Facing the Future:

European Research Infrastructures for the Humanities and Social Sciences

Adrian Duşa, Dietrich Nelle, Günter Stock,
and Gert G. Wagner (Eds.)

GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

Facing the Future:
European Research Infrastructures for the Humanities and Social Sciences

EDITORS:
Adrian Duşa (SCI-SWG), Dietrich Nelle (BMBF), Günter Stock (ALLEA),
and Gert. G. Wagner (RatSWD)

ISBN 978-3-944417-03-5

This book documents the results of the conference *Facing the Future: European Research Infrastructure for Humanities and Social Sciences* (November 21/22 2013, Berlin), initiated by the Social and Cultural Innovation Strategy Working Group of ESFRI (SCI-SWG) and the German Federal Ministry of Education and Research (BMBF), and hosted by the European Federation of Academies of Sciences and Humanities (ALLEA) and the German Data Forum (RatSWD).

Thanks and appreciation are due to all authors, speakers and participants of the conference, and all involved institutions, in particular the German Federal Ministry of Education and Research (BMBF). The ministry funded the conference and this subsequent publication as part of the Union of the German Academies of Sciences and Humanities' project "Survey and Analysis of Basic Humanities and Social Science Research at the Science Academies Related Research Institutes of Europe".

Editing: Dominik Adrian, Camilla Leathem, Thomas Runge, Simon Wolff

Layout and graphic design: Thomas Runge

Contents

Preface

The ESFRI process is of vital importance to the European research landscape. Research Infrastructures developed through European cooperation create opportunities for scientists to participate in the shared knowledge flows of the European Research Area, and thus translate the European idea into everyday life in an exemplary way. Jean Monnet, the father of European integration, once wrote that if he had to begin all over again with European unity, he would start with culture and not with the economy. With this in mind it is not surprising that programmes from among the Humanities and Social Sciences in particular such as SHARE, DARIAH, CLARIN, CESSDA and ESS, have been among the pioneers of the ESFRI process.

The BMBF together with the Strategic Working Group of ESFRI "Social and Cultural Innovation" supports the debate on the further development of the infrastructures in the Humanities and Social Sciences. We are grateful that ALLEA and the Berlin-Brandenburg Academy of Sciences and Humanities in collaboration with the German Data Forum (RatSWD) organized the conference "Facing the Future: European Research Infrastructure for the Humanities and Social Sciences" at the Federal Press Conference in Berlin.

The goal of this debate, in regard to the Humanities and Social Sciences, is to recognize changing infrastructure needs at an early stage, to identify priorities for innovative research activities based on European infrastructures, and to put these issues on the EU Roadmap of Research Infrastructures. The conference in Berlin provided a forum for this debate in a European context and thus contributed to strengthening the position of the Humanities and Social Sciences in the European Research Area – not only by highlighting their significance for the societal challenges in the new EU Research and Innovation Programme "Horizon 2020", but also by playing an active role in shaping ESFRI.

Dietrich Nelle (German Federal Ministry of Education and Research)

A Overview

1 Introduction

Günter Stock (ALLEA), Gert G. Wagner (RatSWD)

European societies are facing tremendous challenges in many fields such as health, migration, demographic change, social inclusion and cohesion, and severe environmental changes. Evidence-based and scientific analyses are a crucial requirement for reacting to these changes and for developing policy solutions. Excellent research environments enable innovative pan-European research – without them, real science and sustainable policies would be impossible. With the current paradigm shift in both the social sciences and the humanities towards data-intensive, cross-disciplinary and explorative scientific discovery, research infrastructures are an indispensable foundation for cutting-edge research in Europe.

Furthermore, Europe also has to avail of the grand opportunities that its cultural and intellectual diversity and richness offer. It is not just in times of crises that it is important to keep in mind that Europe is more than just an economic zone. The understanding of a common European identity is closely tied to the cultural heritage that this continent shares. The humanities investigate and preserve this heritage all over Europe. Currently, the opportunities to link the respective national cultural heritages are often limited by numerous and incompatible data standards. Thus, even the search for an intellectual and cultural Europe is highly dependent on research infrastructures that enable scholarly cooperation.

Horizon 2020, the latest EU funding programme for research and innovation, has recently been launched for the period 2014 to 2020. Since it now connects the fields of research and innovation, its financial resources have been significantly increased. For the first time, the programme includes "Europe in a changing world - Inclusive, innovative and reflective societies" as a major European challenge and thus explicitly codifies the social sciences and humanities (SSH) in the funding guidelines. A further cornerstone for the SSH domain will be the updated roadmap of the European Strategy Forum on Research Infrastructures (ESFRI) in 2016. These measures can be understood as an acknowledgement of the eminent role of SSH research. Scholars of the SSH are now being consulted, for example, as members of expert advisory groups, which in turn creates a certain responsibility that is to be assumed by the SSH research community.

The conference "Facing the Future – European Research Infrastructure for the Humanities and Social Sciences" in November 2013 in Berlin pursued exactly this objective: to strengthen the role of the humanities and social sciences in the future. Social science and humanities experts, representatives from research policy and funding agencies, and research infrastructure coordinators from 19 European countries met in Berlin to discuss the future of the ESFRI process, to identify and discuss emerging research issues and infrastructure needs, and to outline a roadmap for strengthening European research infrastructures in the SSH in the years to come. Which issues, domains and developments in SSH will be relevant in the next ten years? What are the requirements and challenges to research infrastructures and what will they be in the future?

A further objective of the conference was to identify common aims and challenges that bridge the gap between the social sciences and the humanities, and ways of facilitating and fostering European synergies and cross-disciplinary cooperation between the two fields, also in connection with other disciplines.

The conference was initiated by the ESFRI Social and Cultural Innovation Strategy Working Group and the German Federal Ministry of Education and Research. It was jointly hosted by the European Federation of Academies of Sciences and Humanities (ALLEA), which was responsible for ensuring the high-level expertise of the humanities' input in the discussion, and by the German Data Forum (RatSWD), which was responsible for doing so for the social and economic sciences.

The results of the inspiring conference and the publication of the contributions now on hand emphasise the relevance and timeliness of the debate on research infrastructures between the social sciences and the humanities. The conference has given us all a great opportunity to strengthen ties across Europe and to take us another step forward. The successful collaboration between the ESFRI group, the BMBF (German Federal Ministry of Education and Research), ALLEA and the German Data Forum in initiating and organising the conference resulted in this publication, which can be read as a comprehensive view of the state of the art of pan-European interdisciplinary research.

Starting with an introduction on the ALLEA survey on the status quo of the humanities, the first chapter elaborates on the necessity of well-built research infrastructure for the advancement of SSH and what challenges they will face in the future.

Subsequently, four specific areas of both domains were the focus of attention: Administrative Data, Longitudinal Research and Bio-Social Research, Digital Humanities and Digital Communication and Social Media, all of which are highly current and represent the latest developments within their respective fields. Each paper evaluates ambitious infrastructure projects emphasising present challenges and future potential.

Drawing from these discussions, the final chapter (re-)defines the requirements for the next generation of European research infrastructures for the humanities and social sciences. It puts forward answers to the most pressing questions and may provide a new point of departure for on-going discussions on future research infrastructure projects.

We would like to express our sincere gratitude to all authors and conference participants that sparked the vivid discussion that is followed up by this book. Without their crucial input, this publication would not have been possible. In addition, we are grateful that the German Federal Ministry of Education and Research (BMBF) and the ESFRI-strategic working group "Social and Cultural Innovation", namely the chair of the group Adrian Duşa, who supported both the conference and the subsequent publication as they will help to build a sustainable and distinguished European research infrastructure for the SSH domain.

Special thanks go to Simon Wolff and Camilla Leathem for editing the English-language manuscripts and to Dominik Adrian for coordinating ALLEA's contribution to the project. We are also especially grateful to Denis Huschka, Claudia Oellers and Thomas Runge, who did an excellent job the overall coordination of the conference and the publication.

With (almost) everything said, we wish you a pleasant read.

Berlin, April 2014 *Günter Stock and Gert G. Wagner*

2 Research Infrastructures

2.1 Understanding How Research Infrastructures Shape the Social Sciences: Impact, challenges, and outlook[1]

Peter Farago (FORS)

Research infrastructures are the backbone of science. The fact that mature science needs infrastructures is evident to most scholars and observers when talking about physics (e.g., accelerators), astronomy (e.g., observatories), chemistry and pharmaceuticals (e.g., laboratories), life sciences (e.g., biobanks), climate research (e.g., polar research vessels), or information technology (e.g., satellites). It seems to be less obvious for the humanities, although their archives, libraries, and collections of artefacts are the oldest infrastructures of all, dating back to ancient times. When it comes to the social sciences, although the notion of research infrastructures is still unfamiliar to many, research infrastructures exist in a surprisingly large variety of forms and structures, and state-of-the-art empirical research in the social sciences is virtually impossible without them.

Research infrastructures are having profound effects on the ways in which social science research is organised and conducted nationally and internationally. They are opening access to growing volumes of existing data and facilitating their use by forging common documentation standards and technical platforms across which data can move quickly. With an increasing abundance of available data across wide ranges of disciplines and topics, researchers can rely on large data pools to address their research questions.

Further, infrastructures providing large-scale, coordinated, harmonised, international, and interdisciplinary data collections make possible analyses and forms of comparison that were previously out of reach. While infrastructures follow and reflect the research communities that they support, they also contribute to methodological innovation and advances with respect to how data are gathered and used. In addition, research infrastructures are playing an important role in

1 This contribution relies to a large part on a volume on "Understanding Research Infrastructures in the Social Sciences" co-edited by the author together with Brian Kleiner, Isabelle Renschler, Boris Wernli, and Dominique Joye, senior colleagues at FORS, and published in 2013 (Zurich: Seismo). More details on the topic, an extensive literature review, as well as thirteen concrete examples of social science research infrastructures from around the world can be found in the book.

the dissemination of skills, research information, and know-how by way of training and network building within their constituent communities.

This contribution will highlight several ways in which research infrastructures are having a long-term influence on the social sciences. It then turns to a discussion of important challenges in optimising aspects of the relationship between infrastructure and research, and finishes with a look to the future. First, the definition, the key features, and the main components of social science research infrastructures are presented.

Defining research infrastructures for the social sciences

Although the term "research infrastructure" appears with increasing frequency in the social sciences, there is no single accepted definition, and it seems to mean many things to many people. A review of publications, reports, and articles on research infrastructures from the past two decades brings up an extended family of key terms that apply, such as: permanent institutions; long-term projects; best practice and excellence. Moreover, these descriptions shed little light on the core and necessary characteristics of research infrastructures that allow us to distinguish them from other forms of scientific work. Also, the definitions put forward include terms that point to various constellations of technical, operational, organisational, and human features.

It remains a great challenge to provide a definition that is sufficiently comprehensive to include all existing research infrastructures, but at the same time narrow enough to exclude institutions that provide the very basis for research and/or teaching, such as universities, private research organisations, and national statistical offices, and even more so if the definition should also include future developments.

A working definition for research infrastructures for the social sciences might be as follows: they are *durable institutions, technical tools and platforms, and/or services that are put into place for supporting and enhancing research as "public good" resources for the social science community*. The term institution refers in this context to physical or virtual locations, organisations, or networks (loose or formalized).

The challenge in clearly defining research infrastructures may be due to the fact that they are by nature generally invisible. As a substrate on which important

economic and social activities can be developed, we easily disregard infrastructures, even though we use them in our daily lives. Their main mission seems to be "just there" and "ready-at-hand", and they are recognised usually only after they stop working optimally.

Key features

Social science research infrastructures have distinctive features, but they also share to some extent aspects common to all infrastructures, both old and new. The limited but growing literature on the topic distinguishes *five key features* of research infrastructures that are intrinsically interlinked.

First, infrastructures in general provide services and resources as a *public good,* meaningnon-exclusive, non-competitive, and available to all. This also means that the quantity of the service or resource does not diminish with its use: once it has been produced, it benefits all on an on-going basis. It is a matter of processing requests coming from researchers or groups of researchers to make scientific profit out of the possibilities offered.[2] Establishing and maintaining infrastructures involves the coordinated action of a community of interested parties, often across various disciplines or sectors, which are represented by key persons working within established networks who are able to demonstrate their value, synergies, and benefits for funding institutions.

Second, research infrastructures must offer *user-oriented services* corresponding to the needs of researchers. These services can take on various forms, such as data, tools, education and training, and methodological expertise, all aiming at contributing to the advancement of a specific field of science. The nature of these services depends very much on the scientific sector and the research communities involved. Generally, they consist of sets of services and resources that are interrelated.

Third, research infrastructures need to be *durable and stable on a long-term basis* to avoid losing accumulated benefits. Therefore, the establishment and maintenance of infrastructures require effective communication to anchor the infrastructure in public policies and to ensure that policy-makers and the public recognize their legitimacy and benefits to society as a whole. On the user side,

2 As an example, the European Organisation for Nuclear Research CERN offers qualified scientists the possibility to use its instruments, but the application process is competitive and based on an evaluation of requests.

the infrastructure must be able to offer services that are necessary for researchers on a long-term basis, and therefore must provide continuous and stable resources, personnel, platforms, and facilities.

A fourth key feature of research infrastructures is *adaptability* to the changing needs of the scientific community. This can seem somewhat contradictory to institutions that aim to exist on a long-term basis and that must by nature be conservative. However, alterability is fundamental for research infrastructures in order to be able to provide a public good that remains closely aligned with the needs of users, and especially to gain and maintain the support of stakeholders.

Finally, research infrastructures are intrinsically related to the requirements of the scientific *method*, in a way that provides important benefits for the scientific community. By offering transparent and open access to data, research infrastructures support the scientific method by enhancing opportunities for hypothesis testing and replication. In addition, by harmonising standards and by encoding these in practices and tools, infrastructures promote comparability and wider and more efficient use of data toward scientific progress.

Main components

Research infrastructures in the social sciences have several components:

- *Data services* for documenting, preserving, and disseminating data. These can be data collected by individual researchers or research groups, or they might be collected by the infrastructure institutions themselves. In any case, the data are cleaned and prepared for use by scientists. This includes state-of-the-art anonymisation procedures that allow for the distribution of data according to national data protection regulations. Good examples for data services are the member organisations of the European social science data archives consortium CESSDA (www.cessda.org).

- *Collection and harmonization platforms* provide and link data. This includes internationally coordinated surveys that are harmonised ex ante as well as data collections harmonised ex post for comparative purposes. The European Social Survey ESS is a case in point, but also the the Cross-National Data Center in Luxembourg LIS (www.europeansocialsurvey.org, www.lisdatacenter.org).

- *Methodological research* on survey methodology, but also on documenting, archiving, anonymising, accessing, and distributing data is another central element of research infrastructures.

- *Teaching and training* are important activities to promote state-of-the-art techniques and procedures and to introduce researchers to the possibilities research infrastructures can offer them.

These components might be distributed across different institutionsas is the case in the UK. But they might also be combined under the same institutional roof, like in Germany (GESIS) or Switzerland (FORS). In the latter case there is a good chance to exploit synergy potentials optimally.

The impact of research infrastructures in the social sciences

There are several key lines of development that characterise how research infrastructures are reshaping social science at different levels. These include internationalisation, convergence of practice, and the opening and sharing of data and information.

Internationalisation – scaling up the social sciences

Research infrastructures are leading to a greater internationalisation of social science research in a variety of ways. This means that research that used to be confined generally to national contexts is now able to reach wherever its logic requires, especially in cases where national comparisons are crucial to informing theory and addressing policy questions. Infrastructures such as international and national data services are paving the way for easy and open access to social science data, no matter where they may be located. These developments have led to a wider accessibility of data, and to new international alliances, which have to be placed within the context of changing legal frameworks and the creation of new international standards.

Wider access

Individual countries are no longer research islands in the social sciences, and the erosion of national barriers driven by social science research infrastructures means that researchers have easier access to a wider range of data, cutting edge tools, techniques, and know-how. Such access ultimately improves research practice and efficiency.

International alliances

By establishing, expanding, and strengthening cross-national projects, social science research infrastructures are generating new institutional and individual partnerships and productive alliances enabling researchers to gain and exchange experiences, and follow a common agenda with international research partners and experts. In addition to fostering networks, this has the effect of creating shared working vocabularies and common techniques that can be developed and refined by virtue of a greater number of active users through on-going collaboration. In this respect, large international projects have produced a set of standard procedures for scientific surveys that were previously non-existent.

Legal frameworks: orienting internationally

Beyond the bridge-building at technical, conceptual, and linguistic levels, social science research infrastructures are constantly addressing relevant legal and ethical considerations, since the sharing of data across national borders raises a host of questions about confidentiality and intellectual property within diverse legal frameworks. Thus, many social science research infrastructures are on the leading edge of questions of accreditation, anonymisation, consent, ownership, and access to sensitive data within an international context. To allow for the flow of data within and across countries, research infrastructures have been instrumental in ensuring that data protection laws are respected and that data producers and users are informed of their rights and responsibilities.

Combination of data and methods

The gains in efficiency, productivity, and scientific quality within and across disciplines brought about by social science research infrastructures are in large part due to converging data sources, practices, tools, and standards. Common tools ensure easier access to data, and allow for mixing data sources. Methodological innovations and best practices are shared, interdisciplinary platforms are established, and common technical solutions are adopted. This leads to a high degree of standardisation in procedures and classification schemes.

Making data comparable

An important output of social science research infrastructures has been to increase the potential for comparability between countries or between regions within countries. This is made possible by common methodological frameworks within large-scale survey programmes, by international harmonisation platforms, or through international data portals that pool data from different countries. Large survey programmes also allow for regional or intra-national comparative analyses given their sample sizes are large enough. Data archives often make available to researchers data on particular subjects gathered by different projects within the same country.

Use of different types and sources of data

Experience shows that survey microdata are ever more often enriched with other types and sources of data: individual administrative register data, contextual data relating to geographical or political location, biomarkers, interviewer data, and call data are more frequently being made available to researchers. Qualitative and quantitative data and methods are more often used simultaneously in research projects. The same is true for micro and metadata that are supplied to researchers in a more coherent, thorough, and systematic way than before. This has changed the way social scientists work and has led to the diffusion of new analytical and statistical tools. The combination of different types and sources of data also facilitates tackling one of the currently most serious problems of empirical social science, namely declining response rates.

At the same time, more and more data are produced on the individual level, often without even asking the person concerned. Examples include administrative data of all kinds, data produced by using credit cards and other non-cash payment methods, Google-searches, social networking sites, etc. These immense masses of data ("*big data*") can be of value to science. However, the proper use of such data is not a given thing because of specific selection biases, privacy protection rules, private ownership, or technical or legal limitations to access. Nevertheless, the potential is there and could be explored in a much more systematic fashion than has been the case up until now.

Methodological advancement

Social science research infrastructures offer a unique combination of methodological and technical expertise that is disseminated over time and that leads in practice towards a convergence of skills. For example, for large-scale international survey projects, conducting research across different settings promotes innovation and helps to overcome many particular methodological challenges. The exchange and transfer of knowledge between and within partner countries is a natural by-product of such work.

Interdisciplinarity

Large survey platforms were conceived as interdisciplinary programmes in the social sciences from the outset. This has fostered interdisciplinarity, leading to advances and convergences in knowledge across disciplines in terms of methodologies and procedures, but also substantially, with more holistic approaches due to the use of indicators from other fields. In this way social science research infrastructures have been helping to overcome borders between individual disciplines.

Development of technical solutions

Social science research infrastructures have been at the forefront of the development of a wide variety of technical systems that allow for the curation, discovery, and flow of data nationally and internationally. Sometimes called "e-infrastructure", such systems are generally open-source and standardised, and are continuously being improved to meet the needs of researchers. The challenge here is that the tools and technologies used by data services should remain simple and largely diffused, so that easy access to data is ensured from all over the world. The technologies must also be designed to minimise risk of disclosure of individual information with respect to legal frameworks and national laws.

Standardisation

Last but not least, research infrastructures have led to an increase in standardisation in the social sciences. This is especially the case with respect to documentation standards such as DDI, which allow data to be shared and used appropriately for secondary analyses. Also, standardisation of classification schemes like socio-demographic variables, common scales, and missing data treatment open the field for comparative analyses between countries and regions. Standardisa-

tion also allows for a better and more efficient control of procedures and checks, increases data quality, and provides for a more efficient allocation of costs.

Towards more open science

One of the underlying ideas of social science research infrastructures is that science works best when it is done in an open, transparent, and collaborative fashion. Research infrastructures offer data, tools, services, and training that favour openness in scientific practice.

Opening national and transnational access

Social science research infrastructures are leading the way toward overcoming barriers to data access, within and across countries. The result has been concerted and continuous efforts to open access to data and metadata that are increasingly offered to broader audiences, and more frequently with easy and free access through the Internet. The increasing use of English as an international standard for metadata has had a significant impact on transnational access. While data from large-scale measurement instruments are becoming more readily available for researchers via online tools, there still is a clear segmentation for access to data. Especially official statistical data remain difficult to obtain in some cases because of complicated authorization procedures or high fees. Despite the progress made in this respect, new and complex issues relating to data protection, privacy, and research ethics continue to arise in the context of divergent practices.

Changing models of research practice

The work of research infrastructures is leading the social sciences away from the model of one researcher, one project, one dataset and towards a model of commonly produced and shared data on a large scale, used freely by an entire community of researchers. This shift away from small-scale research projects offers several advantages. First, it is of course more cost-efficient as data are paid for once and re-used by many researchers. But more importantly, open access to common pools of data leads to more fair and balanced competition between researchers enabling the scientific community to work on the same material.

Challenges

Social science research infrastructures are currently facing several challenges. The most prominent are the dialectics of continuity and innovation; the tension between open data access and confidentiality; fragmentation, funding, and time-frames issues.

Continuity and innovation

Research projects are usually limited in time and researchers move on to other projects when they finish. In contrast, infrastructures are designed to last in order to provide the raw material for research projects: high quality data, documentation, and tools for storing, accessing, and using data. In the process they acquire know-how in producing and making available large amounts of data that could hardly be gathered by individual researchers. In order to accomplish this important task, infrastructures need to be more stable than research projects.

On the other hand, social science research infrastructures must constantly adapt to the ever-changing needs and conditions of research. Failing to do so would quickly make them obsolete. This means that infrastructures cannot simply follow on the coattails of science, but rather must play an active role, foreseeing new directions and possibilities, and supplying the conceptual and technical expertise needed to go there. Research infrastructures are often on the cutting edge of research and methodological developments, as in the case of large-scale survey projects or international documentation standards like DDI.

Research infrastructures in the social sciences must find the right balance between supporting research communities in a continuous and stable fashion and generating innovation.

Open data access and confidentiality

Sharing data with other researchers is now widely accepted and practiced in scientific research – be it for replicating analyses or for better exploiting rich (and costly) data sources. In the social sciences, data collection (e.g. large surveys) is often too heavy a burden and too expensive to be organised by individual researchers. Well-documented and easily accessible data repositories are a valuable alternative. The growing number of datasets distributed by such institutions shows that there is a large demand for high quality data.

To fulfil this demand, data access has to be open for scientific purposes. However, since most of these data refer to the level of individuals, care has to be taken not to violate the rules of data protection and privacy set out in legal regulations and best practice manuals. Research infrastructures are responsible for enabling controlled access only to data that are anonymised to such an extent that identification of individuals is practically impossible, and only to users who have the qualifications, know-how, and willingness to use the data exclusively for scientific purposes.

However, there are still many conditions that have to be fulfilled to make generalised access possible: persistent identifiers for every dataset; international standards for storage and documentation; powerful and efficient data search engines; authentication, authorisation, and accounting procedures that allow for effective control of data users and data usage. Many initiatives are under way to achieve these goals.

The future of social science research infrastructures will also depend to a considerable degree on the solutions they choose and how successful they will be in securing generalized global access to data for researchers.

Fragmentation, funding, and timeframes

There is considerable *fragmentation* of the landscape of social science research infrastructures, nationally and at the European level. Social science research infrastructures are usually established in isolation, in response to national or otherwise local demand, and are not necessarily coordinated with others. This is related in part to the heterogeneity of the projects led by research infrastructures, but also to path dependency for projects that are integrated later in research infrastructures.

The most obvious factor influencing the development of research infrastructures is *funding*, which is usually shared between different agencies, such as national science foundations, government institutions, universities, and European research programs.

Diverging *timeframes* or differing policies among national funding institutions also have important consequences that can slow down the progression of research infrastructures. For example, regular and repetitive applications involving long assessment periods hamper the development of research infrastructure projects. They are also a source of uncertainty regarding staff and

participation in international projects. Another reason is that in many cases there is no single funding agency devoted to social science research infrastructures at the national level. The question which agency should fund infrastructure-related projects in the long-term has often not been tackled before social science research infrastructures are invested in, and much time and effort is spent on finding arrangements.

Generally speaking, funding institutions should recognise that infrastructures need continuity and financial stability in order to carry out their mandates.

Outlook: integration, coordination, and durability

Social science research infrastructures in Europe have demonstrated their value and will continue to be an integral part of the research landscape. However, there are three main areas where they must be strengthened as a whole in order to be most effective in serving their constituent research communities in the future.

First, there needs to be a better *integration* of social science research infrastructures *into the daily work* of researchers. While the "invisibility" of research infrastructures would indicate that they are functioning smoothly, there are a few areas where the relationship between research infrastructures and researchers could be improved. One area has to do with the tools that are developed for data discovery, access, and documentation. Social science research infrastructures must continue to develop and provide cutting edge, easy-to-use tools that facilitate finding and obtaining relevant data in close collaboration with the researchers who are the ultimate beneficiaries. Users should not have to search multiple sources to find what they are after, and research infrastructures should aim to improve coordination, and to reduce the number of data portals within countries and internationally as much as possible.

However, new solutions are needed to overcome real conceptual, technical, legal, and language-related obstacles, requiring an investment on the part of all interested parties. If data archives encourage researchers to share their data, then they should make standardised data documentation easier. On the other side, researchers must become better skilled in data management and documentation, and should have a better command of issues such as data preservation consent, confidentiality, and anonymisation. Research infrastructures can provide training on this front. Data sharing should become a normal practice

rather than an obligation or after-thought. Finally, in order to provide incentives to researchers for sharing their data, peer-reviewed journals should be encouraged to require citation of data used in publications, and universities should award professional credit and recognition to researchers who have their data cited by their peers in publications.

Second, more *coordination* is needed between infrastructures at the national and international level. Until now, infrastructures have generally developed in relative independence, and usually have established linkages on an ad hoc and a needs basis. Moreover, it is clear that there is great potential for generating more synergies between social science research infrastructures in Europe, especially because they have much in common and could benefit from shared expertise, systems, and tools.

Last, but not least, the *long-term durability* of social science research infrastructures is difficult to imagine without institutional and funding stability. Most infrastructures have mixed funding schemes that include varying shares of project-oriented short-term contributions alongside the basic funds securing their main functions. Long-term funding commitments have often been difficult to obtain, and still depend on the priorities of national funding bodies, on their philosophy, and sometimes on the general economic situation. The current arrangements differ considerably according to country-specific legal and institutional contexts or to international regulations like the European ERIC-statute. Further, social science research infrastructure projects are often in competition nationally with other scientific projects, so that they sometimes cannot move at the same pace as their international counterparts.

All these factors are gradually altering the dynamics of knowledge production in the social sciences and changing the ways in which researchers go about their daily work.

2.2 Challenges for the Humanities: Digital Infrastructures

Gerhard Lauer (University of Göttingen)

Infrastructures are nothing new in the humanities. Since the beginning of modern science, infrastructures have been an essential part of the *ars inveniendi*, the new way of doing research, as formulated by Francis Bacon and others at the beginning of 17th century. In his *Advancement of Learning* of 1605, Bacon himself mentions two institutions that are necessary for modern research: the libraries and careful editions of (canonical) authors.[1] Both are infrastructures in the sense of enabling research. The idea to build a library solely on the principle of its values for scholarship and science, however, still took time to develop. It was not before the enlightenment, not much before Christian Gottlob Heyne, that universities started to organize their libraries exclusively according to scholarly criteria. Before George II opened Göttingen University in 1737, his minister Münchhausen had started to build a library two years in advance because he understood the necessity of infrastructure for modern universities. In the days of Heyne und Münchhausen, research libraries were the main infrastructure. The second infrastructure mentioned by Francis Bacon is the collection of data; at that time in the form of editions.

After Richard Simon published his *Histoire Critique du Vieux Testament* in 1678, these evolved step by step into critical editions. The critical editions of Simon and his contemporaries opened the road for all kinds of carefully collected data with the potential to offer new and unexpected insights into nature and culture. The success story of the humanities in the 19th century relied heavily on these two kinds of infrastructures; libraries and editions. Projects like the 'Corpus Inscriptionum Latinarum',[2] the collection of hundreds of thousands of inscriptions from the Roman Empire initiated by Theodor Mommsen, illustrate the significance of editions for scholarly research. In this long tradition, infrastructures in the humanities have not only the function to secure what is already known, but to open the horizon for new knowledge.

1 Francis Bacon (1605): Advancement of Learning. The Second Book. ed. Hartmut Krech, www.luminarium.org/renascence-editions/adv2.htm.

2 Corpus Inscriptionum Latinarum, cil.bbaw.de.

Theodore Mommsen and his colleagues collected all available data without pursuing a specially formulated research question. Infrastructures are essential for seeking principles and patterns.[3]

The digital data deluge has altered the way scholarship is done. Some speak of a new paradigm, the so called 'fourth paradigm' of data-intensive scientific discovery.[4] However, big data is more than just a metaphor, and computer-based work has been used to handle big data in the humanities since 1949, when Roberto Busa started his edition of the complete works of St. Thomas. In the days of punch cards, big data meant to editing 56 volumes, possible only with the support of Thomas J. Watson, the founder of IBM.[5] Thus if we are looking for a symbolic date for when computing became part of humanities' infrastructures, 1949 might be a good guess. Digital data has been transforming humanities infrastructures even since, albeit initially only on a small scale.

With the rapid growth of the internet and the digitization of millions of books, documents and objects, computing has transformed humanities infrastructures in the 21st century on a large scale. First, it has expanded the depth of historical research. Good examples of this are computer based editions like 'Universal Leonardo',[6] the complete works of Leonardo da Vinci with integrated X-ray and UV analysis of his paintings, or the complete edition of Mozart's work[7] with all available facsimiles and critical comments on his scores. Another example is the recent discovery of a 1200 year old hidden temple around Angkor Wat.[8] Detecting the temple under the trees of the Cambodian jungle was only possible with computer-based light detection and ranging technology (lidar). This kind of technology analyses a single work or place with a fine-grained digital resolution that was hitherto not possible with X-ray methods enabling more than just a handful of experts to look under the surface of Leonardo's paintings. Computer-based infrastructure thus deepens scholarly research. Second, new infrastructures change the speed of cultural analysis. In libraries like the Perseus Digital Library or the Cuneiform Digital Library Initiative,[9] scholars all over the world find their major resources at a scale not imaginable before the digital age. To

3 Bod, Rens (2013): A New History of the Humanities. The Search for Principles and Pattern from Antiquity to the Present. Oxford.

4 Hey, Tony/Tansley, Stewart and Tolle, Kristin (Eds.): The Fourth Paradigm. Data-Intensive Scientific Discovery. research.microsoft.com/en-us/collaboration/fourthparadigm/default.aspx.

5 Corpus Thomisticum, www.corpusthomisticum.org.

6 Universal Leonardo, www.universalleonardo.org.

7 Digitale Mozart Edition, dme.mozarteum.at.

8 Evans, Damian H. et al. (2013): Uncovering archeological landscapes at Angkor using lidar. Proceedings of the National Academy of Sciences (PNAS) 110 (31), 12595–12600.

9 Perseus Digital Library, www.perseus.tufts.edu, and Cuneiform Digital Library Initiative, cdli.ucla.edu.

call this big data is not an exaggeration. The Perseus collections bring together Greek and Roman material including art and archaeological findings, Arabic resources, Humanist and Renaissance texts, German and American editions; ultimately thousands of books and manuscripts. The tradition of scholarly editions has evolved into massive digital libraries. And they include not only texts, but a steadily increasing number of other data, objects, pictures, audio files, films. To give but one example: the European Holocaust Research Infrastructure (EHRI) integrates the huge variety of holocaust archives with diverse types of documentation.[10] Researchers normally have to deal with different archival systems and types of objects. Now, in a virtual research environment like the EHRI, the differences are integrated, enabling research on a vast variety of documentation. From the perspective of virtual research environments, documents, archives and libraries are no longer separate entities: digital infrastructures regroup the order of research institutions. To mention a further example, different types of dictionaries located in different places can now be searched using a one-stop shop and the variety of genres on which the dictionaries are based can be compared to see whether a specific expression is only used in fiction or also in everyday discourse.[11]

This is the "million books" situation, i.e. the moment in history when for the first time millions of books and journals are available, and not only to the happy few. The digital infrastructure enables scholarly research of unprecedented depth and speed. Google with its nearly 30 million books and the much more carefully scanned libraries like Early American Imprints, Early English Books online, Gallica, Deutsches Textarchiv, TextGrid Repository, Verzeichnis Deutscher Drucke not only offer a vast collection of books, but they enable the building of new corpora. It is not the collections or the libraries in themselves or in their sheer size that are new, but the chance to construct corpora for particular research interests. Turning data into corpora is the way in which even the humanities are making increasing use of big data.[12] Computer-based tools and analytical techniques now make it possible to deal with larger and more hetero-

10 European Holocaust Research Infrastructure (EHRI), www.ehri-project.eu; cf. Blanke, Tobias et al. (2013): From Fragments to an Integrated European Holocaust Research Infrastructure. In: Neuroth, Heike/Lossau, Norbert and Rapp, Andrea (Eds.): Evolution der Informationsinfrastruktur. Kooperation zwischen Bibliothek und Wissenschaft. Glückstadt, 157–177.

11 E.g. Digitale Wörterbuch der deutschen Sprache (DWDS), www.dwds.de/.

12 Lauer, Gerhard (2013): Die Vermessung der Kultur. Geisteswissenschaften als Digital Humanities. In: Geiselberger, Heinrich and Moorstedt, Thomas (Eds.): Big Data. Das neue Versprechen der Allwissenheit. Berlin, 99–116.

geneous sets of data out of corpora. Stylometric analytics of literary history,[13] SplitsTrees to map the distribution of fairy tales all over the world,[14] sentiment analysis of affects in culture[15] and visual data analytics to detect new stories[16] are only some of the ways in which scholarly research can analyse cultural evolution on a new scale as long as digital infrastructures are around.

A last point has to be mentioned in conjunction with the challenges of digital infrastructures in the humanities. A major challenge for the humanities is the change in habitual ways of doing scholarly work. Scholars commonly work alone. For most scholars, infrastructure means still libraries and editions. In the digital age, the role and position of scholars are changing. They are becoming more or less part of a group with different expertise: one is able to build the corpus, the other runs the analytics, and a third is able to do the statistics. A division of labour at different scales, depending on the research task, alters the way in which research in the humanities is done. The classical role of scholarly research is one role, but it is no longer the only one, and new interactions between citizen science and the humanities are possible.[17] The infrastructures of the *ars inveniendi* are not history, but are expanded to the extent where humanities become digital humanities without even thinking about being digital.

13 Jannidis, Fotis and Lauer, Gerhard (2014): Burrows's Delta and Its Use in German Literary History. In: Erlin, Matt and Tatlock, Lynne (Eds.): Distant Readings: Topologies of German Culture in the Long Nineteenth Century. Rochester, 29–54.

14 Therani, James (2013): The Phylogeny of Little Red Riding Hood. PLoS ONE 8 (11). e78871. doi:10.1371/journal.pone.0078871.

15 Ahmad, Kurshid (Ed.) (2011): Affective Computing and Sentiment Analysis. Emotion, Metaphor and Terminology. Dordrecht et al.

16 Krstajic, Milos et al. (2013): Story Tracker. Incremental visual text analytics of new story development. Information Visualization 12, 308–323.

17 Hand, Eric (2010): Citizen science: People power. Networks of human minds are taking citizen science to a new level. Nature 466, 685–687.

2.3 Survey and Analysis of Humanities and Social Science Research at the Science Academies and Related Research Institutes of Europe

Camilla Leathem (The Union of the German Academies of Sciences and Humanities)

Introduction to the SASSH initiative

It seems that current financial and monetary difficulties in Europe are overshadowing the issue of a lack of common European identity. 200 years of nation states seem to have suppressed 1800 years of a history shaped by mutual enrichment in politics, science and arts – a European cultural heritage that must be revived in the minds of its citizens. It is for the social sciences and humanities (SSH) to research, explain, propagate and preserve this heritage. While numerous research projects on cultural heritage are conducted on national levels, a pan-European programme on this topic is still lacking. All European Academies (ALLEA) and the Union of the German Academies of Sciences and Humanities thus plan to initiate a European research programme on cultural identity and heritage in Europe led by ALLEA and comprising long-term humanities and social science research. The concept has already met with interest at the European Commission, and our medium-term aim is to formulate a concrete proposal to the European Commission for the necessary funding.

The initiative is driven in part by a paradigmatic supra-regional research network for the SSH: the German "Academies' Programme" run by the Union of the German Academies of Sciences and Humanities. The Union of German Academies is the umbrella organisation of eight academies of sciences and humanities. It comprises over 1900 outstanding scholars from a broad range of academic disciplines and coordinates the "Academies' Programme", one of Germany's most important and comprehensive research programmes in humanities and cultural studies. It coordinates and supports joint basic research projects (e.g. dictionaries, encyclopaedias and editions) between the member academies, promotes the exchange of information and experience between academies, conducts public engagement activities, and organises events on

current issues in academe. Projects run by the Programme contribute to the empirical foundations of cultural heritage research, making it relevant for the present and preserving it for the future.

The SASSH initiative is founded on the conviction that a research programme like the German Academies' Programme at European level would be an asset to the European Research Area, allowing pan-European networks of scholars to address European issues from a European perspective. As a federation of European Academies with a history of excellence in the SSH that brings together 54 Academies in 42 countries from the Council of Europe region, ALLEA offers the ideal framework in which to implement a research programme at European level; a research programme open not only to science academies, but also to all other related research institutes.

The digital dimension of the SASSH initiative

The future coordination of a pan-European research programme requires thorough knowledge of existing research and (e-)research infrastructures. In close cooperation with ALLEA, the Union of German Academies is thus undertaking a pan-European "Survey and Analysis of Basic Research in the Social Sciences and Humanities" at the science academies, learned societies, and related research institutes of Europe (SASSH). Running from August 2013 until April 2015, the project is funded by the German Federal Ministry of Education and Research. The project will not only bring greatly needed transparency to these areas of research, but will investigate and identify opportunities to improve the coherence of the numerous ongoing SSH research projects and activities within and across national borders, including the existing digital tools and infrastructures, ultimately detecting concrete opportunities for a long-term SSH research programme on cultural heritage in Europe. Discovering, interpreting and preserving our European cultural heritage is a societal imperative and in order to do so we need to sustain research infrastructures and to enable open access to data.

Survey content and responses

In December 2013, the linchpins of the project, two surveys, were issued to the science academies, learned societies and related research institutes of Europe. These research institutes had previously declared their willingness to participate, and each runs or is affiliated with numerous research projects in the SSH. The survey responses enable us to catalogue what research is being undertaken, where, and using which digital research tools (hereinafter DRTs). The main survey is directed at research project staff and includes questions on basic project information, potential for collaboration, existing systems of project evaluation, and forms of publication and use of DRTs. More specifically, it addresses paper vs. digital forms of publication and archiving for both research outcome and data and, in the case of digital publishing, the use of open access platforms. It furthermore addresses researchers' willingness to publish all data and outcome on open access platforms in future, the exact DRTs used by researchers to search, discover and read secondary literature, to search, discover and gather data, to analyse data, to collaborate and share data, or to search and discover digital tools. It furthermore investigates the awareness and use of European digital research consortia (e.g. DARIAH, CLARIN, Europeana), and researchers' needs and wishes for future DRTs.

The digital dimension of SSH research being a high priority for the SASSH initiative, there was some concern that knowledge at project level of institutional digital processes like data storage may not be sufficient to reliably answer specific questions, or to answer them at all. In order to gain accurate and broader insights into digital practices across the institution in general, a short survey on the use of DRTs is addressed to IT staff, (digital) library staff, and/or staff of purpose-built digital centres, while also giving project staff the opportunity to complete this survey in addition, where knowledge is sufficient. This survey addresses the availability of DRTs, the most common DRTs in use in the project or institution-wide, membership or use of European digital research consortia like DARIAH, CESSDA, CENDARI and Europeana, attitudes to and forms of digital archiving and publishing, institutional data standards and policies, and the availability of support, training and information on DRTs.

Currently, completed main surveys have been returned by approx. 550 SSH research projects and DRT surveys by approx. 110 institutions, departments or projects from around Europe. The survey is still open and will remain open until the conclusion of the project's analysis phase. Western and northern European countries are particularly encouraged to continue to respond, as many are as yet underrepresented in the survey results.

