IASSIST Quarterly

VOLUME 35 – Number 3 – Fall 2011

Mind the Gap

ESDS

Research Data Centre

RDC-IN-RDC

AddressingHistory

EDINA

iassist

IN THIS ISSUE

3 Papers from IASSIST Conferences

ON PAGE **5**

Editor Karsten Boye Rasmussen

ON PAGE **21**

Membership Information

Online at: iassistdata.org/iq

COLOPHON

IASSIST Quarterly

The IASSIST Quarterly represents an international cooperative effort on the part of individuals managing, operating, or using machine-readable data archives, data libraries, and data services. The QUARTERLY reports on activities related to the production, acquisition, preservation, processing, distribution, and use of machine-readable data carried out by its members and others in the international social science community. Your contributions and suggestions for topics of interest are welcomed. The views set forth by authors of articles contained in this publication are not necessarily those of IASSIST.

Information for Authors

The Quarterly is normally published four times per year. Authors are encouraged to submit papers as word processing files (for further information see: http://www.iassistdata.org/iq/instructions-authors). Manuscripts should be sent to Editor: Karsten Boye Rasmussen.: Email: kbr@sam.sdu.dk

.Announcements of conferences, training sessions, or the like are welcomed and should include a mailing address and a telephone number for the director of the event or for the organization sponsoring the event

Editor

Karsten Boye Rasmussen
Department of Marketing & Management
University of Southern Denmark, SDU
Campusvej 55, DK-5230
Odense M, Denmark
Phone: +45 6550 2115
Email: kbr@sam.sdu.dk

Deputy editor

Walter Piovesan,
Simon Fraser University

Website editor

Stuart MacDonald.
Unviversity of Edinburg

In this issue

Title: Newsletter - International Association for Social Science Information Service and Technology
ISSN - United States: 0739-1137

Online @ http://www.iassistdata.org/iq

Editor's notes

Sharing data and building information
With this issue (volume 35-3, 2011) of the IASSIST Quarterly (IQ) we return to the regular format of a collection of articles not within the same specialist subject area as we have seen in recent special issues of IQ. Naturally the three articles presented here are related to the IQ subject area in general, as in: assisting research with data, acquiring data from research, and making good use of the user community. This last topic could also be spelled "involvement". The hope is that these articles will carry involvement to the IASSIST community, so that the gained knowledge can be shared and practised widely.

"Mind the gap" is a caveat to passengers on the London Underground. The authors of this article are Susan Noble, Celia Russell and Richard Wiseman, all affiliated with ESDS-International hosted by Mimas at the University of Manchester in the UK. The ESDS, standing for "Economic and Social Data Service", are extending their reach beyond the UK. In the article "Mind the Gap: Global Data Sharing" they are looking into how today's research on the important topics of climate change, economic crises, migration and health requires cross-national data sharing. Clearly these topics are international (e.g. the weather or air pollution does not stop at national borders), but the article discusses how existing barriers prevent global data sharing. The paper is based on a presentation in a session on "Sharing data: High Rewards, Formidable Barriers" at the IASSIST 2009 conference. It is demonstrated how even international data produced by intergovernmental organizations like the International Monetary Fund, the International Energy Agency, OECD, the United Nations and the World Bank are often only available with an expensive subscription, presented in complex incomprehensible tables, through special interfaces; such barriers are making the international use of the data difficult. Because of missing metadata standards it is difficult to evaluate the quality of the dataset and to search for and locate the data resources required. The paper highlights the development of e-learning materials that can raise awareness and ease access to international data. In this case the example is e-learning for the "United Nations Millennium Development Goals".

The second paper is also related to the sharing of data with an introduction to the international level. "The Research-Data-Centre in Research-Data-Centre Approach: A First Step Towards Decentralised International Data Sharing" is written by Stefan Bender and Jörg Heining from the Institute for Employment Research (IAB) in Nuremberg, Germany. In order to preserve the confidentiality of single entities, access to complete datasets is often restricted to monitored on-site analysis. Although off-site access is facilitated in other countries, Germany has relied on on-site security. However, an opportunity has been presented where Research Data Centre sites are placed at Statistical Offices around Germany, and also at a Michigan centre for demography. The article contains historical information on approaches and developments in other countries and has a special focus on the German solution. The project will gain experience in the complex balance between confidentiality and analysis, and the differences between national laws.

The paper by Stuart Macdonald from EDINA in Scotland originated as a poster session at the IASSIST 2010 conference. The name of the paper is "AddressingHistory: a Web2.0 community engagement tool and API". The community consists of members within and outside academia, as local history groups and genealogists are using the software to enhance and combine data from historical Scottish Post Office Directories with large-scale historical maps. The background and technical issues are presented in the paper, which also looks into issues and perspectives of user generated content. The "crowdsourcing" tool did successfully generate engagement and there are plans for further development, such as upload and attachment of photos of people, buildings, and landmarks to enrich the collection.

Articles for the IQ are always very welcome. They can be papers from IASSIST conferences or other conferences and workshops, from local presentations or papers especially written for the IQ. If you don't have anything to offer right now, then please prepare yourself for the next IASSIST conference and start planning for participation in a session there. Chairing a conference session with the purpose of aggregating and integrating papers for a special issue IQ is much appreciated as the information in the form of an IQ issue reaches many more people than the session participants and will be readily available on the IASSIST website at http://www.iassistdata.org.

Authors are very welcome to take a look at the instructions and layout:
http://iassistdata.org/iq/instructions-authors

Authors can also contact me via e-mail: kbr@sam.sdu.dk. Should you be interested in compiling a special issue for the IQ as guest editor(s) I will also be delighted to hear from you.

Karsten Boye Rasmussen
December 2011
Editor

Mind the Gap: Global Data Sharing

by Susan Noble, Celia Russell and Richard Wiseman[1]

ESDS

Abstract

In an increasingly globalised world, the importance of collaborative research into international issues, such as climate change, economic crises, migration and health cannot be underestimated. This type of research requires data sharing at a cross-national level. In this paper we discuss the barriers which prevent data sharing on a global scale and the need to address the lack of data awareness. We will explain the licensing issues which restrict the use of data in the ESDS International portfolio by the UK academic community but will also highlight ways in which the ESDS International service helps the non-UK international data research community by identifying free data sources and producing freely available supporting documentation and other materials.

