easily and effectively with colleagues around the world. These functionalities, which are available now, have been part of the e-Research Vision at the beginning of this Millenium. This indicates at what incredible speed these new technologies develop and are takenup in some disciplines.

By and large, the Social Sciences so far have opted for Web 2.0 solutions. The appeal of these Web 2.0 solutions lies in the ease of “ready to use applications”. So far, they seem powerful enough to support most data access and analysis needs in domains. This is currently not the case with sensitive micro data from statistical offices and with panel. Research is underway to include disclosure procedures into data access and analysis systems, which pose particular data protection problems. With increasing data availability and research crossing traditional disciplinary boundaries on global scale, new technologies for large scale data access and high speed computing may be required.

It is up to each scientific community to assess its specific needs and to decide at what speed it wants to move. Sometimes there are latecomer advantages in adopting new technologies, as many detours may be avoided (Schroeder et al. 2007). Nevertheless, it is obvious that a lot of ground laying work needs to be done. A combination of methodological and technical expertise is required to adopt or design and implement the new infrastructures. As has been emphasized in almost all prominent studies quoted, the combination of experts from the social research community working closely with IT specialists is required. Practical experiences from many international projects proof, however, that it is difficult to find the required expertise for limited project lifetimes and that it is even more difficult to keep the additional expertise acquired during the project accessible for further research and development. So, needs assessments, user community studies and capacity building at the interface of social research methodology and computer science are a prerequisite for viable

and sustainable developments. It may be a healthy step to combine future research methodology curricula with modules of what might be called “data science”, which is about data structures, data management, access and interoperability of data bases.

The Open Access Initiatives (e.g. the Berlin Declaration 2003) and the OECD declaration on open access to publicly financed data (OECD 2004) are certainly supportive in creating a culture of data sharing and easing access to information and data, including metadata. The challenges and development needs in e-infrastructure are beyond what a normal research institute can afford to invest in order to keep up on its own with the developments and to cover its long term needs. Forming of Alliances or multilateral institutional co-operation have been solutions of academic self organisation so far. The National Center for e Social Science in UK is an example to create a competence centre designed to serve the social science community in this respect.

Whether future developments will need GRID enabling of social science databases or can be further developed using WEB 2.0 support is currently an open question. The challenges here are seamless integration and interoperability of data bases, a requirement that is also stipulated by internationalisation and trans-disciplinary research.

Progress in e-infrastructure is also dependent on regulatory frameworks (Hahlen 2009) and data policies (e.g. NERC data Policy 2002). Best technical solutions may provide some routines and intelligent algorism to control access to sensitive data. International access, which is technically possible, can be out of question if statistical confidentially or Statistics law prohibit outside use. Last not least, the organisational infrastructure requires sufficient critical mass in terms of expertise, networking capacities and sustainable resources to efficiently support a research community that wants to “take the lead without catching up”.

6. Recommendations

The current stocktaking of socio economic data bases does show again, that impressive amounts of data are available in many fields of research. It would not be surprising, however, that the data base as such is rather scattered, not well integrated and does not lean easily to intra-national or international comparative research or even the combination of different sources for analyses with trans-disciplinary perspective. Apart from harmonizing data on the measurement level non trivial investments would be required to get data bases organised and to get the metadata in place.

Predominantly Social Scientists do not see a particular need to use the Grid technology for e- Social Science developments, as most of their data and computation needs could be handled by the existing Internet capacities. Numerous Internet solutions for access to specific collections, even with local AAA procedures were employed. While many of them provide sufficient user support for their constituency, interoperability of databases and metadata (see Metadata chapter in this book) as well as world wide networked access are rarely possible. There are, however, a few remarkable examples for trans-national data access in virtually distributed data bases.

1) Data Policy and strategic plans for research data management

Some scientific communities have formulated comprehensive **strategic plans** or even published explicit **data policies**. It might be a good starting point to assess needs in international context and to identify challenges, drivers and the blockages for the development of a future German e-infrastructure for the Social Sciences, that would also provide interfaces to and interoperability with leading international networks.