Significance of the survey data for harmonising digital research infrastructures in a pan-European research programme

The major challenge of a collaborative and connective pan-European research programme will be to harmonise digital research practices by drawing together the numerous national and, increasingly, multilateral digital research initiatives. The programme must be based on a digital infrastructure that is sustainable, interoperable, easy to use and comprehensive while also catering to individual subject-specific demands. First and foremost, the data collected using the surveys will be used to compile an overview of the DRTs in use throughout the research lifecycle; from data collection to publishing and archiving, digital library catalogues, digital collections, digital archives and databases, project management tools, document management tools, data-sharing and collaboration tools, communication tools, writing/editing tools, data storage tools, publishing tools, corpora, text analysis tools, image tools and other specific subject-related tools. This will highlight common tendencies and frequently used DRTs on the one hand, and fragmentation, gaps and self-developed DRTs for subject-specific purposes on the other.

A pan-European research programme will additionally rely on sustainable open access to interlinked data sets, repositories and digital collections for integrated searches. It is thus of particular importance to analyse archiving and publishing behaviour at the institutions in question. The survey data will highlight which institutions store research data and research outcome in open access databases or repositories and where they do so, and which do not. It will also show where institutional policies for open access publishing are in place and where they are yet to be implemented, and measure the willingness of researchers to subscribe to full open access publishing of their data and research outcome in future. Open data policies at institutional level are the cornerstone to achieving the kind of comprehensive open data access on which a pan-European research programme will depend.

The implications of data and material stored in a multitude of different repositories also necessitate an overview of existing data standards policies at the respective institutions (where applicable): interlinked data sets require a common query interface, and therewith common coding, formats and schemas. A pan-European research programme will rely on uniform institutional data policies and meta-data standards; using the same standards internationally will markedly increase the possibilities of semantic linking of cultural heritage, making not just the canon, but cultural tradition in its entirety more visible and

translatable in various languages. The data collected using the surveys will help to investigate the potential in existing conditions and structures to achieve this.

Not least, a European programme will also rely on researchers' awareness of the most suitable digital tools across the research lifestyle. This will not only ensure that researchers are working with the best and most suitable tools for their purposes, but will also help to keep fragmentation and repetition to a minimum: the existence – in parallel – of numerous initiatives is not only very expensive, but it inevitably leads to problems in coordination, for example in agreeing on common data standards or on legal issues. A priority for a European research programme may be to contribute to awareness of DRTs and data standards; to promote common platforms, standardised ways of consultation, and, above all, mutual knowledge of DRTs. The survey results will be used to investigate where basic opportunities for raising awareness of DRTs are already in place and where they are not; i.e. where training, support and informative events for DRTs are offered, how regularly, and in what form.

A final major concern, not just for a European research programme, but for all researchers embarking on the digital humanities, is the sustainability of digital infrastructures. Sustainable databases and repositories enable researchers to reliably deposit and/or publish their data and research outcome for the long-term or permanent future while ensuring that familiar and trusted DRTs remain in operation and are not made obsolete by unfamiliar replacements. For the sake of sustainability and uniformity, the SASSH questionnaire surveys the awareness and/or use of European research infrastructure consortia such as CESSDA, DARIAH and Europeana. These consortia and others like them have already made great strides to providing humanities research with a coherent, interoperable e-research infrastructure. Cooperation between them and a European research programme may thus be of great mutual benefit: on the one hand, they ease or have already eased the complex process of harmonisation and offer a ready-made way of ensuring as much consistency, standardisation and interoperability as possible. On the other hand, the scale and means of a pan-European research programme may contribute to the sustainability of these infrastructures and the DRTs affiliated to them by providing them with a transnational platform that promotes and encourages their use. Sustainability and the reliable permanence of the familiar may crucially also reassure scholars who are not yet convinced of the movement towards the digital humanities.

B Special Areas

3 Administrative Data

Peter Elias (University of Warwick)

Administrative data are defined as data which derive from the operation of administrative systems, typically by public sector agencies. They cover activities such as health maintenance, tax and social security, housing, elderly care, vehicle and other licensing systems, educational progress, etc. While such data are not designed for research purposes they often have significant research value, especially when linked to other datasets or to user-generated surveys.

This chapter looks at the ways in which some countries have developed access to such data and derived value from them as research resources. Other countries are now seeking to set up systems, which will provide better access to and linkage between administrative datasets. The chapter also provides the opportunity to discuss how administrative data may be shared between countries and examines the need for distributed research infrastructure to facilitate such data sharing.

3.1 Administrative Data: Problems and Benefits. A perspective from the United Kingdom[1]

Matthew Woollard (UK Data Archive)

This short chapter describes the current state of play in the development of *research and data service* infrastructures for administrative data in the United Kingdom. It provides some background to administrative data for those unfamiliar with it; outlines some of the main problems and benefits of the use of administrative data in research; discusses the process which led to implementation of this infrastructure and then the beginnings of that implementation. It does not consider outcomes of one of the key recommendations of the Shakespeare Report (2013) which suggested that administrative data and other forms of public sector information should be used in the delivery of public services. These initiatives are likely only to enhance the parallel initiatives of the research funders.

What are administrative data?

In the context of this chapter administrative data refers to information collected primarily for administrative purposes. "This type of data is collected by government departments and other organisations for the purposes of registration, transaction and record keeping, usually during the delivery of a service."[2] The three purposes of *registration, transaction and record keeping* are key to the understanding of the differences between administrative data and other 'survey' data, and while these three purposes are often conflated or confused, there are some key distinctions.

Registration is the process by which an entity (usually a citizen or a business) provides some specific information which is to be recorded for reference. The registration of births, deaths and marriages provides a permanent record of these events. Electoral registers provide lists of those who have registered to vote. In the first case the register is used not only as a record of an event;

1 This is an extended version of a presentation given at the workshop 'Facing the Future: European Research Infrastructure for Humanities and Social Sciences', Berlin, 22 November 2013. I am grateful to Libby Bishop, Hilary Beedham and Vanessa Cuthill for their comments on a final draft.

2 Administrative data introduction, www.adls.ac.uk/adls-resources/guidance/introduction/.

it is the primary data which informs a range of published statistics some of which are mandated through legislation to be created. In the second case, it is a legal requirement in the UK to be registered in order to vote in UK elections. In general, digitally held registers either expand vertically – that is new records are added and old records are not generally altered – or, they are replaced.[3]

Transactional, or business management information, is more complex, and a more commonly encountered form of administrative data. The Customer Information Systems of the Department for Work and Pensions (DWP CIS) and Her Majesty's Revenue and Customs are truly transactional data, where information is captured in order for government to carry out its responsibilities, in these cases either to collect taxes or disburse benefits. Transactional data have the tendency to expand both 'horizontally' – the same people pay tax in many years, and 'vertically' – when new people enter the tax-paying regime. Transactional data also may reduce 'vertically' as people die and move out of scope.

Few administrative data are solely *created* for the purpose of record-keeping since technically a record is just evidence of a past event, but as noted above vital events registration is also kept as a permanent record of these events.

For the purposes of this chapter, administrative data are only considered to be data which are created by government, or created for the immediate purposes of government. Some transactional data created by commercial organisations may also be used by government for their activities: these are currently out of scope.

Amongst the key British administrative data sources are:

- Department for Work and Pensions Customer Information Systems[4]
- Her Majesty's Revenue and Custom Customer Information Systems
- Offenders Index/Police National Computer (Ministry of Justice)
- English/Welsh School Census (Department for Education)
- Student Record/Destination of Leavers of Higher Education (Higher Education Statistics Authority)
- NHS Patient Record

3 The process by which corrections are made to these registers of vital events is both complex and fascinating, and demonstrates that even registers which are considered to be 'static' have the possibility of changing.

4 A generic name for the various benefit processing and payment systems used by the Department for Work and Pensions. Frazer 2010, p. 4, includes "the Income Support Computer System (ISCS), the Jobseekers' Allowance Payment System (JSAPS) and the Pension Service Computer System (PSCS) together with the Labour Market System which is used by Jobcentre Plus."

- Inter-Departmental Business Register (Office for National Statistics)
- Hospital Episode Statistics (Department for Health)
- Individualised Learner Records (Department for Business, Innovation and Skills)
- Electoral Register
- Local Authority records, e.g. of social care services
- Housing provider records (including housing associations), e.g. rent and tenant databases

The National Pupil Database (NPD) is a rather hybrid administrative data source since it is the product of linkage of data on pupils from a range of different sources into a single database.

A key example of existing linked administrative data can be found in the Indices of Multiple Deprivation (IMD) which link together a range of datasets, the most notable of which are: health records (NHS), unemployment and tax data (HMRC) and benefits data (DWP) to create an index of deprivation for geographic areas. (DCLG 2011)

Legal status

The legal status of administrative data is key to any understanding of their use beyond their primary purpose. There is not space here to describe this in detail but in an abstract sense there are two key specific areas of legislation which affect the use of these data sources – the Data Protection Act, and also any specific enabling legislation specifying use.

Legal status – electoral registers

There are currently two versions of the electoral registers which are used for different purposes: the 'full register' can only be used for administering elections, in the prevention and detection of crime, the selection of juries, or in checking applications for loans and credit. Credit reference agencies are entitled to buy the full register. The 'edited register' is available for sale to anyone, and can be used for any number of purposes. The 'editing' that takes place is essentially only the removal of individuals who have selected to opt out. This opt out is expressly permitted through the Data Protection Act.

The use and disclosure of the electoral registers is controlled through primary and secondary legislation. That which affects the primary and secondary use, and the supply and sale of both the 'full' and the 'edited' version of any register is the Representation of the People (Amendment) Regulations 2002,[5] but a more recent statutory instrument provides for the disclosure of 'full' electoral registers to the Department of Work and Pensions to allow a comparison against the data which it holds.[6]

Other legal gateways are in place for other administrative data. For example, the use of administrative data to maintain the Inter-Departmental Business Register is governed by a range of legislation including the VAT Act (1994), and the Finance Act (1969). Additionally the Office for National Statistics can use Information Sharing Orders (gained under section 47 of the Statistics and Registration Service Act 2007) to access data outside of these existing gateways, but these can only be used for purposes given in the Order.[7] Many further examples can be found in the Data Sharing Review (Thomas and Walport 2008) which also recommended simplifying the whole of the legal framework for data sharing.

Administrative Data – Benefits

In general, administrative data offer a significant base for analysis since it should cover the universe of relevant individuals – all people in receipt of benefit at any given point in time should be in the DWP CIS. Data collection methods are relatively unobtrusive, and are probably less resented by data subjects than surveys since subjects, in the main, benefit from their transaction with government. Administrative data usually have no attrition with associated problems for inference, and often information is captured in considerable detail, some of which is potentially unknown to the subjects, and would therefore be unable to be reported in a traditional survey. Furthermore, by their very nature, administrative data have the propensity to be highly up-to-date, especially when compared with other data sources.

Obviously for administrative data there is some marginal additional cost to the data creator or owner for reuse, but since, in the main, it is already used for the

5 Representation of the People (Amendment) Regulations 2002 (Statutory Instrument 2002 no. 1871).
6 The Electoral Registration (Disclosure of Electoral Registers) Regulations 2013 (SI 2013 no. 760).
7 An interesting discussion on the use of administrative data by ONS for statistical purposes can be found in the UK Statistics Authority Monitoring Brief 3/12 Creating official statistics from administrative data (16 March 2012).

production of official statistics this cost may be lower than expected. (It is worth noting that in many cases we only know details about these data on the basis of the statistics they are used to produce.)

UK administrative data have also been linked in the past to survey data, but under a wide variety of protocols, and usually with the consent of the survey participants. For example, the Longitudinal Study of Young People in England (LYSPE) has been linked to the National Pupil Database; the Millennium Cohort Study has been linked to hospital registration data, and the English Longitudinal Study of Aging has been linked to health and economic data from both the Department of Work and Pensions and the HMRC (Gray 2009).

Administrative Data – Problems

From a research ethics point of view administrative data are problematic since in general the data subjects have not provided consent for the use of their information whether personal or not. There may also be legal issues with sharing and linking, especially given that there may be multiple data controllers (multiple requirements for access) for any linked dataset. This is a function of having decentralised statistical organisations. A Ministry of Justice response to a Freedom of Information request for access to that part of the Offender's Index relating to people who were dead, and thus out of scope for protection by the Data Protection Act, stated "As a consequence of the *size of the database and the manner in which it is organised* and maintained we are unable to disaggregate any information the database might contain on individuals who are now deceased from the information that relates to the living (my italics)." The manner in which this index may be maintained (i.e., without explicitly reporting deaths) may indeed render the database disaggregatable, but its size and organisation should theoretically not affect this.[8]

From a research point of view administrative data don't cover the entire population: while these data often cover the universe of relevant individuals, non-participants are excluded, which may affect the manner in which the research is undertaken.

Administrative data may also suffer from internal inconsistencies. If a policy change is made which is applied to the underlying data in a programmatic manner, there is the potential that information relating to the period before that

8 'Offenders Index Database': www.whatdotheyknow.com/request/offenders_index_database_2

change is altered without note; if questionnaires alter significantly, there may be breaks in consistency within the data files.[9] The documentation for research purposes may be considered to be poor and the data "less prepared" than survey data prepared for secondary analysis but this may be a critique from researchers who have not encountered material which has not been processed in advance of their use.

The complete accuracy of some data is not even vouchsafed by their departmental owners. One footnote buried in a Freedom of Information Request to the Ministry of Justice states: "... it is important to note that these data have been extracted from large administrative data systems generated by the courts and police forces. As a consequence, care should be taken to ensure data collection processes and their inevitable limitations are taken into account when those data are used."[10]

Finally, researchers have almost no control over the content of administrative data, and the data may not fully meet the needs of the researcher.[11] However, overall, it is fair to say that, generally speaking, the 'content' problems associated with administrative data, are resolvable, and where they are not completely resolvable the potential benefits outweigh the problems.

Administrative Data Taskforce

The Administrative Data Taskforce (ADT) was set up in December 2011, after discussions between various stakeholder groups. The taskforce was comprised of a group of experts from government, research funders and the research community, and chaired by Sir Alan Langlands. Within twelve months the Taskforce published a report (Administrative Data Taskforce 2012) which made a series of recommendations including: the development of an Administrative Data Research Centre in each country of the United Kingdom, an information gateway, a Governing Board, a UK-wide researcher training and accreditation process, a strategy for engaging with the public, the creation of a generic legal

9 See, for example, Iwig et al (2013) or Daas (2009) for many other examples of potential inconsistencies.

10 See the linked Excel spreadsheet on 'Statistics on TV Licence Cases': https://www.whatdotheyknow.com/request/statistics_on_tv_licence_cases

11 Though, sometimes the contents of these administrative systems may spur research in unknown directions. The DWP's Income Support Computer System (ISCS) and Jobseeker's Allowance Payment System (JSAPS) both includes a relationship indicator which allows for an individual to be a polygamous partner. www.gov.uk/government/uploads/system/uploads/attachment_data/file/221378/foi-2185-2011.pdf

gateway for data access, and the provision of sufficient resources to establish and maintain these activities.

The government response was published in mid-2013 stating explicitly that: "Data collected and held by Government is a unique resource. Unlocking that resource has the potential to develop new understanding and insights both between different fields of study and over time" (Department for Business Innovation and Skills 2013). This response also emphasised the importance of "building on existing activities, infrastructure and systems where feasible in developing a new UK-wide approach" and attempting to develop "the infrastructure in a way that maximises the potential benefits to both government analysts and the wider research community", since "both will ultimately benefit citizens" and "ensuring that the full breadth of data sources held in administrative systems where they have analytical value are accessible for research purposes." This hugely positive response also recognised that the proposals made by the Administrative Data Taskforce would deliver a "significant step change in harnessing the full potential of administrative data to illuminate policy options and to monitor progress in policy delivery." The government response also noted that any infrastructure would need to reflect the devolved nature of government across the United Kingdom, and that the Administrative Data Research Centres would be required to provide access to data across these different administrations.

Around the same time, an Administrative Data Taskforce Technical Group was set up to progress these plans. This Technical Group examined and reported on best practice across a range of issues, such as secure access procedures including safe settings, researcher accreditation, research project accreditation and facilities accreditation. This group also made a series of recommendations about the structure of the proposed Administrative Data Research Centres. The group reported in early Summer 2013 (ADT Technical Group (2013)). Simultaneously the Economic and Social Research Council and the government were in discussions, which culminated in David Willets, Minister for Universities and Science formally announcing the ESRC's Big Data Network on 10 October 2013.

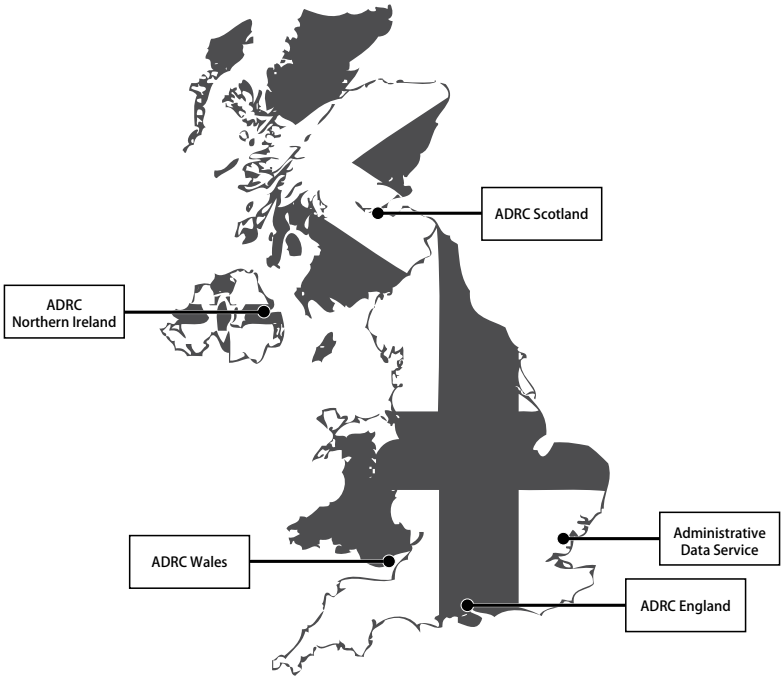


Figure 1: The Administrative Data Research Network

Administrative Data Research Network

It was considered key to the success of leveraging access to administrative data to set up a Administrative Data Research Network (ADRN), with four centres – one each for Scotland, Northern Ireland, Wales and England and a co-ordinating and 'independent' Administrative Data Service. Four Administrative Data Research Centres were required to reflect the devolved nature of some of the government functions in the UK as well as to maximise the key partnerships with the different National Statistical Authorities (which have the opportunities for acting as, or commissioning Trusted Third Parties services) and existing related data investments, especially those related to health research and the Longitudinal Surveys. This distributed network requires both UK-wide coordination and UK-wide governance, to be respectively provided by the Administrative Data Service and the UK Statistics Authority. The concept of a 'network' was considered vitally important by all participants, as different research structures

had different levels of experience and expertise in dealing with the management of researcher access to both administrative and linked data, and leveraging this experience and expertise across the network will be vital to maximise investment in the network.

Administrative Data Service

The contract for the Administrative Data Service was awarded to the University of Essex in partnership with the Universities of Manchester and Oxford in October 2013. The stated vision is to "establish a seamless research infrastructure which places the UK at the forefront of research using administrative data resources". The Administrative Data Service has a series of key coordination activities across the whole of the Network including the cross-Network communications, researcher engagement, public engagement, internal coordination and the provision of a user service. The Administrative Data Service will also have responsibilities for data and linkage commissioning, and critically, engagement with the Data Owners.

The Administrative Data Service is also in the process of setting up cross-sector working groups to develop protocols and standards for all the activities surrounding data access, including research / researcher accreditation, safe settings, data confidentiality and Statistical Disclosure Control. The Administrative Data Service will also have responsibility for the specification of metadata for resource discovery of any administrative data product which may be available more widely than the Administrative Data Research Network. Some of these activities may alter in the coming months.

Citizen participation

Citizen participation is essential, as citizens (data subjects) need to be aware of how data about them are controlled, of how those data about them are being used and for what purpose, and most importantly, of how the results of these analyses benefit society.

At the time of writing the whole of the ADRN's activities are in progress and one of the key achievements is the publication of a specially commissioned report entitled *Dialogue on Data* (Cameron et al. 2014) which explores the views of the public (or data subject) on the use of these types of data for research purposes.

Infrastructure issues

Not all the issues relating the use of administrative data for research have been resolved by these initiatives, and while the problems are known, there are delicate issues which require pragmatic resolution. For example, the Administrative Data Research Network is based on four centres in the different countries of the United Kingdom, and thus the initiatives may become more 'national' than cross-UK. (The National Assembly for Wales, for example, has devolved responsibility for health services, social welfare and education (amongst others) in Wales. The Northern Ireland Assembly has (amongst many 'transferred matters') responsibility for health and social services as well.) It is not clear whether or not researchers in one administration may benefit unduly as a consequence of the arrangements. Will, say, English researchers be able to access administrative data relating to education in Wales as easily as their Welsh counterparts? Sharing data beyond the boundaries of the United Kingdom is another perceived problem which is not currently being addressed since there are more pressing problems at home to be resolved.

It is also not clear how far the existing battery of methods used by social scientists and other researchers will be of value for social science research using administrative data; administrative data are not driven by research issues per se and the social sciences are perceived as methodologically underdeveloped to deal with the special qualities of administrative data. Capacity building in skills for data analysis is required to maximise the benefit of making these data accessible. There are also problems from a data sharing point of view; the existing paradigm across the social science community is to maximise the potential for sharing data across as wide a spectrum of stakeholders as possible. With administrative data the perceived best practice amongst a number of data owners seems to be to destroy data once it has been used for the research it has been created for, leaving little room for validation or transparency.

At the time of writing (March 2014) it is early days of the experiments taking place in the United Kingdom. There remain a number of obstacles to be hurdled, but the speed of development is encouraging. The United Kingdom has lagged behind some countries in Europe in the use of administrative data *within* government and while there have been a number of initiatives outside government these have been limited. The Administrative Data Research Network has the potential to raise the UK to the forefront of activity in this area.

References

All references available 24 March 2014.

Administrative Data Taskforce (2012): Improving Access for Research and Policy. Swindon: ESRC. www.esrc.ac.uk/_images/ADT-Improving-Access-for-Research-and-Policy_tcm8-24462.pdf.

ADT Technical Group (2013): Structuring the Administrative Data Research Network. www.esrc.ac.uk/_images/ADT_Technical_Group_Report_tcm8-26530.pdf.

Cameron, Daniel/Pope, Sarah and Clemence, Michael (2014): Dialogue on Data. Exploring the public's views on using administrative data for research purposes. www.esrc.ac.uk/_images/Dialogue_on_Data_report_tcm8-30270.pdf.

Daas, Piet/Ossen, Saskia/Vis-Visschers, Rachel and Arends-Tóth, Judit (2009): Checklist for the Quality evaluation of Administrative Data Sources. Statistics Netherlands Discussion Paper 09042. www.cbs.nl/nr/rdonlyres/0dbc2574-cdae-4a6d-a68a-88458cf05fb2/0/200942x10pub.pdf

Department for Business Innovation and Skills (2013): Improving Access for Research and Policy. The Government Response to the Report of the Administrative Data Taskforce. www.gov.uk/government/uploads/system/uploads/attachment_data/file/206873/bis-13-920-government-response-administrative-data-taskforce.pdf.

Department for Communities and Local Government (2011): The English Indices of Deprivation 2010. www.gov.uk/government/uploads/system/uploads/attachment_data/file/6871/1871208.pdf.

Frazer, David (2010): Department for Work and Pensions. Statement of Administrative Sources. www.gov.uk/government/publications/statement-of-the-administrative-sources-of-dwp-statistics.

Gray, Michelle (2010): A review of Data linkage procedures at NatCen. www.natcen.ac.uk/media/205504/data-linkage-review.pdf.

HMRC (nd, but 2011): HMRC Statement of Administrative Sources. www.gov.uk/government/uploads/system/uploads/attachment_data/file/252030/cop-admin-sources.pdf.

Iwig, William/Berning, Michael/Marck, Paul and Prell, Mark (2013): Data Quality Assessment Tool for Administrative Data. www.bls.gov/osmr/datatool.pdf.

Shakespeare, Stephan (2013): The Shakespeare Review. An Independent Review of Public Sector Information. www.gov.uk/government/uploads/system/uploads/attachment_data/file/198752/13-744-shakespeare-review-of-public-sector-information.pdf.

Thomas, Richard and Walport, Mark (2008): Data Sharing Review Report. www.connectingforhealth.nhs.uk/systemsandservices/infogov/links/datasharingreview.pdf.

UK Statistics Authority (2012): Creating official statistics from administrative data. Monitoring Brief 3 (12). www.statisticsauthority.gov.uk/reports---correspondence/correspondence/letter-from-sir-michael-scholar-to-rt-hon-francis-maude---administrative-data---16032012.pdf.

3.2 The HMRC Datalab: Sharing administrative and survey data on taxation with the research community

Daniele Bega (HMRC, UK)

In May 2011, HM Revenue & Customs (HMRC) launched the Datalab, a data centre through which, for the first time, UK researchers have been able to access their tax authority's administrative and survey data sources. This initiative allows the analysis of anonymised information in a secure environment, with the aim of producing high quality studies that benefit both HMRC and the research community. This article provides a description on how the Datalab works, highlighting some of the legal and practical challenges this initiative has faced since it was created.

Introduction

HM Revenue & Customs (HMRC) is the UK's tax authority[1], responsible for administering and collecting a wide range of taxes as well as helping families and individuals with targeted financial support, such as tax credits and child benefits. The datasets HMRC holds are among the largest and richest processed by the UK government. The organisation has a unique relationship with businesses and individuals, which is reflected in the scope and depth of the information HMRC processes and creates. All information that the department collects is subject to a strong duty of confidentiality to protect customers' privacy and safeguard the effective operation of the tax system.

In recent years, there has been plenty of discussion around HMRC's information. The increasing interest from the research community and the launch of the transparency agenda in 2010[2] have resulted in a number of requests for HMRC

1 HMRC is responsible for collecting the bulk of tax revenue. The Department manages: Income, Corporation, Capital Gains, Inheritance, Insurance Premium, Stamp and Petroleum Revenue taxes; Value Added Tax (VAT); Excise and Customs duties; Environmental taxes – Climate Change and Aggregates Levies, Landfill Tax and Air Passenger Duty; National Insurance Contributions; Bank Levy; Tax Credits; Child Benefit and the Child Trust Fund; Health in Pregnancy Grant; enforcement of the National Minimum Wage; recovery of Student Loan repayments, statutory payments and provision of the Contracts Finder portal.

2 transparency.number10.gov.uk/

to share more of its information with the general public. The UK government has recognised the benefits of transparency as a way of promoting social and economic growth, and HMRC has been supportive of these aims. Therefore, the department has fully embraced this challenge and has worked to devise solutions to share its data more widely, within the current legal framework.

HMRC's legislation

The legislation that governs HMRC provides a framework for how the organisation is able to share its information. HMRC is a non-ministerial department, established by an act of Parliament, the Commissioners for Revenue and Customs Act (CRCA) 2005, replacing two existing Government departments: Inland Revenue and Customs and Excise. CRCA vested responsibility for the administration of the tax system in Commissioners appointed by the Queen, drawn from the department's top management. HMRC reports to Parliament through a Treasury Minister, the Exchequer Secretary, who is responsible for strategic oversight of the UK tax system, including direct and indirect, business, property and personal taxation.

CRCA prohibits the disclosure of all information held by HMRC in connection with its functions, except in specified circumstances. This statute reflects the importance placed on 'taxpayer confidentiality' by Parliament when the department was created. The effective functioning of the tax authority depends critically on its customers being able to trust that information held on them would be appropriately protected and therefore disclosed only in controlled, limited circumstances.

The legislation also includes additional protection for information relating to an individual or legal entity, in the form of a criminal sanction for unlawful disclosure of identifying information[3]. The prohibition on disclosure does not apply where the exceptions ('gateways') set out in CRCA apply. The main gateways are highlighted in Table 1:

- Where there is UK or EU legislation that permits disclosure ('legislative gateway');
- With the consent of the subject(s) of the information; or
- Where the disclosure is made for the purpose of an HMRC function ('functions gateway').

3 Either disclosing the identity of a taxpayer or allowing their identity to be deduced.

Table 1:

Legislative gateways	HMRC shares information through approximately 250 information gateways with a large number of third parties, including 25 government departments, 50 agencies, the devolved governments, Local Authorities and other countries (e.g. through double taxation agreements). The terms of each information gateway are specific as to the type of information that can be disclosed and its purposes.
Consent	HMRC may disclose information with the consent of each person to whom the information relates. Under the Data Protection Act 1998, consent should be a positive indication of the wishes of the data subject. Consent should also be freely given, fully informed and specific to the circumstances in which it is sought.
Functions gateway	The prohibition on disclosure does not apply to a disclosure that: • is made for the purposes of a function of HMRC; and • does not contravene any restriction imposed by the Commissioners (at present, no restriction has been imposed). HMRC's functions HMRC's functions are the powers and duties of the department's Commissioners and officers set out in CRCA (or in other legislation), primarily the assessment and collection of tax and the payment and management of tax credits. As a statutory department administered by its own Commissioners, HMRC has no common law powers and therefore less flexibility as to what it may do compared to ministerial departments. Ancillary functions Examples of HMRC's ancillary functions include: promoting publicity about the tax system; establishing advisory bodies; entering into agreements; and acquiring and disposing of property. Disclosure for an ancillary function is permitted where there is a sufficiently close connection between the purpose for which the disclosure is made and a core HMRC function.

Table 1: Exceptions to HMRC's prohibition on disclosure

HMRC's data sharing with the general public

HMRC is committed to being as transparent as possible while complying with its statutory duty of confidentiality. It is one of HMRC's functions to publish information that promotes public understanding of its work and increases accountability and public confidence. For this purpose, the department shares its information with the general public through various channels. All information releases take into account the obligation to collect tax and the impact that publication will have on this function. This includes the need to protect sensitive and personal information provided by individual taxpayers in order to encourage openness and promote voluntary compliance.

Table 2 illustrates the different ways in which HMRC currently shares its information with the general public.

Information	Description
National Statistics	A wide range of HMRC's information is released in aggregated form under a programme of publication of Official and National Statistics. This information is regularly released on the department's statistics website[4] and the Office for National Statistics Hub[5]. Statistics cover HMRC's main work from collecting tax to paying out personal tax credit and Child Benefit. In total there are around 100 annually produced statistical products.
Transparency datasets	HMRC publishes information about its performance and activities on its Transparency webpage on www.gov.uk and on www.data.gov.uk, the single online portal for central and local government data. The release of transparency datasets includes information on the department's spending, use of government procurement cards, meetings with external organisations and HMRC's organisational structure.

4 www.gov.uk/government/organisations/hm-revenue-customs/about/statistics
5 www.statistics.gov.uk/hub/index.html

UK Trade Info	The Royal Statistical Society (RSS) award-winning website (www.uktradeinfo.com) provides access to information, guidance and tools relating to trade statistics from the EU-wide Intrastat survey and Customs import and export procedures. This website was designed after extensive consultation with users and offers availability of a wide range of datasets, including capability for data visualisation with time series, charting, mapping and sharing for social media users. HMRC also makes available the datasets used to produce Overseas Trade Statistics and information on importers details as open data.
Freedom of information requests	The Freedom of Information (FOI) Act imposes a duty on public authorities to supply requested information to any applicant. This means that HMRC is legally obliged to release its data in response to FOI requests, with some exemptions. For example information relating to an identifiable individual or legal entity cannot be published. As part of the FOI Act, HMRC also produces a publication scheme that specifies the classes of information which the public authority publishes or intends to publish[6].
Parliamentary Questions (PQ)	HMRC regularly answers questions from Members of Parliament. This information is published on the UK Parliament website[7]
Data Catalogues	HMRC makes available an inventory of its datasets as part of two transparency-related initiatives: the National Information Infrastructure[8] and the HMRC Data Catalogue[9].
Research Data Centres, such as the Datalab	Finally, HMRC releases granular, anonymised information via the Datalab[10], enclaves (such as ELSA – the English Longitudinal Survey for Aging) and the UK Data Services (Survey of Personal Incomes). All these initiatives are subject to safeguards to protect taxpayers' information.

Table 2: HMRC's information sharing with external organisations

6 www.gov.uk/government/publications?departments%5B%5D=hm-revenue-customs&publication_type=foi-releases
7 www.parliament.uk/business/publications/hansard/
8 www.gov.uk/government/publications/national-information-infrastructure
9 www.gov.uk/government/publications/open-data-strategy--2
10 www.hmrc.gov.uk/datalab - soon to be moved to www.gov.uk

The HMRC Datalab

The Datalab is a secure environment where independent research institutions can access anonymised taxpayer and customs data for research and analysis purposes, free of charge. This initiative was launched in May 2011 and forms an integral part of HMRC's commitment to the Transparency Agenda.

HMRC decided to undertake this project in the context of similar initiatives both in the UK and overseas. In the UK, in January 2004 the Office for National Statistics had launched their Virtual Microdata Laboratory (VML)[11], a research data centre providing secure access to confidential business survey data for research purposes. At the same time, in the US, a large body of research relevant to both tax policy (in particular business taxation) and the administration of the tax system had emerged under various arrangements allowing academics to access confidential taxpayer data. A handful of examples included: analysis of the determinants of tax evasion; quantifying compliance and non-compliance; the relationship between compliance and company residency; the impact of tax and other incentives on the self-employed; analysis of how taxes and transfer programmes are assisting low-wage earners.

The idea of setting up the Datalab developed gradually and involved several discussions with leading academic institutions, HMRC data controllers and Ministers. The Datalab was created with the intention of exploiting the potential of administrative data at micro level for new research. By making taxpayer-level data available in a safe way, HMRC aimed to foster new areas of research using improved methodologies, ultimately benefitting society as a whole.

Improving the supply of academic research into tax policy and administration was seen as an important step in developing the evidence base for good policy-making, alongside HMRC's in-house analysis and externally commissioned research. One of the main sources of evidence for enhancing the understanding of HMRC's customers and their behaviour is administrative data on taxpayers' incomes and the taxes they pay. Capturing and preparing such data for analytical purposes makes it possible to derive estimates of how taxpayers respond to incentives and other features of the tax regime.

Since its creation, the Datalab has been very successful. Between May 2011 and October 2013, 26 projects have been approved with about 40 researchers

11 www.ons.gov.uk/ons/about-ons/index.html

accredited. In July 2013, The Economist wrote an article on the Datalab, praising the openness of this initiative: "[The Datalab] could revolutionise the way Britain's economic policies are designed."[12]

Datalab principles

In order to reduce the risks associated with public access to microdata, the Datalab has concentrated on five key principles, modelled on the ONS VML[13] (Table 3):

Aspect	Aim	Criteria
Safe projects	Projects are undertaken for HMRC's functions under CRCA and will not have a risk that might damage HMRC's operation of the tax system.	Projects must have a valid statistical purpose, a demonstrable link to HMRC functions and be carried out by researchers who take ultimate responsibility for the analyses and inferences made.
Safe data	Information on individual taxpayers cannot be directly identified.	Information in the Datalab must be anonymised.
Safe people	Researchers can be trusted.	Researchers must be accredited and associated with an approved research institution; there should be no conflict of interest.
Safe settings	Deliberate and accidental removal of data is not possible.	The IT environment must be inherently secure (that is, preventing the removal of data without the agreement of HMRC officials).
Safe outputs	Approved outputs do not contain any disclosure risk.	Disclosure control methods are designed for Datalab and outputs to prevent identification of taxpayers.

Table 3: Datalab Principles

12 www.economist.com/news/britain/21582011-surprising-findings-new-stash-government-tax-data-mining-leviathan

13 Ritchie, F. (2006): Access to business microdata in the UK: Dealing with the irreducible risks. In: Work session on statistical data confidentiality 2005, UNECE/Eurostat, 239–244.

The Datalab follows HMRC data protection policies and there are restrictions on working practices to safeguard taxpayer confidentiality. In order to be granted access to the Datalab, research must serve one of HMRC's functions under the Commissioner for Revenue and Customs Act 2005. Any output from analysis in the Datalab is subject to statistical disclosure control checks before release. These measures aim to ensure that the outputs are sufficiently aggregated and that no taxpayer's information is directly or indirectly identifiable.

Datalab research is not commissioned by HMRC. Research proposals are submitted by researchers working independently and findings are subject to peer review.

HMRC's model is based on the principle of 'safe people', which is central to this framework. Researchers must come from 'trusted' organisations, where HMRC has assessed there is no conflict of interest. Through a thorough accreditation process, involving service level agreements and a training programme, researchers are made aware of the consequences of abusing this trust. Sanctions apply to any misuse, ranging from withdrawal of access to the Datalab to a potential criminal sentence if researchers deliberately attempt to identify taxpayers and disclose this to others.

Operation of the Datalab

The Datalab team manages the day-to-day running of this centre and is responsible for processing applications and bookings, and providing data support to researchers. A specialised IT group oversees the environment and ensures its operation is in line with HMRC's security requirements.

Decisions relating to technical aspects of the projects are made by the Microdata Release Panel, including output validation and statistical disclosure control checks. This body comprises of HMRC data experts and legal and tax professionals, who provide advice on methodological aspects of the projects and comment on the feasibility of research proposals.

The Datalab Committee sets the strategic direction of this initiative and assesses whether research proposals can be undertaken from a practical and legal point of view. Senior officials in HMRC and HMT are members of this group.

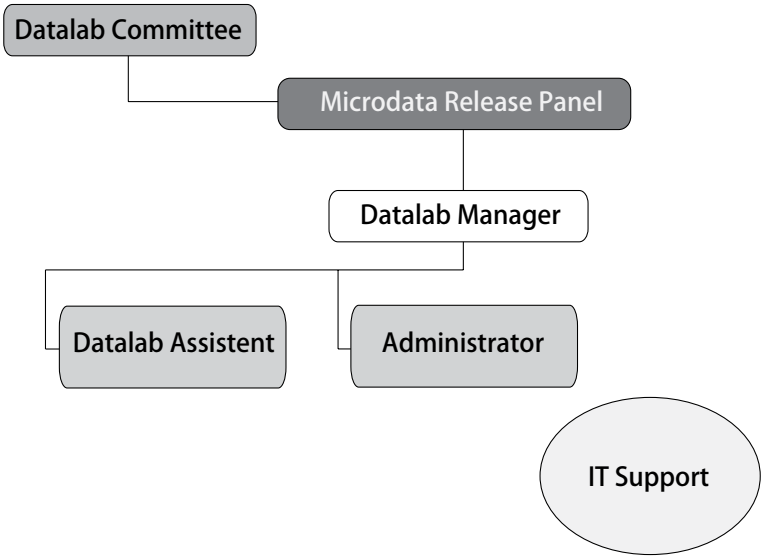


Figure 1: Datalab governance

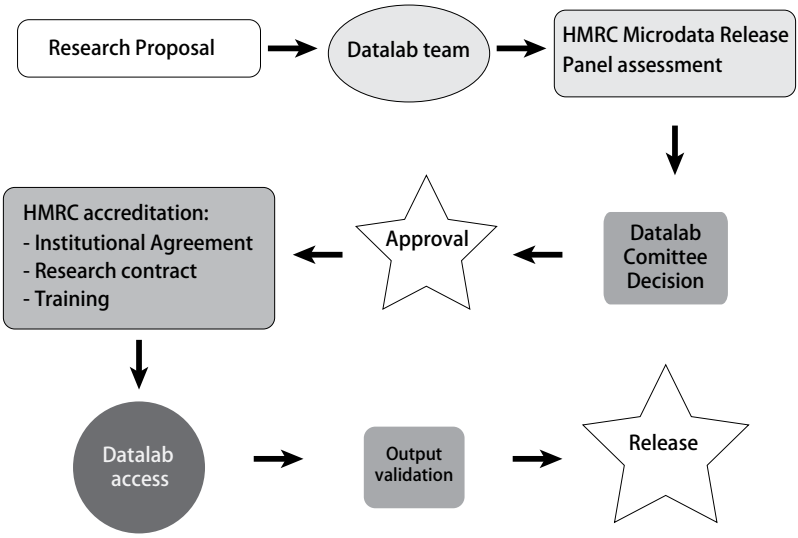


Figure 2: The Datalab in practice

The Datalab in practice

Figure 2 highlights how the Datalab works in practice. Key features of this model are the involvement of the Datalab Committee and the assessment by the Microdata Release Panel, as well as accreditation of the researchers.

Accreditation of researchers

In order to use the Datalab, researchers are required to undertake a training course, delivered by HMRC officials, consisting of four modules:

- Introduction to the Datalab
- Keeping data safe
- Statistical disclosure control
- Bookings and outputs

The aims of the course are to make researchers aware of HMRC's legislation and explain how the Datalab operates, including procedures on how to request outputs and the rules of the IT room. After two years, researchers need to undertake a refresher course.

In addition, users and their institutions are required to sign a service level agreement with HMRC setting out their obligations. The agreement covers the arrangements under which a project can be carried out in the Datalab, the duty of care the researchers must exercise in holding and releasing the information and the consequences of failure to comply with the terms of the agreement.

The Datalab environment

The Datalab can only be accessed on HMRC's premises at Bush House in London. The safe IT room is constantly monitored by HMRC staff and is open 9am to 5pm from Monday to Friday, by appointment only.

The Datalab is continuously in development, to keep hardware and software up to date and expand the datasets available. At the time of writing, the Datalab comprises eight computers (64-bit PCs, 2 with 24GB and 6 with 32GB RAM) on a standalone network on a multi-terabyte server. A wide range of specialist analytical software packages are available for researchers.