The paper will highlight the development of e-learning materials based upon a key socio-economic theme: 'the UN's Millennium Development Goals (MDGs)'. As the current international development planning framework, the eight MDGs are time-bound and quantifiable targets for reducing poverty, improving health and protecting the environment. Intended to raise awareness of the potential use of quantitative international data in research and teaching, additionally, these open access e-learning materials have been developed as an interactive online set of materials to be used as a self-teaching resource.

Keywords: international data, social sciences, macro-economic databanks, ESDS International, Millennium Development Goals, e-learning.

INTRODUCTION

The Organisation for Economic Cooperation and Development (OECD) describe climate change as humankind's central long-term challenge[2] . Addressing it successfully requires the combined efforts of researchers and national governments worldwide. Similarly, the international banking crisis cannot be tackled from a single country perspective alone. These types of transnational problems call for multilateral, multilevel responses combining international and national policy-making supported by evidence-based research. So how can data specialists best support research communities through international data sharing? And how do we aid the development of the statistical capacities required to inform policy-making on a global scale?

The international data which feeds into this type of social science research is produced by intergovernmental organisations (IGOs) such as the International Monetary Fund, the International Energy Agency, OECD, the United Nations and the World Bank. These organizations have a presence in every country in the world, the authority to create international standards and the technical and financial capacity to support the development of national statistical infrastructures. They have long produced high quality, regularly updated time series databanks for their own internal use which typically contain a huge range of macro-economic and social indicators aggregated to national or regional level and collectively cover virtually every country in the world. The academic research community needs access to these unique datasets in order to contribute to and comment on policy responses to global issues. Furthermore, access to these data resources give students the opportunity to work with real world data.

Since the creation of ESDS International no longer is access a barrier ...

ESDS International

In the United Kingdom, ESDS International, a specialist data service of the wider Economic and Social Data Service (ESDS) was established in January 2003 to address the issue of international data access for the UK academic community. Hosted by Mimas at the University of Manchester, the service helps researchers to locate and acquire international survey datasets such as the Eurobarometer and World Values Survey and provides free access to over 35 socio-economic macro datasets from nine different data providers via a common web-based user interface.

Beyond 20/20 Web Data Server, a web-based data dissemination tool provides the common user interface to the ESDS International macro international data. It requires only a standard web browser, is accessibility compliant and can be used to display, subset, visualise and download time series data. Each of the IGO's provides large and complex data files in various formats and the ESDS International developers convert these data files into a format suitable for delivery via Beyond 20/20.

Each of the datasets is accompanied by a differing quantity and quality of explanatory documentation which ESES International collate and re-format into standard supporting documentation, i.e. dataset user guides which are consistent across each of the datasets available in the portfolio irrespective of the IGO data provider. This set of comprehensive supporting documentation includes details of the topics and time range covered, the countries included, the periodicity of the data and links to relevant documentation.

Free access to the service for the UK academic community has been made possible through a series of ground-breaking national data redistribution licensing agreements between each of the data providers and the University of Manchester (which hosts the service).

For the bulk of the licence agreements ESDS International employed the services of Databeuro[3] – a professional data negotiation agent which already had expertise working with IGOs for access to online data and e-books to the academic, government and corporate sectors. Where possible the data re-distribution agreement negotiated with each IGO was based on a model licence. The model licence adopted was very similar to one developed between the Joint Information System Committee and the Publishers Association for the licensing of commercial datasets[4] . It sets out generic terms and conditions of use (e.g. educational use only) but there was also scope for adding additional clauses to reflect special conditions required by different suppliers. As the licence agreement was between the University of Manchester and the IGO, it was necessary to ensure that the University was not exposed to any financial risk by taking on the role of licensing the data on behalf of the entire UK academic community. For this reason, appropriate clauses relating to limitation of liability and dispute resolution were inserted to protect the University of Manchester.

These UK wide redistribution agreements deliver significant savings to the academic community as institutional or individual subscriptions are no longer required and in some instances additional discount was negotiated by agreeing to a five year deail with a single up-front payment in year one. In addition they remove one of the major barriers to use of these datasets in research and teaching by allowing all members of the UK academic community, irrespective of their institution, to access the data through ESDS International, free at the point of use.

The unique licensing arrangements have certainly addressed the issue of data access for the UK academic community, as the service has witnessed an astonishing growth in the UK international research community from a few hundred users at a handful of institutions in 2003 to over 30,000 users representing over 200 different institutions across the UK as of November 2010. However, there is still an issue concerning data access elsewhere in the world, and there the barriers which prevent global data sharing still exist. ESDS International is frequently contacted by non-academic organisations and non-UK academics to request data access or provide support and further information (62 queries logged to our helpdesk between January 2009 and November 2010). In general response to this apparent gap in data provision ESDS International now:

- Provides helpdesk support where possible – e.g. will point user to an alternative source of free data if available – for example, will advise a user from a developing world who to contact to obtain the data.
- Produces comprehensive user guides for each dataset which include details of dataset topics, as well as its geographical and temporal coverage and makes them available to anyone from the ESDS International website.
- Regularly updates a guide to freely available international data and resources.

Since the creation of ESDS International, no longer is access a barrier for instructional data use in teaching, but rather there is a lack of awareness of data resources and their potential for use in a teaching environment. User consultation on this issue identified the following key points regarding the relatively low use of international data in teaching:

A lack of awareness of the potential use of ESDS data for teaching was highlighted. It was felt that the use of international data at an institution was very dependent on pro-active data librarians/ tutors seeking out information about what data is available and promoting it within their institution or department. Librarians appear to play a key role in assisting academic users to identify and access data for research and teaching purposes and should be targeted with information about international datasets.

A lack of teaching materials based on international data – so even when the service is well promoted it is primarily used for post-graduate research and tutors are unsure of how to use the data in a teaching environment. Tutors felt the existence of teaching materials would encourage more extensive instructional use of this type of data. Suggested teaching materials included thematic case studies and teaching datasets (to allow demonstrations of particular estimation and inference methods), and some supporting guidelines and notes to explain how the data might be used. Suggested themes included 'Science, technology and innovation', the 'United Nations (UN) Millennium Development Goals (MDGs)', and 'Environment and economy'.