2) Needs Assessment and Framework Conditions

Like other countries, Germany has the technical infrastructure for modern data services in place. Whether there is the **need** and whether the **regulatory framework** conditions do allow to set up an integrated **German Data Net** has to be assessed. This could best be done by a **working group** that includes experts on methodological, legal and technical issues.

3) Standards on measurement – and metadata- level

Good documentation is one decisive factor for the potential of future data analyses. The Association of German Market Researchers (ADM), the Association of Social Science Institutes (ASI e.V.) and the Federal Statistical Office have agreed on **minimal standards for demographic variables** (Standarddemographie) long ago already to allow for better comparability of measurements across the three sectors. Likewise there are standards for metadata that would allow easier identification of and access to data that is related to the concepts central to the respective research questions. It might be advantageous to follow one meta- data standard, but this is not absolutely required. Nevertheless, to follow at least some metadata standard is a precondition to allow for interoperability at a later stage. DDI is being used by several institutes in Germany already. Working towards wider consensus on adopting metadata standards and agreeing on interfaces is one milestone to the infrastructure highway.

4) Best practice in data management and documentation

Efficient data base management will require close co-operation of researcher networks and data services**.** Best practice has to be communicated to implement metadata capture at the point of data collection already and to cover the whole life cycle from research design via data collection to publication and reuse.

5) Capacity building

Training of researchers in best practice of supplying all relevant information from the research process (cf. the **OAIS model**) and training of data professionals should be oriented towards what could be named "data science" in future curricula. The substantial investments in sound data bases need to be based on best methodological, data management and IT expertise. This is hard to find on the labour market in this combination and equally difficult to combine in research teams, simply because there is a serious lack of professionally trained people in this field. Data management, documentation and access could become one module "**data science**" in studies of social research methods. There is a huge market in demand of these skills –such as social and market research, insurance companies, media centres and media archives, data providers etc.

6) Research funding should also cover data management

It is not always easy to assess the relevance of data for future needs. Nevertheless, a vast uninspired "omnium gatherum" should be avoided. At least reference studies and data collections that lean to comparability over time or space should be properly documented for further use. This is a non trivial and labour intensive phase in the research process.

Frequently the data management to create high quality data bases requires a lot of methodological and technical expertise. This should be acknowledged by funding authorities and evaluation committees, which tend to honour the analyses but not the investment in preparing the data for it. So future funding of data collection should include **a line on data management and documentation.** Likewise **evaluation criteria** should also include whether data bases have been created following **methodological and technical best practice**.

7) Technical developments

Whether current institution specific data portals, remote access to individual data bases, product catalogues in integrated literature and **data portals** like SOWIPORT or networked solutions with **central data repositories**, e.g. the **DRIVER development on global level**, or

even **Data Grid solutions** are the needs of the future has to be assessed with a mid term and a long term perspective.

8) e- Infrastructure Competence Center for the Social Sciences

The Open Access Initiatives (e.g. the Berlin Declaration 2003) and the OECD declaration on open access to publicly financed data (OECD 2004) are certainly supportive in creating a culture of data sharing and easing access to information and data, including metadata. The challenges and development needs in e-infrastructure are beyond what a normal research institute can afford to invest in order to keep up on its own with the developments and to cover its long term needs. Forming of Alliances or multilateral institutional co-operation have been solutions of academic self organisation so far. The National Center for e Social Science in UK is an example to create a competence centre designed to serve the social science community in this respect.

References:

Atkins, D.E. et al. (2003): Revolutionizing Science and Engineering through Cyberinfrastructure. Report of the National Science Foundation. Blue-Ribbon Advisory Panel on Cyberinfrastructure. January 2003, 5 and 12 .

Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities, Max-Planck-Gesellschaft, Berlin 2003.

Berman, F. and Brady, H. (2005): Final Report: NSF SBE-CISE Workshop on Cyberinfrastructure and the Social Sciences.

Building the e-Infrastructure. Computer and network infrastructures for research and eduction in Europe.