Datasets available at the moment include[14]:

- Compliance Perceptions Survey
- Corporation Tax returns
- HMRC Customer Survey
- Pay as you earn (PAYE) data
- Self-Assessment
- Stamp Duty Land Tax
- Survey of Personal Incomes (Public Use Tapes)
- Tax Credits
- Trade Statistics
- Value Added Tax

HMRC uploads information based on the requests received by researchers. For this purpose the Department releases a data catalogue[15] on the HMRC website and allows applicants to make requests for data that is currently not lodged on the Datalab.

Datalab publications

An example of organisations that have used the Datalab is available in Table 4.

Research Institutions
Oxford University
London School of Economics
Nottingham University
Imperial College
Warwick University
Institute for Fiscal Studies
The Department for Business Innovation and Skills
Essex University

Table 4: Institutions that have used the Datalab

14 A number of lookup tables and metadata are also provided.

15 www.gov.uk/government/uploads/system/uploads/attachment_data/file/89260/implementation-plan-catalogue.xls

A number of academic papers have already been produced using the Datalab. A few examples are listed below:

- Corporation Tax in the United Kingdom, Said Business School Oxford University (2011)[16]
- The Elasticity of Corporate Taxable Income: New Evidence from UK Tax Records, Said Business School Oxford University (2013)[17]
- Housing Market Responses to Transaction Taxes: Evidence from Notches and Stimulus in the UK, London School of Economics (2013)[18]
- Optimization Frictions in the Choice of UK Flat Rate Scheme of VAT, London School of Economics (2013)
- The Investment Effect of Taxation: Evidence from a Corporate Tax Kink, London School of Economics (2013)[19]

The Department also allows collaborations with overseas organisations where a UK-based institution is willing to sponsor.

Conclusions and next steps

The Datalab has been a very successful initiative that is continuing to grow. New challenges lie ahead for HMRC in this evolving landscape. A number of data sharing initiatives are currently being undertaken in the UK which will likely have an impact on the way the Datalab operates.

In the summer of 2013, HMRC undertook a consultation on 'sharing and publishing data for the public benefit'[20]. This consultation generated many helpful, informative and constructive responses on whether HMRC should increase the scope for the department to share non-identifying information (that is, information that does not relate to identifiable individuals or legal entities), and on proposed safeguards.

16 www.sbs.ox.ac.uk/sites/default/files/Business_Taxation/Docs/Publications/Reports/corporation-tax-in-the-uk-feb-2011.pdf
17 areas.kenan-flagler.unc.edu/conferences/2013cfea/Documents/EstimatingElasticity.pdf
18 economics.stanford.edu/files/Kleven9_24.pdf
19 www.sbs.ox.ac.uk/sites/default/files/Business_Taxation/Events/conferences/symposia/2013/brockmeyer.pdf
20 www.gov.uk/government/consultations/sharing-and-publishing-data-for-public-benefit

The creation of the UK Administrative Data Network[21] in October 2013 has also prompted the Datalab to develop strategies to collaborate with other newly formed Administrative Data Service and Administrative Data Research Centres and consider new mechanisms for sharing information with the research community.

Finally, the future operation of the Datalab will take into consideration the development of international data sharing initiatives, such as the Data Without Boundaries[22] programme. HMRC will aim to work in partnership with tax administrations and data research centres across the world to share expertise and keep up to date with progress in this area.

21 www.esrc.ac.uk/_images/ADT-Improving-Access-for-Research-and-Policy_tcm8-24462.pdf
22 www.dwbproject.org/

3.3 International Access to Administrative Data for Germany and Europe[1]

Stefan Bender, Anja Burghardt, and David Schiller (IAB)

Introduction

In the last years, access to research data has made a lot of progress in EU countries. Despite developments such as "Data without Boundaries", transnational access to confidential microdata remains complicated and needs improvement.

The first part of the paper describes the modes of (international) access to highly sensitive German administrative labour market data and how this international access is expanded within the "research data centre in research data centre" (RDC-in-RDC) approach. In the second part, we give a broader view of international access to official microdata in the EU. Starting with a brief overview of the EU-funded project "Data without Boundaries" (DwB), we present a possible roadmap for international access in Europe and beyond.

International access to German microdata

The German "research data centre movement" is a fairly recent development (see KVI 2001 or Bender et al. 2011) with little more than 10 years of experience. Other countries, which often have less stringent or very different data protection legislation, have a much longer tradition in operating research data centres (RDCs). Nevertheless, Germany is a very interesting example, because it progressed from almost no access to providing systematic access to microdata in less than 15 years. The German Data Forum (RatSWD) has and will play a decisive role in this development (www.ratswd.de/en).

This independent body of empirical researchers and representatives of important data producers has succeeded in opening and improving access to existing data, as well as creating an increased synergy between researchers and data producers. By the end of 2013, the RatSWD had accredited 27 RDCs.

1 This work made possible by the European Commission Seventh Framework Programme (grant agreement no. 262608) funded project "Data without Boundaries" (DwB).

Access through the FDZ of the BA at the IAB

German administrative labour market data is of great value to research in the fields of economics, sociology, and related disciplines. In order to provide access to such data the Research Data Centre (FDZ) of the German Federal Employment Agency (BA) at the Institute for Employment Research (IAB) was established in 2004 (fdz.iab.de/en.aspx).

The FDZ facilitates access to survey and administrative labour market data for non-commercial empirical research. Data are collected from the social security notifications and the internal processes of the BA as well as from surveys carried out by the IAB. The combination of administrative data with other data (like surveys, commercial data, big/smart data) will significantly expand research possibilities. To this end, the German Record Linkage Center (GRLC) was established in 2011 as a service centre which uses record linkage techniques to link different data sources and produces innovative research data (www.record-linkage.de). Funded by the German Research Foundation (DFG), the German Record Linkage Center has a service facility located at the FDZ and conducts research on technical solutions mainly at the University of Duisburg-Essen. Among the provided services are, for example, offering individual support during the planning and realisation stages of data linkage projects, conducting data linkages as contract work, and updating and maintaining the record linkage software Merge Tool Box (Schnell et al. 2004).

There are different ways to make microdata accessible via the FDZ. Those means of access correspond with different security measures which are enforced to protect data sources with different levels of disclosure risk. For non-disclosing data sources such as Campus/Public-Use-Files, free download of the data is possible. Scientific-Use-Files are more detailed and are already being used for many research projects; those data files are available for controlled download for specific time-restricted, non-commercial research projects. The highly detailed data sources that are needed for a sophisticated research project with a high impact level are only accessible with strong restrictions. One option for the researcher is to come to a location of the data holder (e.g. Nuremberg) and to work with the data in a secured room (on-site access or guest stay). The other option is to use job submission for indirect access from a distant location. When using job submission, the user sends his or her inquiries to the data holder. The calculations are carried out on the servers of the data holder and the results are sent back after output control.

The Research Data Centre in Research Data Centre Approach

The highly detailed data of the FDZ can only be stored in the facilities of the BA in Nuremberg due to data security reasons. It is therefore, necessary for researchers wanting to work with these data sources to travel to the location of the data holder. The same applies not only to the FDZ, but to every confidential data source all over Europe.

The central idea of the Research Data Centre in Research Data Centre (RDC-in-RDC) approach is to enable data access from other RDCs or institutions (guest-RDC) which share comparable standards as the RDC where the data are actually stored (data-RDC), but which are located at different sites. In this case, it does not matter whether the guest-RDCs are located in Nuremberg, Germany or abroad, because all RDCs have a common standard for accessing data. The only difference is that the guest researcher's room is not at the local data-RDC (for instance in Nuremberg) but at another guest-RDC. The guest-RDCs can be any institution that fulfils the FDZ security requirements (Bender and Heining 2011). The guest-RDC is responsible for physical access control to a secure room and makes sure only researchers with a valid contract are able to access the facility. In addition, the guest-RDC staff monitors the researcher's activities in the secure room and safeguards adherence to the code of practice. The secure room is a separate room made specifically for accessing confidential data. It is equipped only with devices for accessing the remote data sources; access to any other source of information is not possible. From the secure room within the guest-RDC, a secured and encrypted connection is established to the secured servers behind the firewalls of the data-RDC. Within this secure remote desktop environment, researchers are able to access the secure servers of the data holder from the guest-RDC and to work with the confidential research data (browse, modify, run calculations, get output). The microdata stays on the secured servers at all times and only a live stream of the used graphical user interface is transferred to the device (screen) of the user.

By the end of 2013, the RDC-in-RDC-network consisted of eight guest-RDCs (see figure 1). This number is continually increasing with guest-RDCs, for example, at the UKDA at the University of Essex and Harvard University.

In the future, this approach could be used as a template and starting point for a network of guest-RDCs (Safe Centres). Such Safe Centres, based on mutually agreed security standards, could be located in every bigger region. Accordingly, researchers would not have to travel far. A secure network could connect those Safe Centres to different data-RDCs that are using the Safe Centres as access points without setting up access points by themselves (Brandt and Schiller 2013).

International access to decentralized European microdata: The "Data without Boundaries" (DwB) project

Funded through the European Commission Seventh Framework Programme (grant agreement no. 262608), "Data without Boundaries" (DwB) works on improving social science research in Europe. The project focuses on discussing, describing and promoting concepts, solutions and frameworks. 29 partners from the European Statistical System (10 National Statistical Institutes or statistical departments), the Council of European Social Science Data Archives (11 Data Archives) and the research community (7 universities and 1 private company involved in methodological research) are working together.

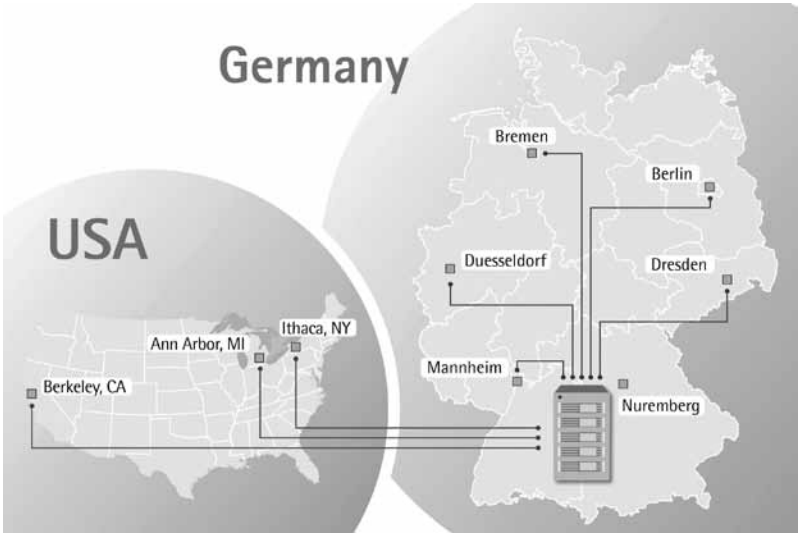


Figure 1: The RDC-in-RDC network of the FDZ (end of 2013)

Currently, comparative research projects based on microdata from different countries have to go through a process of multiple accreditations and deal with quite different technical and methodological environments. There exists a wealth of official microdata which is currently underused and siloed inside national, legislative, technical, and cultural borders. These can only be overcome by cooperation and political will. The main goal of DwB is to establish equal and easy access to official microdata for the European Research Area within a structured framework that equally distributes responsibilities and liabilities. The work of DwB will result in concepts and improvements for an European research accreditation process and a Europe-wide distributed remote access to confidential microdata of national datasets. DwB also takes part in the discussion about metadata standards (SDMX/DDI) with the aim of establishing a single point of information on research data in Europe.

Under the umbrella of DwB, it is possible to address the demand for a comprehensive and easy-to-access research data infrastructure in Europe which will enable cutting-edge research and reliable policy evaluations. DwB is in close contact with and aims to accommodate the needs of existing infrastructures such as the Council of European Social Science Data Archives (CESSDA) and the European Statistical System (ESS).

A roadmap for international access to decentralized European microdata

DwB produces evaluated concepts and pilots which are derived from discussions and project work, but because of its conceptual character, there will be no real implementation and therefore no live and running improvements towards comprehensive and easy access for researchers using the European research data infrastructure. The next logical step is the implementation of the findings of DwB within a European framework.

Such a framework was suggested in the DwB Work Package 4 (improving access to microdata). This framework is designed not only to connect researchers to different RDCs all over Europe, but also to connect researchers and project teams all over Europe. The conceptualization of this European infrastructure, called the European Remote Access Network (Eu-RAN), also incorporates models and proposals of other work packages in DwB such as a European Service Centre as an information platform for research data in Europe.

Future: European Remote Access Network (Eu-RAN)

The Eu-RAN will bring together researchers and research groups from all over Europe with data sources from all over Europe (Schiller 2013). The Eu-RAN is divided into the following main components: access points for using the network; the Single Point of Access (SPA) which comprises additional services, and the secured remote access network itself (see figure 2).

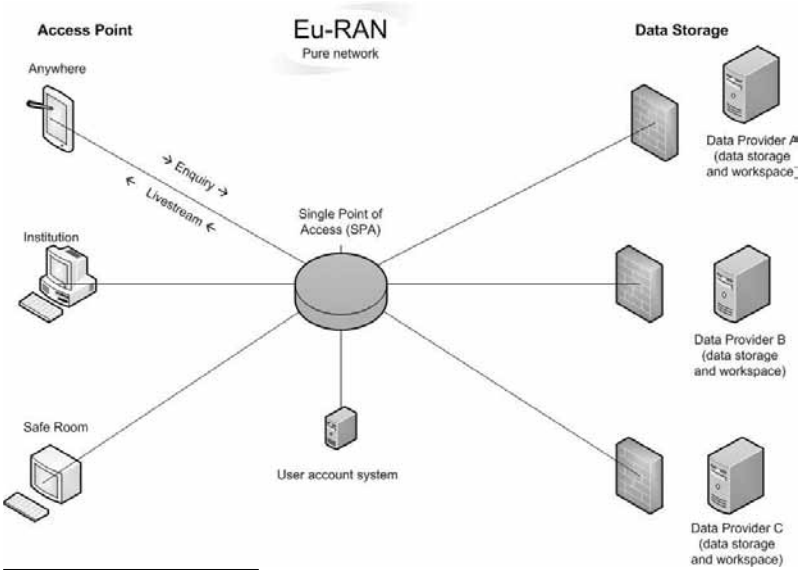


Figure 2: The Eu-RAN architecture

Regarding access points, the Eu-RAN will support different security levels. One is that of Safe Centres and the other are restricted access from universities or research institutes. In addition unrestricted access from anywhere is also possible. The latter allows only restricted access and limited functions of the Eu-RAN (for example use of communication tools or to browse metadata) without direct access to the restricted data. The Eu-RAN infrastructure will be able to support different needs depending on the security level of the accessed data.

The main task of the SPA is to check user authentication. When accessing Eu-RAN from one of the access points, the enquiries of the user are routed to the SPA where the authentication of the user (including location of access point) is checked. After being logged into the system, the user then has access to the data and services his or her contract is valid for.

Access points, SPA and data storage servers are connected by a secured remote access network. By using encrypted tunnels through the internet, all information are secure at any point of time. The data itself (confidential research data, outputs and information about the users) are not moved, they remain behind the firewalls on secured servers. Only enquiries from the access devices and pictures for the graphical user interface on the access device are transmitted through the encrypted tunnels. When using the RDC-in-RDC approach, a closed and secure working environment for research on confidential data is in place.

Beside the technical network, an organisational network of partners (data-owner, data-holder, data-user) is required and the Eu-RAN infrastructure needs to be based on contracts, agreements and, above all trust, between the partners.

Future: Single Point of Access (SPA) and incorporated service hub

There are already a number of remote desktop solutions being used in Europe (Report on the state of the art of current SC in Europe 2012; DwB deliverable 4.1). Those solutions enable access to the data sources of one specific data owner (National Statistical Institute, Data Archive, etc.). The establishment of Eu-RAN pools those existing solutions and enables users to access different data sources from a Single Point of Access (SPA) (see figure 2). The power of the central node, the SPA, is only fully utilised if support services for researchers are offered. Such services are controlled by a service hub working as a central device that host numbers of different services. Only a few of the potential services will be mentioned in order to give an impression of the service hub (see also figure 3).

One of the services will be the web portal that functions as the graphical user interface to use the whole Eu-RAN infrastructure; an additional service is the user account management system that deals with the authentication of users and where users can manage their contracts and data holders can manage the access rights of the user. In addition, tools supporting research, like editors, statistical packages, wikis, calendars; (data) documentation and interfaces are provided services within the SPA.

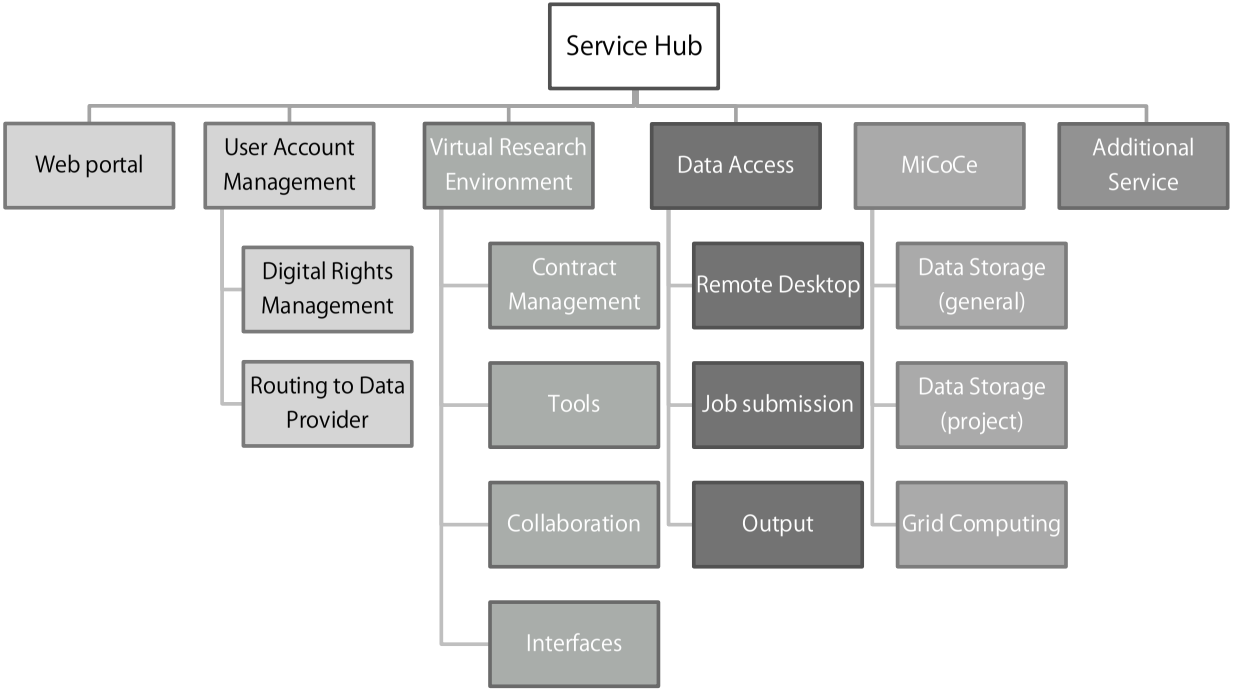


Figure 3: Services of the service hub

When working with data, good data documentation is required. Therefore, a European Service Centre for official statistics, as proposed by DwB (Report on concept for and components of European Service Centre for official statistics 2012; deliverable 5.1), should be hosted by the service hub. In the context of transnational access, two services (or groups of services respectively) are of crucial importance in order to enable collaborative work of researchers at different locations, and comparative analysis with data from different data sources. Those two crucial services are the Virtual Research Environment and the Microdata Computation Centre. Naturally, all of these services will not be usable without Eu-RAN running in the background.

Virtual Research Environments (VRE)

In general, Virtual Research Environments (VRE) are web portals providing services to users. Those services are connected to underlying databases. VREs can be technical secured and access can be restricted, if need be. Within such virtual environments researchers are able to use – for example – editors, calendars, wikis, forums or statistical software for their daily work. VREs are built to carry out scientific research in a community and they can be used as a platform for exchange between different disciplines or countries (Allan 2009; Carusi and Reimer 2010). By offering a central stored, digital working environment, a possibility for standardization and archiving (reproducibility of research) is opened up in the back end of the VRE infrastructure. For example program codes, generated in the VRE, can be re-used to replicate results and to build new research upon those results; software tools used in the VRE can be connected to documentation standards; and the complete project working material can be archived for further use.

While VREs are in general made to support collaborative work of research teams, also access to sources of confidential microdata can be incorporated into the VRE. For example, the above mentioned job submission or remote desktop solutions are examples.

Microdata Computation Centre for de-centralized data sources (MiCoCe)

In Europe, most of the confidential microdata have to be stored in the country, where the data was collected. Even if there is no explicit regulation, security requirements force data holders to keep the microdata in its country of origin (Tubaro et al. 2013). The RDC-in-RDC approach and other solutions, based on secure remote desktop connections, allow analysing data even from locations

abroad. There is a huge demand on a European level to analyse data from different countries simultaneously. The challenge is therefore to enable analysis with distributed data sources and, at the same time, without having to move the data. The solution to this problem could be to set up a Microdata Computation Centre for de-centralized data sources (MiCoCe). In the MiCoCe, enquiries to the distributed data sources will be sent from a central point and the single results per data set will be combined to a final result. According to that, no disclosing microdata will be moved in the MiCoCe, only non-disclosing (part-)results are moved to a secured central node. Solutions can come from the areas of statistics, grid/parallel computing and federated databases.

Summary and Outlook

Data access in Germany is now well established because there was and is a strong will among data providers, researchers and ministries (sponsoring and legal support) to offer access to highly sensitive microdata. By establishing the German Data Forum as an institution for supporting and developing the data infrastructure, Germany now serve as a blueprint for other countries.

German administrative data, as provided by the FDZ of the BA at IAB, is a powerful resource of knowledge discovery that can already be enriched by linking it to survey data, commercial or big data. The RDC-in-RDC approach, pioneered by the FDZ of the BA at the IAB, has shown that it is possible to grant transnational access to researchers outside Germany, despite the legal restrictions.

Although there are solutions like the RDC-in-RDC approach or secure remote access, transnational access to microdata is still in its infancy. Here, Data without Boundaries plays an important role. Under the umbrella of DwB, it is possible to address the demand for a comprehensive and easy-to-access research data infrastructure in Europe which will enable cutting-edge research and reliable policy evaluations.

But DwB is only the starting point for transnational access to microdata. The concept of a European Remote Access Network that builds the basis for transnational research in Europe will unleash the power of European data for cutting edge research in Europe by offering different security levels to serve the needs of data holders all over Europe, and by providing tools like virtual research environments and the Microdata Computation Centre that support transnational collaborative research teams and enable the use of distributed data sources.

Furthermore there should be a continuous dialogue ultimately leading to a roadmap for international access to sensitive microdata. Infrastructures like the Eu-RAN, VREs and the MiCoCe will support future research. Having an infrastructure in place that can deal with different kinds of data (from survey data to administrative data to big data) will also secure the future of research in the social sciences.

The next logical step is to make the proposed concepts and models come to life. The established culture of communication between archives, NSIs and the research community will be the basis for further development and will guarantee real implementation of the proposed infrastructures. This infrastructure will be developed from the technical current state of play in Europe and is aiming on the future needs of European research.

Establishing a trustworthy and functioning organisational network, which is supported by technical solutions, will help build the environment required for high-quality research with sensitive microdata in Europe and beyond.

References

Allan, R. (2009): Virtual Research Environments: From Portals to Science Gateways. Oxford: Chandos Publishing.

Bender, S./Himmelreicher, R./Zühlke, S and Zwick M. (2011): Access to Microdata from Official Statistics. In: German Data Forum (Ed.): Building on Progress – Expanding the Research Infrastructure for the Social, Economic and Behavioural Science, Budrich UniPress, Opladen & Farmington Hills, MI, 215-230.

Bender, S. and Heining, J. (2011): The Research-Data-Centre in Research-Data-Centre Approach: A First Step Towards Decentralised International Data Sharing. IASSIST Quarterly 35 (3), 10-16.

Brandt, M. and Schiller, D. (2013): Safe Centre Network – Need for Safe Centre to enrich European research. In: Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, Ottawa.

Carusi, A. and Reimer, T. (2010): Virtual Research Environment – Collaborative Landscape Study.

KVI – Kommission zur Verbesserung der informationellen Infrastruktur zwischen Wissenschaft und Statistik (2001): Baden-Baden: Nomos Verlagsgesellschaft.

Report on the state of the art of current SC in Europe (September 2012): www.dwbproject.org/deliverables (deliverable 4.1).

Report on concept for and components of European Service Centre for official statistics (April 2012): www.dwbproject.org/deliverables (deliverable 5.1).

Schiller, D. (2013): Proposal for a European Remote Access Network (Eu-RAN) – main components. In: Joint UNECE/Eurostat work session on statistical data confidentiality, Ottawa.

Schnell, R./Bachteler, T. and Bender, S. (2004): A Toolbox for record linkage. Austrian Journal of Statistics 33 (1–2), 125–133.

Tubaro, P./Cros, M. and Silberman, R. (2013): Access to Official Data and Researcher Accreditation in Europe: Existing Barriers and a Way Forward. IASSIST Quarterly 36 (1), 22–27.

4 Longitudinal Sciences and Bio-Social Sciences

John Hobcraft (University of York)

Understanding human behaviours and disentangling pathways to lifetime outcomes requires longitudinal data. There is growing evidence that early life experiences, including stress, poverty, and parenting have profound and lasting influences on later life outcomes. Increasingly there is recognition that the interplays between these lifetime experiences 'outside the skin' and biology are important for such understanding, including the roles of neuroscience, genomics and biomarkers for stress and disease. This chapter will evaluate the importance of ensuring widespread availability of comparable rich longitudinal data across the life course for Europe and the potential benefits of integrating biological measures into such studies.

4.1 Success – but Sustainability? The Survey of Health, Ageing and Retirement in Europe (SHARE)

Axel Börsch-Supan (MEA)

The Survey of Health, Ageing and Retirement in Europe (SHARE) is a distributed research data infrastructure for social scientists including demographers, economists, psychologists, sociologists, biologists, epidemiologists, public health and health policy experts, who are interested in population ageing, one of the important trends and grand challenges of the 21st century. SHARE's main aim is to provide data on individuals as they age and on their environment in order to analyse the process of individual and population ageing in depth.

Development

SHARE is a unique and innovative multidisciplinary and cross-national panel database of microdata on health, socio-economic status, and social and family networks of more than 80000 individuals aged 50 or over (Börsch-Supan et al. 2013). SHARE was created as a response to a communication by the European Commission calling to "examine the possibility of establishing, in co-operation with Member States, a European Longitudinal Ageing Survey". While its development process started only in 2002, SHARE has become one of the crucial pillars of the European Research Area.

Eleven countries contributed data to the 2004 SHARE baseline study. They are a balanced representation of the various regions in Europe, ranging from Scandinavia (Denmark and Sweden) through Central Europe (Austria, France, Germany, Switzerland, Belgium, and the Netherlands) to the Mediterranean (Spain, Italy, and Greece). Further data were collected in 2005/06 in Israel. Two new EU member states – the Czech Republic and Poland – as well as Ireland joined SHARE in 2006 and participated in the second wave of data collection in 2006/07. The survey's third wave, SHARELIFE, collected detailed retrospective life histories in thirteen countries in 2008/09. In 2010, the fourth wave – including a new social network module based on a name generator approach –

also included Estonia, Hungary, Portugal, and Slovenia. This adds up to 19 European countries contributing to the survey and collecting data for the fifth wave in 2012.

SHARE is harmonized with the U.S. Health and Retirement Study (HRS) and the English Longitudinal Study of Ageing (ELSA). Studies in Korea, Japan, China, India, and Brazil follow the SHARE model. Its scientific impact is based on its panel design which is able to grasp the dynamic character of the ageing process. SHARE's multidisciplinary approach provides us with a comprehensive picture of the ageing process. Rigorous procedural guidelines, electronic tools, and instruments ensure an ex-ante harmonized cross-national design.

Covering the key areas of life, namely health, socio-economics, and social networks, SHARE includes a great variety of information: health variables (e.g. self-reported health, health conditions, physical and cognitive functioning, health behaviour, use of health care facilities), bio-markers (e.g. grip strength, body-mass index, peak flow; and piloting dried blood spots, waist circumference, blood pressure), psychological variables (e.g. psychological health, well-being, life satisfaction), economic variables (e.g. current work activity, job characteristics, opportunities to work past retirement age, sources and composition of current income, wealth and consumption, housing, education), and social support variables (e.g. assistance within families, transfers of income and assets, social networks, volunteer activities) as well as social network information (e.g. contacts, proximity, satisfaction with network). Researchers may download the SHARE data free of charge from the project's website at www.share-project.org.

SHARE started as a predominantly centrally financed enterprise. This was crucial for harmonization across all member states. Data collection for waves 1-3 was funded primarily by the European Commission through the 5th and 6th framework programmes. Additional funding came from the U.S. National Institute on Aging. The SHARE data collection received additional national funding in Austria, Belgium, France, Ireland, and Switzerland. The SHARE data collection in Israel was funded by the U.S. National Institute on Aging, the German-Israeli Foundation for Scientific Research and Development, and the National Insurance Institute of Israel.

SHARE is part of the ESFRI (European Strategy Forum on Research Infrastructures) roadmap process. With the start of the fourth wave, SHARE became the first ERIC (European Research Infrastructure Consortium). National funding

is now prevalent, albeit with substantial support from the European Commission's DG Employment, Social Affairs and Equal Opportunities to the four new countries. Central tasks and coordination, however, are financed by the German Ministry for Science and Education, the European Commission, and the US National Institute of Aging. Maintaining such central funding remains important as SHARE's main aim is to provide harmonized data across all member countries in order to enable valid cross-national comparisons.

Management

SHARE is a collaborative effort of more than 150 researchers across Europe organized in multidisciplinary national teams and cross-national working groups. Management of such a distributed research infrastructure has its challenges and requires well-defined communication channels and institutions.

SHARE is coordinated centrally at the Munich Center for the Economics of Aging (MEA), which is part of the Max Planck Institute of Social Law and Social Policy, and is directed by the author of this article. Substantial central tasks were allocated to Italy (such as the creation of derived economic variables, sampling weights, and imputed variables) and the Netherlands (such as all software development).

Country team leaders form the backbone of the project. They include representatives from all disciplines. They are responsible for ensuring scientific excellence in their field, the timely execution of interim and final deliverables involved in the work packages of SHARE, and the adherence to standards and procedures in each country. The assembly of country team leaders prepares all relevant scientific and budget decisions that affect more than a single country.

The assembly proposes a small core management group or management board that advises the coordinator in all central, strategic, and financial questions. They include the four area coordinators of the SHARE-ERIC for social/family networks, health, healthcare, and economics, in order to represent the interdisciplinary breadth of the project.

The main governing body is the SHARE-ERIC Council. The governments of all ERIC member countries send out representatives whose voting rights are based on the extent of their financial commitments. All other SHARE countries may participate as observers. The Council decides on the budget and approves

the scientific action plan as well as the key personnel (the management board including the coordinator) as proposed by the assembly of country team leaders.

SHARE has also established a scientific monitoring board. Its main task is to monitor the scientific quality of the project. It includes eminent scientists from Europe and the US and represents all the disciplines included in SHARE (sociology, health, and economics). One member is a principle investigator of the U.S. Health and Retirement Study who ensures close co-operation with this parallel survey, and offers additional advice and guidance. A network of advisors (including three Nobel laureates) helps to maintain and improve the project's high scientific standards.

Success

Since the first public release of SHARE data in April 2005, SHARE has attracted more than 3,200 registered users with an unbroken, more than linearly increasing trend. SHARE only counts official applications on the project's website that result in the delivery of a SHARE data set. Students, research assistants and co-authors usually do not register as users themselves and instead get access to the data through their instructor's or main author's license. The actual number of users is therefore much higher. Although users include mainly scientists from Europe, researchers from the US are now the second largest user group after Germany – before Italy and the Netherlands.

In addition to four comprehensive volumes of first results from the SHARE baseline, longitudinal, and retrospective waves (Börsch-Supan et al. 2005, 2008, 2011, 2013) – which have been complemented by several national collections of findings – more than 850 articles based on SHARE data have been published in peer-reviewed journals and books as of December 2013. It is not only the sheer number of studies conducted within the four years after the first data release that is impressive, but also the quality of publications. One indicator that may be used for such an assessment is the number of articles published in journals covered by the renowned Social Science Citation Index. This currently amounts to over 350.

SHARE has produced some novel and surprising results. While it is well known that the countries in the European North have higher wealth and incomes than in the South, SHARE data has revealed many other north-south divides. In spite of the differences in longevity, individuals of both sexes from the North are

significantly healthier than those in the South (Avendano et al. 2005). Family networks also differ. Against many prejudices, intergenerational exchange is as frequent in the North as it is in the South – but it involves more financial transfers and less time (Hank 2007; Brandt/Haberkern and Szydlik 2009; Deindl and Brandt 2011; Brandt 2013; Brandt and Deindl 2013).

Retirement is usually seen as positive for individuals as they receive income support without the necessity to continue working, enabling them to enjoy more leisure and relieving them from stress at work. SHARE research, however, has also uncovered less pleasant side effects. Cognition – measured mainly by memory abilities such as delayed word recall – declines during retirement, and even faster for early retirees. This remains true even when one corrects for the fact that early retirees are on average unhealthier and have lower cognitive abilities than later retirees. This controversial finding has sparked an entire new strand of literature (Adam et al 2007; Bonsang et al. 2010; Rohwedder and Willis 2010; Mazzonna and Peracchi 2012; Börsch-Supan and Schuth 2013). An internationally comparable data set such as SHARE is essential for this research because it contains instruments such as the eligibility age for early and normal retirement, or similar institutional characteristics, that allow detecting causal pathways.

A genuine sense of community and intellectual exchange is created by the SHARE user conferences which are organized centrally for every wave on an international level. In addition, SHARE countries have also held local user conferences. Several networking activities are work packages sponsored by the European Commission. A central aim of these networking activities is to increase integration in several dimensions: among the participating countries, among the many disciplines involved in SHARE, and between users and designers. One formal work package develops and maintains standards and procedures for efficient communication within the network. Since SHARE has become very large, this is an important task: standards and procedures cannot simply be copied from purely national models. The SHARE data have become very complex. SHARE provides detailed synopses and concordances across member countries as a service for our users. A user-friendly database includes the imputation of missing variables, and the addition of process-generated geo-coded and environmental variables, data quality indicators, meta-statistics and para-statistics.

SHARE offers training to users, feeds back user reviews to the database managers, and maintains an external expert board that evaluates user access and services. The “user is client” philosophy is essential for the SHARE infrastruc-

ture. Even more than other European data collection efforts (such as EU-SILC or the European Labour Surveys), supporting research is the main task of SHARE, and the type of data collected is determined by the researchers – both members of the SHARE consortium and the user community at large.

SHARE also harmonizes training of interviewers across member countries using innovative training tools and control survey execution in the member countries according to the standards set in the harmonized training. We train professionals from survey agencies in our ICT-based survey technologies, which has significantly pushed the latest state-of-the-art forward in all SHARE member countries.

SHARE provides very fast and unlimited data access through our own data dissemination system. In addition, our membership in the CESSDA consortium of European micro-data archives, another ESFRI project, safeguards long-term stability and adherence to international data dissemination standards.

Sustainability?

So far, SHARE has been successful. We have expanded from 8 countries in the initial FP5 application to now 20 countries. We have enlarged the samples, attracted ever more users, and created a large number of peer-reviewed publications. The ERIC regulation 723/2009 has given SHARE a solid legal and governance structure. The management structure is now firmly established in the SHARE-ERIC statutes and has worked well for this distributed research infrastructure, which needs to balance central coordination with the cultural and institutional diversity of 20 member countries – a management task that is quite different from running a telescope or a bio-medical laboratory in a single place. The common and simplified procurement rules under the ERIC regulation are a great help in running SHARE efficiently across different legislations.

The ERIC construction is novel. As with all new developments, we are learning. The key idea to fund pan-European research infrastructures in a completely decentralized fashion has not worked out fully for SHARE. Not even half of the SHARE member countries have joined the ERIC and only two have committed to the long-term financing aim of three waves at a time. In wave 5, only 15 countries participated. The reasons for this are manifold. Some countries lack the appropriate funding lines: in Denmark, for example, SHARE fell between the cracks of medical research and socio-economics. Some countries are not inter-

ested in international projects and prefer to spend their scarce funds on national projects. The economic crisis has made matters worse. Economists predict that when a common public good has to be financed de-centrally, countries tend to underinvest and to free ride. This is exactly what happened.

The implications for SHARE are far-reaching. It jeopardizes the financial foundation due to high fixed costs, thereby creating a vicious circle: every country leaving SHARE makes it more expensive for the remaining countries which then have an even larger incentive to leave. The researchers in SHARE have spent an inordinate time to limit the centrifugal forces which undermine their research capabilities and create inefficiently high costs. Since SHARE observes the same individuals over time, countries which drop out are systematically destroying the initial investment made. All this has severely damaged the quality of SHARE as a research tool.

To provide a concrete example: Greece, due to the austerity measures, is unlikely to continue funding SHARE exactly at the point where analysts in the Commission and researchers all over Europe are particularly interested in understanding the effects of the crisis on the Greek population. Unlike many other infrastructures, SHARE needs every country independent of whether the country itself is interested in SHARE – Europe is interested in all of these countries because they are the very object of our research.

Distributed research infrastructures such as SHARE tend to provide a large European added value over and above the value for the member states. To be sustainable, the funding model of such an infrastructure needs to be aligned with the added value emanating from this infrastructure. In distributed infrastructures, this requires the European Commission to take on a greater role in funding than initially envisaged during the ESFRI process which provided that funding comes exclusively from the member states. In concrete terms, the European Commission as the political entity above the centrifugal forces of the member states needs to fund international coordination, which harnesses these centrifugal forces, as well as a small core survey which determines that all countries of interest will be represented at least for the sake of international comparisons.

SHARE is set to expand to all EU member states. In addition, we will strengthen its interdisciplinary approach by broadening the set of biomarkers to measure health as objectively as possible in such a large population study; we will validate its economic measures with process-generated administrative data; and we

will deepen SHARE's measures of social and family support networks to better understand the social developments in this unprecedented period of demographic change. The ERIC regulation 723/2009 is an excellent starting point for the further evolution of European research infrastructures in general, and for the achievement of SHARE's scientific aims in particular. SHARE is grateful to all participants in this evolutionary process.

References

Adam, Stéphane/Bonsang, Eric/Germain, Sophie and Perelman, Sergio (2007): Retirement and cognitive reserve: a stochastic frontier approach applied to survey data. CREPP working papers 2007/04, HEC-ULg.

Avendano M.,/Aro, A. R. and Mackenbach, J. (2005): Socio-Economic Disparities in Physical Health in 10 European Countries, In: Börsch-Supan, A./Brugiavini, A./Jürges, H./Mackenbach, J./Siegrist, J. and Weber, G. (eds.): Health, Ageing and Retirement in Europe – First Results from the Survey of Health, Ageing, and Retirement in Europe. Mannheim: Mannheim Research Institute for the Economics of Aging (MEA).

Bonsang, Eric/Adam, Stéphane and Perelman, Sergio (2010): Does retirement affect cognitive functioning? ROA Research Memorandum 2010/1, Maastricht University.

Börsch-Supan, A./Brugiavini, A./Jürges, H./Mackenbach, J./Siegrist, J. and Weber, G. (eds.) (2005): Health, Ageing and Retirement in Europe – First Results from the Survey of Health, Ageing and Retirement in Europe. Mannheim: Mannheim Research Institute for the Economics of Aging (MEA).

Börsch-Supan, A./Brugiavini, A./Jürges, H. et al. (eds.) (2008): Health, Ageing and Retirement in Europe (2004–2007). Starting the Longitudinal Dimension. Mannheim: Mannheim Research Institute for the Economics of Aging (MEA).

Börsch-Supan, A./Brandt, M./Hank, K. and Schröder, M. (eds.) (2011): The Individual and the Welfare State. Life Histories in Europe. Heidelberg: Springer.

Börsch-Supan, A./Brandt, M./Litwin, H. and Weber, G. (eds.) (2013): Active Ageing and Solidarity between Generations in Europe. First Results from SHARE after the Economic Crisis. Berlin: DeGruyter.

Börsch-Supan, A. and Schuth, M. (2013): Early retirement, mental health, and social networks, In: Börsch-Supan, A./Brandt, M./Andersen-Ranberg, K./Litwin, H. and Weber, G. (eds.): Active Ageing and Solidarity between Generations in Europe: First Results from SHARE after the Economic Crisis, De Gruyter, Berlin.

Börsch-Supan, Axel/Brandt, Martina/Hunkler, Christian/Kneip, Thorsten/Korbmacher, Julie/Malter, Frederic/Schaan, Barbara/Stuck, Stephanie and Zuber, Sabrina (2013): Data Resource Profile: The Survey of Health, Ageing and Retirement in Europe (SHARE). International Journal of Epidemiology, doi: 10.1093/ije/dyt088.