Access problems were highlighted by a number of tutors - some commented that the registration process was too complicated to get all their students registered prior to using the datasets in a lesson. Feedback about the provision of support services indicates that courses should be made available via the web, since face to face courses are not always convenient to attend due to timing and location.

In 2007, ESDS International started to address these issues with the launch its first major learning resource entitled, "Countries and Citizens".

This self-guided training resource, including online tutorials, activities, study guides and videos, is designed to show how to combine socio-economic data from country-level aggregate databanks (macro data) with individual-level survey datasets (micro data). This resource has proved extremely popular with an average 20,000 page views per year since its release.

The extraordinary level of usage combined with a growing demand for entry level statistical learning resources fuelled the decision to develop a new online teaching resource based on the MDGs. In addition, the development of a new learning package based on the MDGs aligns well with ESDS International's strategy to promote and encourage the use of international databanks in teaching and research and helps to fulfil its responsibility to raise statistical capacities.

The UN Millennium Development Goal e-learning materials

The UN's MDGs offer a wealth of freely available, cross-national, policy relevant data. The goals themselves have formed the policy framework for international development through public investment, capacity building, domestic resource mobilization and the targeting of official development assistance. Progress towards each goal is measured by a series of quantifiable indicators. For example, *Millennium Development Goal 5: Improve maternal health* is measured by the following quantifiable indicators for each country:

- Maternal mortality ratio
- Proportion of births attended by skilled health personnel
- Contraceptive prevalence rate
- Adolescent birth rate
- Antenatal care coverage (at least one visit and at least four visits)
- Unmet need for family planning

While the MDG data is freely available, there are still many potential data users who are simply unaware of how this global data resource could better inform their research and teaching. In order to rectify this issue the ESDS International service has developed a suite of e-learning materials based upon the socio-economic theme of the UN's MDGs. The suite has been designed as an interactive online set of materials to be used as a self-teaching resource intended to raise awareness of international data resources and help students develop their own data handling skills using real world data. The online materials are targeted at postgraduates and undergraduates mirroring the current user profile of the ESDS International service which counts almost 75% of its registered users as postgraduate or undergraduate students.

The learning package is available as an open access resource via the ESDS International data service. This fact meant the course had to cater to users with differing levels of expertise in statistical analysis techniques and/or awareness of international data resources.

Implementation

A number of possible implementations to satisfy the requirements for the MDGs learning package were considered. For example, we looked into developing a blended approach (i.e., face–to-face and virtual elements) but decided we did not have the resources to dedicate to this. In addition, it was thought that there would be some key benefits in developing an entirely online e-learning package, for example, it would be available to any institution at any time and users can pick and choose elements from the package for inclusion in their specific learning environment.

The e-learning package was developed as web-based materials rather than within a Virtual Learning Environment (VLE) such as Blackboard, as ESDS International is a national data service and, as such, must ensure its resources are available to users from any UK institution no matter what VLE their institution supports.
Consultation with a number of key data users who use international data in the classroom contributed to the identification and development of the e-learning package contents. The content was produced by Susan Noble and Celia Russell, Richard Wiseman developed the e-tutorials based on Camtasia Studio 5. Susan Noble also developed the framework and programming for the resource.

The MDG e-Learning Resource

The e-learning materials have been designed primarily to help learners explore the United Nation Millennium Development Goals. In addition, they guide participants through using the ESDS International data service and provide learners with an increased understanding of data availability and potential data use for research and teaching. The materials comprise four main sections:

1. Guide to ESDS International - provides a comprehensive overview of the international data service including the topics, time range, frequency and geographical coverage of the data provided.
2. Guide to the MDGs - describes the background to the MDGs, the relationship between the goals, targets and indicators, identifies the data sources and how progress towards the goals is measured.
3. Activities section - provides step-by-step activities enabling learners to, for example, search the ESDS International data portfolio or carry out a simple data visualisation task. All activities are available as e-Tutorials.
4. Links and resources - contains links to relevant UN MDG related interfaces such as UN MDG Gapminder and a CommonGIS interface to a subset of MDG data.

The resource is based on a blend of instructional and constructivist elements, allowing students to discover the content by a variety of means. For example, in the guides, the target material is structured into a formal hierarchy, enabling the learner to work through sequentially. Alternatively, students can tackle real tasks using the data by attempting related activities at the end of the section.

One particularly successful element of the e-learning materials is the e-tutorials. These are video screen capture demonstrations built using Camtasia Studio 5. Five e-tutorials were developed collectively demonstrating how to access, search, source, visualize and cite cross-national data.

The resource uses real-life examples drawn from the MDG indicators as a basis for the interactive activities. For example, in the data access e-tutorial, students look for data that supports two of the indicators for MDG Goal 5: improve maternal health. By basing the activity on authentic data gathered for a genuine purpose (maternal deaths represent the greatest indicator of inequity between rich and poor women) students discover for themselves the worrying fact that in many countries there has been little improvement in maternal death rates over the past two decades. These hands-on activities, which also explore poverty and hunger (Goal 1) and women in parliament (Goal 3), are key constructivist elements of the resource. As they move through the activities, students learn how to access, visualise and cite real MDG data by taking part in these authentic tasks. They can also discover how to take action in support of the MDGs by exploring the 'What you can do?' area of the Links and Resources section.

Future work for these materials could include the development of teaching datasets, including simple data analysis tasks. Further, the materials could be made available as a re-usable Content Package which can be downloaded into an institutional VLE.

Conclusions

Within this paper, we have discussed the barriers that prevent international data sharing, relevant data access and licensing issues and mentioned ways in which these barriers can be reduced.

The UK model for providing access to international data through national licensing agreements has shown to be a very effective way of increasing the use of cross-national datasets and is one that can be replicated elsewhere.

Traditionally most intergovernmental organisations make some of their data freely available, but large portions are only available at a cost. This trend is reversing. In recent years the United Nations and the International Labour Organisation have made data freely available. In April 2010, the World Bank publicly released all of their data. The removal of the key barrier of cost is, of course, beneficial, as a larger amount of people will be able to access the data and indeed the World Bank has highlighted this increased opportunity to turn data into knowledge[5].