De Roure, D./Jennings, N.R. and Shadbolt, N.R.(): The Semantic Grid: A Future e-Science Infrastructure.

Engel, B. (2009): Commercial transaction surveys. (In this volume).

ESFRI - European Strategy Forum on Research Infrastructures (2006): European Roadmap for Research Infrastructures. Report 2006, p. …

Foster, I. and Kesselmann, C. (1999): The Grid: Blueprint for a New Computing Infrastructure, Morgan Kaufmann Publisher Inc.

Foster, I./Kesselmann, C. and Tuecke, S. (2000): The Anatomy of the Grid.

Gampe, J. (2009): Biomarkers II. (In this volume).

Hahlen, J. (2009): Creating a sound institutional foundation for the informational infrastructure. (In this volume).

Hey, T. (2008): The Corporate Vice President for Technical Computing, Microsoft: "eScience, Semantic Computing and the Cloud: Towards a Smart Cyberinfrastructure for eScience" auf der 38. Jahrestagung der Gesellschaft für Informatik: Informatik 2008 – Beherrschbare Systeme – dank Informatik, 08.-13.09. München.

King, G. (2007): An Introduction to the Dataverse Network as an Infrastructure for Data Sharing. In: Sociological Methods and Research, Vol. 32, No. 2 (November, 2007), 173-199.

Lane, J. (2009): Administrative transaction data. (In this volume).

Meulemann, H. (Ed.) (2004): Erwin K. Scheuch. Infrastrukturen für die sozialwissenschaftliche Forschung. Gesammelte Aufsätze. Bonn.

Mochmann, E. (2001): Infrastruktur für die Komparative Sozialforschung in Europa. In: Hasebrink, U. and Matzen, Ch. (Eds.): Forschungsgegenstand Öffentliche Kommunikation. Funktionen, Aufgaben und Strukturen der Medienforschung. Symposien des Hans-Bredow-Instituts, Band 20. Baden-Baden/Hamburg, 161-173.

Natural Environmental Research Council (NERC) (2002): Data Policy Handbook Version 2.2, 2002.

Neuroth, H./Kerzel, M. and Gentzsch, W. (Eds.) (2007): Die D-Grid Initiative.

Nie, N. (1989): Model vs. data driven science and the role of the ICPSR in the progress of the social sciences. Adress delivered to the 27th Annual Conference of Official Representatives of the ICPSR. Ann Arbor, Michigan.

OECD (Ed.) (2004): Science, Technology and Innovation for the 21st Century. Meeting of the OECD Committee for Scientific and Technological Policy at Ministerial Level, 29.-30. January, Paris - Final Communique.

Rockwell, R.C. (1999): Data and statistics: empirical bases for the social sciences. In: Kazangicil, A. and Makinson, D. (Eds.): World Social Science Report, 156-166.

Scheuch, E.K. (2003): History and visions in the development of data services for the social sciences. In: International Social Science Journal 177, 385-399.

Scheuch, E.K. and Mochmann, E. (1972): Sozialforschung. In: Merten, P.(Ed.): Angewandte Informatik. Berlin/New York.

Schnell, R. (2009): Biomarkers I. (In this volume).

Schroeder, R./Den Besten, M. and Fry, J. (2007): Catching Up or Latecomer Advantage? Lessons from e-Research Strategies in Germany, in the UK and Beyond. Paper presented at the German e-Science Conference 2007. Baden-Baden.

Serenate Report on Final Workshop results IST-2001-34925. Study into European Research and Education Networking As Targeted by eEurope. SERENATE Deliverable no. D19. Authors: Vietsch, K. and Martin, J. (Eds.), Cavalli, V./Dyer, J./Robertson, D./Saugstrup, D./Scott, M. and Wood, Sh., Date: 27 December 2003, 14f.

Taylor, J. (2001): Presentation given at UK e-Science Meeting, London, July.

Tanenbaum, E. and Mochmann, E. (Eds.) (1994): Integrating the European database: Infrastructure services and the need for integration. In: International Social Science Journal, Vol. 46, 1994, Nr. 4 (142), 499-511.