Brandt, M. (2013): Intergenerational help and public assistance in Europe: A case of specialization? European Societies 15 (1), 26–56.

Brandt, M. and Deindl, C. (2013): Intergenerational Transfers to Adult Children in Europe: Do Social Policies Matter? Journal of Marriage and Family 75 (1), 235–251.

Brandt, M./Haberkern, K. and Szydlik, M. (2009): Intergenerational help and care in Europe. European Sociological Review 25 (5), 585–601.

Deindl, C. and Brandt, M. (2011): Financial support and practical help between older parents and their middle-aged children in Europe. Ageing and Society 31 (4), 645.

Hank, K. (2007): Proximity and Contacts Between Older Parents and Their Children: A European Comparison. Journal of Marriage and Family 69 (1), 157–173.

Mazzonna, Fabrizio and Peracchi, Franco (2012): Aging, cognitive abilities and retirement. European Economic Review 56 (4), 691–710.

Rohwedder, Susann and Willis, Robert J. (2010): Mental retirement. Journal of Economic Perspectives 24 (1), 119–138.

4.2 Generations and Gender Programme: A Research Infrastructure For Analyzing Relationships over the Life-Course

Anne H. Gauthier, Tom Emery (NIDI)

Introduction

The GGP is a cross-national research infrastructure that was established in 2001 and which aims at understanding how the lives of individuals evolve over the whole life course, from young adulthood to older ages (more information can be found on our website: www.ggp-i.org). It furthermore aims at understanding the ways in which various factors, such as public policy, affect family life including the relationships between generations and between genders. It is a research infrastructure built on the principle of open access. It provides comparable micro-level data from 19 countries as well as related contextual data. In this paper, we first provide an overview of the GGP including its key features (i.e. what is the GGP?) and its capabilities (i.e. why do we need a GGP?). We then provide examples of some of its scientific accomplishments as well as its potential in terms of answering emerging research questions. We then discuss the way forward including our strategic plan through to 2020.

What is the Generations and Gender Programme?

In a nutshell, the GGP is best defined as a "harmonized, large-scale, longitudinal, cross-national panel study of individuals & contextual database". It is a longitudinal panel study covering the whole life course from 18 to 79 years of age. It collects both retrospective information on topics such as fertility, family formation and dissolution, as well as prospective information collected through subsequent waves of the survey, allowing us to see changes in people's lives over time. It is also a large-scale project involving data collection from about 10000 individuals per country (including both men and women). Such large sample sizes are necessary to study specific population subgroups such as migrants or people at the extreme ends of the income distribution, as well as to capture a sufficiently large number of life-events for statistical analyses.

The GGP is also a cross-national project currently covering 19 countries with data harmonized in a large database for cross-national comparisons. Moreover, 12 of these 19 countries have carried out subsequent waves of data collection (on the same individuals) allowing us to see changes over time in a variety of contexts. Finally, the micro-level data are also complemented by a contextual database providing information about policies and the economic environment at the regional and country-level that may affect individuals.

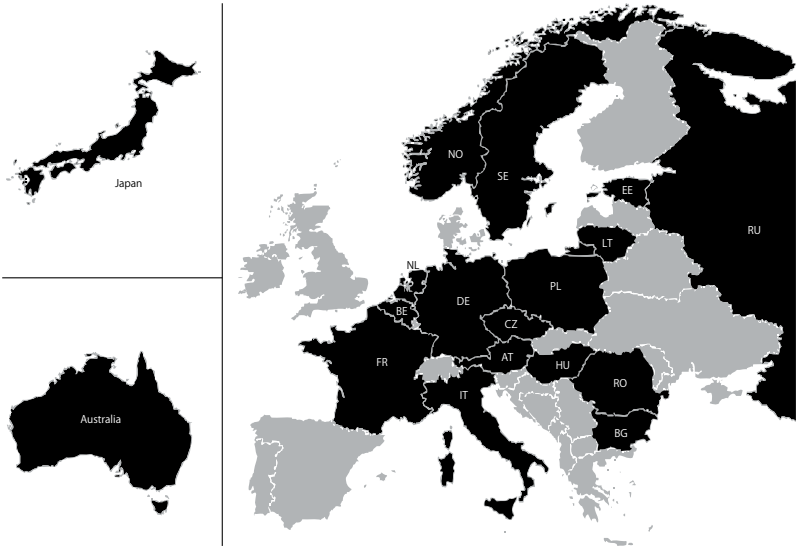


Figure 1: Participating Countries in the Generations and Gender Programme[1]

The GGP covers a wide range of topics and collects data on: fertility and partnership histories, transition to adulthood, work-family balance, gender relations and gender division of housework, intergenerational exchange including informal and formal care, well-being and health, grandparenthood, and economic activity and retirement.

As a research infrastructure, the GGP is built on the principle of open access. Micro-level data can be downloaded directly from the web after researchers have been granted access through a centralized registration and accreditation

1 Note: Australia, Austria, Belgium, Bulgaria, Czech Republic, Estonia, France, Georgia, Germany, Hungary, Italy, Japan, Lithuania, Netherlands, Norway, Poland, Romania, Russia, Sweden

process. Meta-information and online analysis is possible for anyone through the NESSTAR system. Our Contextual Database contains information on more than 200 harmonized indicators, and tracks population trends and policy changes in 60 countries over the past 40 years. Both as a stand-alone tool for analysis and as a supplement to the individual-level database, this dataset is a powerful analytical component of the infrastructure which enables us to understand individuals' relationships and personal histories in the context of policy developments and social change.

The number of registered users for the GGP micro-level data has increased rapidly over the years and has now exceeded 1,800. Users come from a large number of social science disciplines and from more than 30 different countries, and include both young and more established scholars. The GGP appears on the roadmap of the Netherlands and Norway, and is on the path to inclusion in France. It is governed by a consortium board of twelve academic institutions and research institutes from 10 countries, and a Council of Partners with representatives from 34 countries, an Advisory Board, and an international coordination team located at the Netherlands Interdisciplinary Demographic Institute in The Hague.

Why we need the Generations and Gender Programme

The increasing complexity of individuals' life-courses

To illustrate the essence of the GGP, let us first introduce Sylvia. She was born in 1955. She finished high school and became a secretary at the age of 18, met her future husband that year, got engaged at the age of 19, married at the age of 20, and had her first child at the age of 21. She went on to have a total of three children and lived happily ever after (Hicks, 2008). What is notable is that all of her key life transitions were concentrated early in life and within a very short time period. Her life story resembles that of many other women born about the same time. In our jargon, we say that her life story was standardized in that it followed a standard sequence and timing of events (Billari and Wilson 2001; Elzinga and Liefbroer 2007).

Now, let's contrast this to the life story of her middle daughter, Julia, who was born in 1978. Julia studied longer than her mother and eventually graduated with a post-secondary degree at the age of 23. While she was still a student, Julia had left home to live with friends at the age of 19, something her mother

has never done. She then moved in with a boyfriend, ended up having a child with this partner at the age of 28, and eventually married the father of her child at the age of 30. What is very clear here is that Julia's key life transitions were much less concentrated in time than those of her mother. While her mother finished school, got married, and had her first child all within a 3-year period, Julia had an interval of 11 years between leaving home and having her first child. In technical terms, Julia's life story was de-standardized in that it followed a much less standard sequence and timing of events.

So, why do these two stories matter? They matter because they reflect different sets of norms and opportunities associated with different decades and different cohorts of adults. They also matter because they have very large consequences for the context in which children are born and they have consequences for the relationships between generations and between genders. This is precisely what the GGP is about.

Findings to date

The scientific accomplishments of the GGP are numerous. The GGP has contributed important knowledge on the changing context of parenthood and childbearing, such as the question of who has children outside of marriage. For example, analyses with GGP data have supported the long held belief that having a first child outside marriage is more prevalent among those with lower levels of education. Amongst the lower education groups in the Netherlands 45% of births occur outside marriage. Yet amongst those with higher levels of education, just 29% of births occur outside marriage (Perelli–Harris et al. 2010). However, in some countries, such as France, this is no longer true and it is in fact those with higher education who are more likely to have a child outside of marriage. Whether or not this pattern will spread to other countries, and whether or not it is influenced by the legal and institutional framework affecting families, are key research questions that researchers are looking to answer with the GGP. Answering these questions will enable us to understand what marriage and parenthood mean in the 21st century. Why do we get married? Why do we have kids? And what have the two got to do with each other?

Another example comes from our research on intergenerational relationships. The GGP has been used to show how loneliness in older ages is more prevalent in Eastern than Western Europe (Gierveld and Van Tilburg 2010). This is attributable to the greater health and wealth of older generations in Western Europe and the extent to which it helps them combat loneliness. The GGP has

also revealed that older generations are not just vulnerable but also play an important part in supporting younger generations. In some countries, such as Hungary, grandparents providing child care support was found to be important for young women who want to return to work after having a child. Yet in other countries, like the Netherlands, this didn't affect the woman's decision to work (Aassve et al. 2012). Future research will be able to examine whether this is because of culture, policy or other factors. These are just two of the many ways in which the GGP has demonstrated the complexity and diversity of relationships between generations as well as the need to consider this diversity in different countries.

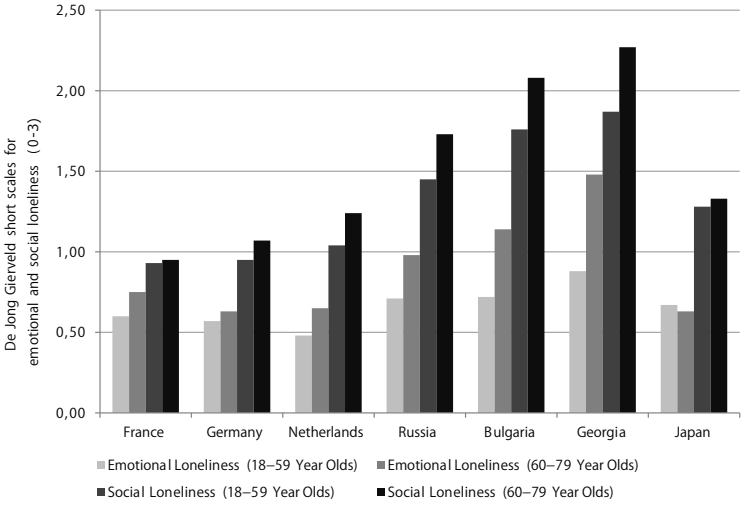


Figure 2: Loneliness amongst Older & Younger Persons in 7 Countries[2]

2 Source: Adapted from Gierveld, J. D. J. and Van Tilburg, T. (2010): The De Jong Gierveld short scales for emotional and social loneliness: Tested on data from 7 countries in the UN generations and gender surveys. European Journal of Ageing.

Future Research Questions

As a research infrastructure, the GGP will be essential for answering emerging scientific questions. In particular, there are two key questions that are going to be pivotal to our work in the coming years. The first one concerns the short- and long-term impact of the economic crisis. Since the crisis unfolded in 2008, European governments and the media have paid much attention to the fate of the unemployed, especially among young adults. However, we have little information on the long-term impacts of the crisis on the life-course of individuals. By continuously tracking young adults through subsequent waves of the GGP, we will be able to see the consequences of the crisis on the life-course of individuals and of their families. For example, to what extent has the experience of the crisis forced people to postpone having children, or prevented them from having children at all? And what does unemployment and a delay in leaving home mean for later relationships between young adults and their ageing parents? These are some of the questions that we hope to answer with future waves of the GGP.

The second key aspect concerns the long term effects of childhood and youth experiences. Research has shown that disadvantages in childhood and early adulthood have consequences in later-life. For example, having a child at a young age or outside a stable relationship has been shown to affect one's health, wealth and well-being much later in life (Lucas 2007). What is not known, however, is whether the impact of such negative effects weaken or accumulate over time. Moreover, some social scientists suspect that these negative effects of early life events may vary across countries as a result of different institutions and policies (Peters and Liefbroer 1997). This is because, in some countries, those in challenging circumstances are better supported by welfare arrangements than in others. With data from 19 countries, the GGP is a leading source of information on how best to support individuals through many of life's challenges because it follows individuals in a wide variety of circumstances.

The way forward

Since its inception in 2001, the development of the GGP has strongly relied on the commitment of the participating institutions and been funded from the bottom up. At the national level, participating institutions have put major efforts into fundraising in order to implement data collection at the national micro- and macro-level. Most funding is being provided either by national governments, statistical offices, or national science foundations. In addition, some institutions

have invested considerable funds in data collection from their own resources. At the international level, the coordination and development of the GGP was funded exclusively by participating Consortium Board institutions until 2007. In 2007, a grant totalling € 230k for the 2007–2008 period from EU-DG Employment to the UNECE, which was coordinating the programme at the time, allowed for an acceleration of the programme development. The EU-DG Research Grant 'Design Studies for Research Infrastructures' within the 7th Framework Programme in 2009 signified a major change in the tempo of development and led to a rethinking of the programme's long-term strategy. This EU-FP7 Design Study totalled € 2M for the 2009–2012 period and has been used to assess the state of the programme's methodological components and the preparation of a blueprint for the future of the GGP.

A challenging funding environment

Maintaining a research infrastructure is expensive. It includes high data collection costs at the national level as well as substantial coordination costs at the international level. The costs of data collection in particular have increased over the years. To give an example: conducting one GGP wave of face-to-face interviews among 10000 respondents in a high-cost country like the Netherlands cost over € 1.5M. Despite such considerable costs, many participating institutions were able to raise funding for their national surveys in the past. However, the economic crisis has made it increasingly difficult to secure funding for new waves of the GGP. In addition, several countries that have shown serious interest in participating in the GGP have not been able to raise sufficient funds to turn this interest into actual participation. Although the success of national fundraising does not only depend on the costs of the infrastructure (but also, for instance, on its perceived importance), the cost element is critical. Looking ahead, the GGP aims to continue collecting data in the 19 participating countries every three years. It is also the intention to expand the programme to new countries. The goal is to ultimately establish the Generations and Gender survey in all 28 EU member states. To achieve this ambitious aim and to secure a sustainable future for the GGP, it is necessary to consider cost efficiency measures in data collection.

Introducing Web Surveys

Many efforts were made in the Design Study to evaluate the current design of the Generations and Gender Survey and to suggest changes that could make it more cost-efficient. The main change is the decision to move from face-to-face surveys to web-based surveys. This measure reduces non-response, attrition, and can be more effective in gaining insights into individuals' personal relationships and attitudes. It is also estimated that such a change will decrease the costs of data collection per country by about one third, a considerable reduction in the funds required to conduct the survey. This shift does however create challenges as well as opportunities and the GGP has invested and continues to invest in ways of tackling the problems brought about by web surveys, such as selection and mode effects which reduce the comparability between responses given over the web and face-to-face.

A Sustainable Infrastructure

Future waves of the GGP will also be completed and processed using a standardized, centralized, highly efficient data collection process. This system, standard within ESFRI Social Science projects, will enable participating countries to reduce data collection costs further, will improve the timeliness and quality of data releases, and prepare the infrastructure, upon which the GGP is based, for the future. There are many parts of the data collection process that could be centralized and therefore reduce costs for individual countries. These include questionnaire testing, harmonization of measures and production of accurate and comprehensive documentation. The GGP has made great strides in this area but there are still considerable returns to increased standardization, centralization and economies of scale. These efficiencies will reduce data collection costs for individual countries and thus increase the sustainability of the programme as a whole. These measures will also enable the GGP to meet the standards for a European Research Infrastructure with regards to accessibility, documentation and legal frameworks, and, hopefully, facilitate inclusion in the ESFRI Roadmap and constitution as an ERIC.

Conclusion

The GGP, through its longitudinal coverage of the whole life-course, occupies a central position as a research infrastructure. It is an essential tool to allow a better understanding of the increasingly complex life-course of individuals and family life, as well as their cross-national differences and similarities. The GGP is committed to providing data that fit Europe's research strategy as outlined in Horizon 2020. With abundant information on two of its key themes – health, demographic change and well-being, and inclusive, innovative and secure societies – and its wide coverage of European countries, the GGP is ideally suited to provide scientifically informed and policy-relevant answers to key societal questions. In FP6 and FP7, many social science projects – e.g. MAGGIE, MULTILINKS, REPRO, NEUJOBS – and ten ERC grants used or are using GGP data. With the release of a significant number of additional longitudinal datasets in the coming years and the realisation of the planned developments outlined above, it is expected that the GGP will be used even more in projects funded by Horizon 2020. Key steps have been made, and will continue to be made, to ensure that the GGP is a research infrastructure which meets the highest technical standards in order to answer some of the most pressing questions in the social sciences.

Bibliography

Aassve, A./Arpino, B. and Goisis, A. (2012): Grandparenting and mothers labour force participation: A comparative analysis using the Generations and Gender Survey. Demographic Research 3, 53–84.

Billari, F. C. and Wilson, C. (2001): Convergence towards diversity? Cohort dynamics in the transition to adulthood in contemporary Western Europe. Max Planck Institute for Demographic Research, Working Paper, 2001-039.

Elzinga, C. H. and Liefbroer, A. C. (2007): De-standardization of family-life trajectories of young adults: A cross-national comparison using sequence analysis. European Journal of Population/Revue européenne de Démographie 3-4, 225–250.

Gierveld, J. D. J. and Van Tilburg, T. (2010): The De Jong Gierveld short scales for emotional and social loneliness: Tested on data from 7 countries in the UN generations and gender surveys. European Journal of Ageing 2, 121–130.

Hicks, P. (2008): The Olivia Framework: Concepts for Use in Finely-Grained, Integrated Social Policy Analysis. Queens University School of Policy Studies, Working Paper.

Lucas, R. E. (2007): Adaptation and the set-point model of subjective well-being: Does happiness change after major life events? Current Directions in Psychological Science 16, 75–79.

Perelli–Harris, B./Sigle - Rushton, W./Kreyenfeld, M./Lappegård, T./Keizer, R. and Berghammer, C. (2010): The educational gradient of childbearing within cohabitation in Europe. Population and Development Review 4, 775–801.

Peters, A. and Liefbroer, A. (1997): Beyond marital status: Partner history and well-being in old age. Journal of Marriage and the Family 59, 687–699.

4.3 Elfe: A multidisciplinary birth cohort including biological collection

Jean-Louis Lanoë (Elfe)[1]

Scientific key issues and national context

Experiences made during the first years of life, or even in utero, are critically important for the composition of the future adult. The development of a child, its socialization, is a complex process which unfolds in constant interaction with the natural and social environment that the embryo, the fetus, the child, and the adolescent grows up in.

In recent decades, this environment has changed dramatically: changes in dietary habits, exposure to new pollutants, decreased physical activity, blended families, extended schooling, increased female employment, job insecurity, etc. Research in these areas is complex due to the large number of factors and their potential interactions. The most suitable method of analysis is the cohort study, that is to say, the creation of a large sample of children who are followed throughout their development until adulthood.

Such cohorts have been used for decades in many countries. In the early 2000s, French public bodies (such as the National Council for Statistical Information), several different reports (on health and nutrition) and governmental departments (mostly health and environment) started to push for the launching of a birth cohort in France. This resulted in the establishment of two projects. The first was set up by a research unit operated jointly by the French National Institute for Demographic Studies (INED) and the National Institute for Health and Medical Research (INSERM). It followed a multidisciplinary approach with an emphasis on socio-demographic and epidemiological/health inequalities that allows analyzing many different aspects of children's lives.

1 Assisted by C. Zaros, X. Thierry, M.-N. Dufourg, C. Bois, A. Rakotonirina, B. Geay, and M. A. Charles (all Elfe)..

At the same time, the French Institute for Public Health Surveillance (InVS) was asked to design a study, as part of France's National Environmental Health Plan drawn up in 2004, for the purpose of identifying the effects of the exposure to various types of pollution on children's health.

This lead to two projects with overlapping objectives and similar methodology competing with each other and it appeared necessary to merge them for the benefit of all concerned. From this time on, a single study covered the fields of health, environmental health and social sciences. A target figure of 20000 children seemed necessary to ensure a successful outcome (i.e. one in forty of all French births in 2011). The French Longitudinal Study for Children Elfe, as a merger of the two projects, was born.

One cohort and a multidisciplinary approach

The first step of building up the cohort was to call for research proposals to involve the scientific community in setting-up the project. Sixty research teams (almost 100 research projects and 400 researchers) from universities, public-sector research institutions, and health agencies were finally selected to take part.

Working Groups (see Annex 1) bringing together researchers of the same or neighbouring fields were set up to define and develop the scientific objectives and suitable methodologies. The coordinators of these groups constitute the Scientific Project Group which governs the Elfe project together with the Scientific Council.

Among the questions guiding the research are, for example: 'At what age should food be diversified?', 'What influence do food preferences have on children's health?', 'What is the impact of environmental pollutants on children's health and development?', 'What are the effects of child care on young children's future relationships with other children, their integration in kindergarten and language acquisition?' 'What are the economic and socio-cultural factors which influence success at school?', 'What is the influence of computer use, sports or cultural activities on the physical and intellectual development of children?

As shown in the diversity of questions, the originality of the Elfe study consists in considering the child's environment in the broadest sense of the term.

Elfe was started in April 2011, following a pilot study in 2007. A joint unit INED-INSERM-EFS (French National Blood Service), created in 2010, is in charge of the operationalization and management of the cohort. Elfe has the endorsement of the Ministry for Higher Education and Research, the Ministry for Labour, Employment and Health, and the Ministry for Ecology, Sustainable Development, Transport and Housing. It is mostly funded by the 'Investments for the future EQUIPEX 2011 program'.

Recruitment strategy, inclusion criteria and data collection

The children of the current survey were recruited at birth in 320 maternity units which were randomly selected from the 542 maternities that existed throughout metropolitan France in 2011. Recruitment was carried out in four waves of 4 to 8 days between April 1 and December 5, 2011. Of the 25 days included in the study, 12 match those of the Permanent Demographic Sample of INSEE (National Institute for Statistics and Economic Studies) so that, for almost half of the cohort, it will be possible to compare the participating children to the general population of children born on the same dates.

To be included, children had to be born on one of the days covered. Their mothers were required to be over 18 years old, able to sign an informed consent form, and to continue living in metropolitan France for a minimum of 3 following years. Only single or twin births could be included and pregnancies had to be at least in week 33. EPIPAGE 2, an epidemiological study on low gestational age that we are closely collaborating with, started also in 2011.

More than 18024 families including 288 families with twins agreed to participate (51% of eligible mothers contacted in maternities).

During hospitalisation after delivery, each mother was asked to answer a face-to-face questionnaire, fill out an auto-questionnaire, and give permission to collect data from her medical record. Biological samples, such as maternal and cord blood, breast milk or baby stool were collected from a subgroup (Annex 2).

Families and children follow-up

From the age of 2 months to 2 years old

Parents were required to answer a CATI questionnaire when the children reached the age of 2 months. The questions focus on the household, extended family, income, parental occupation, housing conditions (including exposure to pesticides during pregnancy), parental and infant health, and child care.

Between 2 and 10 months, parents complete a questionnaire on the children's diet diversification. Families are contacted again for a telephone interviews on the first and second birthday of their children. The interviews at 2 months, 1 year and 2 years follow the same protocol: initial contact with the mothers, in-depth interview of the 'referential' parent (most often the mother), and a shorter but systematic interview of the cohabitating or not cohabitating parent (most often the father).

From the age of 3 years and a half to 8 years old

Children are surveyed once again when they reach the age of 3 and a half. All families also answer a short CATI interview and only a sub-sample will be visited home by an investigator who collects biological samples amongst those for whom samples were already collected in maternities and the 'Picture Similarities'-part of the British Ability Scale (BAS) will be used.

An internet game for children at 4 years of age and a newly developed Elfe telephone interview at 5 years of age have both been tested in the pilot study.

The procedure for the following years has yet to be agreed upon with the working groups.. These now include new research teams following a call for proposals in 2012 which focus on children over 5 years of age. A provisional timetable is already available.

'Family response rate for telephone surveys

The response rate for telephone surveys, including partially completed interviews with at least one parent, was 90% after 2 months, 80% after 1 year and close to 77% after 2 years,based on a preliminary analysis of the first and second waves, for the initial population of 18024 families. After 2 months, if both parents were living together with the children, the response rate among father

was good (85%), but quite low (16%) among non-cohabiting fathers which was partly due to the lack of contact information. After 1 year, figures were not very different (82.7% and 20.8%, respectively). However, we noted an increase in the amount of partially completed questionnaires (4% after 2months and 10% after 2 years) which may be due to the length of the interviews. One of the main reasons for non-response is that it can be difficult to reach some parents due to relocation moves and/or changing telephone numbers. These difficulties occur despite efforts to continuously update contact information by the IS department of the Elfe team and in-depth research by the institute in charge of the interviews. In the ongoing survey, if the mother cannot be reached, interviewers are trying to contact the father (if any).

Regarding the three telephone surveys, if the 'referential' parent could be reached and was able to answer, the response rate was around 90%.

More than 87% of those who participated in the survey after two months also participated in the 1 year survey. However, a quarter of the families who were not interviewed initially participated at 1 year. Non-response does not automatically mean attrition. In view of the survey participation after 2 months and after 1 year, we have the following figures:

Families completing:

- both surveys: 78%
- the 2 months survey only: 13%
- the 1 year survey only : 2%
- no survey: 7%

At the end of the 2-year-survey, a simplified and short paper or internet questionnaire is now being sent to families that cannot be contacted. This questionnaire is also sent out if only one of the two parents expressed temporary or permanent refusal, or could not be re-interviewed after a successful contact. This questionnaire comprises 20 questions and aims essentially at keeping in touch with families, updating contact information and getting some statistical variables that are of interest.

Passive follow-up: A partial substitution for non-participating families

Medical information (ambulatory and hospital care) on both the mothers (during their pregnancies) and the children (since they were born), made available by the Social Security's Medical Insurance Scheme, will be linked with data collected through the surveys.

Independently of its specific interest, geo-localization of parent' dwellings, of nurseries, of kindergartens, of schools will be done and might be used later as a substitute for missing information on environmental matters.

As was mentioned before, almost half of the children included in the cohort are part of the 'Permanent Demographic Sample' of INSEE. This will allow getting in-depth information about their families after being interviewed for the national cenus, whether they are taking part in Elfe or not.

The Environment and Child Health International Birth Cohort Group

Elfe is actively participating in Environment and Child Health International Birth Cohort Group together with scientist and managers of four other birth cohorts. the JECS from Japan has already started, while the NCS from the US and the Chinese SBC are still in pilot phase. The German environmental cohort is still in the planning phase. The main objective of the group is collaboration and ongoing exchange, and to harmonize environmental exposure assessment by questionnaires.

So far exchanges mostly concerned questions of measuring exposure of common interest (mercury, pesticides, phthalates, PCBs, PBDEs, etc.), definitions of a framework to provide complete information about the source of each question, questionnaires' timing and mode of administration. The harmonization process was exacerbated by the different timing of the cohorts, different exposure outcomes, the lack of internationally validated questionnaires, cultural specificities, and differences in terminology.

Access to data

The Elfe data base will be the first component of a future research platform (RE-CO-NAI) which will include different French birth cohorts (EPIPAGE 1/2, EDEN, etc.).

The data collected by Elfe, presently consisting of the data collected in maternities and in the 2 months survey, will be available internationally to all public or private research teams. Third parties that do not provide public research will be able to gain access if their projects cooperate with public research institutions. Foreign researchers can get access through framework agreements in consortia or specific conventions.

All research teams, whether or not they were represented in the initial working groups, must submit a request to the Data Access Committee (CADE) to gain access.

There may be a demand for data and/or biological materials collected under the general Elfe protocol (analysis projects), or a demand to collect additional data or biological materials (satellite projects).

Researchers who were a part of the initial working groups and took part in designing the general protocol in collaboration with the Elfe unit have exclusive access to the data for a period of 18 months, starting when data are available on the platform.

The data access requests must be done via a website[2] which includes all the necessary documentation and allows choosing the variables and/or biological samples needed. The platform will also allow selection of individual data sets, but also in the form of thematic blocks. Researchers will be encouraged to seek access by thematic blocks.

Requests for identifying or indirectly identifying data will require prior approval of the French Data Protection Commission (CNIL). Any request for access to data will be logged. The application requires the following information: general information about the applicant, a brief description of the specific objective, scientific issues and methodological aspects of the project as well as a list of requested variables.

2 www.pandora.vjf.inserm.fr/public

The investigation of the demand by the CADE involves the following steps:

- Admissibility of the application by checking the completeness of the elements to be included in the application
- Technical review by the Data Access Committee (feasibility, relevance of the requested data to the objectives of the project)
- A priori no evaluation of the projects' scientific quality will be done, as far as they already have been granted funding.
- If the research objectives or methods of implementation may raise questions of an ethical nature, the Elfe Ethics Group will be consulted. The INED Ethics Committee may also be involved in second appeal
- Technical remarks, scientific and methodological will be forwarded to the project leader with favorable or unfavorable opinion of the CADE.
- The maximum length for this procedure should not exceed three months from the time the CADE receives the first application package.

Conclusion

The multidisciplinary approach of the Elfe cohort has a number of obvious advantages: it will, provided that the research teams will mobilize in this sense, develop interdisciplinary works, for example in the field of epigenetics. Otherwise, remaining only in the field of multidisciplinarity, in-depth questionings in the many areas covered will allow researchers to deal with sets of data rarely disciplinary available or, at best, very briefly documented. Work on the socialization of children when being able to take into account their health, their housing characteristics, exposures they encountered should contribute to enrich issues for original results. Similarly, addressing children's health, regardless of the disease or problem concerned, with particularly rich data on the socioeconomical characteristics of the families and the environment in which children grow up, their life style can only improve knowledge in this specific area.

However, these benefits have a cost. Regardless of time spent (and sometimes of difficult discussions) to achieve the trade-offs between the different topics to be addressed, multidisciplinary objectives result in numerous questionings, in various forms and often long. This may discourage some families from participating despite multiple actions developed to keep in touch with them, to inform them and to reward ‚their children of their' participation'.

Annex 1: List of the 'working groups' and their key issues

Environmental health	Social sciences
Chemical expositions	Demography, family
Physical expositions	Socialisation
	Economy, poverty
	School
Health	**Transversal**
Perinatal period	Diet, nutrition
Growth, reproduction	Psychomotor development
Accidents	Physical activity
Health care, infections	
Respiratory diseases	
Mental health	
Paediatric	
Sleep	

Annex 2: Participation and samples

- 211 maternities were selected for biological sampling mainly according to their birth rate, their distance to a bio-bank and their non-participation to the French Network of Placental Blood.
- 75% (159/211) accepted to participate to this part of the project with important regional discrepancies (participation rate from 0% to 100%).

The biological sampling started with the second wave (in July 2011) and allowed to collect:

4000	Maternal venous blood samples	1300	Cord tissue samples
3200	Maternal urine samples	2900	Meconium samples
5000	Maternal hairs samples	2900	Stool samples
2000	Maternal milk samples	800	Cord blood samples in Paxgene tubes
3900	Cord blood samples		

5 Digital Humanities

Sandra Collins (Digital Repository of Ireland),
Jacques Dubucs (SCI-SWG)

ALLEA and the European Academies constitute a unique pan-European knowledge base that is trusted, non-partisan and long-term. The Academies therefore have an important contribution to make to debates regarding sustained digital infrastructures, the achievement of long-term durable digital preservation, and the societal responsibility for preservation of our digital cultural heritage - and we welcome the dialogue and engagement this conference has generated.

Digital Humanities data can be rich and complex, non-standardised in format, without common or consistent metadata and ontologies, and can be subject to complex rights issues. Consensus and best practice regarding digitisation and metadata standards for common usage, that still retain the richness of different disciplines and data types, could enable open access to Humanities data, and facilitate data exchange and sharing between the wealth of archives, repositories and libraries across Europe.

Sustaining research infrastructures which have achieved excellence and wide use, together with open access to research data, are the two lynchpins upon which accelerated and enhanced discoveries, best return on public investment, re-use for education and cross-sectoral use, and research validation rely.

The Digital Humanities session included three wonderful speakers who highlighted the value of, and the opportunities and challenges facing our work. They spoke with deep insight, experience and humour. There remains more work to do, both scholarly pursuits but also the practical implementation of best practices, and continued advocacy and cross-disciplinary collaboration, so that the value of Digital Humanities is understood and embedded into international programmes such as Horizon 2020 and the Research Data Alliance, and the state of the art continues to advance at the fast pace we have set ourselves.

5.1 Research Infrastructures in the Humanities: The Challenges of 'Visibility' and 'Impact'

Milena Žic Fuchs (University of Zagreb, Croatian Academy of Science and Arts)

'Facing the Future'

Research infrastructures in the humanities, just like those belonging to other domains of research, are currently undergoing various assessments and it can be expected that they will be assessed to an even greater extent in the future. A crucial question that arises is whether assessment parameters used for RIs in other domains of research can be readily, 'applied' to Humanities RIs, or whether considerations specific to the domain of the humanities have to be taken into account. One parameter of assessment that is being debated more and more is 'impact', a many-layered concept that can be interpreted and broken down into manifold variables. The concept of 'impact' in the humanities is very closely linked to the notion of 'visibility' of RIs, again a concept implying a number of possible layers of interpretation.

The aim of this paper is to, at least in part, reflect on 'visibility' and 'impact' and their inter-connectedness within the context of future assessments of Humanities RIs, particularly those that are of, or may aspire to, pan-European relevance in the future.

How RI assessments develop

It is interesting to note how assessments of RIs have been developing during recent years. Illustrative in this respect are the various stages of assessing projects of the ESFRI Roadmap 2010, stages that indicate directions in which assessing RIs can and are developing. In brief overview of developments, a few landmark reports are worth mentioning.

In 2011, the ESFRI Implementation Group was established with the aim of supporting ESFRI projects in order to reach a high level of the implementation goal that was set up by the Innovation Union Flagship Initiative. As can be seen

from their report entitled *State of Play of the Implementation of the Projects on the ESFRI Roadmap 2010*,[1] basic criteria were identified to pinpoint those projects which could be considered to be under 'implementation'.

The process of assessment was taken a step further in August 2013, and a more detailed report on the various stages of development of ESFRI projects was published: *A High Level Expert Group Report on the Assessment of the Projects on the ESFRI Roadmap*.[2] This report covers, in quite extensive detail, the assessment of 35 projects on the ESFRI Roadmap and is based on an Assessment Matrix comprised of six modules: Cost and Financial Structure, Governance and Legal Structure, Stakeholder and Financial Commitments, Human Resources and Project Management, User Strategy and Risk.[3] It should be noted that assessment of the ‚science mission' was not a part of this report.

Together with representatives of the European Commission, ESFRI has gone a step further and has set up an Expert Group whose aim is to determine indicators on the basis of which 'pan-European relevance' can be assessed. At the time of completing this paper it is known that the report has been finalized, but as yet has not been published. It will be interesting to see which indicators the Expert Group has proposed for the evaluation of pan-European relevance, especially from the perspective of Humanities RIs on the ESFRI Roadmap.

The above overview of the directions that RI assessments have been taking within the wider ESFRI context indicates the need to look ahead and to try to see how Humanities RIs can respond to the challenges that lay ahead, especially those that can be subsumed under the complex notion of 'impact'.

1 ESFRI (2012): State of Play of the Implementation of the Projects on the ESFRI Roadmap. Report of the Implementation Group to the ESFRI Forum. ec.europa.eu/research/infrastructures/pdf/esfri_implementation_report_2012.pdf

2 Calvia-Goetz, A. et al. (2013): A High Level Expert Group Report on the Assessment of the Projects on the ESFRI Roadmap. ec.europa.eu/research/infrastructures/pdf/jd-final-aegreport-23sept13.pdf

3 The author of this paper was one of the members of the High Level Expert Group on assessing the projects on the ESFRI Roadmap.

'Visibility' and related issues

The notion of RIs in the humanities is often still a vague concept for researchers in other domains, and consequently for those partaking in assessments. The low level of 'visibility', and, one could add, recognition of Humanities RIs, is also often reflected in National Roadmaps of RIs across Europe.

The lack of awareness of the existence and importance of humanities RIs can have negative repercussions in humanities research itself, but also in the context of inter-/multi-/transdisciplinary lines of research that are at present gaining momentum, as is evident within the framework of Horizon 2020. On the one hand, articulating research agendas and questions that will hopefully lead to answering the grand/societal challenges that mankind faces is difficult to envisage without input from the social sciences and humanities. On the other hand, humanities research has in many funding schemes been on the periphery of funders' attention, often considered as, 'less important' or, 'less crucial'. The movement towards multidisciplinary research, especially research directly connected to the grand/societal challenges, and its underpinning, global dimension', will hopefully bring the humanities and social sciences to the forefront of major research topics. Thus it is not surprising that the current European Commissioner for Research, Innovation and Science Máire Geoghegan-Quinn, speaking about the role of SSH in Horizon 2020, stated in Vilnius (September, 2013) at a major conference entitled *Horizons for Social Sciences and Humanities* that:

> *"... the social sciences and humanities are anchored at the heart of Horizon 2020."*[4]

And in order to achieve the high aims set by Horizon 2020 and to come to a deeply-rooted grounding of SSH in the grand/societal challenges, it goes without saying that one of the major prerequisites is the enhancement of the 'visibility' of existing Humanities RIs as well as those we find in the social sciences. It is also necessary to identify and fill gaps, in the sense of establishing RIs that are necessary in addressing specific issues inherent in the grand/societal challenges, and work towards collaboration between research domains by identifying those Humanities RIs necessary for contributing to specific research topics. In a nutshell, this implies filling in gaps where specifically oriented data does not exist, but also connecting data where it does exist but lives a life of its own in an unconnected place. Concentrated efforts in such directions could provide a better foundation not only for future research but could also bring new direction into the RIs landscape.

4 See *Vilnius Declaration – Horizons for Social Sciences and Humanities*. horizons.mruni.eu/

In the context of the above, it is worth noting that major efforts should be made in increasing the 'visibility' of Humanities RIs, and that these efforts should be accelerated. A case in point, or an illustration, as to how difficult it is to increase the 'visibility' of Humanities RIs, comes from MERIL (*Mapping the European Research Infrastructure Landscape*)[5], a project of the ESF (*European Science Foundation*) funded by the European Commission through Framework Programme 7.

From the very beginnings of MERIL, both the humanities and social sciences have been an integral part of the concept behind establishing such a portal. However, work on the MERIL portal has shown over the years that, despite the fact that National Data Intermediaries (NDIs) were provided with elaborated categories for the inclusion of SSH RIs, concrete proposals for inclusion in the portal were few and far between. During regular workshops for the NDIs together with the Members of the Steering Committee of MERIL, at least part of the problem was identified. What surfaced was that SSH RIs do not appear on National Roadmaps and are subsequently not put forward for inclusion on the MERIL portal. This reflects the low 'visibility' of Humanities RIs, and what is more, the low level of their recognition, within national RIs communities as well as funders at national level. It is interesting to note that one of the arguments put forward for not including them in proposals for the MERIL portal was that they were ‚national', and hence did not fulfil the requirement of MERIL that the RIs showcased on the portal had to be, '... of more-than-national relevance ...'.[6]

This claim in many cases reflects the lack of understanding of the relation between the 'content' of a Humanities RI and its users, and, following this, its status. Namely, the 'content' of an RI may be national, but its relevance can at the same time go far beyond what is 'national' and be seen not only as of European relevance, but of global relevance as well.[7] A case in point are, for example, language corpora in which the ‚content' is linguistic data of a specific language, be it English, Czech, German or Croatian. However, the users, be they philologists, linguists or others, are spread all over Europe as well as globally, and this brings what may seem to be a 'national' RI into a completely different perspective.

5 For the history and the MERIL portal see portal.meril.eu/. The author of this paper has been a member of the Steering Committee of MERIL since its beginnings in 2010.

6 portal.meril.eu/

7 See Žic Fuchs (2014: 159-160) on the national 'relativity of concepts' in determining audiences for Humanities Research outputs. This notion is also applicable to quite a number of Humanities RIs.

The MERIL Steering Committee together with the NDIs is putting in extra efforts in raising the level of understanding of RIs in the humanities and is pushing for the inclusion of Humanities RIs in the MERIL portal. This is by no means an easy endavour for it implies a learning process and also aims to change, sometimes deeply-rooted, misconceptions.

Major efforts to raise awareness of the existence of RIs in the humanities, and hence their 'visibility', can be attributed to the former Standing Committee for the Humanities (SCH) of the ESF. In 2011, the SCH published a Science Policy Briefing entitled *Research Infrastructures in the Digital Humanities*.[8]

8 www.esf.org/fileadmin/Public_documents/Publications/spb42_RI_DigitalHumanities.pdf. The SPB was the result of intense efforts of the RI Expert Group of the SCH, chaired by Professor Claudine Moulin. The work of the RI Expert Group was supported by input and collaboration of more than fifty European scholars working on Humanities RIs as well as those from the US and Japan. The final document was revised after comments from external peer reviewers.