We have also introduced e-learning materials based upon the UN's Millennium Development Goals, and highlighted how this approach reduces the gap between the richness of available data resources and their uptake. A central aspect of the new resource is that it is completely based on genuine data collected to monitor progress towards international development. Real world data is often messy, in particular from the developing world where "exceptions are the rule"[6] . For this reason in many teaching environments data is pre-fabricated, outliers removed, missing data omitted, resulting in sufficient data for a problem-free analysis. However, outliers provide interesting insights into the issue at hand and managing missing data is a skill often required in the outside world. Handling real world data gives students important skills they can use in their later careers. Moreover, the pre-fabrication of data disengages students from the real-life aspects of the task.

The MDG indicators monitor ongoing processes and, as such, are constantly being modified, updated and extended. As the Millennium Development Goal e-learning materials described here sit on top of the UN's own MDG indicator data, they are well-suited to accommodate this changing data environment. In other words, by delivering learning resources through live web interfaces, underlying data can evolve naturally, thus engaging students with the changing world they see around them.

ESDS International's teaching and learning resources, including the UN Millennium Development Goal e-learning materials can be found at: http://www.esds.ac.uk/international/resources/learning.asp

Notes

1. Susan Noble, Dr Celia Russell and Richard Wiseman. All of ESDS International, Mimas, University of Manchester, UK. (contact at: susan.noble@manchester.ac.uk, celia.russell@manchester.ac.uk, richard.wiseman@manchester.ac.uk).

This work was presented in session 'C2: Sharing Data: High Rewards, Formidable Barriers' at the 2009 IASSIST conference.

2. Angel Gurría, Secretary-General, OECD. Opening speech at the OECD Forum 2008: http://www.oecd.org/dataoecd/25/5/40762048.pdf.
3. Databeuro - http://www.databeuro.com/.
4. JISC and Publishers Association Model Licence - http://www.jisc.ac.uk/aboutus/committees/workinggroups/disbanded/standardlicensing/report.aspx
5. Dr Eric Swanson, World Bank (2010) 'Working with International Development Data', University of Manchester: http://www.esds.ac.uk/international/news/news.asp#16dec10a
6. Levine, Joel H, 1993 Exceptions Are the Rule: An Inquiry into Methods in the Social Sciences Boulder, Westview Press,

The Research-Data-Centre in Research-Data-Centre Approach

A First Step Towards Decentralised International Data Sharing

by Stefan Bender, Jörg Heining[1]

RDC-in-RDC

Abstract

Remote data access, defined as the ability of a researcher to access and evaluate restricted micro data via a secure internet connection from his home desktop computer at any time, has not been implemented by a German Research Data Centre (RDC) so far. Privacy regulations and especially the problem of access control are reasons why German RDCs are not able to offer restricted data via remote data access to the research community. By initiating the "RDC-in-RDC" approach, the Research Data Centre (FDZ) of the German Federal Employment Agency (BA) at the Institute for Employment Research (IAB) in Nuremberg, Germany, aims to bring data access in Germany closer to the ideal perception of remote access. The basic idea is to allow remote data access from designated institutions with comparable standards at locations other than Nuremberg. In a first step, access to BA and IAB data will be granted from four sites in Germany and one site in the US. Moreover, the RDC-in-RDC approach represents a change of paradigms in two respects. First, data access will be decentralised and the FDZ literally brings its data closer to the researchers. Second, data of the FDZ will be accessible from abroad so the dissemination of micro data will be no longer restricted to national borders. The RDC-in-RDC approach may therefore be regarded as a first step towards remote access in Germany and may also represent a blue print for an intensified international data sharing.

Keywords: micro data, remote data access, international data access

Introduction

Fostered by the rapid developments in technologies and methodologies, statistical institutions and authorities have experienced a growing demand for high-quality micro data by both the scientific community and policy-makers over the past years. Despite the fact that the dis-semination of micro data for scientific purposes is part of their legal mandate, the preservation of the confidentiality in the data (i.e. to prevent the disclosure of single entities) stands above all when outsiders are granted access to micro data by the statistical authorities. In order to ensure privacy for individuals and to serve the needs of the scientific community, statistical authorities usually apply a combination of different access strategies (see Lane et al. 2008). These strategies may include for example the approval of projects by (statistical) authorities and/or scientific boards, the training of researchers, the anonymisation of data or the establishment of 'safe' settings for on-site use. Widely-used examples of these strategies are Public Use Files for off-site use or Research Data Centres (RDCs) and secure data enclaves in order to allow on-site analyses of confidential micro data.

The most efficient and for researchers most convenient type of off-site use is remote data access defined here as means by which an approved researcher may access restricted micro data for her approved project via a secure internet connection (see Grim et al. 2009 and Hundepool et al. 2009). She is able to do all preparations of the data and analyses off-site but the restricted micro data never leave the safe setting of the statistical authority or an RDC. After the program codes of the researcher have been processed with the data, the outputs are screened and sent back. Depending on the prevailing national data protection acts, remote access systems may even allow researchers to actually see the underlying data.

Statistical authorities in many countries have undertaken efforts to make micro data accessible for the scientific community and to establish remote data access systems over the past years. In contrast to other countries in Europe, North America or Oceania which have already succeeded in the implemetation of remote access systems, however, Germany still lags behind this development. Legal Concerns still prevent the implementation of such access ways to confidential data. Besides the varying stages of development of remote access sys-tems, the diversity of national data protection legislations leads to considerable differences in the scope of performance and services provided by the particular data access sytems. The systems may be limited in terms of statistical analysis tools or only provide data access to a limited number of, or only parts of, data products. Moreover, remote access to micro data is usually restricted to national borders. As pointed out by Ahmad et al. 2009/2010, the limited enforceability of contractual terms and penalties abroad, virtually restricts data access to resident researchers due to high transaction costs for non-residents.

With the Research-Data-Centre in Research-Data-Centre (RDC-in-RDC) approach the Research Data Centre (FDZ) of the German Federal Employment Agency (BA) at the Institute for Employment Research (IAB) in Nuremberg, Germany tries to overcome the existing legal barriers and to bring micro data access in Germany closer to the ideal perception of remote access. The basic idea of this approach is to allow data access from designated national and international institutions with comparable standards as the FDZ site in Nuremberg. By using a secure internet connection, researchers can link to a server and access the whole scope of micro data available for on-site use in Nuremberg.