This publication covers not only an overview of the history of RIs in the domain of the humanities – dating back as early as 3rd Century B.C., to the *Mouseion* (a cultural centre, university and library) founded in Alexandria – but also gives an overview and typology of Humanities RIs up to the present, with special emphasis on what we find within the wide context of Digital Humanities. In order to show the diversity of what can be found under the heading of Digital Humanities alone, we list the four primary levels of RIs identified:

- Physical infrastructures: Collections of physical objects
- Digital data infrastructures: these comprise single-sited or interconnected data repositories, spread over several institutions/countries
- E-infrastructures: network and/or computing facilities spread over various institutions and/or countries – examples include GRID computing, cluster computing, cloud computing and the networks that connect them
- Meta-infrastructures: conglomerates of independent RIs, residing in different institutions/countries with different data formats and data structures (i.e., resulting from different activities) yet connected using compatible metadata formats or processes, thus enabling access to different data archives

Another important push towards making Humanities RIs more visible has come from ESFRI (*European Strategy Forum on Research Infrastructures*), or to be more precise, the inclusion of two 'pan-European' infrastructures on the ESFRI Roadmap – CLARIN (*The Common Language Resources and Technology Infrastructure*) and DARIAH (*Digital Research Infrastructures for the Arts and Humanities*).[9]

The advent of Digital Infrastructures of different kinds in the humanities, and especially the 'pan-European' CLARIN and DARIAH, offers scholars new and productive ways of exploring old questions and opening up new ones. Apart from making accessible our cultural heritage in digital form, the Digital Humanities also open up new front lines in the sense of developing questions aimed at a better understanding of the complex, many-layered implications of the grand/societal challenges and possible effects on the individual, societies and the world at large.

9 DARIAH: www.dariah.eu/, CLARIN: www.clarin.eu/

‘Impact‘ and related issues

1 On the basis of the above mentioned efforts, one could come to the conclusion that the ‘visibility’ of Humanities RIs is slowly increasing, and to a certain extent this is true. However, at the same time, the rate of assessing RIs can be said to be accelerating, and this brings to the forefront the questions: is the process of making Humanities RIs more visible (and more recognized) in step with the rapidly increasing rate of assessment exercises appearing not only for those RIs on the ESFRI Roadmap, but also those found at national level, on national roadmaps.

In order to accelerate the process of making Humanities RIs a full-fledged and recognized part of the RI landscape, the notion of ‘impact’ has to be taken up and integrated into the way RIs present themselves not only to policy makers and funders but also to the research communities they are linked to. It should not be forgotten that RIs are ‘built’ and set up to facilitate and generate research with the aim of taking it to higher levels in terms of both the quantity and especially of the quality of research results and outputs. And it is precisely this fact that is often not stressed enough, or is even understated in the case of Humanities RIs.

Measuring the ‘impact’, as a multi-faceted outcome of research based on a research infrastructure, may well be one of the future directions and developments in assessing pan-European as well as national RIs. How measuring ‚impact‘ will be defined and implemented in the future remains to be seen, since ‘impact’ as a concept and an outcome is not easily defined, nor can it be readily elaborated. However, the focal, underpinning question or element of ‘impact’, is how do RIs foster concrete research and how do they provide the basis for concrete research results. More precisely, what is their true role in achieving answers to research questions and challenges. This question becomes even more relevant if one accepts the claim that the academic community of humanities scholars in Europe is in fact extremely numerous. Unfortunately, concrete data pertaining to this claim is still not available, although, on the basis of common sense, overviews of the different communities representing the long list of humanities disciplines would indicate that it is true.

Thus, assessing the ‘number of users’ per discipline within the humanities can again be seen as a ‘relative concept’, especially since cross-disciplinary research is becoming more widespread. A starting point could be to estimate the number of users per discipline connected or using a concrete RI. From such a viewpoint

one could gain deeper insights into how different disciplines harvest the benefits RIs provide for fostering concrete research. A more diversified view of the 'user structure' would also indicate to what extent Humanities RIs are 'visible' to the different segments of what certainly seems to be a huge community of scholars in the domain of the humanities.

The second general issue pertaining to 'impact', one often raised in assessment exercises, is that of how pan-European RIs in the humanities can transcend national level databases of different kinds. Namely, in RI assessment environments, one sometimes hears the query whether pan-European RIs in the humanities are simply agglomerations or collections of national databases that result in forms of 'networks'. Some RI specialists question whether such 'networks' are in fact research infrastructures in the true sense of the word, and following this line of thought they may question whether and how Humanities RIs provide the basis for achieving added value in research.

All of the questions and dilemmas mentioned above indicate that it is of the utmost importance to showcase 'impact' of Humanities RIs in research, and simultaneously make great efforts to increase their visibility and recognition. The question is, of course, how can this, often dual, goal be achieved.

2

In July 2013, young researchers in the Digital Humanities launched a Manifesto[10] in which they advocate the necessity of bringing closer the flourishing digital practices and their institutional acknowledgement, or more precisely, making research within Digital Humanities an integral part of promotions and of applications. They clearly state in the Manifesto:

> "The **widening gap between flourishing digital practices and their institutional acknowledgment** represent a threat for the academic community as a whole and for young scholars in particular, since it casts uncertainty on their future as research professionals." (emphasis mine)

The young researchers that instigated the launch of the Manifesto are in my opinion quite right in advocating a wider perspective in assessing research performance and pushing for more concentrated efforts in articulating the need for including, "digital outputs in the evolution of scholars, whether it be for promotions or in job applications."[11] Namely, so-called digital outputs are not

10 *Young Reserchers in the Digital Humanities: A Manifesto*. dhdhi.hypotheses.org/1855.

11 This line of thinking can also be found in the *San Francisco Declaration on Research Assessment* (DORA), am.ascb.org/dora/. This declaration instigated by researchers from the so-called hard sciences stresses the following: >>

simple ‚technical devices', but their 'construction' implies extensive knowledge of the content at hand, as well as the know-how of setting it up in such a way that it can provide a well-grounded basis for its users. Raising awareness of the necessity of including new forms of scholarly outputs, such as the creation of databases, development of technical tools, dynamic bibliographies, wikis, etc., is to my mind not only a necessary form of recognition for the important work and endeavours of (often) early career researchers, but a move in this direction would also be *a step forward in raising the 'visibility' and showcasing the research infrastructures one finds in the humanities domain today.*

Academic recognition of digital practices in both the humanities and social sciences would at the same time help to stress the important role of SSH in the above mentioned inter-/multi-/transdisciplinary trends in research, for RIs are not simply the basis for research, but can also open up new vistas in cross-domain research. In this sense it is worth quoting the following from the Manifesto:

> "The Humanities and Social Sciences are a vital component of human culture and offer an essential insight into the world in which we live. The Digital Humanities reflect the transition of the Humanities to the digital age. However, they do not only bring with them **new technical means**, but also **new forms of knowledge creation and dissemination within, across and outside academic disciplines**." (emphasis mine)

3 As already mentioned, 'number of users' can be one of the parameters used in RIs; both those that come under the rubric of being pan-Europeans as well as those that are labelled 'national'. However, 'number of users' may not fully reflect the impact a RI has in terms of concrete research results and innovative findings. One of the ways of showcasing the research outcomes or results is by keeping track not only of the number of Ph.D.s that have grown out of a specific RI, but also of the topics researched. The question thus arises of whether we know, for instance, how many Ph.D.s have been based on the data provided by CLARIN and DARIAH? It would be beneficial for future assessments to be able to demonstrate ‚impact' by keeping track of Ph.D.s related to a specific research infrastructure. In Croatia, for instance, there are two large

"... for the purpose of research assessment consider the value and impact of **all research outputs (including databases and software)**." (emphasis mine)

Also see *EuroScientist*, February 2014, on various issues pertaining to digitally-enhanced research, as well as its potential evaluation.

language corpora[12] used extensively by Croatian linguists as well as Slavic scholars world-wide. Quite a number of Ph.D.s have been based on either or both these corpora, but the (more or less exact) number is not known, nor are the titles showcased on introductory web-pages. Since so-called 'national' RIs also undergo assessment, if nothing else for funding purposes, this information could prove a beneficial indicator of impact for funders.

Systematic showcasing of journal articles, books/monographs (especially important in the humanities), in which the research itself is based on a RI, especially a pan-European one, clearly shows how a RI provides a basis and added value for assessing the impact of concrete research results. Only one example follows, but a highly illustrative one: a paper by Enhard Heinrichs submitted to the linguistic journal *Lingua* (Special Issue) showing how CLARIN has made it possible to solve a major linguistic question pertaining to an age-old issue in German linguistics. It is worth quoting extensively from the abstract submitted to *Lingua*:

> "The historical development and linguistic environments for auxiliary fronting in German is an old research question in German linguistics, dating at least as far back as Grimm's famous *Deutsche Grammatik* (Grimm 1891) ...
>
> The use of historical and synchronic corpora, which include relevant levels of linguistic annotations, makes it possible to track the syntactic environments of this construction and to witness syntactic change across time ...
>
> Such evidence has not been available until very recently, due to the unavailability of electronically corpora with sufficient amounts of data and linguistic annotations. The availability of such corpora as part of the Common Language Resources and Technology Infrastructure (CLARIN) has made it possible to fill this gap.
>
> **A crucial aspect of the linguistic investigation made possible as part of the CLARIN infrastructure concerns the interoperability of the treebank and DTA resources mentioned above. Since all resources involved share a common layer of part-of-speech annotation, using the same STTS tagset for German, it becomes possible to search for the same patterns in all resources and thus track linguistic change over more than four centuries.**" (emphasis mine)

12 Croatian National Corpus (www.hnk.ffzg.hr/cnc.htm) and Croatian Language Repository (riznica.ihjj.hr/)

The above showcases an example of a research result that shows clearly the necessity for pan-European RIs and, what is more, *it shows how they transcend the misconceptions that they are simply networks of national level databases*. More precisely, it illustrates the 'impact' of the research potential of a pan-European RI in the humanities. Moreover, articles such as the above, printed in *traditional journals*, can raise awareness among those linguists who are not familiar with, for instance CLARIN, and can subsequently induce new research based not only on extensive linguistic data, but data which provide different possibilities of interoperability. This in turn provides new possibilities for researching long-standing questions and dilemmas which have in some cases baffled linguists for decades or even centuries.

A few more steps towards 'Facing the Future'

It is well-known that predicting the future may be a risky exercise. However, in my opinion, a few points may be brought up in the context of 'impact' in future assessments of RIs. First and foremost, it is a well-known fact that humanities scholars very often work alone, but the research landscape is and has been changing during the last couple of decades. In this sense we have been witnessing the appearance of 'larger research projects'. Taking this into consideration, a future parameter of 'impact' could be the development of complex humanities-based projects in which larger groups of scholars directly link their research to pan-European Humanities RIs.

Following the already mentioned trends towards inter-/multi-/transdisciplinary research one can readily envisage the setting up of databases focused on specific research topics reflecting the multi-sided nature of the grand/societal challenges. Research infrastructures that would be geared towards achieving synergy for high-level multidisciplinarity could be expected to be viewed as having a high level of 'impact'.

And last but not least, in a very futuristic vein, one can envisage interactive research infrastructures, which would be easily accessible to researchers from all domains, globally. Needless to say, much has to be done within the whole RI landscape in order to achieve such a goal. However, future 'interactive RIs' would in many ways achieve the ultimate as far as 'impact' is concerned.

Instead of a conclusion

It is very difficult to write a conclusion on what is, in part, a vision of what may develop in the future. The aim of this paper was to, at least in some respects, shed light on the notions of 'visibility' and 'impact' in their inter-relatedness within possible future assessments. Although quite a lot of attention has been paid to inter/multi/transdisciplinary research, there is one essential feature of humanities research that should not be forgotten, and which Martin Wynne clearly articulates in his 2013 article *The Role of CLARIN in Digital Transformations in the Humanities*:

> "To argue for a digital humanities which is primarily concerned with the accumulation and analysis of data, and which has goals of promoting the economy, and other specific social or political goals (such as fighting terrorism) would do a disservice to the humanities. As Stanley Fish and others have noted, **there is a priority of the humanities to stand up for its own traditional values today**. (*emphasis mine*) ... We need to follow the sciences in deciding priorities, adopting standards, reducing complexity and variety, but only as pragmatic measures to promote shared facilities and infrastructures. At the same time, we need to avoid the promotion of an excessively data-driven, empirical and scientistic view of the humanities, and continue to defend the traditions of qualitative research in the humanities, and pursue the humanities for their own sake."

Despite new trends, despite possible future developments of new RIs, room has to be left for the further development of existing RIs, as well as for the appearance of new ones, that would primarily be geared towards answering long-standing questions within the disciplines of the humanities themselves. This goal should never be let out of sight.

References

Heinrichs, E. (paper submitted): Using Large CLARIN Corpus Data to Trade Syntactic Change: The Case Study of Auxiliary Fronting in German. submitted to Lingua.

Wynne, M. (2013): The Role of CLARIN in Digital Transformations in the Humanities. International Journal of Humanities and Arts Computing 7, 89–104.

Žic Fuchs, M. (2014): Bibliometrics: Use and Abuse in the Humanities. In: Blockmans, W./Engwall, L. and Weaire, D. (eds.): Bibliometrics, Use and Abuse in the Review of Research Performance. London: Portland Press, 105–114.

5.2 The Humanities and Social Sciences Confronted with the Challenges of Interdisciplinarity and Big Data

Philippe Vendrix (CNRS)

The Sciences at a Turning Point

For the humanities and social sciences, recent years have been the scene of radical upheaval, provoked essentially, but not exclusively, by three scientific orientations or practices. The first of these is the construction of new tools inducing the establishment of large technical infrastructures and the refounding of the fundamental principles of philology, and more generally, heuristics. Grouped under the generic designation "digital humanities", these tools were initially conceived for the particular needs of individual disciplines: sociology, literature, the arts, etc. They led to the creation of vast national and international programmes, the progressive development of common languages (TEI being a notable example) and the definition of new competences based on the interpenetration of two fields of knowledge: one the one hand, the humanities (designated in French universities by the acronym "SHS": *Sciences de l'homme et de la société*); on the other, technology and information science. The second innovation took the form of a global movement, sustained by both civil society and political decision makers, which promoted the idea of "*sciences en société*". Researchers consequently found themselves obliged to position their work within the vast and complex network of society, of societies; a society of knowledge being understood to owe its existence to its capacity to demonstrate its vitality. The list of scholarly publications no longer sufficed – though it remained, and indeed remains indispensable, contrary to what has sometimes been suggested – as proof of scientific production. In other words, the scientist is now required to conceive his research project within a judiciously composed kaleidoscope of intellectual originality, pedagogical scope, societal impact and cultural anchorage. Researchers and decision makers alike currently find themselves confronted with the panoply of exigencies implied by research's societal dimension. Finally, there is the combined response to these two changes: to the confrontation of the digital humanities and the new socio-economic demands. Conceivers of research programmes today not only face new challenges, but

also considerably increased financial risks necessitated by setting up ambitious projects whose scope reaches well beyond the working framework within which they was hitherto used to operating (until recently, for most branches of the humanities, rather rudimentary technological resources, exploited to precise ends by a few investigators, sufficed to produce publishable results). In order to provide guarantees (necessarily relative) of success with respect to these new exigencies, interdisciplinarity appears to be a promising way forward.

Interdisciplinarity necessitates that researchers themselves pose some important questions: it would be useless to dismiss these from the debate. But from the moment that each researcher, whatever their discipline, acknowledges their own contribution to science, there is nothing standing in the way of a fruitful dialogue. Incidentally, the picture broadly outlined above is not intended to throw into shadow other initiatives, such as the critical study of fundamental works. Everything is a question of scale: there are individual, collective and collaborative projects, to use generic terminology that needs to be nuanced. The difference between these three levels could be understood to revolve around the three challenges mentioned above (the researcher addresses a community of readers, just as collaborative projects suppose, due to the means expended, a certain socio-economic or socio-cultural impact).

Interdisciplinarity is not the only thing that raises essential questions. The digital humanities, besides revising philological practices, question rights – notably those pertaining to the access of knowledge, its diffusion and its appropriation by the public. The cases of music and rights linked to the creative arts clearly illustrate this legislative issue: the most popular music, that which is most fundamentally anchored in the social practices of contemporary society, is also that which is excluded, and radically so, from all scientific enterprises, but which could, however, lay claim to the digital humanities a strong societal impact and an interdisciplinary approach (a “hit” is more than just a tune).

Nobody today can claim to be able to formulate the question that will suddenly and unanimously orient the community of researchers towards an ideal solution. There is no more an ideal solution than there is an ideal question. On the other hand, inquiries supported by European and national organisations and initiatives undertaken by researchers keen to tackle the three main challenges – big data, societal impact and interdisciplinarity – connect regularly and productively. This article will detail the objectives of one of these initiatives – *Intelligence des Patrimoines* – with an aim to providing an idea of how the challenges are currently being met.

A Project Combining Numerous Fields of Knowledge

The *Intelligence des Patrimoines* project is situated at the intersection of a series of social, economic, political and cultural issues related to the notion of heritage on a national and European scale. The beginning of the 21st century is characterised, in France as in Europe, by a sharp rise in heritage themes. The growing anxiety of European societies, caused by their uncertain identities and future, has been accompanied by the emergence of a new perception of their cultural and natural heritage. This has given rise to a clamour of political, economic, social, cultural and heritage-linked claims, which make clear the general desire to preserve this heritage. On a local level too, the multiplication of associations, museums and festivals related to heritage also testifies to the social embedding of this notion. A movement of similar amplitude is equally taking hold in the domain of nature conservation. There is no longer any doubt that we have entered a new era, the Anthropocene, in which our environment is subject to changes on all scales, from the most local to the most global, principally due to perturbations induced by man. For this reason, the sustainability of our development, that is, the compatibility between current development and that of the future, has emerged as a particularly imposing requirement. The three pillars of sustainable development are clearly economic, environmental and social equity. Its attainment also relies on a multidisciplinary approach which *Intelligence des Patrimoines* aims to facilitate and reinforce. The capacity to provide a response to the new questions raised by highly complex and integrated socio-ecological systems depends on inventing new ways of addressing and treating them, new institutions and training programmes.

In view of this context, *Intelligence des Patrimoines* proposes an original approach to heritage by weaving together the scientific, cultural and social aspects, and by adopting new scientific approaches in order to understand and to develop new social, cultural and political practices, as well as to valorise and transfer these results. The aim is thus to combine varied scientific competences in order to analyse the notion of heritage in all its complexity, to study the multiplicity of its uses and to reveal new perspectives in research, training and economic development.

Combining different scientific epistemological approaches represents a very real challenge that we wish to meet. It is customary, for example, when one wants to evaluate ecosystem services (services offered to man by ecosystems), to call on economics. While this raises multiple questions, it is perfectly viable in certain cases: for example, when quantifying the contribution of polli-

nators in food production, or the role of forests in the filtration of drinking water. However, the apprehension of certain other ecosystem services, notably cultural, but also those linked to the conservation of biodiversity or natural milieu, is far more difficult to quantify. This does not mean that these questions are less at the heart of our project. When monetary quantification is possible, the figures speak for themselves. New methods, in particular psychological, can be introduced in order to quantify intangible things – the beauty of a landscape, for example, or the conservation of sacred forests –, without necessarily seeking to monetise that which cannot be monetised.

The academic world is the ideal place for encouraging the convergence and inter-fecundation of approaches, thanks to its diversity, its balanced structure and its fundamentally exploratory and innovative way of proceeding: a new world calls for new solutions produced by a new ferment. This can only happen under certain conditions, necessitating an evolution of the academic world, especially since time is short. Let us mention:

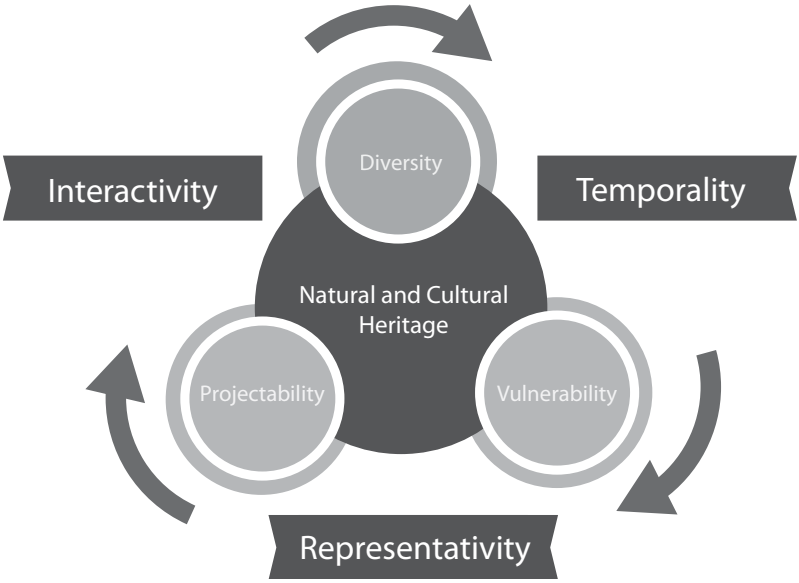
1. The creation of supple and easily maintained links with NGOs, enterprises, territorial organisations and other stakeholders, often in the form of new institutions created to serve as interfaces.

2. The establishment of effective policies based on evidence rather than *a priori* judgements, and a new type of research, distinct from the habitual duo "fundamental research – applied research". This new type of research aims, in the tradition of fundamental research, to pursue excellence, but by applying itself to objects and to the nexuses of conflicts identified by the community as a whole.

3. The development of means of valorising and of welcoming partners situated close to the laboratories.

The *Intelligence des Patrimoines* project is built on the redefinition of traditional approaches to natural and cultural heritage. If the science of materials and the humanities abundantly collaborate and continue to do so, if natural science has regular recourse to mathematics, bringing together specialists in living organisms, environmental questions, man and society, technology, information and communication into a single entity incontestably remains an original ambition in the European research context. The innovative character of this interdisciplinary project aligns itself with an increasingly pressing desire to develop new approaches to heritage.

The scientific project is constructed on the identification of a series of site-based research projects characterised by a growing number of epistemological, technical and scientific issues, stimulating enquiry in the humanities and social sciences (SHS), the science of living organisms (SDV) and the information and communication technologies domain (STIC). Founded on the experience of certain collaborations already well underway between partnering laboratories, *Intelligence des Patrimoines* aims to extend and amplify these overlapping enquiries. It will take into account existing competencies which constitute the base of future development in order to call forth original, innovating projects in all the concerned domains.



Interplay of this double triad illustrated by the systemic schema of the project

Given the breadth of natural and cultural heritage, three points need to be immediately stressed:

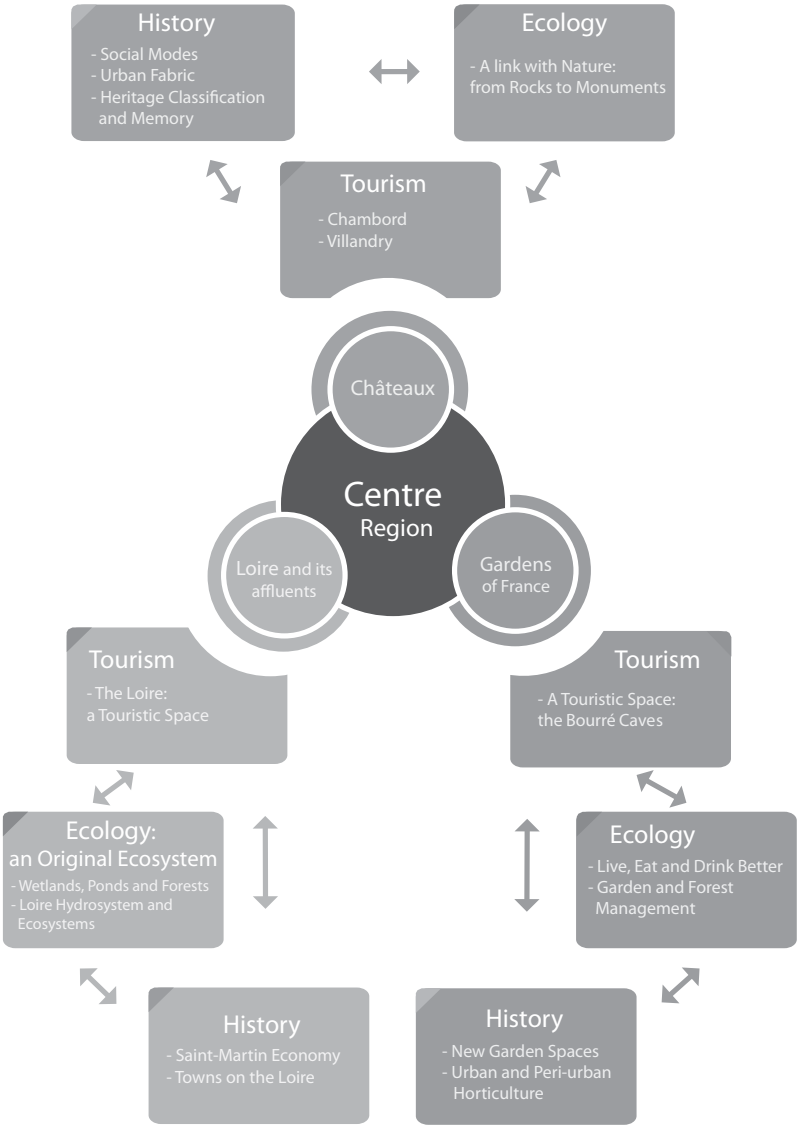
- their diversity;
- their vulnerability;
- their projectability.

Diversity, vulnerability and projectability, however, can only be thoroughly appreciated as pertinent notions when viewed from the triple perspective of:

- temporality;
- representativity;
- interactivity.

Our way of proceeding consists of assessing natural and cultural heritage by bringing into play the tension provoked by the concomitance of factual observations and analytical concepts. While it is true that pre-visibility is not strictly factual, taking it into account in a global procedure induces dynamism. Heritage is at once past and present. By acknowledging its diversity and vulnerability, it becomes apparent that its future also needs to be taken into account. Through collaborative work, *Intelligence des Patrimoines* thus seeks to create new models for analysis and predictability which, after being applied to heritage in the Centre Region, could easily be adapted to other places in France and the rest of the world. In addition, the concomitance of cultural and natural heritage will facilitate the study of the value of heritage independently of its nature and adoption of the multidisciplinary approach necessary for their own sustainability and that of the region's development.

Natural and cultural heritage covers a vast field of objects, the apprehension of which, from the point of view of this project, necessitates a process of identification. Following the example of the epistemological framework that constitutes the foundation of *Intelligence des Patrimoines*, it is by way of a double spiral of concepts that the areas of intervention have been defined. The first spiral highlights three strong elements of the Centre Region's identity: its agricultural land, its historical monuments and the Loire and its tributaries. Each of these elements can then be viewed from three complementary perspectives: history, ecology and tourism. The diagram on the next page synthesises the **interaction of the two spirals**.



Interaction of the two spirals

Three field projects could be rapidly initiated. These field projects not only cover a wide geographical area within the Centre Region, they also mobilise a large range of indicators, meaning that the impact of this resolutely innovative project would be easy to assess. The field projects are also conceived as illustrations of the both the attractiveness of the project and the potentially exportable (other sites in France, Europe and elsewhere) analytical models.

- Field project 1 : **The National Domain of Chambord**
- Field project 2 : **Eat, Drink and Live Better**
- Field project 3 : **Ecosystem and Hydrosystem of the Loire**

From the Field Project to Big Data Management

Beyond the research projects, *Intelligence des Patrimoines* will substantiate its claim to being interdisciplinary by setting up an infrastructure regrouping all the collected data. The construction of such an infrastructure necessarily involves concerted reflection on the nature of the different corpora to be interrelated and the ways and means of exploiting these corpora while respecting the project's aim to confront natural and cultural heritage.

Heritage Characterisation: en Masse and in its Particularities

Considerable advances made by the TICs (Technology Innovation Centres) and the development of Next-Generation Sequencing (NGS) already enable us to acquire complex data automatically and at high-speed. Where **cultural heritage** is concerned, the project will target an ensemble of data drawn from distinct corpora: excavation data, visual data, audio recordings, prosopographic databases, musical corpora and textual corpora. All have been the object of intense inquiry over the last decade. In the field of food studies, for example, we now dispose of video recordings issued from surveys and oral archives have been created. Archaeologists, likewise, have made great leaps in the creation of archives, including, on the one hand, non reproducible primary data (recordings, 2D and 3D graphics, photos, etc.) and, on the other, collections of artefacts and vestiges of the material cultural of past and present societies. The creation of a digital database dedicated to the Loire valley châteaux is being prepared on the basis of the referencing system of the Rihvage archival and bibliographic database. The digitising, referencing and online publication of the vast

corpus of works and archives conserved in the Centre Region will be pursued and valorised using advanced treatments developed by the TIC in liaison with the digital humanities (automation, interoperability, stability and data access). This work concerns both textual corpora (BVH) and musical corpora (Ricecar), both of which occupy a prominent place in today's worldwide network of online resources.

On the natural heritage front, the project will particularly target biodiversity – a fundamental element of the natural heritage richness of a territory. Priority will be accorded to adaptive traits – notably in relation to climatic changes and anthropic impacts – and resistance. These characteristics will be evaluated by means of a wide range of measurements (phenotyping and genotyping), collected at high speed (using automated and structured data capture and image analysis) and under different conditions (the characteristics of which will equally be acquired and included in the database), in order to appreciate the effects of environmental changes. Different ways of valorising this biological heritage, drawing on the characterisation and screening of natural biological resources (production of new molecules; development of new environmental surveillance indicators; development of new bioprocedures), will be researched and characterised. The project will also study the interactions between organisms and their milieu, as well as interactions between organisms – a keystone of ecosystem and agricultural-system functioning, a determining factor in the majority of ecosystem services and useful, in certain cases, for developing disruptive bio-eco technology based on the valorisation of this natural biological heritage. As the origin of the numerous ecosystem services and the majority of emerging illnesses, insects will be the object of studies dedicated to assessing and monitoring their biodiversity (as well as that of associated micro-organisms), their role in agricultural systems and natural ecosystems, and their adaptability in the face of environmental restrictions (climatic changes, biological invasions and the evolution of management practices). These studies also aim to contribute to the development of tools for ecological engineering with societal interests (pollution control and site clean-up). They will benefit from new biodiversity inventory methods combining various forms of innovative technology (biomimetism, ADN taxonomy and high-speed sequencing), which will enable the creation of an ADN barcode reference library.

Forest resources will be studied in terms of their plasticity with regards to climatic changes. Monitoring will combine in-situ (forests) and ex-situ (controlled-situation climatic rooms) observations. Animal genetic resources, selected and preserved in various regions (CRB Anim project), will be bred under

different conditions and characterised in detail (fine phenotyping) in order to identify favourable alleles (resistance, adaptability), valorise these genetic resources and contribute to the development of regional produce (AOC cheeses). At ground level – soil, sediments, subsoil and underground water systems –, we will structure the characterisation, conservation (notably, by setting up a data bank of strains) and valorisation of environmental bacterial heritage with major biotechnological potential.

Data Analysis: Interoperability and Semantic Mediation

In compliment to the analysis of the acquired big data, the identification of indicators and their application also implies being able to deal with the largest amount of available data possible, drawn from all the many facets of the region's heritage. Due to the diversity and the variety of the traits under consideration, this will necessitate using **interoperable databases** and **semantic mediation**. Establishing the interoperability of our databases will be one of the goals of the project. It will rely on common coding and protocols at a technical level, *on shared formats and schemas at the syntactic level, and on a consensus concerning what should be represented and how at the semantic level*. Semantic interoperability is one of the major challenges presented by data valorisation in the humanities. It also plays an increasingly important role in the study of the various different scales of living heritage, from genes to milieu. Finally, it contributes to the development of cartographic interfaces (web mapping) – particularly important for applications destined for the tourist industry. A first application of this type will be created for the Centre Region's pond heritage, with an aim to fully exploiting its touristic value. If technical and syntactic interoperability will require precise frameworks in the form of technical protocols and shared, rigorous syntactic formats, the quest for a consensus in terms of semantic representation (knowledge engineering) – extremely useful and a source of much knowledge –, cannot hope to achieve a completely common, completely normative framework. The resulting heterogeneity is classically dealt with by data integration, a complementary research domain. This process combines data collected from different sources in a model of the domain of application (called a global schema), thus permitting the data's exploitation by means of a common query interface. Semantic mediation consists in conceiving the global schema in the form of an ontology and describing the "mappings" between this ontology and each of the data sources. Conceptual representation of a domain, fruit of a consensus, and equipped with automatic inference capacities, the ontology plays the role of integrator.

In this way, different characteristics of genetic resources from very different species can be found to possess similar physiological mechanisms (growth, behaviour, resistance), and be conjointly used to identify conserved genes, once the heavy work of elaborating these correspondences, both at the level of the traits and at that of the conditions of the milieu. Semantic mediation is the system that enables the sources to be queried via the global ontology. One of the major points of interest presented by this recent approach – still being developed by a number of national and international research projects – resides in the autonomy it accords the sources. Besides providing centralised access to an ensemble of existing resources, this semantic web approach not only allows data to be formally represented, but also the sense that we attribute to it, in the form of ontologies, thereby enabling it to be automatically exploited. It thus offers a framework for inventing and constructing new models and query tools for very diverse resources.

This approach will be applied to cultural heritage in order to identify the specific data of each corpus: chrono-thematic indexation of video archives and linguistic analysis for the food corpora; dematerialisation of archaeological and archaeozoological reference collections and semantic mapping of existing indexes using CIDOC-CRM (conceptual reference model for cultural heritage data) ontological structuration. The semantic model enabling interoperability will take into account the structuration of the different types of data and the relations between them in order to retrieve rich and "contextualised" information. This interrelated approach to data will lead to a new understanding of the region's cultural heritage. One of the main campaigns of actions will be the digitisation of cultural heritage archives, either already assembled or in the process of being assembled (manuscripts and printed texts; images; objects; oral, sonorous and multimedia archives; soil archives, which constitute the primary data in archaeology and archaeozoology, etc.). Taken as a whole, these archives represent a vast corpus of disseminated data (stored in ancient archives, in the soil and subsoil), that we aim to render interoperable with the help of semantic language. In certain disciplines, digitisation is already underway, but requires reinforcement in order to avoid being called into question; in others (like history and history of art), it needs to be rapidly initiated in order to make up for lost time. The objective, in the medium term, is to renew the processes of data treatment and publication and, more generally, the way of working with the cultural heritage of the Val de Loire. Firmly embedded in the framework of regional programmes and linked at national level to the Huma-Num TGIR (*Très Grande*

Infrastructure de Recherche), the ambition of this project is to offer, for the first time in France, digital access to heritage data and a new scientific comprehension of cultural heritage in the Centre Region.

Regarding animal and vegetal resources, the identification of mechanisms explaining the most remarkable characteristics will be accelerated by the access provided by the semantic web to the ensemble of available data concerning similar characteristics. These studies will facilitate the selection of organisms presenting resistance and/or adaptability characteristics. Once the experimental proof of this approach is established, the methods developed will be transferred to selected organisms and, following this, to the concerned industries (agricultural and industrial). In liaison with the creation of a bank of environmental bacterial strains – a bio/geochemical database centralising the georeferencing of samples –, the physico-chemical characteristics describing original samples and the microbiological and molecular characteristics of conserved isolates will be structured.

Conclusions

Intelligence des Patrimoines can hope to offer firm guarantees of success no more and no less than any other project can. It does, however, dispose of undeniable assets: an exceptional field of investigation (it suffices to refer to the National Domain of Chambord); teams issued from numerous laboratories who have devoted months to the elaboration of a federating project; solid financial support. The stakes are thus elsewhere. On the one hand, it will be necessary to demonstrate the scientific validity of a resolutely interdisciplinary approach while avoiding the pitfall of simple addition, tiresome from all points of view. On the other, it will be necessary to prove the efficiency of the tool, i.e. an infrastructure uniting data of a highly diverse nature. In addition to these two stumbling blocks, it will be necessary to spread this mode of approach, through cooperation, to other projects, other territories, actuated by similar dynamics, but involving different forms of inquiry. If this wager is successful, it may well provide a response to the challenges that society and science present us with today.

5.3 Open Access to Bibliodiversity
Issues surrounding open digital publishing infrastructures in the humanities and social sciences

Marin Dacos[1] (OpenEdition)

In the era of the cloud, software as a service (SAS), big data and global digital giants, the debate surrounding European initiatives on digital research infrastructures seems unavoidable[2]. Awareness of this issue dates back to 2006 and was American-led.[3] Since then, European actors have also woken up to the issue, in particular thanks to the ESFRI roadmap,[4] which put digital technology on the agenda. In terms of the humanities and social sciences (HSS), it is apparent that the resources mobilised are modest, and disproportionate to the academic stakes. In this respect, the Strategy Report on Research Infrastructures (2010) is particularly enlightening. If we take the construction costs of European ESFRI infrastructures in all disciplines, we find that the humanities and social sciences represent only 1% of these costs.[5] And yet, the stakes – of constructing, mobilising, reusing, interconnecting, conserving, disseminating and developing data, results and publications in the HSS – are high,[6] both from an academic point of view (exploiting digital technology to enhance the cumulative nature of results) and a societal one. It is no coincidence that H2020 points to culture as playing a structural role in Europe's development.

1 Founding director of OpenEdition, CNRS, Aix-Marseille University, EHESS, Avignon University, CLEO UMS 3287, 13284 Marseille, France.

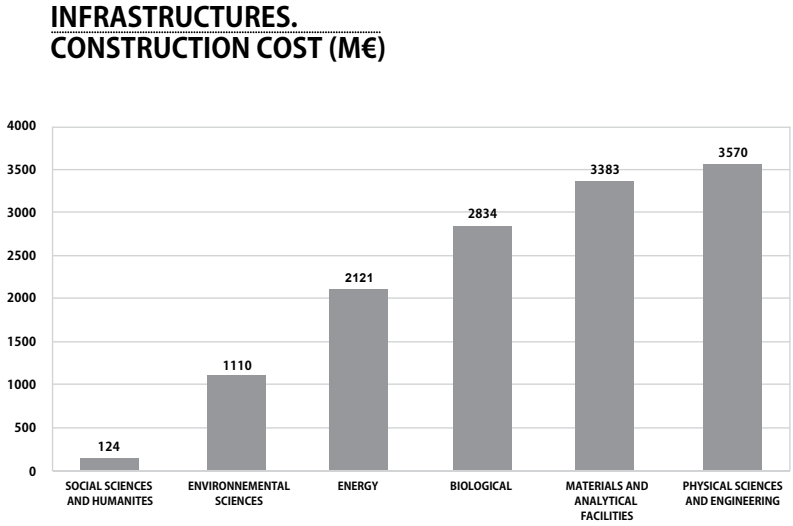
2 This text has been translated from the French by Helen Tomlinson.

3 American Council of Learned Societies (2006): Our Cultural Commonwealth. The Report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences. New York, 51.

4 ESFRI (2006): European Roadmap for Research Infrastructures. Report 2006. Luxemburg: European Commission. ec.europa.eu/research/infrastructures/pdf/esfri/esfri_roadmap/roadmap_2006/esfri_roadmap_2006_en.pdf

5 ESFRI (2011): Strategy Report on Research Infrastructures – Roadmap 2010. Publications Office of the European Union. ec.europa.eu/research/infrastructures/pdf/esfri-strategy_report_and_roadmap.pdf

6 For more information, see Dacos, Marin (2013): Cyberclio. Vers une Cyberinfrastructure au cœur de la discipline historique. In: Clavert, Frédéric and Noiret, Serge (ed.): L'histoire contemporaine à l'ère numérique / Contemporary History in the Digital Age, 29–43.



Construction costs

Bibliodiversity vs. monoculture

Let's pause and consider what is specific about Europe compared to other economically developed continents. One thing to emphasise is its cultural and linguistic diversity, summed up in a remark attributed to Umberto Eco: "The language of Europe is translation."[7] If this linguistic diversity impedes academic communication between neighbouring countries, it is ultimately a historic opportunity that assures Europe's distinctiveness. The European continent is a condensed version of global cultural diversity, that is to say, the diversity of the Web itself. The World Wide Web created by Tim Berners-Lee is not – as some might think – an anglophone expanse. Linguistic diversity has become a characteristic of the ever-expanding Web.[8]

Taking this principle further, the International Alliance of Independent Publishers advocates cultivating and reinforcing what is known as bibliodiversity.[9]

7 www.laviedesidees.fr/Le-multilinguisme-est-un-humanisme.html

8 Le Crosnier, Hervé and Vannini, Laurent (ed.) (2012): Net.lang. Réussir le cyber-espace multilingue. Caen: C&F éditions, 477. cfeditions.com/NetlangFR/.

9 www.alliance-editeurs.org/bibliodiversity?lang=en

Let's briefly define bibliodiversity as the diversity of the publishing and cultural sectors, a notion akin to that of biodiversity in biology. In culture as in ecology, diversity is a source of ecosystems' creativity and robustness. Bibliodiversity can be contrasted with monoculture, a term which, in French, applies to both the cultural and agricultural sectors. Bibliodiversity concentrates the heuristic potential of the diversity of languages and cultures. It does not entail a desire to construct a tower of Babel in which peoples, and more particularly researchers, are unable to communicate among themselves. Bibliodiversity could sit very easily with the use of a common spoken language, English as it happens, or more exactly Globish, that lingua franca that proves perfectly conducive to communicating ideas, but perhaps less so to thinking in a language other than one's mother tongue.[10]

"Core journals disease"

In academia, bibliodiversity corresponds to a diversity of languages, a diversity of disciplines, a diversity of types of publication, and a diversity of publishing actors. It runs counter to a certain image of the academic sector, one that is almost exclusively monolingual (dominated by English); focused on a few disciplines (with investment concentrated on science, technology and medicine); biased in favour of one type of publication, the article (at the expense of the book, nonetheless paramount in the humanities and social sciences, and of new publishing forms that are emerging with the Web); and prone to putting its eggs in a few rare baskets, those publishing oligopolies we all know and which impose their diktats on libraries.[11] Let's not mince our words: a monoculture centred on a single impact factor[12] flies in the face of bibliodiversity. If we are not careful, it will seriously impoverish academic endeavours and disciplinary diversity.

10 Dacos, Marin (2013): La stratégie du Sauna finlandais. Les frontières de Digital Humanities. Essai de géographie politique d'une communauté scientifique, 13.

11 See for example: thecostofknowledge.com/

12 Campbell, P. (2008): Escape from the impact factor. Ethics Sci Environ Polit 8 (5-7). doi:10.3354/esep00078.

According to the Web of Science, the HSS do not exist

A brief look at Thomson Reuters's Web of Science (WOS) confirms this hypothesis. If Thomson Reuters is to be believed, the WOS is a major reference, or "The world's most trusted citation index", notably because it includes the crème de la crème of worldwide academic literature ("covering the leading scholarly literature"). Jean-Claude Guédon, fifteen years ago, was already at pains to point out the sterilising nature of the notion of the "core journal".[13] Let's take a look at the situation of the francophone humanities and social sciences in 2014, the sector I know best. In 2013, the Web of Science included the film magazine Positif, which can be bought in newsagents and offers in-depth reporting on cinematic news, but which is in no way whatsoever an academic publication. Similarly, the WOS indexes Historia, a popular history magazine that is sold in railway stations and is a favourite of my thirteen-year-old son, but which publishes no research articles. Unfortunately, these are not isolated examples. On the contrary, 99% of France's thousand most preeminent journals are entirely absent from the WOS, with a few (arbitrary) exceptions. Annales, a journal founded by Marc Bloch and Lucien Febvre in 1929, and which gave rise to the eponymous and internationally renowned Annales School,[14] is entirely overlooked by the WOS. The same goes for Études photographiques, an international reference in its field. I could go on listing these flagrant oversights. And yet, it is the WOS that presides when it comes to defining the impact factor.

The slogan "Covering the leading scholarly literature" is therefore bogus. This is not for want of requesting, politely and repeatedly, that Thomson add France's leading journals to its index. In its dealings with non-English speakers, the WOS's attitude smacks of arrogance, scorn and – some would claim – geographical, linguistic and disciplinary protectionism.

The core of the HSS is thus absent from the WOS. Consequently, research organisations that, whether by resignation, inertia or lack of resources, have grown accustomed to using the impact factor as a fundamental – if not unique – evaluative mechanism, have become inclined to think that the humanities and social sciences do not exist. (I won't dwell here on the other problem raised by

13 Guédon, Jean-Claude (2001): In Oldenburg's Long Shadow: Librarians, Research Scientists, Publishers, and the Control. In: Association of Research Libraries: ARL Membership meeting proceedings. www.arl.org/storage/documents/publications/in-oldenburgs-long-shadow.pdf.

14 Burguière, André (2006): L'École des Annales. Une histoire intellectuelle, Paris, Odile Jacob; Burke, Peter (1991): The French Historical Revolution. The Annales School 1929-89. Stanford University Press; Dosse, François (1987): L'Histoire en miettes: des Annales à la « nouvelle histoire », La Découverte.

the impact factor,[15] i.e. the definition of an indicator measuring journal impact, which is transferred by an artificial transitivity to the article, and then to the author of the article.) This problem has been identified and is widely recognised.

Will the humanities and social sciences opt for public appeal?

It is difficult to say whether the lack of HSS publications in the WOS explains the marginality of humanities and social-science budgets in European infrastructures, or whether it is lack of investment over the last few decades that explains the inability of the HSS to break the glass ceiling that is the academic sector's all-powerful indicator.

The necessary social turn

The humanities and social sciences have doubtless not understood the magnitude of the social turn taken by research funding, something other disciplines fully grasped a long time ago. The latter have been able to make extremely abstract questions about the infinitely small or the infinitely large both appealing and more or less comprehensible. The dazzling examples that are NASA and CERN, both of which have made the public aware of and interested in their research, should raise questions about the manner in which the humanities and social sciences position themselves with regards to social issues, and about the potential import of their research for our contemporaries. It seems far easier to explain the importance of studying the workings of societies involved in the "Arab Spring",[16] social attitudes to ageing,[17] processes of social exclusion,[18] and the socio-historical processes governing the creation and evolution of languages,[19] than it is to explain the Higgs boson. And yet, there remains a lingering modesty, or sometimes even snobbery, within the humanities and social sciences that prevents them from reaching out to the public or taking the time to explain their research and results beyond a select circle of colleagues.

15 See on this subject: sparceurope.org/citations/.

16 Gobe , Éric (2012): Un printemps arabe? L'Année du Maghreb VIII. DOI: 10.4000/anneemaghreb.1370.

17 Moulaert, Thibauld and Viriot Durandal, Jean-Philippe (2013): De la notion au référentiel international de politique publique. Le savant, l'expert et le politique dans la construction du vieillissement actif. Recherches sociologiques et anthropologiques 44 (1). DOI: 10.4000/rsa.904; Voléry, Ingrid and Legrand, Monique (2012): L'autonomie au grand-âge: corporéisation du vieillissement et distinctions de sexe. SociologieS. sociologies.revues.org/4128.

18 Messu, Michel (2010): Les politiques publiques de lutte contre la pauvreté. Variation sur l'approche française. Forum Sociológico 20. DOI: 10.4000/sociologico.152.

19 Hombert, Jean-Marie (2005): Aux origines des langues et du langage. Paris: Fayard.

The humanities and social sciences are a global success (but they don't always know it)

In reality, whole swathes of the humanities and social sciences have stepped out of their ivory tower and successfully presented their research to the general public. The Digital Public Library of America (DPLA), and Hypotheses, a platform for research blogs,[20] demonstrate that it is possible to reach out to the public – not only in English, but in a multilingual environment too.

The homepages of the DPLA and Hypotheses. Both make a special effort to editorialise content, i.e. by showcasing a handful of items selected from among thousands of resources, and enhancing content through eye-catching titles, design and images.

Hervé Théry's maps showing the geographic distribution of content disseminated on the France-based OpenEdition portal – which still includes a majority of content in French – indicate that its readership is not limited to France or even to the French-speaking world. A similar trend would probably emerge if we studied the geographical location of readers of Redalyc,[21] Scielo,[22] or Persée.[23] Webometrics' ranking of academic portals also points to the significance of humanities and social-science portals.[24]

20 Disclaimer: I am the director of the Centre for Open Electronic Publishing, which develops the OpenEdition portal, itself comprising four platforms (including Hypotheses). The other three platforms are Revues.org, Calenda and OpenEdition Books.

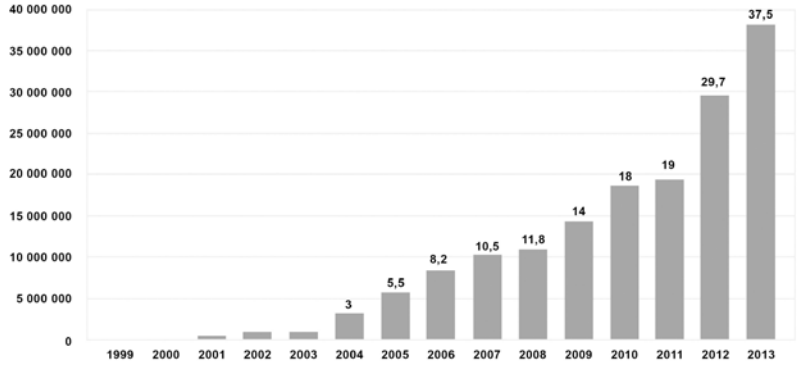
21 redalyc.org/

22 scielo.br/

23 persee.fr/Web/guest/home

24 repositories.Webometrics.info/en/top_portals

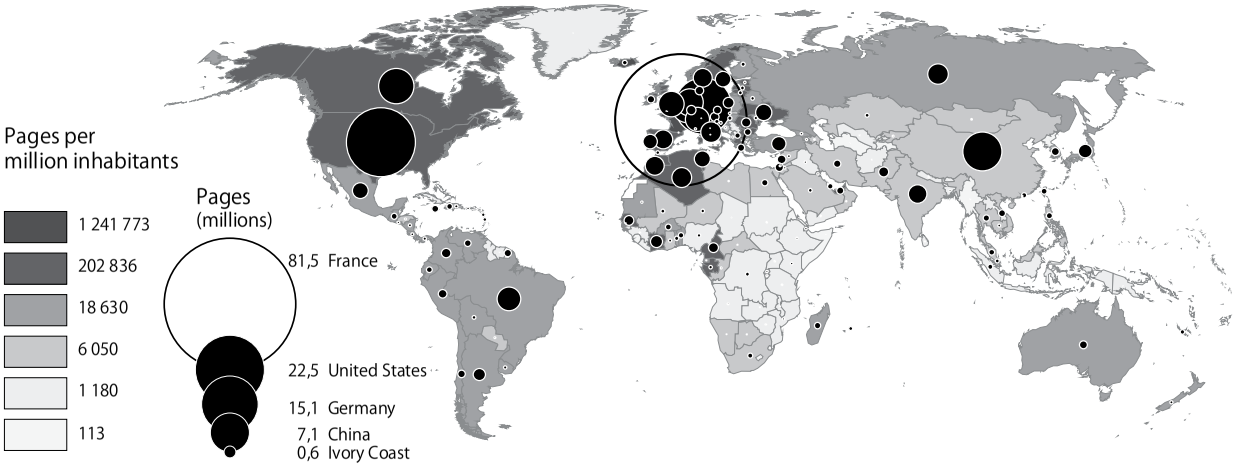
In absolute terms, the number of visits and visitors to the major humanities and social-science portals demonstrates that this kind of content is of interest far beyond the confines of academia. OpenEdition's consolidated figures (minus the "noise" of search engine bots) indicate massive and rapidly rising consultation rates. The 37.5 million visits per year to OpenEdition correspond to 20 million unique visitors, no doubt researchers, lecturers and students, but also journalists, professionals, pupils and citizens. Who said that the humanities and social sciences appeal only to that endangered species: the cloistered Sorbonne scholar?



OpenEdition. Number of annual visits by the million.

OpenEdition (2013) - all platforms

Revues.org, OpenEdition Books, Calenda, Hypotheses

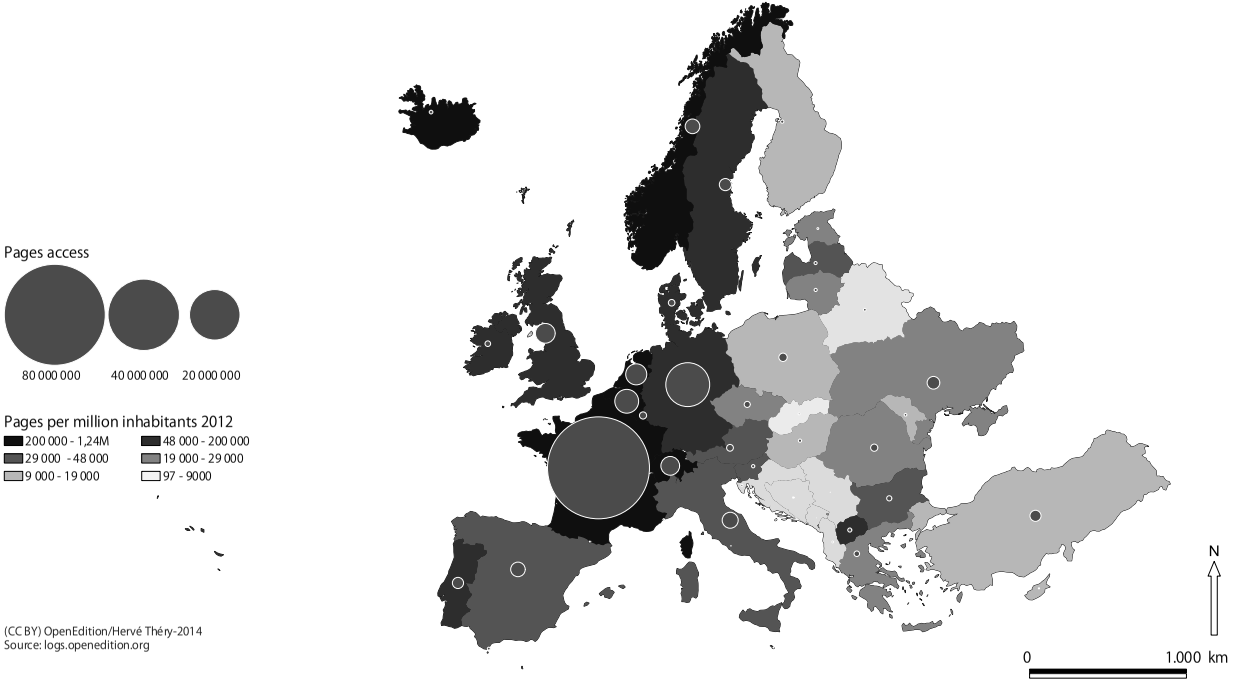


Made with Philcarto * http://philcarto.free.fr

Source: logs.openedition.org

OpenEdition (2013)

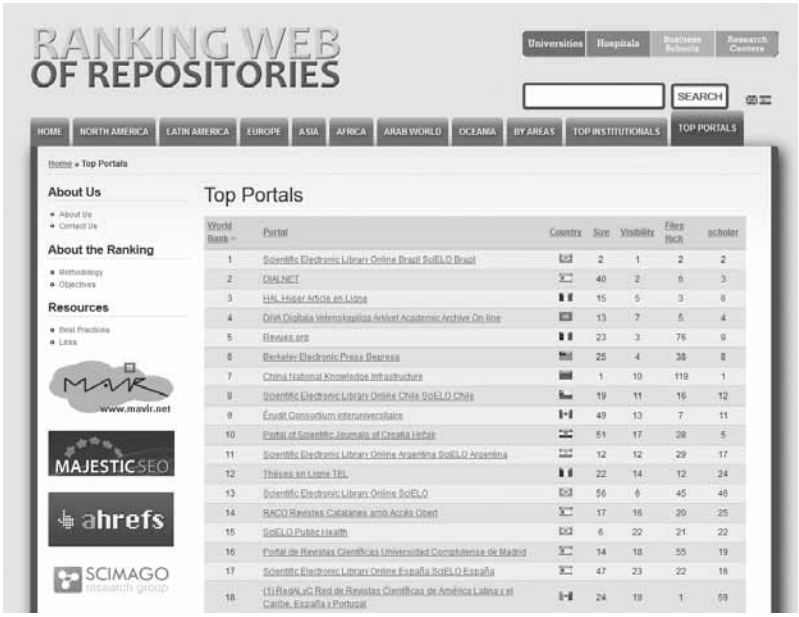
Revues.org, OpenEdition Books, Calenda, Hypotheses



Source: logs.openedition.org

Attenuating the "Matthew effect" through open access

It is up to us, and us alone, to convince university, national and European authorities that publishing is at the very heart of the relationship between HSS research and society. Publishing is the missing link in the chain, affected by the decades-long drop in book sales and yet still very much alive thanks to the Web in general and open access in particular. To maximise the efficiency of research funding, it is clear that research results must be disseminated in open access. If that raises obvious questions in terms of funding mechanisms for digital publishing, the gain in terms of impact is such that the question now is not whether we should commit to open access, but how, and how quickly.[25]



The Webometrics ranking for 2013 (repositories.webometrics.info/en/top_portals)

25 See Suber, Peter (2012): Open Access. Cambridge, Mass: MIT Press; Swan, Alma (2012): Policy guidelines for the development and promotion of open access. Paris: UNESCO; Willinsky, John (2006): The access principle: The case for open access to research and scholarship. Cambridge Mass.: MIT Press; Bailey Jr., Charles, W. (2010): Transforming Scholarly Publishing through Open Access: A Bibliography. Digital Scholarship.

Open access is both a way to reach out to a wider public and a historic opportunity to end the "Matthew effect",[26] a process by which the strong get stronger and the weak weaker. In economics, measures are taken to combat this process, which is not conducive to balanced competition. That is the rationale for the existence of antitrust laws. In academia, open access can be seen as an opportunity to foster a new geographical balance of power and the emergence of new actors. Are not the top five worldwide portals, according to Webometric's Ranking Web of Repositories, based outside Anglo-America? In rank order, they are are Brazil (Scielo), Spain (Dialnet), France (HAL), Scandinavia (DiVA) and France (Revues.org). It is highly likely that historically powerful actors, overrepresented in the WOS and underrepresented in the world of open access, will follow suit and establish a foothold within this new competition. Europe would be wise to wake up to the fact that this gives it important leverage with which to promote its research. And that it must stop prevaricating, because the window of opportunity – during which it can position itself for the long term – will not last forever.

Being convincing

Open access is not enough to cultivate a stronger relationship between science and society. We must strive to make our disciplines appealing and comprehensible. That does not mean simplifying and impoverishing our research by stripping it of its technical sophistication, vocabulary and scientific rigour. We can, however, take inspiration from highly abstract and technical disciplines such as those pursued at NASA and CERN, both of which have adopted a clear and deep-rooted policy commitment to communicating their results to the public. This takes careful planning.

Adopting a similar policy in the HSS will require genuine European infrastructures for digital publishing, ones that exploit the already sizeable volume of content available, something which in itself acts as a gateway to better understanding the complexity of the world we live in.

This conclusion simply extends the Manifesto for the Digital Humanities (2010), which called for the creation of digital infrastructures ("We call for the creation of scalable digital infrastructures responding to real needs. These digital infra-

26 Merton, Robert K. (1968): The Matthew Effet. Science 159 (3810), 56–63. See also Rossiter, Margaret W. (2003): L'effet ~~Matthieu~~ Mathilda en sciences, Les cahiers du CEDREF 11. cedref.revues.org/503.

The Manifesto for the Digital Humanities (2010), written at THATCamp Paris, has been translated into many languages.

structures will be built iteratively, based upon methods and approaches that prove successful in research communities."[27]).

Public investments in this area are a prerequisite, but they are not the whole story. We must immediately begin improving access to – and public understanding and perception of – the contribution made by the humanities and social sciences. This area abounds with solutions,[28] which vary depending on the objects of study, target publics and timescales involved. Geographic information systems and cartography are assets allowing us to produce new synthetic and easily digestible knowledge forms. The digital humanities as a whole have been experimenting for a few years now with more accessible visualisation and writing tools. Geographers have led the way by developing an expertise in cartographic language. More generally, content editorialisation will be at the heart of each and every one of these strategies, allowing information to transcend disciplinary boundaries and ensuring that, in the HSS, the relationship between science and society is more like a marriage than a divorce.[29]

27 tcp.hypotheses.org/411

28 See for example 4humanities.org/

29 See Dacos, Marin (2012): Vers des médias numériques en sciences humaines et sociales: Une contribution à l'épanouissement de la place des sciences humaines et sociales dans les sociétés contemporaines, Tracés. Revue de Sciences humaines 12. DOI: 10.4000/traces.5534.

6 Digital Communication and Social Media

Ranjana Sarkar (DLR-PT)

Discussion in Panel "New Forms of Data for Research in Humanities and Social Sciences"

The panel was part of the conference Facing the Future: European Research Infrastructure for Humanities and Social Sciences, initiated by the Social and Cultural Innovation Strategy Working Group of ESFRI and the German Federal Ministry of Education and Research, and hosted by the European Federation of Academies of Sciences and Humanities (ALLEA) and the German Data Forum (RatSWD). Invited panelists were the humanities and social science experts Markus Strohmaier, Scientific Director of Computational Social Science, GESIS–Leibniz Institute for the Social Sciences, Cologne; Diana Maynard, Senior Researcher at the Department of Computer Science, University of Sheffield Natural Language Processing Group; Franco Niccolucci, Director of VAST-LAB at PIN-University of Florence and ARIADNE Project Coordinator; Laurent Romary, Inria, Directeur de Recherche and DARIAH Director, who all contributed with their expertise to the success of the panel.

Introduction

The major challenge for future infrastructures, tackling "New Data", is just as much content-based as it is technologically driven. A notable impact is being made by communicative evolution based on digitization, networked communication and related social changes. These developments can be identified on multiple levels ranging from interaction changes and dramatic restructuring processes of organisations and institutions to systemic developments on a global level. Concerning research on communication from a Social Sciences and Humanities perspective, this means massive changes and rather disruptive innovation processes.

The Results of the Panel concerning Contents of New Media

The main thesis of the panel discussion was that the digital world is tracking the social world more and more closely. This enables us to use computation to discover patterns, build models, validate social theories and learn about societies. Challenges for Research Infrastructures/Computational Infrastructures resulting from this development are:

- More data means that analyzing large amounts of data is necessary
- Fuzzy data: imprecise and noisy data have to be cleaned up
- New kinds of data are processed: real-time sensor streams and web data
- Correlations: understanding how and why New Data correlate with societal phenomena
- Data acquisition of social media is tricky: hashtags are problematic, no standard search tools available; recall is likely to be low, issues of scale, dynamic nature.
- No existing real standards for sharing data: user profiles, friendship formations, linked content, what metadata should / shouldn't be included, etc.
- Infrastructural support for data storage layer: current methods combine Storm + Hadoop, but this is still rather unknown territory.
- Contextual preservation: Twitter as source of New Data is a transient medium and many messages are conversation replies – which is research relevant
- Lack of simple, easily accessible, integrated tools: for social scientists, information scientists, etc.

The main results of the panel concerning the technological challenges for the research community are that early standardization and interoperability is crucial for New Data in SSH-Research.

Summary of Panel II

According to the discussion within the panel "New Forms of Data for Research in SSH", the crucial requirements and most urgent needs for research in "New Data" concern standardization and interoperability. Standardization and interoperability supports defining methods or models to facilitate the exchange of lexical data, the pooling of heterogeneous lexical data and the interoperability between software components. Standardization and interoperability further builds the basis for constructing repositories, search engines, online presentation. In terms of science and research it guarantees comparability of results and conceptual coverage of lexical databases.

6.1 Challenges and Opportunities for Computational Social Science

Markus Strohmaier, Maria Zens (GESIS)

The field we would like to sketch out – Computational Social Science (CSS) – is an emerging area of research situated at the intersection of Social and Computer Sciences. According to the Computational Social Science Society of the Americas (CSSSA), the subject matter can be outlined as "The science that investigates social phenomena through the medium of computing and related advanced information processing technologies." The two-fold orientation of CSS towards algorithms and Social Sciences might prove beneficial for both disciplines. On the one hand, CSS reaches out to offer means for processing large amounts of data to the Social Sciences and, on the other hand, takes hypotheses and theories from the Social Sciences to arrive at meaningful models of social behavior which can be applied to and tested against large data sets taken, for example from social media.

Data-induced opportunities and challenges for CSS

The formation of CSS responds to a situation in which interactions in the digital world generate and shape social structures in a novel way and, in doing so, provide social research with prolific new data sources. The increasing integration of the World Wide Web in our daily lives already has created massive volumes of social data, i.e. data about humans' everyday behavior and social interactions in the real world. Such social data opens up exciting new opportunities as well as challenges for computer and social scientists to work towards a new and deeper quantitative understanding of complex social systems. At the same time, the increasing availability of such social data has led to new types of and directions for research. In our contribution to this volume, we will discuss these and other related issues from both a Computer Science and a Social Sciences perspective. We will give examples from recent and current research that illustrate how the use of new and large data sets can support the Social Sciences in analyzing socio-political phenomena such as mobility issues, interpersonal communication, and the structures of political discourse.

Currently, the Social Sciences appear to be a discipline in distress; as the sociologists Mike Savage and Roger Burrows put it some years ago: "both the sample survey and the in-depth interview are increasingly dated research methods, which are unlikely to provide a robust base for the jurisdiction of empirical sociologists in coming decades" (Savage and Burrows 2007: 885). The possible turn-out of this "crisis" is that we are about to encounter a data-driven advancement of a well-established discipline and an ever deepening and serious collaborative intent on all sides.

The confluence of Social Sciences and Computer Science seems natural in a situation in which both sides are in need: computer scientists need to make social sense of "big data" and social scientists require tools to handle new amounts (and the new quality) of data that go beyond their traditional ways of collecting, structuring, and evaluating.

Having stressed the challenges in quantity and the social novelties posed by transactions and communicative interactions on the web, one has to bear in mind that data-driven progress is nothing new in the history of the Social Sciences. A brief detour into history might illustrate that, although the kinds of data might be new, the fact of fruitful co-operation is surely not.

The paradigm: Technology-driven advancement in processing social data

Herman Hollerith (1860–1929), the son of German immigrants to the US, is an early example of how technology-driven innovation at the intersection of Social Sciences and what became Computer Science can be established. In his work on "An Electric Tabulating System" (1889), which formed the basis for his PhD at Columbia one year later, Hollerith developed a crucial foundation for the advancement of Computer Science during the 20th century. Hollerith, who made use of previous technological knowledge gathered by the textile industry (Jacquard loom techniques), developed a mechanical tabulator based on punched cards to rapidly tabulate statistics from millions of pieces of data. His machines were able to tally not only overall numbers, but also individual characteristics and even cross-tabulations; he invented the first automatic card-feed mechanism and the first key punch. The prime use case for his invention was social data: Hollerith built calculating machines under contract for the US Census Office; in 1890 Hollerith machines were first used to tabulate US census data, and, subsequently, in many more censuses in various countries. Due to

his invention, processing time could be decreased enormously; while it had taken about eight years to tabulate the 1880 US census, the 1890 census using his machines took – figures differ on this point – only one year or even less. Hollerith's firm "Tabulating Machine Company" merged with others to become the "Computing Tabulating Recording Company", which was later renamed "International Business Machine Corporation" (IBM).

One might reasonably dispute the grounds for progress – whether the advancement in technology fostered efficiency for the administration or, vice versa, whether administrative needs induced the development of technology. However, with his invention Hollerith provided the means to leverage the processing of administrative data on the US society and in many more states.

Bearing this landmark in mind and being in the midst of a new deluge of digital data that we are struggling to make social sense of, we should ask what are the respective 21st century challenges and opportunities? How should "modern Hollerith machines" be structured and where could they be deployed? We will give a few examples to try and find preliminary answers to these questions.

New kinds of data on macro, micro, and meso scale: Mobility as a use case

First, we will look at the opportunities that mobility data offer the Social Sciences. Human mobility in societies is one of the key issues that can be addressed with large data-sets generated from GPS or cell-phones. We will take the three-level-approach familiar to Humanities and Social Sciences and look at the impact of new kinds of data in this field at the macro, meso, and micro levels. On the macro scale, GPS-data or check-in-data from social media applications like Foursquare, Gowalla or Twitter provide information about the locations, destinations and travel modes of people and give us a kind of "big picture" of mobility. These data are increasingly available and being used for research. Cheng et al., for instance, analyzed the use of location sharing services by investigating 22 million check-ins by 220000 users (Cheng et al. 2011). What is interesting from a Social Sciences point of view is that the authors not only studied spatio-temporal mobility patterns, but also analyzed the correlation of social status and mobility behavior. In fact, they themselves regard this aspect of their research as "one of the more exciting possibilities raised by the social structure inherent in location sharing services" (ibid.: 87).

On the meso level, the Amsterdam real time project can be mentioned. The project dates from 2002: Amsterdam residents were asked to use a tracer unit with GPS, the real-time visualization of the collected data showed lines against a black background and hereby constructed a (partial) map of Amsterdam based on the actual movements of real people. The project was carried out in conjunction with artists and became part of the exhibition "Maps of Amsterdam 1866–2000" in the Municipal Archive of Amsterdam. Almost a decade later and with advanced technologies, Calabrese et al. conducted a case study on Rome ("Real Time Rome"), which also made use of real-time mobility data and was developed for the Tenth International Architecture Exhibition of the Venice Biennale. (Calabrese et al. 2011) For this project a monitoring system using a variety of tools was deployed to grasp urban mobility from cell-phones and locational data from the public transport system. Among other issues, the density of people using mobile phones at historic sites or during events was measured and visualized. The distributions revealed dynamics of movement during the course of the day, as well as hot spots for tourists or event gatherings. The practical impacts of such services cover the optimization of urban planning, of services and traffic information, the reduction of inefficiencies and the support of efforts to put public transportation where the people need it.

For the significance of new data on the micro scale we would like to turn to an experimental project in which Radio Frequency Identification (RFID) was used to monitor the dynamics of communicative interaction between people (Cattuto et al. 2010). At the core of this group's work is the aim to present micro level interpersonal relations at high resolution, but with the clear objective to provide tools that can be scaled up. Sensing tiers with unobtrusive RFID devices were embedded into conference badges to sense face-to-face interactions and spatial proximity of participants as well as the duration of contacts. One of the results was a power law distribution with identifiable "super-connectors", the "crucial actors in defining the pattern of spreading phenomena", who "not only develop a large number of distinct interactions, but also dedicate an increasingly larger amount of time to such interactions" (ibid.: 5).

"Digital society": data sources for analyzing political discourse and social action

Shifting from the granularity of data to the topical areas relevant for the Social Sciences, one has to recall the double function of the new social media as sources of data and social arenas in their own right. In their introduction to the special issue of "Sociology" on the relationship of new technologies and society, Linda McKie and Louise Ryan point out: "New technology is not simply capturing but actively constituting social interaction"(McKie and Ryan 2012: 6). Therefore, both the structures of "digital society" and the reflections of social behavior or political discourse in the digital world are of interest.

In what can be regarded a seminal paper for Computational Social Science, Lazer et al. (2009) show a visualization of political conversations in the blogosphere around the 2004 US election and made this a prime example for how existing socio-political theories can profit from examining vast data. They displayed the network structure they found in a community of political blogs, where this structure – with a clear distinction between the liberal and the conservative camps – reflected the political map very closely.

The impact of web technologies on political events has been discussed in, for example, the context of the so-called Arab Spring, when communication through social media channels played an important role for both the diffusion of alternative political information and the organization of the protest movement. Starbird and Palen looked at the uprising in Egypt and analyzed the most influential Twitter users during the revolution in early 2011, with influence being measured by the number of retweets and followers. They found individuals, bloggers, journalists, and mainstream news channels among the most influential actors. Yet first up on the crucial days in January 2011 is the internet activist Wael Ghonim (Starbird and Palen 2012). In our own research (conducted with Lichan Hong at Xerox Parc) on political conversations on Twitter (about 100 million tweets) during the Egyptian revolution 2011, we found that the hashtag "#jan25" – which denotes the first day of the protests – was a top-trending hashtag prior to the actual date.

These findings show, that social media might serve as an additional channel for the dissemination of mainstream information, but also as alternative channels for independent journalists and activists; basically, they are an open communication space for governments, traditional media, social movements, and dissidents alike. Social media are important in the competition for political

hegemony and interpretation, which becomes evident when they are subject to regulation, censorship, and surveillance.

Finally, we would like to mention research on the representation of the German parliamentary elections 2013 in social media, which is work in progress at GESIS. We analyze twitter accounts of electoral candidates from the various parties; we try to apply models of communication taken from the Social Sciences to the conversations on Twitter in the run-up to the election and look at topical conversation practices (via hashtagging) and structural conversation practices such as mentioning and re-tweeting. In so doing, we want to investigate the similarity between parties as measured by hashtags, the foci (both topical and structural) of partisan communication and the stability of conversation practices (i.e. measure focus shifts at certain points in the electoral time-line).

Challenge: providing computational infrastructures

With these examples, we have tried to highlight both the data side of CSS – with respect to modelling on the macro, the meso, and the micro levels – and the Social Sciences side of CSS – with respect to their contribution to revealing behavioral patterns in socially and politically relevant realms. Since the digital world is tracking the social world more and more closely and specific forms of a "digital society" emerge, CSS sets out to use computation to discover patterns, build models, validate social theories and learn about societies. Further to that, we have to address data management issues, archival issues, and legal issues concerning privacy and data protection.

The efforts, of course, go beyond the ones we could mention and take various angles – data mining and processing, sociology, political science, network analysis etc. –, but the main challenge for research infrastructures is to provide computational infrastructures for dealing with (1) more data: for analyzing large amounts of data, (2) fuzzy data: for cleaning up imprecise and noisy data, (3) new kinds of data: for processing real-time sensor-streams and web data, (4) correlations: for understanding what (in addition to why).

This brings us back to Herman Hollerith and his technological achievement. To specify and build what we may call "modern Hollerith machines" is a major challenge for Computational Social Science. There already exists a range of tools and platforms for extracting big data streams (Hadoop, Mahout, SAMOA, S4, Storm, R, WEKA, MOA – cf. De Francisci Morales 2013); what is needed are

algorithms and clustering techniques that focus on social structures and expand the horizon of data modelling from space and time complexity to include social complexity.

Challenge: computation-focused social theories

From the point of view of the Social Sciences and their "crisis", the challenges are mainly attached to the large amounts of and the uncontrolled quality of the new data at stake. These data have the advantage of being easily accessible, but they do not meet the traditional standards of social science research. They are not intentionally collected under ceteris-paribus-conditions, but "found data". They are often far from being representative, and suffer from single channel and self-selection biases. In short: using transaction data from social media for Social Sciences research requires new, robust methods of data collection, cleansing and evaluation. Moreover, Social Scientists should take on the challenge to further include computation-focused social theories, e.g. network theories.

Opportunity: CSS – more than adding up expertise

The confluence of Social Sciences and Computer Science merges expertise: Computer Science offers the ability to process large data sets and provides algorithms and methods of data mining. The Social Sciences contribute their knowledge of social theories, methods, data collection, and relevant issues. This means more than simply adding up the things that we knew before. While working together on Computational Social Science issues, both knowledge systems are being transformed by the opportunities of analyzing new amounts of digital information with regard to social systems.

References

Calabrese, F./Colonna, M./Lovisolo, P./Parata, D. and Ratti, C. (2011): Real-Time Urban Monitoring Using Cellular Phones: A Case-Study in Rome. IEEE Transactions on Intelligent Transportation Systems 12 (1), 141–151.

Cattuto, C./Van den Broeck, W./Barrat, A./Colizza, V./Pinton, J. F. and Vespignani, A. (2010): Dynamics of Person-to-Person Interactions from Distributed RFID Sensor Networks. PloS One 5 (7), e11596.

Cheng, Z./Caverlee, J./Lee, K. and Sui, D. Z. (2011): Exploring Millions of Footprints in Location Sharing Services. ICWSM, 81–88.

De Francisci Morales, G. (2013): SAMOA – A platform for mining big data streams. Proceedings of the 22nd international conference on World Wide Web companion, 777–778.

Hollerith, H. (1889): An Electric Tabulating System. The Quarterly, Columbia University School of Mines, Vol. X , No. 16, 238–255.

Hollerith, H. (1890): In connection with the electric tabulation system which has been adopted by U.S. government for the work of the census bureau. Ph.D. dissertation, Columbia University School of Mines.

Hollerith, H. (1894): The Electric Tabulating Machine. Journal of the Royal Statistical Association 57 (4), 678–682.

Lazer, D./Pentland, A. S./Adamic, L./Aral, S./Barabasi, A. L./Brewer, D., [...] and Van Alstyne, M. (2009): Life in the network: The coming age of computational social science. Science 323 (5915), 721–723.

McKie, L. and Ryan, L. (2012): Exploring Trends and Challenges in Sociological Research, Sociology 46 (6), 1–7.

Savage, M. and Burrows, R. (2007): The coming crisis of empirical sociology. Sociology 41 (5), 885–899.

Starbird, K. and Palen, L. (2012): (How) will the revolution be retweeted? Information diffusion and the 2011 Egyptian uprising. Proceedings of the ACM 2012 conference on computer supported cooperative work, 7–16.

6.2 Challenges in Analysing Social Media

Diana Maynard (University of Sheffield)

Introduction: The importance of social media analysis

Information, thoughts and opinions are shared prolifically on the social web these days, with approximately 72% of online adults using social networking sites. The amount of data now available in the form of social media (tweets, Facebook posts, etc.) is constantly growing, and forms an incredibly rich source of information for research in NLP, social science and information science, among other disciplines.

One popular misconception about social media is that it is not useful because it consists mainly of trivia, such as discussion of pop music, TV shows and the minutiae of people's daily lives. The 10 Twitter accounts with the highest number of followers include 7 pop stars and 2 social media sites (although Barack Obama comes in at number 4). Clearly, there is a lot of mindless drivel on social media. However, there is plenty of evidence to support the fact that useful information can be gleaned even from such trivia. For example, the Germtracker tool [6] derives accurate real-time epidemiological information from tweets in order to predict who might get flu, to identify restaurants with a high risk of food poisoning, and so on. It works by analysing the role of interactions between users of social media on the real-life spread of disease. Social media is also becoming a critical means of communication and information in times of emergency. Clearly, the amount of information that can be made use of is huge, but the problem is that we still need good real-time analysis tools to help process this data.

Social web analysis is all about the users who are actively engaged and generate content. This content is dynamic, reflecting the societal and sentimental fluctuations of the authors as well as the ever-changing use of language. Social networks are pools of a wide range of articulation methods, from simple "Like" buttons to complete articles, their content representing the diversity of opinions of the public. User activities on social networking sites are often triggered by specific events and related entities (e.g. sports events, celebrations, crises, news articles) and topics (e.g. global warming, financial crisis, swine flu).

Alongside natural language, a large number of the interactions which occur between social web participants include other media, in particular images. While we shall not discuss multimedia analysis here, suffice it to say that this introduces an added dimension to the problem. Textual analysis can be enhanced by combining information from images, sound and videos to deal with ambiguity or add contextual information, but multimedia analysis is also extremely problematic. Similarly, textual analysis tools can assist with multimedia information extraction by reducing the search space for image matching. Current research in projects such as ARCOMEM [4] is investigating the integration of these technologies, but this is very much ongoing research.

One particular subtask of social media analysis which is currently in enormous demand is opinion mining, also known as sentiment analysis. This can take a variety of forms, but the principle is generally the same: namely, to find out what people are thinking. Opinion mining is big business: the usefulness of understanding customer reviews and so on is important for companies, but there are many other uses, such as tracking political opinions, tracking the influence of public mood on stock market fluctuations, studying the distribution of opinions in relation to demographics, and so on. Understanding what events can cause people to change their opinions, who the opinion leaders are, and how opinions change over time are all key areas for study.

Challenges for NLP

Despite the current plethora of work on social media analysis, there are still many challenges to be faced. For example, traditional opinion mining approaches that focus on product reviews and so on are not necessarily suitable, partly because they typically operate within a single narrow domain, and partly because the target of the opinion is either known in advance or at least has a limited subset of possibilities (e.g. film titles, product names, companies, political parties, etc.).

The interesting nature of social media data is precisely what makes it also so challenging for NLP. It is fast-growing, highly dynamic and high volume, reflecting the ever-changing language used in today's society, and reflects current societal views. This makes it a wonderful source of material for opinion mining tools, but this specialised use of language, along with the size and dynamicity of the data, is also precisely the cause of many problems.

Typically, social media text is rich in abbreviations, slang, domain-specific terms, and spelling and grammatical errors. Standard NLP techniques, which are used to analyse text and provide formal representations of surface data, produce lower quality results on these kinds of degraded text. For example, shortened or misspelled words, which are very frequent in this kind of informal style, increase the variability in the forms for expressing a single concept; use of hashtags causes problems for tokenisation; and lack of context in microposts such as tweets gives rise to ambiguities.

The quality of the text affects not only term and entity recognition, but all the linguistic processing components in the pipeline, such as tokenisers, POS taggers and so on. Degraded performance on any of these components is likely to have a negative effect on any other components which rely on these. The higher up the chain the error occurs, the worse the knock-on effect will be. It is not easy simply to retrain components on appropriate texts, because variation and errors are not consistent: for example, there are no typical words which are not capitalised correctly, and although some misspellings and typos are more common than others, this is not consistent enough to be very useful. In recent experiments [2], the GATE-based IE tool ANNIE dropped from 87% F-measure to below 40% when applied to tweets, and other tools such as the Stanford NE recogniser performed even worse. Typically, social media also contains extensive use of slang, irony and sarcasm, which are hard to deal with and complicate tasks such as opinion mining. Efforts are currently being made to develop Twitter-specific NER tools such as TwitIE [1], TwiNER [3], and T-NER [5], but performance is still far from ideal, and development is held back by lack of standards, easily accessible data and common evaluation frameworks.

Infrastructure challenges for the research community

Aside from the linguistic complexities of analysing social media, there are also many fundamental infrastructural challenges which need to be addressed by the relevant research communities. First, data acquisition of social media is tricky. It is not easy to actually find the material wanted: hashtags are problematic since they are not standardised and exhibit both ambiguity and polysemy; there are no standard social media search tools available, and thus recall is likely to be low.

Second, there are no real existing standards yet for sharing such data: for example, the representation of user proles, friendship formations, linked content, and

which metadata should or should not be included. This gives rise to issues of contextual preservation: Twitter is a transient medium and many messages are conversation replies - it is clearly detrimental to take those out and discard all the context, but it is not clear how to include the context either.