In a first step, FDZ data may be accessed from four RDC sites of the Statistical Offices of the Länder[2] in Germany. Moreover, a fifth site at the Michigan Center for the Demography of Aging (MiCDA) Enclave[3] at University of Michigan's Institute for Social Research (ISR) in Ann Arbor, Michigan, USA represents the international component of the RDC-in-RDC approach.

The sole aim of the project is not merely the facilitation of access to FDZ data in Germany or the US. It is intended to gather experiences and by doing so, to build expertise among statis-tical authorities, researchers and data protection officials in decentralised ways of data access. The RDC-in-RDC approach may not be equivalent to remote data access, but it may serve as stepping stone, especially for countries where legal concerns are still hindering the establishment of decentralised access ways to restricted micro data. Moreover, due to its international aspect and the insights gained from it, this project may be beneficiary for statistical authorities all over the world. It may represent a blue print for data sharing beyond na-tional borders.

The paper is organized as follows: Section 2 provides a short overview of the international state of developments with regard to remote access, as well as a brief description of both the German situation and the FDZ. The technical implementation of the RDC-in-RDC approach is sketched in section 3. Finally, section 4 concludes.

Applications of Remote Data Access

International developments

The German "research data centre movement" is quite a recent development (see KVI 2000 or Bender et al. 2009). Other countries, often with less stringent data protection legislation, have a longer tradition of operating RDCs and have already implemented remote data ac-cess systems or are currently working on it. Moreover, some countries have also taken first steps towards international data sharing. Several examples of remote data access systems are briefly described in what follows.

One of the oldest remote data access systems for micro data is the Lissy System of the Lux-embourg Income Study (LIS)[4]. The project began in 1983 and was extended to include the Luxembourg Employment Study (LES). The main aim of Lissy has always been to make mi-cro data of a large number of countries available for comparative social research. Lissy is a fully automated system running 24 hours a day, seven days a week. The users submit their statistical requests under the form of specific statistical package programs (SPSS, SAS, Stata) via the internet mailing system or a secure graphical user interface. LISSY will automatically process jobs and return the outputs to the e-mail address given during the registration process.

Although not operated by a national statistical authority, the IPUMS-International Project (Integrated Public Use Microdata Series)[5] is another old example of a remote data access system. IMPUS is a collaboration of the Minnesota Population Center, National Statistical Offices, and international data archives. It was set up in 1999 in order to obtain frequency counts from diverse censuses that are in compliance with data protection regulations. Data are made available through a data extraction system in which users select the variables and samples they desire. They download the data and analyze them on their local computer.

The Cornell Restricted Access Data Center (CRADC) was established 1999. As part of the Cornell Institute for Social and Economic Research (CISER) in Ithaca, NY, the CRADC provides secure access to confidential research data. Researchers of the Cornell University can acquire, house, and use restricted data in CRADC's secure computing environment. After signing a CRADC data user agreement, researchers can access confidential data hosted at CRADC by using either a Windows terminal services client, a terminal services client software or a remote desktop client[6].

Besides the IMPUS-International Project and the CARDC another example of a non-governmental institution providing access to confidential micro data via remote data access systems in the US is the National Opinion Research Center (NORC)[7]. NORC is private entity and located at the University of Chicago. The NORC data enclave runs a remote data access system which mainly provides access to firm data of several governmental and non-governmental data producers and collectors, including the Annie E. Casey Foundation or the U.S. Department of Agriculture among others.

Several governmental authorities in the US have successfully established online systems providing researchers with frequency counts and tabulations, too. In this context, the Data Analysis System (DAS) of the National Center for Education Statistics (NCES) stands out. Besides tabulations the researcher may also calculate simple covariance analyses online[8]. The implementation of an own remote access system is currently also considered by the US Census Bureau. In collaboration with external experts from academia the so called Microdata Analysis System (MAS) is planned, offering access for limited statistical analyses on full Census micro data sets (see Foster et al. 2009/2010).

Statistics Denmark first disseminated micro datasets to researchers in 1986 under an "in-house researcher arrangement". In 2001, remote data access was introduced and 55 access points had already been set up by the end of 2003[9].

Statistics Netherlands, too, has a long tradition of making micro data available to researchers (since the early 1990s). After the demand by researchers for on-site access had reached a very high level, remote data access was introduced in 2006 (OnSite@Home). Researchers can access Dutch micro data by means of some special software which is installed on a regular desktop computer, located in a separate and lockable room at the researcher's institution. By 2009, this special software has been installed on 45 terminals, one even located in Italy. (see Hoeve 2009/2010).

The Australian Bureau of Statistics also operates a remote data access system (RADL), which was set up in April 2003 (see Tam et al. 2009/2010). The RADL system works in three steps. Researchers submit their programs via a secure website, where they are first checked for illegal commands. If this check finds no such commands the program is run and the out-come is automatically checked. There is an additional audit process in which output is manually inspected to ensure that the analysis using the micro data does not violate any legal regulations[10]. Since 2009, the Australian Bureau of Statistics and Statistics New Zealand are providing mutual access to anonymised micro data by using the RADL system. Australian data may be accessed in New Zealand and vice versa (see Upfold et al. 2009/2010 and Tam et al. 2009/2010).

Statistics Sweden has had a remote data access system (MONA) since 2005. This system provides researchers with the possibility of remote access from any computer with Internet access[11]. The MONA system is based on communication between a terminal server and a terminal client. By using a secure internet connection users access a terminal server where they can start applications remotely. For more extensive processing a batch environment is available.

In the United Kingdom, two remote data access applications are currently operated (see Ritchie 2009/2010). The Virtual Microdata Laboratory (VML) by the Office of National Statistics (ONS) allows ONS and governmental staff access to micro data through their desktop computers. Researchers from other institutions use designated thin terminals at government offices instead. To overcome this disadvantage for non-ONS and non-governmental staff, the Secure Data Service (SDS) hosted by the UK Data Archive has become fully operational in 2010. The SDS enables safe and secure remote access for approved researchers to the data of the British Household Panel Survey. The SDS operates using thin-client and Citrix technologies, whereby data are available only via a controlled network (see Wright 2009).