Third, there is still limited infrastructural support for the data storage and acquisition layer. The sheer volume of data gives rise to numerous issues of scale which are not supported by current physical infrastructures: trying to search several terabytes of data is still difficult, and often only readily supported by large organisations, who may not allow access to the data by others. All this makes both competition and collaboration difficult.

Data sharing is not straightforward either: constraints by Twitter and other social media sites make it difficult to share corpora, and only ad hoc solutions exist rather than a single sharing mechanism. Even getting tweets via their IDs is not easy, due both to constraints on Twitter download rate limits, and the fact that the task is still non-trivial with large datasets. Tools such as Gardenhose are also problematic: since only 10% of tweets are released this way, linking information gets lost, i.e. you cannot guarantee that all parts of a conversation thread will be maintained.

Fourth, there are problems with analysing dynamic data effectively. Crawling of sites such as Twitter is impossible in the traditional sense, due to legal restrictions, and crawling via APIs to extract the data is not ideal, as discussed above. Batching processes mean that the data is already old by the time it is available, and while streaming solutions such as Storm (potentially combined with Hadoop), Yahoo S4 and Amazon Kinesis are available, all these processes still have the issue of breaking the relationship links in a network: it is not clear how to relate new information as it comes in, how to decide what goes in a batch, and how to deal with things like entity, author and topic co-reference.

Issues for humanities and social science research

It seems that there is a lack of simple, easily accessible, integrated tools: not just for NLP research but also for other communities such as social sciences, humanities and information science. These communities are keen to analyse social data, but often do not have the means to collect and make sense of it. While there are numerous text analysis tools such as GATE[1], OpenNLP[2], UIMA[3], and so on, they are not widely known by other communities, and it is not necessarily easy to integrate them with existing visualisation tools such as Pajek[4], NetMiner[5] and NodeXL[6]. Even existing data collection and visualisation tools are not always well known to these communities, and are not adaptable enough. There is also a lack of communication between the different research communities, so that NLP research on social media analysis is often not widely known about in other communities such as social science.

We thus propose moving towards the development of a common shared framework for collecting, analysing, and visualising data. Ideally, this would contain modular pipelines enabling people to interchange modules for e.g. collection, analysis/processing and graphing/visualisation.

Conclusions

In summary, we propose a number of areas in which effort could be focused by various communities specific to social media: data acquisition and streaming approaches to analysis; the creation of standards for sharing this data; the development of a common shared framework for collecting, analysing, and visualising data, the production of standardised evaluation datasets, and finally better collaboration between the communities of NLP, social science, humanities and information scientists.

Research on social media in all these fields is a hot topic, but it is still rather fragmented, and there are still many unsolved problems hindering principled progress. These problems will only increase as social media gains importance and as more tools are developed. The time is therefore right for more integrated and collaborative efforts to resolve the major bottlenecks.

1 gate.ac.uk
2 opennlp.apache.org
3 uima.apache.org
4 pajek.imfm.si
5 www.netminer.com/index.php
6 nodexl.codeplex.com

References

[1] Bontcheva, Kalina/Derczynski, Leon/Funk, Adam/Greenwood, Mark A./ Maynard, Diana and Aswani, Niraj (2013): TwitIE: An Open-Source Information Extraction Pipeline for Microblog Text. In: Proceedings of the International Conference on Recent Advances in Natural Language Processing. Association for Computational Linguistics.

[2] Derczynski, L./Maynard, D./Aswani, N. and Bontcheva, K. (2013): Microblog-Genre Noise and Impact on Semantic Annotation Accuracy. In: Proceedings of the 24th ACM Conference on Hypertext and Social Media.

[3] Li, Chenliang/Weng, Jianshu/He, Qi/Yao, Yuxia/Datta, Anwitaman/Sun, Aixin and Lee, Bu-Sung (2012): Twiner: Named entity recognition in targeted twitter stream. In: Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, 721–730.

[4] Maynard, Diana/Dupplaw, David and Hare, Jonathon (2013): Multimodal Sentiment Analysis of Social Media. In: Proceedings of Thirty-third SGAI International Conference on Articial Intelligence (AI-2013 Workshop on Social Media Analysis, Cambridge, UK.

[5] Ritter,A./Clark, S./Mausam and Etzioni, O.(2011): Named entity recognition in tweets: An experimental study. In: Proc. of Empirical Methods for Natural Language Processing (EMNLP), Edinburgh, UK.

[6] Sadilek, A./Kautz, H. and Silenzio, V. (2012): Modeling spread of disease from social interactions. In: International AAAI Conference on Webblogs and Social Media (ICWSM), 322–329.

6.3 The ARIADNE approach to Digital Cultural Heritage

Franco Niccolucci (VAST-LAB)

Introduction

ARIADNE (Advanced Research Infrastructure for Archaeological Dataset Networking in Europe, www.ariadne-infrastructure.eu) is a European Integrating Infrastructure project that addresses the fragmentation of archaeological datasets throughout Europe and fosters the use and reuse of archaeological data through the interoperability of existing repositories. The ARIADNE partnership consists of 24 partners from 16 European countries.

The research infrastructure concept is familiar in various scientific domains in which the laboratory component plays a key role and complex instruments may be optimally used through shared access. In the humanities, infrastructures nowadays mainly consist of digital resources and of the services that enable access, use and reuse of such accumulated data. Although the concept of collaboration is not new to archaeology, where investigation builds on previous work by other researchers, modern national borders, academic traditions, different languages and the diversity of standards in use for archaeological documentation have led so far to datasets that are insulated from each other. Nowhere in the humanities can the well-known metaphor of 'information silos' be better applied than in archaeology. Nevertheless, in the archaeological community there is a strong demand for integration, and for underlining the need of novel tools to overcome the current difficulty of accessing, let alone re-using, the wealth of data created since digital technologies made their way in the body of the discipline. Especially with the progress of visual digital documentation, nowadays including digital photos, 3D models and more, the issue of 'big data' is also becoming a problem for archaeological informatics. Laboratory activity is creating an additional quantity of data produced by instruments that output digital results.

Until now, raw data were not destined for scientific publication. Most scholars used to produce and collect data for their own use, and distilled their data

analysis into a synthetic interpretation that was communicated to the research community in the traditional way of articles and monographs. On the other hand, the importance of corpora has never been ignored. Massive collections[1] of complete sets of information concerning specific topics, providing details as close to the originals as possible, and reference collections of artefacts useful for classification have been created by researchers since the 19th century to organize their discoveries and to relate them to the body of existing knowledge.

Nowadays, data publication is a must. It is feasible because technology makes data publication straightforward. It is necessary because such data may still contain valuable but unexploited information for other researchers. It is demanded because such data are not private property, but almost always produced with public money. However, accessing data may be difficult, and combining data from different sources may be cumbersome, if not impossible. Data re-use is therefore a high priority challenge on ARIADNE's agenda.

The ARIADNE contribution: first steps

As a first task, ARIADNE has undertaken a reconnaissance of what it is going to deal with. For this purpose, it is developing a Registry of archaeological digital resources, including datasets, thesauri and gazetteers. It will start from the ARIADNE community, adopting a model developed by a project task force named the ACDM (ARIADNE Catalogue Data Model), which describes datasets using an extension of the international standard DCAT. The ARIADNE Registry will be open for contribution and consultation to every institution and researcher. Registry operations started in early 2014, when partners began uploading the information about the datasets owned by them and accessible by the public. So far, a preliminary survey among partners about their digital assets led to the following interesting results.

The ARIADNE partnership manages data from a total of 20 countries, including more than 1500,000 database records, about half of which are contained in an RDBMS, and about 40000 'grey literature' documents, most of which are excavation reports. The latter often belong to collections of reports, which

1 For example the Corpus Inscriptionum Latinarum (CIL) created by Mommsen dates back to 1853 for the Berlin-Brandenburg Academy of Sciences and Humanities. It collects Latin inscriptions and is still updated today. Another example, again of German origin, is the catalogue of Roman amphora types created in mid 1800 by Dressel, whose typology is still used today. Both these repertoires, as most of the other reference collections, are now digital and available on-line.

may include the outcomes of scientific analyses, images, and other multimedia. A substantial number of GIS are also present.The respective archaeological datasets contain data about excavations, sites, settlements, burials, finds, and objects specific to particular regions and periods. There are several scientific datasets with archaeological sciences data, notably an international dendrochronology archive with some 50,000 measurement series of tree-rings. As regards standards, most partners use metadata schemas that are compatible with international standards, but are often customized or extended for local use. Controlled vocabularies relate more to local or specific needs, as can be reasonably expected, with no inter-language crosswalk. In conclusion, there is nothing unexpected about fragmentation, but there is also no insuperable obstacle to integration.

Work has also started to establish the context in which the research community wants to place the integrated services that ARIADNE will provide. A survey of all partners and a large number of other stakeholders provided just under a thousand confirmed contacts. About half of them responded, providing useful information for a comprehensive appreciation of the community needs and expectations, and for the design of the ARIADNE services.

Transnational activities have also started, publishing the call for 2014 Summer Schools, one year earlier than planned in the original description of ARIADNE work. The schools will address the creation and management of new datasets, dealing with legacy datasets and facing the challenges of 3D. The latter is a topic that Horizon 2020 has reaffirmed as "[having a] key role ... [and able to] offer new perspectives to researchers and new understandings to citizens, research users and the cultural and creative industries". The second year of ARIADNE activity, 2014, will be the period in which, according to the project work plan, the project profile is consolidated with the definition of the infrastructure design and of the services design, including the development and assessment of an initial test implementation of the latter.

The ARIADNE contribution: the Innovation Agenda

One of the most important results that ARIADNE will deliver in 2014 is the Innovation Agenda, to be presented to stakeholders at the mid-term event at the end of the present year. A few preliminary considerations are already available for discussion. They include the ARIADNE remit in the body of digital humanities, evidence the interrelation of its work with other disciplines, and the principal fields the Action Plan will concern.

The existing background

Research in the humanities is highly fragmented and, traditionally, individualistic. The growth of computer use still reflects these characterstics, creating a myriad of individual datasets with little standardization, let alone integration. As computers were considered until not too long ago as the equivalent of "household appliances"[2], datasets were developed independently of each other and were often designed to serve individual research purposes. In the library domain, and in general wherever texts are the matter of investigation, data organization soon showed its potential for text searches and text analysis, both of which rely on digitized texts and on a well-ordered approach[3]. This led to an early adoption of computerized methods and to standardization, paving the way for modern universal digital libraries. Where, in contrast, the matter of study was more tangible, as in archaeology and conservation, scientific methods and techniques were soon adopted, but usually still in an ancillary role, and often with a do-it-yourself approach[4]. Computers were first used here for text descriptions of records, and, with much less importance, as the outcomes of scientific analyses, generally incorporated in the investigation results only through a concise summary report. ICT was not infrequently used just as a drunkard uses a lamppost: for support, not illumination.

Nowadays, times are changing. The need to utilize digital technologies to use and re-use accumulated data, to integrate as yet dispersed repositories into EU-wide research infrastructures to be used by researchers as humanities laboratories, has become a priority for the humanities as well. Unfortunately, the methods, the underlying technology and the related expertise cannot be straightforwardly borrowed from other domains, not even from the closely tied

2 As denounced e.g. in D'Andrea, A. and Niccolucci, F. (2001): Informatica archeologica: Alcune istruzioni per l'uso, Archeologia e Calcolatori 12, 199–210.

3 The Dewey library classification system dates back to some 150 years ago.

4 Pollard, M. and Bray, P. (2007): A bicycle for two, Ann. Rev. Anthropology, 36, 245–259.

sector of social sciences, which is divided from the humanities by a methodological gap. The latter include disciplines based on texts and images as expressions of human intellect and creativity; on the connection of different historical evidence stored in archives; and finally on the study of the remains of material culture. On the other hand, the social sciences consist of the study of the behaviour and conditions of humans. The keyword for the social sciences is statistics; for the humanities, it is semantics.

Such a methodological gap mirrors the peculiarity of the techniques and the technological tools required for turning humanities into e-humanities, where data integration is a compulsory first step. This specificity of the humanities and its needs was well interpreted by the ESFRI Roadmap when establishing DARIAH as a specific strategic RI for the humanities and for cultural heritage.

Clustering and integrating actions

Standardization

Standardization is a primary requirement for integration. Thanks to the work required to provide digital content to Europeana and to the pressure for data integration, use and re-use, awareness has greatly increased in the humanities community about the need of standardization, and has paved the way for a research-focused demand for interoperability. Initiatives already exist for establishing a common understanding on this subject, which will be particularly beneficial to investigations adopting a holistic approach based on material culture as well as on text-based sources. This is the case, for example, for classical archaeology, for the history of science and technology, and for many studies in social history.

Within the domain of cultural heritage and archaeology, CIDOC CRM[5] has established itself as the reference domain ontology. It is being continuously updated and extended to cover all the needs of integration, notably within the ARIADNE project. Large repositories in highly reputed cultural institutions such as the British Museum are moving to the CRM. The mapping of other 'local' standards to the CRM is also well developed, and coverage of other 'hard' sciences is underway. The long-standing collaboration of CIDOC CRM with FRBR-OO[6] led

5 CIDOC (International Committee for Documentation of ICOM) CRM (Conceptual Reference Model), www.cidoc-crm.org, is the ISO 21127:2006 standard for cultural heritage documentation. It has extensions for archaeological documentation and other related domains.

6 FRBR-OO (Functional Requirements for Bibliographic Records-Object-oriented) >>

to the establishment of the International Working Group on FRBR/CRM Harmonisation, bringing together representatives from both communities with the common goals of expressing the FRBR model with the concepts, tools, mechanisms, and notation conventions provided by the CIDOC CRM, and of aligning the two object-oriented models with the aim to contribute to the solution of the problem of semantic interoperability between the documentation structures used for library and museum information. Harmonisation between archival standards such as EAD[7] and ISAD(G)[8] has been explored in research papers, but not yet exemplified in practice. Initiatives for harmonising TEI[9] with CIDOC CRM have been undertaken by the TEI Ontologies SIG.

In conclusion, there are a large number of initiatives that aim to provide a conceptual glue to link different semantic schemes tailored to the needs of different disciplines in the broader field of humanities. Such initiatives need to be structured and to reach a mature stage.

Data management and dataset integration

These activities fall within the remit of existing integrated infrastructure projects that provide solutions to domain-specific integration demands. As such they have special needs and different approaches. However, they share common goals such as the creation of a registry of datasets based on a data model describing datasets within each domain; the incorporation of thesauri and authority files addressing multilingual issues in data integrating activities; and models for integrated services based on advanced ICT tools, such as 3D visualization interfaces for dataset management and output.

www.ifla.org/node/2016, is a formal ontology intended to capture and represent the underlying semantics of bibliographic information and to facilitate the integration, mediation, and interchange of bibliographic and museum information, developed by the International Federation of Library Associations and Institutions (IFLA).

7 EAD (Encoded Archival Description), www.loc.gov/ead, is an XML standard for encoding archival finding aids maintained by the Society of American Archivists, in partnership with the Library of Congress.

8 ISAD(G) (General International Standard Archival Description), is a standard defined by ICA (International Council for Archives) for elements that should be included in a finding aid register for archival documents produced by corporations, persons and families. It is maintained and documented by the ICA Committee on Descriptive Standards, www.icacds.org.uk/eng/standards.htm.

9 TEI (Text Encoding Initiative) is a consortium which develops and maintains a standard for the representation of texts in digital form, addressing the needs of the humanities, social sciences and linguistics.

Material culture as a key component of the humanities

Tangible cultural heritage is a key component of European identity. Accumulated research data and datasets created for management purposes but containing invaluable information also from a research perspective, form together a rich and largely unexploited resource. Dealing with tangible substance is a distinctive tract of this sector of the humanities field. Bridging the tangible and intangible components of cultural heritage is a challenge for data services and future Virtual Research Environments (VRE) for the humanities.

When dealing with tangible heritage, availing of virtual 3D replicas is a key component of data management and, as already mentioned, an acknowledged pillar for the research of tomorrow. Applications in the cultural heritage domain have already been explored within previous EU-funded projects such as EPOCH, 3D-COFORM and various FP7 STREPs. As for all technological contributions, keeping strong connections with the humanities is the only way to acknowledge the inseparability of the tangible and intangible components of cultural heritage.

Virtual Research Environments

As mentioned above, collaborative research does not belong to the DNA of humanities scholars. Nevertheless, challenges like the "big data issue"that today affect the sector[10] no less than other scientific domains, require IT-based teamwork and hence naturally promote VRE.

VRE for the humanities are difficult to design and to create due to the nature of the matter investigated, ranging from a most intangible substance such as human thought to the strongly tangible remains of the past studied by archaeologists and restorers. VRE must include collaborative research environments that are aware of the individual study tradition. They need to provide cutting-edge IT tools to simulate in silico experiments that cannot be done in vivo, as is the case for restoration or for scholarly re-enactment of the past. They must enable the use of 3D virtual replicas and advanced scientific visualization whenever tangible matter is involved. They must enable the semantic manipulation and synthesis of the vast amounts of heterogeneous data made available by improved access to large-scale repositories. They must be based on a methodology that is accepted by the scholarly community as an integral part of the disci-

10 Niccolucci, F. and Richards, J. (2013): ARIADNE: Advanced Research Infrastructures for Archaeological Dataset Networking in Europe. International Journal of Humanities and Arts Computing 7 (1-2), 70–88, especially 73/74.

plinary toolbox. In other words, VRE in the humanities are today's challenge for the domain. However, work here does not start from scratch, and action may rely on various modules already available and build on tools that have been fruitfully used until now, but not yet assembled into an integrated environment.

Intellectual Property and the humanities

In the arts and humanities there is a tension between universal access to culture and the legitimate rights of cultural workers that parallels the tension between public open access and use of heritage and the protection of the owning community's rights on its exploitation. This often concerns huge commercial business and is therefore no easy topic to deal with. The traditional individualism of scholarly work makes open access for research a rather sensitive topic. There is a need to carry on studies on how an advanced use of ICT impacts on IPR[11] in the humanities. This may start from work already done on sectorial topics[12] and move on towards a global approach that reconciles protection with the openness of culture.

Conclusion and future work

In the first year of work, ARIADNE met acceptance in the scientific community well beyond the most optimistic expectations. Policy managers and scholars, not only in the cultural domain, now acknowledge that archaeology and heritage-related research need to incorporate digital technologies in their disciplinary body, and integrating activities such as those envisaged by ARIADNE are strongly needed. This is grounds for optimism about the feasibility of and the expected support for the future work as outlined above.

11 Intellectual Property Rights

12 See for example the work done within the 3D ICONS EU project on IPR management for 3D models of cultural heritage: Spearman, M./Emslie, S. and O'Sullivan, P.: D7.2 IPR schemes, available from www.3dicons-project.eu/eng/Resources.

6.4 Sustainable Data for Sustainable Infrastructures

Laurent Romary (Inria, DARIAH)

Abstract

DARIAH, the Digital Research Infrastructure for the Arts and Humanities, is committed to advancing the digital revolution that has captured the arts and humanities. As more legacy primary and secondary sources become digital, more digital content is being produced and more digital tools are being deployed, we see a next generation of digitally aware scholars in the humanities emerge. DARIAH aims to connect these resources, tools and scholars, ensuring that the state-of-the-art in research is sustained and integrated across European countries.

To do so, it is important to understand the actual role that proper data modelling and standards could play to make digital content sustainable. Even if it does not seem obvious at first sight that the arts and humanities would be fit for taking up the technological prerequisites of standardisation, we want to show in this paper that we can and should integrate standardisation issues at the core of our DARIAH infrastructural work. This analysis may lead us to a wider understanding of the role of scholars within a digital infrastructure and consequently on how DARIAH could better integrate a variety of research communities in the arts and humanities.

DARIAH – A digital infrastructure for the arts and humanities

In recent years there have been two major trends that have directly impacted on the establishment of DARIAH as an e-Research infrastructure in the humanities:

- A remarkably growing interest for digital methods in nearly all research domains in the humanities at large[1];
- The development of generic eScience initiatives, usually anchored on strong political activities at national and European levels[2].

In this context, DARIAH faces a double challenge of a) possible difficulties with focussing on precise objectives in terms of service provision because of the large number of communities to address at the same time and b) to spend most of its energy on liaising (or concerting) with other ongoing (maybe ephemeral) projects and/or political bodies which see their own activities as related to ours.

To circumvent these difficulties we can identify some core strategic orientations based, on the one hand, on the essential steps in the digital scholarship workflow and, on the other hand, on suggesting a strong data oriented perspective for DARIAH, which may help us identify where we have a real role to play and where we need to collaborate with others. Whereas we acknowledge that the technological context is also an important factor to consider, and will indeed appear at several points in our presentation of the institutional landscape, we do not provide any specific background analysis here.

The history of DARIAH began in January 2006[3] when representatives from four European institutions[4] met to identify how they could join efforts in providing services to the research communities they served, with a strong focus on the humanities. The idea behind this initiative was to move towards a consortium of institutions which would ensure the long-term sustainability of the underlying infrastructure and a strong political voice towards the EU. Each institution played a role in coordinating or developing digital services in the humanities at national level, and could thus speak from a national perspective.

1 See the growing success of the Digital Humanities conferences (adho.org/conference).

2 The latest development of which is the overarching Research Data Alliance (rd-alliance.org).

3 Just a few weeks earlier, the first meeting of Clarin took place at the same location in the Headquarters of the CNRS in Paris, grouping together the previously existing Parole and Telri networks.

4 Sheila Anderson, director of AHDS; Peter Doorn, director of DANS; Laurent Romary, director for scientific information at CNRS; Ralf Schimmer, representing Harald Suckfuell, in charge of scientific information for the Max Planck Society.

Within a hazardous context in which the idea of going digital is not necessarily mainstream in the humanities, DARIAH has managed to move forward to a stage where it is about to become one of the most stable components in the eHumanities landscape. Still, this should not prevent us from analysing the reasons why it is so complex to establish an infrastructure for the humanities; a problem that can be construed along the following lines of tension:

- A research infrastructure in the humanities should be able to provide concrete short-term services that may lend it scholarly recognition;
- At the same time, it should have a clear vision of its general objectives that will guide the evolution of the infrastructure over the years;
- It should gain institutional support for both aspects and demonstrate that it matches the strategic objectives of its funders;
- It should elicit how much it complements local initiatives to provide technical support to researchers;
- It must show its value for money in the sense that scholars do not see the infrastructure as consuming budget that would otherwise go to research.

These elements potentially apply to all scientific domains. Still, the humanities represent an even more complex environment because, on the one hand, of their highly fragmented scholarly structure, and, on the other hand, of their low technical literacy. Besides, the humanities are usually subject to comparatively low budgets, which leaves even less leeway for dedicating funding to infrastructural activities. Whereas DARIAH has managed to gain institutional recognition at European and national level, it is its capacity to relate to this complex community of users that will be a real measure of its success.

A user-oriented view on DARIAH

In the short term, DARIAH will have to provide simple services that correspond to the expectations of its users. By users, we mean the now quite large community of scholars who have to deal with digital content[5], regardless of whether they master the technical background related to the creation or management of these resources.

The sufficiency with which services fulfil expectations will rely a great deal on the level of digital awareness that scholars actually have, which in turn may change rapidly in the coming period. We will thus have to face the difficult situation of responding to changing needs, as well as having to deal with a very heterogeneous community ranging from early adopters of digital techniques to completely computer illiterate scholars.

In this context, simple services can be characterised by the fact that, on the one hand, they can easily be adapted to new usages and new demands, and, on the other hand, they are closely anchored on the basic processes related to the scholarly research process, seen here from the point of view of working with digital data or sources.

In the remaining section we will briefly go through what we think are the essential aspects of the scholarly research process and identify the services that DARIAH should prioritize accordingly[6].

Finding and quoting digital sources

The most important step for introducing a virtuous digital circle in the humanities research process is to provide scholars with the means to identify and locate existing digital sources that they can explore, study, and finally reference in their own research. To help achieve this, DARIAH works on deploying services along the following lines:

- Discovery portals that acts as single entry points to existing online resources[7];

5 Usually because they benefit from a research grant where they engaged themselves in delivering digital content or applying digital methods.

6 See also: "Reinventing research? Information practices in the humanities", Research Information Network Report, April 2011. www.rin.ac.uk/our-work/using-and-accessing-information-resources/information-use-case-studies-humanities.

7 See the exemplary service provided by Isidore at CNRS (www.rechercheisidore.fr)

- Recommendations on optimal web searchability (e.g. what to provide access to, which entry points, in the context of sitemaps, for instance) to be widely disseminated within the research communities in the humanities, but also to funding agencies for them to integrate these in their call for projects;

- Interfacing in such portals of exemplary resources and archives in targeted scholarly domains (this could be based on the direct output of national and European initiatives such as EHRI[8] or CENDARI[9]) to foster the use of online resources;

- Recommendations concerning the citation of sources in the humanities, combining appropriate reference to the source as well as to its creator.

Creating and annotating digital content

The second important step in going digital is to be able to create one's own digital assets out of existing primary analogue sources, or annotate (resp. enrich) existing digital sources. In this domain, DARIAH prioritizes the provision of services that help scholars to quickly learn how to work autonomously in a digital environment. In particular, we need to focus on the following core services:

- Guidelines for the elementary creation of digital sources ("starter set") – together with appropriate reference examples[10];

- Provision of editors in a box that point to a reduced set of environments that can be directly installed or used online to create relevant scholarlydigital content;

- Advertise and/or organize training workshops all over Europe so that scholars or newly hired students can be trained and gain quick autonomy.

These services should be strongly articulated with the standardisation strategy we will delineate later in this paper.

8 www.ehri-project.eu

9 www.cendari.eu

10 In the case of textual resources, we would for instance point to the TEI by example page (tbe.kantl.be/TBE/) and contribute to its maintenance.

Preserving and disseminating content

Once digital assets have been created, it is essential that researchers are not left wondering how to make them widely accessible while ensuring that the resources will be trustfully used and cited. To this end, the DARIAH short-term agenda includes the following priorities:

- Provide transparent services to facilitate the unique identification of researchers. In this domain, we should take an early part in the Orcid initiative but also encourage the deployment of national initiatives for researcher identification[11];
- Provide an online service for research asset PIDs. In this context, we should strengthen our relationship with EPIC[12] and DataCite;
- Provide recommendations on a core set of meta-data they have to apply in their resources to make them useful and citable for other researchers (identification and documentation of the source, sampling strategy, description of the digitization added-value, proper identification of responsibilities and affiliations)
- Provide recommendations on simple licensing schemes to be applied in digital assets. Basically, we should advocate a simple CC-BY license for all publicly funded projects to which no further constraints apply (cf. open access discussion below);
- Offer an early service for archiving and hosting generic digital resources (images, XML transcriptions). This should not only be implemented through an archive-in-a-box strategy, but also by offering real hosting services (e.g. XML database farms)

Additional service related to publications

Although scholars may not request it from the outset, DARIAH needs to provide the necessary expertise concerning the management of publications in the humanities. We thus recommend that the following aspects be pursued at an early stage of the creation phase of DARIAH:

- Provide advice (even proselytise) on open access and in particular the early deposit of scholarly papers in a publication repository;

11 See for instance the IdRef service at ABES in France (http://www.idref.fr)

12 EPIC – the European Persistent Identifier Consortium; http://www.pidconsortium.eu

- Recommend appropriate editorial platforms for the creation of new journals or the migration of existing ones towards scholarly models;
- Provide a critical study of existing scientific social networks and in particular identify their actual capacity to relate to publication archives.

Overview of short-term priorities

We understand that DARIAH can benefit scholars by offering modest but targeted services. DARIAH should also be able to boast this modesty to external actors (members, EU) and show how it is part of a long-term strategy to develop an infrastructure for the humanities.

The adequate provision of a sound portfolio of such needs-oriented services will facilitate the development of more ambitious digital humanities environments. In particular, such basic services should be thought of as preliminary building blocks in the creation of more elaborate virtual research spaces[13] based on a more data-oriented perspective, as outlined in the next section.

A data-oriented view for DARIAH

Towards a stable perspective for DARIAH

Contrary to the short-term strategy, the long-term vision of DARIAH should somehow go beyond a purely user-centric view. Indeed, given the speed at which technological awareness is presently evolving, it is nearly impossible to anticipate what scholars will actually request from a digital infrastructure in the humanities over the next five years alone. In this context, our duty is to create a sound and solid background that is likely to ensure the stability of digital assets in the long run, but also the development of a wide range of as yet unanticipated services to carry out new forms of research on these assets.

This data-centred strategy echoes various reports and statements that have been issued recently, in particular "Riding the wave"[14], which has placed the management of scientific data very high on the EU commission's agenda. This report stresses the importance of a long-term strategy concerning the manage-

13 Cf. Romary, Laurent (2012): Scholarly Communication. In: Mehler, A. and Romary, L.: Handbook of Technical Communication de Gruyter. hal.inria.fr/inria-00593677.

14 ec.europa.eu/information_society/newsroom/cf/itemlongdetail.cfm?item_id=6204

ment of scholarly data in all disciplines, which comprises both technical aspects (identification, preservation), editorial aspects (curation, standards) and sociological aspects (openness, scholarly recognition).

In this section, we go even further by considering that a data-centred strategy for DARIAH will secure a long-term vision both in terms of the deployment of future services andin the way we organise our collaborations with other initiatives, in particular in the cultural heritage domain. To do so, we outline the role of digital surrogates in digital humanities as a core concept for data management and explore the actual consequences of such a vision.

NOTE: We will speak henceforth of primary sources as covering all types of documents or information sources that may be used as testimonial information to support research. This wide notion typically covers objects such as manuscripts, artefacts, sculptures, recordings, statistical data, observations, questionnaires, etc.

Surrogate–Definition

We define a surrogate here as an information structure intended to identify, document or represent a primary source used in scholarly work. Surrogates can take a wide variety of forms ranging from metadata records, scanned images of a document, digital photographs, transcriptions of a textual source, or any kind of extract from or transformation[15] of existing data.

The notion of a surrogate is at the core of digitally based scholarship since it is intended to act as a stable reference for further scholarly work, as a replacement for – or complement to – the original physical source it represents or describes. By definition, it should always contain some minimal information to refer to the source(s) upon which it is based.

In turn, a given surrogate can act as a primary source for the creation of further surrogates, for instance with the purpose of consolidating existing information or creating complex information structures out of different sources.

As a consequence, a network of digital surrogates will reflect the various steps of the scholarly workflow where sources are combined and enriched up to the point that the results can be further disseminated to a wider community. Indeed, we do not anticipate a flat space of digital surrogates, but a complex data space integrating the various evolutions that such surrogates may encounter.

15 E.g. the spectral analysis of a recorded speech signal.

In the remaining sub-sections we will analyse the consequences of having surrogates at the centre of our perspective concerning digital humanities, and contemplate the impact of this on our delivery of services.

Data management issues

A coherent vision on a unified data landscape for humanities research should be based upon a clear policy in the domain of standards and good practices. In particular, DARIAH should not only make strong recommendations as to which standards may optimize the sharing and use of digital surrogates in research activities, but it should also contribute to shaping the standardisation landscape itself by supporting participation in corresponding working groups and organisations.

Acknowledging the fact that other communities of practice (publishers, cultural heritage institutions, libraries) may have different agendas and practices in the domain of standards, we should also endeavour to define interoperability conditions between heterogeneous worlds (e.g. EAD – TEI relationship).

Finally, we need to assess the consequences of an extremely widely distributed network of potential data sources, ranging from individual scholars to major national libraries. Providing guidance to individual users as to how one can navigate and use digital assets in such a heterogeneous data landscape will be a major challenge for DARIAH. To this end, the evolutionary surrogate model outlined above will be essential in defining conditions aggregating identifiers, versions and enrichments of digital assets.

Technical issues

Whereas the data landscape will heavily rely on third party providers (cf. political issues below), the development of a data-based strategy for DARIAH will impact on some of our technical priorities in the short term as well as the long term. We can outline the three levels where DARIAH should invest specific efforts as follows:

- Define a repository infrastructure for scholarly data where researchers can transparently and trustfully deposit their productions. Such an infrastructure should be in charge of maintaining permanent identification and access, targeted dissemination (private, restricted and public) and rights management. In this context we should identify

the optimal level of centralization that allows efficiency, reliability and evolution[16];

- Spend meaningful effort on defining and implementing standardized interfaces for accessing data through such repositories, but also through third-party data sources. The objective of such interfaces must be to make it easy to derive simple services in the domains of threading, searching, selecting, visualising, importing data;

- Experiment with the development of agile virtual research spaces based on such services that allow specific research communities to adopt their own data-based research workflow while being seamlessly integrated in the DARIAH data infrastructure[17].

Licensing issues – open access strategy

The evolution of the digital humanities towards a complex and interrelated data landscape will require a strong policy concerning the legal conditions under which each data asset will actually be disseminated. To tackle such issues, there are indeed two different, but probably complementary, points of view:

- Theideological factors in the debate provide that each scholarly production financed by means of public funding is in essence a public good[18]. This should lead us to defend a generalised open access strategy for all scholarly productions;

- A pragmatic view, informed, for instance, by the experience of the genomic domain, acknowledges that it is unpractical, even impossible, to do data-based research within a data landscape bearing heterogeneous reuse constraints and/or licensing models.

All in all, the core reasons why we have no choice but to work towards an open data space are well identified and boil down to the issues of[19]: more efficient scientific discovery and learning, access for other researchers – and the wide

16 Cf. for a discussion of possible models: Romary, Laurent and Armbruster, Chris (2010): Beyond institutional repositories. International Journal of Digital Library Systems 1 (1), 44–61. hal.archives-ouvertes.fr/hal-00399881.

17 See Romary, Laurent (tbp): Scientific information.

18 Which in the humanities strongly overlap with the notion of "scientific good" (as opposed to the case of bio-medical research for example).

19 Freely adapted from a personal communication from Trish Groves, Deputy editor, BMJ (British Medical Journal). Note here that although the words used are clearly referring to hard sciences, they seem to perfectly fit what we could dream of in the human sciences.

public – to raw numbers, analyses, facts, ideas, and images that do not make it into published articles and registries, better understanding of research methods and results, more transparency about the quality of research, greater ability to confirm or refute research through replication.

To achieve this in the humanities, DARIAH should provide guidance on two complementary aspects:

- Advocate an early dissemination of digital assets, explaining that the fear of compromising academic primacy should be put in perspective with the potential gain in extra citation to the data itself;
- Encourage the systematic use of a Creative Commons license CC-BY, that basically supports systematic attribution (and thus citation) of the source.

To take a further example from the genomic field, CC-BY should be preferred to less restrictive (e.g. CC-0) licenses, since attribution lies at the centre of the academic process, and of course to more restrictive ones, which are either inapplicable ('share-alike') or prevent a wide use of the digital asset ('non-commercial').

Besides, DARIAH should apply this scheme to itself in such a way that all documents and data produced specifically within DARIAH (or DARIAH affiliated projects) should be associated with a CC-BY licence. DARIAH should also contribute to large scale negotiations with cultural heritage partners (libraries, museums, archives, or representatives thereof) to ensure global agreements through which the lightest possible licensing schemes are applied to the data made available to scholars.[20]

Political issues

The global strategy put forward above concerning the management of digital assets/surrogates in the humanities is by far too complex to be dealt with within DARIAH alone. It is of strategic importance that we articulate our activities in this domain in strong collaboration with the various actors of the data continuum we have identified. In particular, we need to consider to what extent potential data providers (cultural heritage entities, libraries or even private sector

20 To cite here the final conclusions of the High Level Expert Group on Digital Libraries, under the auspices of commissioner Reding: "public domain content in the analogue world should remain in the public domain in the digital environment."

stakeholders such as Google) could become partners in creating the seamless data landscape we are all dreaming of. Such partnerships should be articulated along the following lines:

- General reuse agreements[21] that would systematically apply when scholars require access to sources available from data providers, comprising usage in publications, presentation on web sites, integration (or referencing) in digital editions, etc.;
- Definition of standardized formats and APIs that could make access to one or the other data provider more transparent;
- Identification of possible scenarios in which the archival location of versions of records is clearly identified and, by the same token, enrichment mechanisms are contemplated[22].

Role of standards

The main issue in defining a policy about standards is to understand what they actually are. Standards are documents informing about practices, protocols, artefact characteristics or data formats that can be used as reference for two parties working in the same field of activity to be able to produce comparable (or interoperable) results. Standards are usually published by standardisation organisations (such as ISO, W3C or the TEI consortium), which ensure that the following three requirements for standards are actually fulfilled:

- Expression of a consensus: the standard should reflect the expertise of a wide (possibly international) group of experts in the field
- Publication: the standard should be accessible to anyone who wants to know its content
- Maintenance: the standard is updated, replaced or deprecated depending on the evolution of the corresponding technical field

Standards are not regulations. There is no obligation to follow them except when one actually wants to produce results that can be compared with those of a wider community. This is why a standardisation policy for DARIAH should include recommendation as to which attitude the scholarly communities could or should adopt with regards to specific standards.

21 We should take as a background document "The Europeana Licensing Framework", issued in 2011, see creativecommons.org/weblog/entry/30609.

22 For example, TEI transcriptions made by scholars could be archived in the library where the primary source is situated.

The preceding characteristics outlined for standards put a strong emphasis on the role of communities of practice and the corresponding bodies that represent them. Ideally, a good standard reflects the work of a relevant community and is maintained by the appropriate body. This is exactly with the case for the Text Encoding Initiative for text representation standards and, to a lesser extent, for EAD, whose maintenance is taken up by the Library of Congress with support of the Society of American Archivists.

Because there is no obligation to use a given standard, it is essential to provide potential users with a) awareness about the appropriate standards and the interest to adopt them, and b) the cognitive tools to help them identify the optimal use of standards through the selection and possibly customisation of a reference portfolio. In our experience with working with numerous projects (including those cited in this document) that were in the need of adopting existing standards, there was always an initial phase in which scholars should be made aware of some core standards that are systematically related to the definition of interoperable digital objects. We call these core standards a standardisation survival kit (SSK) and outline in Table 1 a first group of such standards. As we will see later in this document, the SSK should be part of several concrete actions for DARIAH in the domain of education and interaction with funding agencies.

An important aspect in this dissemination strategy is that projects should be told to refrain from defining their own local formats and instead first demonstrate that their needs are not covered by the wide varieties of already existing initiatives in the digital humanities landscape. This is also why DARIAH should avoid taking any specific lead in the definition of new standards[23], but should have a pro-active role in helping communities to participate in standardisation activities where they exist. Such a strategy will also contribute to the actual stabilisation of existing conceptual and technical knowledge within ongoing projects, as well as providing a channel for the wider dissemination of the corresponding results.

23 In this respect we should strongly depart from the strategy adopted in Clarin with infrastructure-internal format developments such as TCF or CMDI.

ISO 639 series	Codes for the representation of languages and language families
ISO 15924	Codes for the representation of scripts
ISO 3166	Codes for the representation of country names
IETF BCP 47	Standard for encoding linguistic content, combining ISO 639, ISO 15924 and ISO 3166
ISO 10646, Unicode	Universal encoding of characters
ISO 8601	Representation of dates and times
XML recommendation	Provides the basic technical concept related to XML documents

Table 1: Outline of a standardisation survival kit

Recommendations

The preceding sections could potentially lead to many possible action points for DARIAH. At this stage, we can boil these down to the following concrete recommendations:

- Define a basic curriculum on data modelling comprising awareness about digital surrogates, meta-data, versioning, multiple publishing, annotation and re-use
- Re-design the schema registry activity to focus on designing data models and formats toolkits for research projects
- Define and maintain a Standardisation Survival Kit that corresponds to the baseline of an awareness and recommendation activity on standards
- Support and coordinate (VCC2 and VCC4) standard awareness workshops targeted at specific scholarly communities
- Encourage DARIAH members to allocate means for their participating institutions to contribute to standardisation activities

Conclusion

DARIAH should contribute to excellence in research by being seminal in the establishment of a large coverage, coherent and accessible data space for the humanities. Whether acting at the level of standards, education or core IT services, we should keep this vision in mind when setting priorities as to what will impact the sustainability of the future digital ecology of scholars. Above all, such a strategy should directly influence the way we will advocate DARIAH towards funding or supporting institutions, and also how we will manage our collaboration schemes with other initiatives in Europe and worldwide.

C Future

7 Future Strategies and Directions

7.1 Report from the DASISH SSH Workshop, Gothenburg, 4–5th October 2013

Hans Jørgen Marker[1] (SND)

The DASISH SSH workshop was held with the purpose of exchanging information and enhancing coordination between infrastructures and projects active in the social sciences and humanities[2]. Their future perspectives were the central theme of the workshop.

On the first day of the workshop the five central infrastructure meta-projects were presented: CESSDA, CLARIN, DARIAH, ESS, and SHARE and from the projects present: DASISH, ADRIADNE, CENDARI, CHARISMA, DwB, EHRI, and InGRID.

On the second day the presentations were followed up by general discussions about the challenges the projects and the infrastructures have to face in the future. A central purpose of these discussions was to exchange information and views between the participants in the hope that it will facilitate future coordination between them.

The five infrastructure projects have some longevity as they have completed, or are near to completion, of the establishment of legal entities, whereas the projects by the very nature of things will end at a known date. The cooperation between the involved project partners may continue and the resources created through the projects have to be maintained.

Horizon 2020 will be able to facilitate future cooperation between the project partners and between the infrastructure projects. For this reason the concrete implementation of Horizon 2020 is naturally of the highest interest. However it was noted that some cooperation might go beyond the SSH domain and that interdisciplinary cooperation should be encouraged where most salient.

This paper is a result of the discussion that took place during and after the workshop. It identifies some of the areas in which the workshop participants found future activities and cooperation were needed the most.

1 With excellent help from the workshop participants.

2 The invitation for the workshop was extended to the list of Networks of RIs funded under FP7 as Integrating Activities (ec.europa.eu/research/infrastructures/index_en.cfm?pg=ri_projects_fp7)

Infrastructures are different

There is a lot of diversity in the social sciences and humanities and it is not possible to find a single definition of a proper research infrastructure in the social sciences and humanities; it is rather a whole spectrum of different requirements. This entire spectrum has to be taken into consideration. There are two basic distinctions: infrastructures are either based on a very large initial investment and modest maintenance cost or they have been granted a modest initial investment and subsequently have much higher maintenance costs. Infrastructures in the social sciences and humanities are often of the second type. This has important consequences to how the infrastructures work and – more importantly – how sustainability is achieved. Some SSH infrastructures are based on the provision of shared tools, others on the provision of shared data, and yet others on the joint exploitation of data resources. With this range of different types of support for SSH, infrastructures need to be flexible and multifaceted to ensure that the SSH can be effective in answering great societal challenges.

The research infrastructures in the social sciences and humanities clearly lay a different emphasis on the research aspect and the infrastructure aspect. All of them aim at improving accessibility, sharing and interoperability of software architectures and solutions. However, some are to a higher extent more data oriented while others are focus on providing tools and services for the research community. This has implications for the support they require.

The wide range of project types - and thus of EC funding schemes (Integrated Activity capacity-building projects, RI projects, cluster projects, etc.) – clearly serves the goals set for the next generation of programmes. However, efforts should be made to allow for a greater integration of projects with each other which there is a lack of in their current generation. This is of particular relevance to the hardly-existing formal/institutional links between IA (Integrated Activity, former FP6 I3) projects and RI projects. Potential pathways from IA projects to RI projects should also be further explored.

This lack of integration can also be found in the horizontal links between IA projects: In some cases, they are conspicously working on identical issues (e.g. metadata quality management, accreditation processes, legal issues, access conditions, etc.) and may advocate different solutions to such cross-cutting issues – thus being detrimental to their integrated capacity-building role. Although these projects should coordinate their efforts as a matter of course, it

cannot be expected that such coordination is as efficient as possible considering (1) their limited resources and (2) the continuous multiplication of such projects.

This said, some communalities remain:

- There is a move to bigger tools and their use across disciplines.
- Interdisciplinarity facilitates creativity, for instance, through cooperation with infrastructures outside of the social sciences and humanities such as life sciences or other clusters. At the same time, new opportunities for collaboration within the SSH cluster have been identified.
- There is an interest in open data, linked open data, and Big Data.
- Social data (e.g. Facebook) creates new possibilities for data collections and research.

Joint registries for Centres and Services and networks of Social Sciences and Humanities Centres

The need to establish a federation of trusted centres for research data that store, manage, preserve and give access to data in a trusted way is widely recognised. Trusted centres must fulfil certain criteria and undergo regular assessment to ensure that their policies are adequate. This is especially required as a consequence of 'policy-based archiving' which is another item on the 'common cluster activities' list. They will also need to provide certain standard entry points to support data access, monitoring, etc. For example, to support citation and reproducible science, reliable and continuous services are needed. To provide access to confidential data, security has to follow agreed standards.

This development has been precipitated by the RDA-WDS (Research Data Alliance/World Data System) initiative and this is the reason that some people associated with WDS/ICSU started an initiative to draft a common worldwide registry with human and machine readable information with an agreed upon structure to allow automated procedures as it is known from computer networking:

- Funders want to ensure that the data they funded will be stored in "trusted centres"

- There are some centres in the social sciences and humanities that are able to play such a role as trusted centres in a worldwide network:
    - CLARIN now has about 15 certified/almost certified centres
    - CESSDA has a network of 23 trusted centres, certification is on-going
    - DARIAH has some well-prepared national centres (DANS, ADONIS, etc.)
    - There are strong libraries ready to fulfil the requirements
- However, there are regional differences in Europe:
    - There are less centres in the South-Eastern part of the continent
    - Many of the centres in the humanities have insufficient resources to live up to future requirements.

The social sciences and humanities should participate in the WDS/ICSU initiative and it must be ensured that there are enough resources available to adapt all social sciences- and humanities-centres to participate in this world wide registry. Broad implementation is important for visibility, recognition and so fourth. There is need for more funds to close existing financial gaps and to remain competitive. There is a close relation to the next paragraph, "policy-based" archiving, which describes ways to help centres meet future challenges brought about by increasing amounts of data.

Policy-based archiving

Policy-based archiving is one of the essential requirements to establish trust in data providers in the long run. All policies that are applied by a centre should be based on explicit and declarative statements, which are then turned into executable, certified procedures. In the future there will be no other ways to assess the quality of centres. This issue is being addressed in the DASISH project as well as in CESSDA, but more work is needed. There is also an initiative to create a registry of accepted policies for different tasks such as preservation, replication, curation, giving access, etc.

Data centres in social sciences and humanities are presently working to provide explicit policies. A starting point for expressing basic policies is the Open Archival Information System (OAIS). However, we need to go far beyond OAIS to meet the needs emerging from using concrete architectures and solutions.

Sustainability

Social sciences and humanities need long-term commitment. Sustainability is therefore a very important issue that has various dimensions.

Sustainability of resources created by projects is a rather obvious issue. In some cases the output of a project is something that is dealt with by existing infrastructures – such as the five SSH ESFRI infrastructures, data archives or libraries. But in many cases the created results constitute big challenges. Take for example a web resource presenting the legal and ethical rules presently in force in the social science and humanities domain in Europe. Such a resource will very rapidly be reduced to a historical document of limited interest if it is not maintained properly.

Moreover, particular attention should be paid to not multiply the number of research infrastructures when even the existing ones may not be stabilised yet (notably in terms of funding) for the reasons described in part 1. Though we agree that the focus should be on a competitive and world-class European Remote Access (ERA) based on strong and innovative research infrastructures, current obstacles and challenges related to the consolidation of the "existing" must not be left aside or overlooked.

Software tools and services represent a specific challenge. Although there is no escaping the fact that continuous maintenance costs money, paying attention to software sustainability can actually minimise costs. The subject should also be considered important for aspects such as:

- Verifiability of results: some results need the original tools to be reproducible
- Persistency of knowledge, cost-effective training of students and PhDs.
- Maintenance of created tools and updates of generated resources (e.g. databases)

- Limited funding – most of the existing SSH infrastructures do not have permanent funding, but live on a project-like funding scheme with no assurance that funding will continue. Sustainability must be secured in order to avoid waste of investment.

- In social sciences a key challenge is the regular collection of data across as much of the ERA as possible. Ensuring the sustainability of data collections for infrastructures like SHARE and the ESS is critical and the transition to ERIC statutes is causing considerable challenges in this respect.

- In the humanities a broad coverage in terms of e.g. languages is similarly necessary in order to live up to the basic idea of infrastructures.

Besides applying better software development methodologies, there are also organizational strategies that can help. First of all enlarging the user-base of a tool should be considered as it also increases the possible funding base. Therefore, domain specific, software registries with a purpose of sharing knowledge about specific tools are important means to facilitate this. Those registries should also encourage feedback from its users and sufficient resources should be made available to check on entered tool information. Another organisational strategy to promote sharing tools is the use of broker organizations; such an organization has (domain specific) expertise on the available tools and is able to broaden their user-base by attracting new user-groups.

VRE: Virtual Research Environment

VREs are software applications that can integrate (existing) tools and services for a community's research workflow. It allows sharing of data and services and is a one-stop shop for specific research workflows. They can be either specific tools with considerable logic packaged into a single programme or they are themselves flexible research infrastructures that can be modelled to a specific workflow.

Different groups and projects have been developing and are using such VRE (-like) functionality e.g.: TextGrid (DARIAH), or they are developing infrastructure use-cases that come close to it, like CLARIN's tools (WebLicht+VLO+Annotation) and of course the NESSTAR software and other CESSDA products which have been providing some of these features already for decades. The DwB project

underlines the importance of multi-level VREs for an ERA-network that will allow researchers to work together on confidential microdata.

Some important aspects of a VRE:

- Collaborations of (also virtual) organizations or users to access services and resources
- Virtual storage and remote services
- Registries for services and resources via metadata and PIDs: the (internal) administration
- Interoperability between services and data
- Data from the web (annotation and processing)
- Transparent archiving and publication of end product data
- For the sustainability of VRE software one has to rely as much as possible on existing, general infrastructure services rather than making specialised services within the VRE.

Crowd sourcing

Crowd sourcing is not entirely new. However, large-scale crowd sourcing is and it will change how research is undertaken in many areas. Citizens will be able to actively participate as producers and consumers and the role of researchers will change.

The tools and infrastructure to support such an increased interest in crowd sourcing are not yet in place. Therefore, new projects are creating new tools that are virtually identical to existing tools at the moment. We are in need of an infrastructure that provides the framework for crowd sourcing projects so that these projects can concentrate on their core subject.

Providing such an infrastructure will create new possibilities but it will also result in big challenges. Therefore, the following requirements have to be ensured:

- Quick action to prevent future data losses due to amateurish and unsustainable developments
- Involvement of small and medium sized enterprises, as they know how to do cross-platform programming for mobile devices and how to design user friendly apps.
- Design studies that address the issue of representativity in crowd sourcing studies and prevent waste of resources by non-scientific approaches.
- A strong European initiative that will ensure that the EU will have a strong hold on the infrastructure level.

Big Data

There is a growing number of ways to access administrative and social media data and to take advantage of new data collection technologies. The SSH domain is shaping these exciting new opportunities. But we still need the following:

- Support for design studies that facilitate innovation in this area
- Effective support for collaboration with the private sector
- Flexible cluster projects that allow true interdisciplinary work
- Close involvement of NSIs in collaboration with the CESSDA network
- Close involvement of Eurostat

Furthermore, this domain might also be highly relevant as it creates pathways and incentives for a better involvement of commercial enterprises – and a greater integration of the private sector with the broader research community – which is partly unsatisfactory in the current generation of programmes.

Workshop participants (27 people)

DASISH:	8 representatives
ADRIADNE:	2 representatives
CENDARI:	1 representative
CHARISMA:	2 representatives
DwB:	2 representatives
EHRI:	2 representatives
InGRID:	1 representative
SHARE:	2 representatives
ESS:	1 representative
DARIAH:	1 representative
CLARIN:	1 representative
CESSDA:	1 representative
Europeana/The European Library:	1 representative
Facing the Future (conference organization team):	4 representatives

7.2 Better Transnational Access and Data Sharing to Solve Common Questions

Julia Lane (American Institutes for Research),
Victoria Stodden (Stanford University),
Stefan Bender (IAB),
Helen Nissenbaum (New York University)

Introduction

Science has always required a community to study big questions. While the Manhattan Project and the moon landing are the most famous recent examples, scientific discourse has always required scientists replicating and validating each other's work. Indeed, scientific discourse has been central to the development of many statistical procedures in their own right – astronomers' attempts to measure astronomical distances led to the development of Gaussian statistics (Stigler 1988).

Social scientists are increasingly being asked to answer equally big questions. A partial, non-exhaustive list might include such challenges as increasing employment and reducing unemployment; addressing the problems posed by ageing populations; understanding the human dimensions of climate change; improving food and energy security; building a better understanding of how to support science and foster innovation (Elias and Entwisle 2012).

Yet the data world has changed. As has been abundantly noted elsewhere, the key distinction between the data world that existed twenty years ago and the world now: data are less likely to be collected purposively and with a clear legal mandate. This has diminished the leadership role held by statistical agencies. Indeed, while an edited book on statistical disclosure protection issued in 2001 featured five chapters authored by statistical employees, a parallel book to be published in 2014 features none (Doyle et al. 2001; Lane et al. 2014). As a result, data access is threatened – decision makers have many fewer resources on which to draw in making decisions about whether or not to release data, and consequently may well err on the side of caution, harming our ability to answer important societal questions. We need an international research agenda to address many of the key legal, operational and statistical issues.

A legal research agenda

The change in the ways in which data are collected means that the legal framework within which data is used, namely the framework within which data are accessed and social science research conducted, has changed. These changes include the authority to collect data as well as the ownership and stewardship of data.

Each individual now produces data that are potentially useful for research as part of their everyday participation in the digital world. There has always been a lack of clarity in legal guidance stemming from a lack of clarity in who owns the data – whether it is the person who is the subject of the information, the person or organization who collects data (the data custodian), the person who complies, analyzes or otherwise adds value to information, the person who purchases interest in data, or society at large. And the lack of clarity is exacerbated because some laws treat data as property and some treat it as information (Cecil and Eden 2003). But the new types of data make the ownership rules even more unclear: data are no longer housed in statistical agencies, with well-defined rules of conduct, but are housed in businesses or administrative agencies. In addition, since digital data can be infinitely lived, ownership could be claimed by yet to be born relatives whose personal privacy could be threatened by release of information about blood relations.

Trust clearly depends on people's views on privacy, but these views are changing rapidly (Nissenbaum 2011). Nissenbaum notes that it is increasingly difficult for many people to understand where the old norms end and new ones begin, as "default constraints on streams of information from us and about us seem to respond not to social, ethical, and political logic but to the logic of technical possibility: that is, whatever the Net allows." Yet there is some evidence that people do not require complete protection, and will gladly share even private information provided that certain social norms are met, similar to what Gerber reported in 2001 (Gerber 2001). There are three factors that affect these norms: actors (the information senders and recipients, or providers and users); attributes, especially types of information about the providers, including how these might be transformed or linked; and transmission principles (the constraints underlying the information flows).

When statistical agencies were the main collectors of data, they did so under very clear statutory authority with statutory protections. For example, Title 26 (Internal Revenue Service) and Title 13 (Census Bureau) of the US code provided

penalties for breaches of confidentiality, and agencies developed researcher access modalities in accordance with their statutory authorization. The statutory constraint on agencies such as IRS and Census makes it clear that the agencies, as data producers, should take "reasonable means" to protect data, although these reasonable means are not defined.

The statutory authorization for the new, technology-enabled collection of data is less clear. The Fourth Amendment to the Constitution, for example, constrains the government's power to "search" the citizenry's "persons, houses, papers, and effects." State privacy torts create liability for "intrusion upon seclusion." Yet the generation of big data often takes place in the open, or through commercial transactions with a business, and hence is not covered by either of these frameworks. There are major questions as to what is reasonably private, and what constitutes unwarranted intrusion (Strandburg 2014). Data generated by interacting with professionals, such as lawyers and doctors, or by online consumer transactions, are governed by laws requiring "informed consent" and draw on the Fair Information Practice Principles (FIPPs). Despite the FIPPs explicit application to "data," they are typically confined to personal information, and do not address the large-scale data collection issues that arise through location tracking and smart grid data (Strandburg 2014). *We need an international research agenda to examine important issues such as who should be asked to provide informed consent for the use of big data. Is it possible to transform the responsibility for data access from the person (informed consent) to the data producer or disseminator ("responsible use")? How much does it matter who gets and uses the data, what kind of data are provided and how access is provided?*

The existing statutory frameworks are increasingly irrelevant and potentially harmful for future research. We need to develop and advocate for a consistent legal framework for the collection and subsequent use of big data on human beings.

A Statistical Research Agenda

An eloquent description of statistical confidentiality is "the stewardship of data to be used for statistical purposes" (Duncan/Elliot/Salazar-González 2011). Statistical agencies have been at the forefront of developing that stewardship community in a number of ways. First, on the job, training is provided to statistical agency employees. Second, in the United States, academic programs such as the Joint Program on Survey Methodology[1], communities such as the Federal Committee on Statistical Methodology[2], and resources such as the Committee on National Statistics[3] have been largely supported by the federal statistical community. But the focus is almost exclusively on developing methodologies to improve the analytical use of survey data, and to a lesser extent, administrative data. Nothing similar exists to train scientists in developing an understanding of such issues as identifying the relevant population and linkage methodologies.

The risk of reidentifying individuals in a micro-dataset is intuitively obvious. Indeed, one way to formally measure the reidentification risk associated with a particular file is to measure the likelihood that a record can be matched to a master file (Winkler 2005). If the data include direct identifiers, like names, social security numbers, establishment ID numbers, the risk is quite high. However, even access to close identifiers, such as physical addresses and IP addresses can be problematic. Indeed, HIPAA regulations under the Privacy Rule of 2003 require the removal of 18 different types of identifiers including other less obvious identifiers such as birth date, vehicle serial numbers, URLs, and voice prints. However, even seemingly innocuous information makes it relatively straightforward to reidentify individuals, for example, by finding a record with sufficient information such that there is only one person in the relevant population with that set of characteristics: the risk of re-identification has been increasing due to the increased public availability of identified data and rapid advances in the technology of linking files (Dwork 2011). With many variables, everyone is a population unique. Since big data have wide-ranging coverage, one cannot rely on protection from sampling (Karr and Reiter 2014). Indeed, as Ohm points out, a person with knowledge of an individual's zip code, birthdate and sex can reidentify more than 80% of Netflix users, yet none of those are typically classified as personally identifiable information (Ohm 2010).

1 www.jpsm.umd.edu/jpsm/
2 www.fcsm.gov/
3 www7.nationalacademies.org/cnstat/

These observations highlight a broader conceptual challenge. Big data have created new practices that have radically disrupted information flows including privacy issues (Baracas and Nissenbaum 2014). In particular, it may no longer make sense to protect specific fields, such as name information, or insist on anonymity to avoid the ethical concerns raised by the big data paradigm. In practice, because records are now so comprehensive, they subvert the very meaning of anonymity. Insights drawn from big data can furnish additional facts about individuals without any knowledge of their specific identity. Indeed, as Baracas and Nissenbaum point out, in the commercial world, name information is too noisy to be useful – all the information needed to identify individuals is derived from other sources.

Developing different approaches

There are new approaches that are being developed in response to privacy concerns. There is a great deal of research that can be used to inform the development of such a structure, but it has been substantially siloed into separate activities in different research areas - statistics, cyber security, cryptography4 – as well as a variety of different practical applications, including the successful development of secure remote access data enclaves. There has also been a great deal of research on the features of reproducible science, particularly in the computer science and legal community.

A research agenda can be built that draws on encryption approaches such as differential privacy, incremental privacy or homomorphic encryption (Dwork 2014). Alternatively, or in parallel, the agenda could examine the value of establishing boundary organizations, like privacy markets, data banking, bonding/liability systems to manage privacy on behalf of individuals (Willbanks 2014). One intriguing approach, which applies an approach of "radical honesty" towards data contribution, acknowledges upfront the tension between anonymization and utility, and the difficulty of true de-identification. It provides a commons-based framework for reuse: it attempts to recruit individuals who understand the risks and uncertainties of making their data available for reuse. The attractive feature of such an approach is the goal to create reusable and redistributable open data, and it leverages cloud-based systems to facilitate storage, collaborative reuse, and analysis of data. These frameworks include "open consent", "portable consent" and interoperable consent (Willbanks 2014).

4 See, for example, www.lrdc.pitt.edu/schunn/cdi2009/home.html

A parallel research agenda could be built around new technical approaches such as the development of institutional controls that provide users with more control over their data, and permit large scale interoperability for data sharing between and among institutions. These controls should include responsive rules-based systems of governance and fine-grained authorizations for distributed rights management (Pentland et al. 2014). Alternatively examine the potential to institute access control and information flow policies, use media encryption, attribute based encryption or secure multiparty computation (Landwehr 2014).

The removal of key variables as "PII" is no longer sufficient to protect data against reidentification. A proposed research agenda would be to develop new technical approaches to ensure privacy. Another would be to develop markets for privacy rather than leaving privacy protection to non-experts, and create a different compact between researchers and study units in place by creating a cohort of individuals who are willing to serve as guinea pigs.

Conclusion

Big data does not only change the way in which data are collected and generated, but should change the way in which we think about statistical inference and the way in which we think about privacy. If we are to be serious about answering important social questions, and use these new types of data for the public good, we must solve the access problems. The questions are too hard to be solved by an individual researcher and by constructing one-off, unreplicable datasets. We need transnational datasets and transnational access.

We need a transnational approach that combines forces to develop new models of confidentiality protection, to address the legal challenges and to identify a broad range of solutions.

References

Baracas, S. and Nissenbaum, H. (2014): The Limits of Anonymity and Consent in the Big Data Age. In: Lane et al. (2014).

Cecil, J. and Eden, D. (2003): The Legal Foundations of Confidentiality. In Key Issues in Confidentiality Research: Results of an NSF workshop. National Science Foundation.

Dwork, Cynthia (2011): A Firm Foundation for Private Data Analysis. CACM.

Doyle, P./Lane, J./Theeuwes, J. and Zayatz, L. (2001): Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies. Amsterdam: Elsevier.

Duncan, G. T./Elliot, M. and Salazar-González, J. J. (2011): Statistical Confidentiality: Principles and Practice. Springer. books.google.com/books?id=7CPW8-oooX0C

Dwork, C. (2014): Differential Privacy: A Cryptographic Approach to Private Data Analysis. In: Lane et al. (2014).

Elias, P. and Entwisle, B. (2012): Changing Science: New Data for Understanding the Human Condition. Paris.

Gerber, E. (2001): The Privacy Context of Survey Response: An Ethnographic Account. In: Doyle, P./Lane, J./Theeuwes, J. and Zayatz, L. (2001): Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies. Amsterdam: Elsevier.

Karr, A. and Reiter, J. P. (2014): Analytical Frameworks for Data Release: A Statistical View. In: Lane et al. (2014).

Landwehr, C. (2014): The Operational Framework: Engineered Controls. In: Lane et al. (2014).

Lane, J./Stodden, V./Bender, S. and Nissenbaum, H. (Eds.): Privacy, Big Data, and the Public Good: Frameworks for Engagement. Cambridge University Press.

Nissenbaum, H. (2011): A Contextual Approach to Privacy Online. Daedalus; Journal of the American Academy of Arts & Sciences.

Ohm, P. (2010): Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization. UCLA Law Review.

Pentland, A./Greenwood, D./Sweatt, B./Stopczynski, A. and De Montjoye, Y.-A. (2014): The Operational Framework: Institutional Controls. In: Lane et al. (2014).

Stigler, S. (1988): The History of Statistics. Harvard University Press.

Strandburg, K. J. (2014): Plain View, Third Parties, Surveillance, and Consent: Legal Approaches to Data making in the Big Data Context. In: Lane et al. (2014).

Willbanks, J. (2014): The Analytical Framework: Portable approaches to informed consent and open data. In: Lane et al. (2014).

Winkler, W. (2005): Re-identification Methods for Evaluating the Confidentiality of Analytically Valid Microdata. U.S. Census Bureau.

8 A Common Agenda for the European Research Infrastructures in the Social Sciences and Humanities

Adrian Duşa (SCI-SWG), Claudia Oellers, and Simon Wolff (RatSWD)

Research in the social sciences and the humanities (SSH) in Europe is currently facing a historic turning point. The social sciences and the humanities have been included in the European Union's new framework program for research and innovation Horizon 2020 and embedded across the societal challenges the program seeks to address. The integration of both disciplines is a sign for the increasing recognition and the essential role SSH plays in addressing the societal challenges and the great number of pressing issues Europe faces today and in the near future. These include employment, demographic change and ageing populations, migration, poverty, climate change, food and energy security, European cohesion and cultural diversity.

The pan-European research infrastructures (such as the ESFRI SSH infrastructures CESSDA, CLARIN, DARIAH, ESS, SHARE) were a necessary precondition for SSH to take on this important role. Today, these infrastructures provide the necessary means for innovative research. The existing wealth and diversity of European research infrastructures and the complex issues they address is reflected in this book's contributions.

The "Facing the Future" conference was held to reflect upon what has been achieved, and what the requirements for the European SSH research infrastructures will be in the future, not only to meet the ambitious challenges set out by Horizon 2020, but also to stay internationally competitive and to strengthen pan-European research initiatives in the long-run. Moreover, it was an important aim of the conference to identify present and future challenges for European research infrastructures from the perspective of their users, the SSH research community, and to formulate a common agenda, for both the social sciences and the humanities with regard to the advancement of research infrastructures at the European level.

The conference participants from the social sciences and humanities (researchers, policy makers, funding agencies, scientific research infrastructure coordinators) were able to find much agreement and common ground regarding future challenges and infrastructure needs. In the course of the conference, five essential challenges were identified as key towards strengthening European infrastructures and thus ensuring that the social sciences and humanities will be able to fulfill their important role in the future. 1. Ensuring sustainability and establishing permanent/sustainable institutions. 2. Facilitating research cooperation and interdisciplinarity. 3. Tapping new sources of (big) data. 4. Safeguarding data protection at all levels of research. 5. Increasing the visibility of SSH research infrastructures in their respective fields and for the general public. Clearly, these challenges are inherently interrelated and the following synopsis cannot consider every aspect in full detail.

1. Ensuring Sustainability of Research Infrastructures in the Social Sciences and Humanities

Across the board, sustainability of research infrastructures was considered a crucial issue. In the current European framework, sustainability of research infrastructures is at risk due to the lack of sufficient and long-term funding (funding only for one survey wave, or funding that only covers implementation/set-up costs) as well as the lack of cross-national cooperation, and the fragmentation between the European member states (diverging time frames and funding rounds, differences in national science policy). The introduction of the ERIC legal framework created the basis for pan-European projects; however, it did not install a sustainable European funding scheme. Solutions need to ensure sustainability of European research infrastructures by establishing them as permanent institutions at the European level.

While there are differences between infrastructures, some focusing more on providing data, while others focus more on providing tools and services, the common challenge is turning them into sustainable institutions. By definition, research infrastructures in SSH must be designed to be durable, stable and reliable. At the same time, they operate at the forefront of scientific innovation and must constantly evolve and adapt to technological progress, advances in methodology, and, most importantly, the needs of their users in science and research. In this dialectic of continuity and innovation, as Peter Farago puts it in his article, research infrastructures must continue to find the right balance.

Social science research infrastructures that provide primary research data, for example major, cross-national and longitudinal surveys, rely on broad participation and long-term funding to achieve their full potential of comparability over time and across national borders. While the initial costs for setting up these infrastructures are relatively small, they have to be consistent in the long-run. Large-scale social surveys have become invaluable sources for monitoring social change which can only fulfill their function when built on a foundation of valid, complete and comparable data. Similarly, the preservation and long-term archiving of Europe's cultural heritage in Europe and the provision of these high-quality data to research form the backbone of humanities research that requires serious long-term commitment. Long-term preservation of European cultural heritage and the leading role of Europe in the humanities will depend on the capacity to continue to build and sustain these infrastructures (see Niccolucci in this book) over time.

The experiences of the Survey of Health, Ageing and Retirement in Europe (SHARE) project (see Börsch-Supan in this book) illustrate the problems of fragmented funding schemes for pan-European projects and the dropping-out of participants due to the lack of funding at the national level. Ageing is a very different process in various European regions, and research of this issue demands high quality data from all those regions. Typically, it is those countries in which research on political, economic and social changes is particularly urgent that are not included in the European research effort. Failure of European projects or individual countries dropping out due to the fragmentation and lack of funding is a crucial issue that has to be prevented in the future in order to achieve a real European added value for these research infrastructures.

Moreover, there is a strong need for scientific leadership beyond single committed individuals in order to be able to advance research infrastructures. In a funding environment where resources are allocated according to competition between various stakeholders, public funding institutions require strong partners in the scientific community who represent their constituency and formulate needs and requirements.

2. Facilitating research cooperation and interdisciplinarity

Interdisciplinarity is essential to meet the complex challenges of the 21st century such as labor markets, ageing populations, climate change, food and energy security, or cities and well-being. Such cutting-edge research will increase the need for interdisciplinary research by default in order to achieve robust results because they transcend academic disciplines as well as national borders. Addressing the "real life problems" which make up the societal challenges described by Horizon 2020 requires an integrated, interdisciplinary approach between very diverse scientific domains.

Research infrastructures facilitate and foster interdisciplinarity by design, promoting cooperation and harmonization as well as strengthening cross-national projects. An essential prerequisite to making interdisciplinarity work is the comparability of data over time, across national borders, and between scientific disciplines. Comparative analysis lies at the heart of the European infrastructures in the social sciences. Research infrastructures have been successfully working on increasing comparability between regions and countries through common methodological frameworks, international harmonization platforms and international data portals. But there is still much effort required to enhance interdisciplinary comparability of data on a trans-European level.

There is unanimous agreement on the benefits of cooperation. However, it remains a huge challenge in light of the extreme heterogeneity of the existing research infrastructures. Most research infrastructures are established in response to a demand for specific data on a national level and are not coordinated with other infrastructures from the onset. More coordination between research infrastructures is needed at all levels – not least to avoid a loss of efficiency and a waste of resources. In order to fully exploit the potential of this diversity, by creating synergies between infrastructures instead of doubling institutions, coordination among them has to be improved at the national and international level. Some efforts have started with the DASISH project, funded by the European Commission through the FP7, which seeks to reach compatibility between the five projects of the ESFRI roadmap for the Social and Cultural Innovation area, in order to find common solutions to common problems (see also Marker in this book).

The project Data without Boundaries (see Bender in this book) has been successful in by coordinating existing infrastructures, the Council of European Social Science Data Archives (CESSDA) and the European Statistical System (ESS), and is making access to official research microdata easier than ever by overcoming the need for multiple accreditations when requesting access to comparable datasets.

Another pervasive challenge in this area is establishing common standards between infrastructures and disciplines. A prerequisite for linking different kinds of data are common, harmonised standards for data documentation. In the humanities, especially, there is still much room for improvement regarding standardization. In the future, these efforts will have to be extended further to facilitate linking different types of research data from different sources and disciplines.

3. Big Data: Tapping New Data Sources for Research

New forms and sources of data are emerging everywhere in the Digital Age. Today, wittingly or unwittingly, individuals produce vast amounts of data just by being online, going shopping, or using a mobile phone; personal data are shared in social networks such as Facebook or Twitter. The quantity as well as the quality of these new data sources is challenge in itself.

However, while not all of these data are scientifically significant, there are many hidden sources of data that could potentially have enormous value for innovative social research and which previously could not have been collected or were not available (see Maynard and Strohmaier/Zens in this book). The potential of these new data sources is higher when they are linked to ("traditional") survey data. Linking data from various sources to other data, such as private sector data (commercial data, tracking data, and satellite imagery), internet data (social media) or biological data (mental/physical measures, biomarkers, genomics), will be of growing importance and has to be facilitated by infrastructures in the future. This will also require science to increasingly turn to partnerships with private actors for gaining access to data that is not produced by science itself. Issues that have to be addressed are how to provide access to this data under simultaneous consideration of the legal issues related to personal information contained in the data, the question of consent for reuse and the safeguarding of data protection, quality control for unprepared and undocumented data, replicability and durability (see Marker in this book for detailed requirements).

On the whole, it will become more important to establish access to relevant new data sources for research. Where specific and relevant research data does not yet exist, it is important to fill in the gaps. However, efforts should be continued to open up and connect already existing data repositories to research infrastructures. An important step would be to set up a central, international (European) centre for cross-disciplinary research on new forms of data and the establishment of European data service centres. Especially in the UK, pioneering progress has been made in making a wealth of administrative data available to research which could serve as a model to other countries. Administrative data have an enormous potential for social science research because they cover a majority of the population in great detail, are already prepared in many ways for scientific use, and are relatively up-to-date (see Woollard and Bega in this book).

4. Data Protection, Confidentiality, and Research Ethics

One of the most pervasive challenges to research, particularly in the social sciences, is safeguarding data protection at all levels of research without obstructing innovative research. The different regulations of data protection standards on a European level remain a challenge for cross-national cooperation. The planned general European data protection regulation will be a step towards harmonizing national standards in the EU.

In the age of big data and increasing voluntary disclosure of private information, researchers have to reflect on the changing nature of privacy and confidentiality. Who do you ask for consent when tapping sources of big data? As mentioned in the last paragraph dealing with new data sources, data protection in this changing environment will require an international research agenda to reflect on these issues and find new ways of dealing with a new data situation (see Lane et al.).

In the future, it will be essential to find the right balance between data acquisition and data protection which is always a question of finding the right balance. Research infrastructures will continue to play an important role in establishing best practice of data protection and research ethics.

5. Promoting Research Infrastructures and Increasing Visibility

Among the greatest challenges to the SSH research infrastructure community is to increase the visibility of research infrastructures, integrating them further into the daily work of researchers, and enhancing the way their benefits are showcased towards the general public.

First, visibility of research infrastructures has to be further improved by promoting their benefits to their potential users, the researchers in the social sciences and humanities, and to encourage sharing, using, and finding research data by using existing infrastructures. The notion of "one researcher, one project, one dataset" is gradually being superseded by a culture of sharing, cooperation and re-use of data, but there is still much work to be done. While there has been considerable progress in promoting a culture of data sharing in the social sciences, there is still a lack of knowledge and acceptance among researchers in the humanities. Especially in the latter, the development towards setting up infrastructures and the methodological transition to Digital Humanities has been met with some scepticism. However, the main reason for reluctance to engage in data sharing is the same for both disciplines: missing incentives for researchers for investing time and energy into preparing their data for secondary use. The hard work involved with producing, and also sharing, high-quality data is currently not being appropriately rewarded by journals, universities and funding agencies with professional credit. Researchers involved in setting up research infrastructures perform a service of great value to the entire scientific community and should not fall behind the reputation of colleagues due to this important commitment. Technical difficulties, however, are also an obstacle to sharing data. This should be facilitated by making data management and documentation as easy and user-friendly as possible. The same applies for research tools which should be designed in close cooperation with their users in research to ensure usability. Further developing international standards for documentation and the use of persistent identifiers will also lead to greater recognition of data production as an important scientific achievement of its own.

However, a simple lack of knowledge about what is being offered is also often a reason why researchers do not make use of the resources research infrastructures provide. These have to be better integrated into the daily work of researchers. To this end, finding suitable data for a research project should be made as easy as finding other information on the web. This need for efficient

and easy-to-use tools needs to be addressed in the near future. Moreover, data portals should be organised in a more centralized fashion in order to avoid searching across multiple sources.

The second aspect of visibility is that SSH research infrastructures need to be more visible towards funders and policy-makers. The integration of the social sciences and the humanities into the framework of Horizon 2020 is a great step forward, but, more generally, SSH are usually not deemed as essential as the natural sciences. Moreover, SSH will soon be required to produce concrete data on the impact of publicly funded research infrastructures in view of future project assessments, as Žic Fuchs points out in her contribution to this book. Gathering detailed information on the added value produced by research infrastructures, while necessary, might also enable SSH to showcase their success on the basis of specific information.

The third visibility challenge is to promote the benefits of research infrastructures beyond the field of science. Towards the greater public, the natural sciences have been very successful in justifying enormous investments in e.g. the Large Hadron Collider or space programs by showcasing their merits and touching upon people's imagination. Without doubt, the social sciences and the humanities are producing research results that are equally relevant to many, but ways have to be found to better communicate and showcase their results to the public. Since most research is publicly funded, citizen participation is essential. Especially in the social sciences, where vast amounts of personal information from citizens are gathered for research, it will be increasingly necessary to make the public aware of how and to what extent personal data are gathered, how the data is processed and put to use, and, ultimately, why this will advance progress in science and society.

Not least due to the European-wide establishment of research infrastructures, the European social sciences and humanities are world-leading and will continue to play a crucial role in analyzing the societal challenges facing Europe in the future. Their capacity to do so will depend on the capacity of the scientific community, research funders, and European policy makers to find good and innovative ways to meet the challenges for European research infrastructures in the future.

Authors

Bega, Daniele	HM Revenue and Customs, UK
Bender, Stefan	German Institute for Employment Research (IAB)
Börsch-Supan, Axel	Munich Center for the Economics of Aging at the Max-Planck-Institute for Social Law and Social Policy
Burghardt, Anja	German Institute for Employment Research (IAB)
Collins, Sandra	Digital Repository of Ireland; ALLEA Working Group "E-Humanities"
Dacos, Marin	Centre National de la Recherche Scientifique (CNRS), Aix-Marseille University: Cléo UMS 3287
Dubucs, Jacques	Social Science and Innovation – Strategic Working Group (SCI-SWG); French Ministry of Higher Education and Research
Duşa, Adrian	Social Science and Innovation – Strategic Working Group (SCI-SWG); University of Bucharest; Romanian Social Data Archive (RODA)
Elias, Peter	University of Warwick, UK; Social Science and Innovation – Strategic Working Group (SCI-SWG)
Emery, Tom	Netherlands Interdisciplinary Demographic Institute (NIDI); Generations & Gender Programme (GGP)
Farago, Peter	Swiss Centre of Expertise in the Social Sciences (FORS), Lausanne
Gauthier, Anne	Netherlands Interdisciplinary Demographic Institute (NIDI); Generations & Gender Programme (GGP)
Hobcraft, John	University of York, UK; Social Science and Innovation – Strategic Working Group (SCI-SWG)
Lane, Julia	American Institutes for Research, US
Lanoë, Jean-Louis	French Longitudinal Study of Children (Elfe)
Lauer, Gerhard	University of Göttingen, Centre for Digital Humanities, DE; Union of German Academies of Sciences and Humanities; ALLEA Working Group "E-Humanities"
Leathem, Camilla	The Union of the German Academies of Sciences and Humanities

Marker, Hans Jørgen	Swedish National Data Service (SND); Council of European Social Science Data Archives (CESSDA); Data Service Infrastructure for the Social Sciences and Humanities (DASISH)
Maynard, Diana	University of Sheffield, UK
Nelle, Dietrich	German Federal Ministry of Education and Research (BMBF)
Niccolucci, Franco	VAST-Laboratory, PIN-University of Florence, IT
Nissenbaum, Helen	New York University, US
Oellers, Claudia	German Data Forum (RatSWD), Business Office
Romary, Laurent	French Institute for Research in Computer Science and Control (inria); Digital Research Infrastructure for the Arts and Humanities (DARIAH)
Sarkar, Ranjana	Project Management Agency at the German Aerospace Centre (DLR-PT)
Schiller, David	German Institute for Employment Research (IAB)
Stock, Günter	ALLEA; Berlin-Brandenburg Academy of Sciences and Humanities; Union of German Academies of Sciences and Humanities
Stodden, Victoria	Stanford University, US
Strohmaier, Markus	GESIS – Leibniz Institute for the Social Sciences, DE
Vendrix, Philippe	Centre National de la Recherche Scientifique (CNRS), Université de Tours: CESR UMR7323
Wagner, Gert G.	German Data Forum (RatSWD); Berlin University of Technology; German Institute for Economic Research (DIW Berlin); Max Planck Institute for Human Development, Berlin
Wolff, Simon	German Data Forum (RatSWD), Business Office
Woollard, Matthew	UK Data Archive; University of Essex, UK
Zens, Maria	GESIS – Leibniz Institute for the Social Sciences, DE
Žic Fuchs, Milena	University of Zagreb, Croatian Academy of Science and Arts

List of Abbreviations

ALLEA	European Federation of Academies of Sciences and Humanities
ARIADNE	Advanced Research Infrastructure for Archaeological Dataset Networking in Europe
BMBF	German Federal Ministry of Education and Research
CENDARI	Collaborative European Digital Archive Infrastructure
CESR	Centre for Advanced Renaissance Studies , FR
CESSDA	Council of European Social Science Data Archives
CHARISMA	Cultural Heritage Advanced Research Infrastructures
CLARIN	Common Language Resources and Technology Infrastructure
Cléo	Centre for Open Electronic Publishing
CNRS	French National Centre for Scientific Research
DARIAH	Digital Research Infrastructure for the Arts and Humanities
DASISH	Data Service Infrastructure for the Social Sciences and Humanities
DIW Berlin	German Institute for Economic Research
DwB	Data without Boundaries
EHRI	European Holocaust Research Infrastructure
Elfe	French Longitudinal Study of Children
ERIC	European Research Infrastructure Consortium
ESF	European Science Foundation
ESFRI	European Strategy Forum on Research Infrastructures
ESRC	Economic and Social Research Council, UK
ESS	European Social Survey
FORS	Swiss Centre of Expertise in the Social Sciences

FP7	Seventh Framework Programme for Research of the European Union
GESIS	German Social Science Infrastructure Services – Leibniz Institute for the Social Sciences, DE
GGP	Generations and Gender Programme
HMRC	Her Majesty's Revenue and Customs, UK
Horizon 2020	Eighth phase of the Framework Programmes for Research and Technological Development of the European Union
IAB	German Institute for Employment Research
InGRID	Inclusive Growth Research Infrastructure Diffusion
Inria	French Institute for Research in Computer Science and Control
MERIL	Mapping the European Research Infrastructure Landscape
NIDI	Netherlands Interdisciplinary Demographic Institute,
NSD	Norwegian Social Science Data Services
RatSWD	German Data Forum
RI	Research Infrastructure
RODA	Romanian Social Data Archive
SCI-SWG	Social and Cultural Innovation Strategy – Working Group of ESFRI
SHARE	Survey of Health, Ageing and Retirement in Europe
SND	Swedish National Data Service