Statistics Canada introduced the so called "Real Time Remote Access" (RTRA) in 2010. RTRA is partially based on the RADL model developed by the Australian Bureau of Statistics. Researchers will submit their requests through a secure portal to a protected server located on the secure Statistics Canada network. After a check for forbidden commands, the syntax will be processed with the data. Disclosure control will be automated as well as notifications to the submitting researcher (see Goldmann 2009/2010).

Also in 2010, the French remote access centre CASD (Centre d'Accès Sécurisé Distant aux donnés) became operative. Designed and developed by the National Institute of Statistics and Economic Studies (INSEE), CASD provides access to household data in France. The CASD system is exceptional since it is a hardware-based solution using the so-called SD-Box (patent pending). After being installed in the researcher's institution, the SD-Box provides a secure biometric access between the researcher and a secure server hosting confidential data. About thirty research projects in France and one project in the United Kingdom have already used CASD (see Gadouche 2011).

The implementation and operation of remote access systems is not limited to national authorities or organisations. Comparable developments are also taking place on the transnational level. To access micro data sets of the European Union (EU)/Eurostat researchers still have to visit the safe centre of Eurostat in Luxembourg. In order to facilitate data access the Essnet-project "Decentralised Access to EU Microdata Sets"[12] was established. The idea is to develop decentralised access by which a researcher from a certain EU member state can use European datasets in his member state. The concept of research data centres which has been already realized in some European countries as well as (the concept of) the safe centre of Eurostat could be examples for a decentralised access to European micro data sets.

The Essnet-project has showed first results of allowing access to European micro data in safe centres (on site). It included the methodology, guidelines and requirements which are essential to implement an access to European micro data in safe centres in the member states. A follow-up project is planned. For an overview on all these activities on the European level see Bujnowska et al. 2009/2010.

The Data without Boundaries (DWB) project is another transnational initiative which started in May 2011 and is funded by the 7th Framework Programme for Research and Technological Development (FP7) of the European Commission. The project brings together data archives, national statistical institutions and universities. The objective of DWB is to develop an integrated model where the best solutions for micro data access are available irrespective of national boundaries but are flexible enough to fit national arrangements. Hence, DWB aims to achieve standardization and harmonization of micro data access methods as a concerted effort on a European scale.

Situation in Germany

Access to restricted micro data stemming both from administrative processes and surveys was rather limited in Germany until ten years ago. In 2001, the Commission to Improve the Informational Infrastructure by Cooperation of the Scientific Community and Official Statistics (Kommission zur Verbesserung der informationellen Infrastruktur zwischen Wissenschaft und Statistik, KVI) finally recommended the foundation of research data centres for public producers of micro data in Germany (see KVI 2001).

The establishment of RDCs at the Statistical Offices, the German Pension Insurance Fund and the Federal Employment Agency (Bundesagentur für Arbeit, BA) resulted in standardised access methods for restricted data collected by the Federal Statistical Office, its regional offices and by the labour and social security administration (see Bender et al. 2009).

German RDCs currently provide two methods of access to restricted or weakly anonymous micro data: either by controlled remote execution or on-site use at the premises of the RDC. Controlled remote execution is a limited mode of remote data access. It means that external researchers send evaluation programs to the RDC, where RDC employees conduct the evaluations, check the results for compliance with data protection regulations and send the tested results to the researcher. In contrast to remote data access, remote execution is general not automated. Hence, it may be regarded as a sequence of single tasks conducted by RDC employees rather than as an integrated and automated system as operated by Statistics Canada or the Australian Bureau of Statistics.

Remote execution is inefficient in two respects. First, as the researchers have no direct contact with the data, they sometimes program "blindly". As a consequence, programs have to run several times until the desired evaluation is obtained. Second, the level of support required from the RDC staff is high. Research visits for on-site use at special separate workplaces for guest researchers at the RDC avoids these problems as the researcher has direct access to the data. However the researchers have to travel to the RDC, a growing number of them even from abroad. This often entails high travel and accommodation expenses.

Although not strictly forbidden by law, the implementation of true remote data access systems which provide access to weakly anonymised micro data in Germany was hindered by the concerns of data protection officers so far. Their main concerns focus on the additional information available from the internet and on the question of access control. Since remote access systems allow data access from outside a safe environment like an RDC, the usage of additional information cannot be controlled. Nowadays, a vast amount of (additional) information is easily accessible via the internet for everyone. It is almost impossible to prevent information from the internet from being used to disclose a single entity in the data. As a consequence, only absolutely anonymised data may be accessed by remote access systems in Germany (see Schaar 2009). Moreover, when processing confidential data outside an RDC via a remote data access system, the problem of how to ensure access control arises.

It is questionable whether in the environment of the researcher's home office or workplace only the approved researchers will have access to the confidential data. Since the mere visual inspection of the data represents a transmission according to German law (§ 67, subpar-agraph 6, number 3, second sentence of the Social Code X), an individual, for example a family member may get unauthorized access to confidential micro data by just glancing on the computer screen at the home office of the approved researcher.

The Research Data Centre (FDZ) of the German Federal Employment Agency (BA) at the Institute for Employment Research (IAB)

When the FDZ was founded in December 2003, there had been no systematic access to social data up until that point. Following a positive evaluation by the German Council for Social and Economic Data in April 2006, the FDZ was permanently established as an independent research data centre of the BA at the IAB. An evaluation by the German Council of Science and Humanities in 2007 confirmed that the FDZ was an internationally unique institution: "The Research Data Centre (focusing on methods and data access) is an internationally visible, indispensable service institution, unique in Europe and a prime example to other institutions, possessing large datasets of scientific importance." (Wissenschaftsrat (German Council of Science and Humanities) 2007, p.55)

The FDZ prepares individual datasets developed in the sphere of social security and in employment research and makes them available for research purposes – primarily for external researchers. With its website (http://fdz.iab.de), the documentation and working tools available online, and its workshops and users' conferences, the FDZ makes it easier for external researchers to work with the datasets. The micro datasets available at the FDZ include the IAB Establishment Panel, the Sample of Integrated Labour Market Biographies (SIAB), the BA Employment Panel (BAP), the Establishment History Panel (BHP), the Linked-Employer-Employee Data from the IAB (LIAB) and the panel study "Labour Market and Social Security"' (PASS) among others.
The FDZ serves not only the national but also the international market. One important step towards internationalisation in 2007 was releasing web pages in English and having almost all of the data documentation translated (see Bender et al. 2009). In addition to this, members of the FDZ have given numerous talks on the FDZ, the projects of the FDZ and the available data at international conferences and foreign universities. As a consequence, the number of users from abroad constantly increased over the past years. Several of these international data users also participated in one or more of the four users' conferences orga-nized by the FDZ.

Besides these activities the FDZ is also involved in several international projects focusing on the creation of new data products or the further development of data access ways. For the project BLUE-Enterprise and Trade Statistics[13] (BLUE-ETS) the FDZ cooperates with the University of Southampton and the Italian Institute of Statistics (ISTAT) in the development of better test data for complex Linked-Employer-Employee-Data. These new test data not only reproduce the structure and content of the original and confidential data, they also share the same statistical properties. Due to the higher resemblance of this new kind of test data with the original data, researchers can prepare their program codes for remote execution in a much more efficient way.

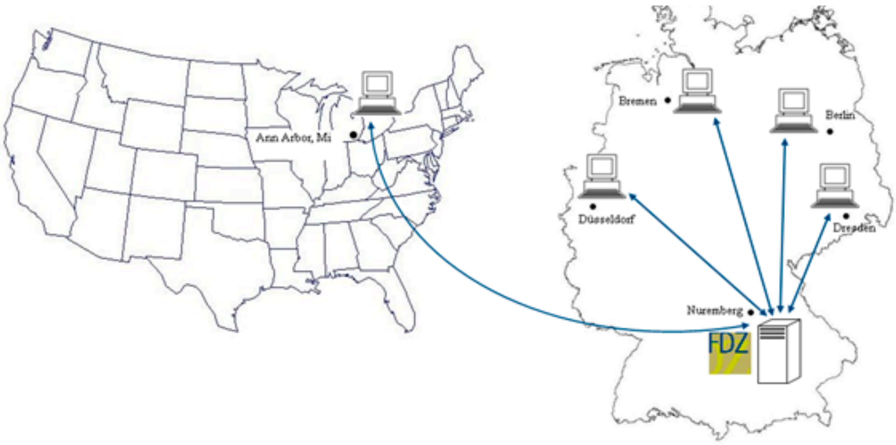
The FDZ is also (co-)organizer of the "Workshop on Data Access" (WDA). Representatives of several national and international RDCs meet at WDA to discuss new developments and to exchange practical experiences. Three workshops have been held already. Other international projects of the FDZ include the Data without Boundaries (DwB) project as described above and, of course, the RDC-in-RDC approach.

Implementation of RDC-in-RDC

The central idea of RDC-in-RDC approach is to enable data access from other RDCs or institutions (called "guest-RDC" in the following) which share comparable security standards as the RDC (called "data-RDC" in the following) where the data are actually stored, but which are located at different sites. In doing so it does not matter whether the guest-RDCs are located in Germany or abroad. The data are accessed in a similar way to the on-site use at the RDCs. 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. In the pilot project the FDZ is the data-RDC. The guest-RDCs can be institutions which fulfill the security requirements of the FDZ. These include all German RDCs[14] as well as comparable institutions in other countries.

In order to improve access to the data of the BA and the IAB in Germany, the RDC of the German Statistical Offices of the Länder and the FDZ are working together on this project. The Statistical Offices of the Länder in the federal states of Berlin/Brandenburg, Bremen, North-Rhine Westphalia and Saxony are participating in the project as pilot locations. Data access for researchers abroad is to be improved by means of cooperation between FDZ and the MiCDA data enclave at University of Michigan's Institute for Social Research (ISR) (see figure 1).

In the following sections various aspects regarding the "RDC-in-RDC" mode of data access are explained in more detail, in particular the division of tasks between the data-RDCs and the guest-RDCs and issues concerning the technical implementation.



Applying for data access

The work of the German RDCs is influenced by different legal framework conditions (Social Code and Federal Statistics Act). For instance, by legal definition the data available at the FDZ are so called social data. The dissemination of social data is regulated by the Social Code (Sozialgesetzbuch - SGB). On the other hand, data from the Statistical Offices are not defined as social data and are made available on the basis of the Federal Statistics Act (Bundesstatistikgesetz - BStatG). Because of this difference in the legal definitions, access procedures differ, too. When applying for social data, for example, researchers have to out-line to what extent the project is related to the social security system in Germany. This is definitely not necessary when applying for data of the Statistical Offices since they are by definition no social data. It will therefore be very difficult to standardise the respective access mechanisms or to transfer these different regulations between the RDCs. As a consequence it is necessary for users to continue to submit their applications for data access to the data-RDC.

At the FDZ all external researchers continue to submit an application for data access in accordance with the Social Code . After the application has been approved by the German Federal Ministry for Labour and Social Affairs (Bundesministerium für Arbeit und Soziales - BMAS), the FDZ concludes a data access agreement with the institution of the user and the user itself. In this agreement the user undertakes to comply with the data protection regulations recorded in the agreement and to bear the consequences stipulated by German law if the agreement is breached.

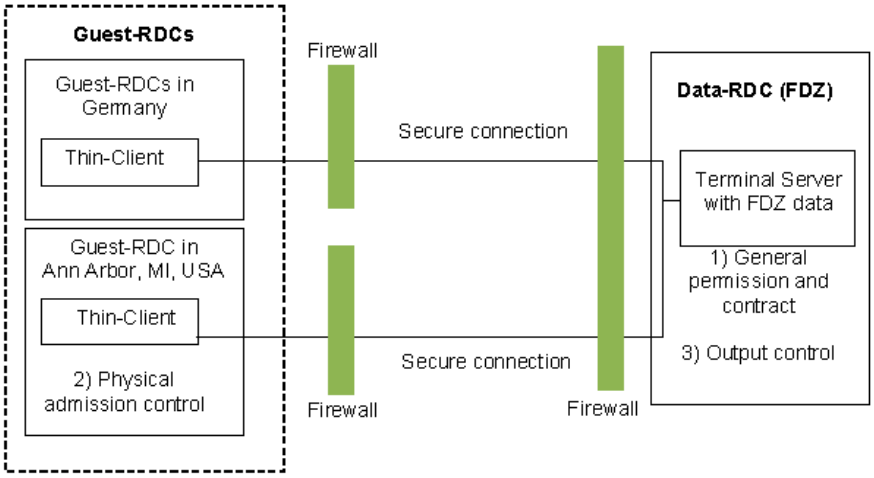


Figure 2: Scheme of the RDC-in-RDC approach. While the conclusion of the use agreement and the output control take place at the data-RDC, physical access control is necessary at the guest-RDC abroad

Access to the authorised data

In order to enable access to the requested data from a guest-RDC it is necessary to develop a new technical concept: the requested data can be accessed from dedicated workstations at the guest-RDCs (see also Figure 2). For this, the same security criteria must be fulfilled at the guest-RDC as apply at the data-RDC. Data access should occur via a secure data line. From the dedicated workstation at the guest-RDC, the researcher logs onto a server of the data-RDC using the remote desktop function and a password. A researcher working at the guest-RDC has the same access rights as the researchers conducting analyses at guest workstations in the data-RDC. In this case, the researcher obtains access to certain servers and to certain directories within the local guest network on these servers. He or she is thus not given the opportunity to intrude into the home network of the data-RDC. Similar to a research visit at the data-RDC, the researcher at the guest-RDC can only look at results on the computer screen.

It must be guaranteed that only authorised users work with the data at the guest-RDC. At the guest-RDCs located in Germany, the supervision is to be performed by employees of the guest-RDC whose data security expertise is regarded as equivalent to that of the staff at the data-RDC. The task of the staff at a guest-RDC is essentially data access control, i.e. they ensure that only the individuals named by the data-RDC gain access to the data. For data protection reasons the employees of the guest-RDC themselves do not gain access to the data, nor may they access the guest researchers' directories. Hence, a researcher may only print results or transmit them electronically after approval by a member of staff of the data-RDC.

A different solution for access control has to be found for the guest-RDCs which are located outside Germany. According to the requirements made by the data protection experts at the BA/IAB, the US site has to be supervised by trained on-site FDZ employees in order to ensure access control and to guarantee the compliance of German data protection regulations in the US. Since the hours of work in an US site differ considerably from the local office hours in Nuremberg, trained employees of the FDZ with administrator rights are required in order to maintain regular operation, too.

Technical implementation for data access control

In the context of the RDC-in-RDC approach a thin client solution using Citrix software will be used for data access control and data access restrictions (see Figure 3). Other methods, for example biometric authentication, webcam monitoring or hardware authentication are possible and used internationally (on this issue see also Grim et al. 2009 or Rowland 2003), but bear either technical disadvantages or, in case of webcam monitoring, are not compatible with German law.

Within this technical solution, normal PCs are turned into so-called thin clients[15] . Thin clients may be regarded as conventional computers which are only able to perform a limited scope of tasks. Thus, for instance, all possibilities for external copying (onto USB, CD-ROM, DVD), for Internet access (including wireless access) or for printing are disabled, the existing software is restricted and access to data is also limited. By running special software (for example Citrix), the thin client guarantees that a user only uses the approved drives and directories and also prevents him/her from being able to install additional programs. This is a component of the common solution for remote data access in other countries.

Researchers connect from the data-RDC via a Citirx Access Gateway to a terminal server which stores the data. Access to this terminal server is provided by an encrypted SSL connection (see figure 3).

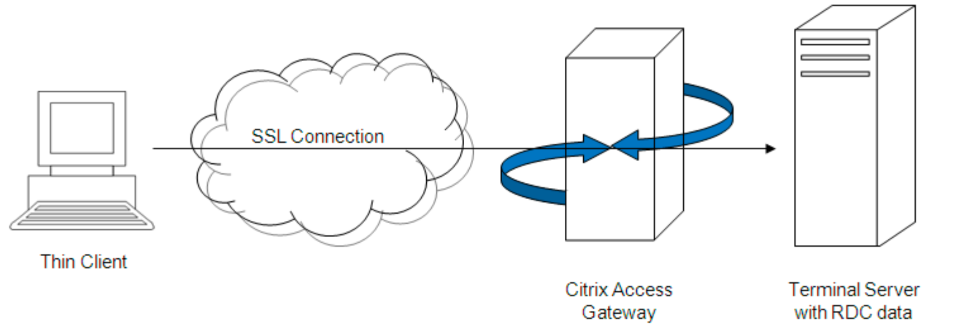


Figure 3: Technical implementation of the RDC-in-RDC approach.

Output control and transmission

Output control continues to be performed at the data-RDCs for several reasons. First, the employees of the guest-RDC have no legal right to access the data. In addition, different legal framework conditions apply for the different RDCs, which influences the monitoring of output for statistical confidentiality. The regulations for monitoring statistical confidentiality can therefore not easily be standardised. Therefore, control remains at the data-RDC and the data-RDC transmits the monitored output to the respective researcher.

Conclusion

Remote access is regarded as an efficient and convenient method of data access which has already has been implemented in several countries such as the United States, Sweden or the Netherlands. Due to legal restrictions, Germany still lags behind this development.

By initiating the RDC-in-RDC approach the FDZ aims to bring data access in Germany closer to the ideal perception of remote access. By allowing data access from designated institutions with comparable standards but locations other than Nuremberg. In a first step, access to BA and IAB data will be granted from four sites in Germany and one site in the US.

Moreover, the RDC-in-RDC approach represents also a change of paradigms in two respects. First, before the implementation of the RDC-in-RDC approach, researchers had to come or connect to a RDC in order to access sensitive micro data. Now, "access is distributed rather than data" (Ritchie 2009/2010, p. 113). By establishing a decentralised way of data access the FDZ literally brings data access to the researchers. Second, data access will also be possible for non-resident researchers. Thus, the dissemination of micro data will be no longer restricted to national borders.

The successful implementation of the RDC-in-RDC approach may not only serve as a stepping stone for statistical authorities in Germany on their way to remote data access. It may also serve as a role model for other countries with a comparable state of development in terms of micro data access. Even countries with well established remote data access procedures may benefit from the experiences gained by the RDC-in-RDC approach. Because of its international dimension, it may represent a blue print for shifting data access beyond national borders leading to intensified international data sharing.

References

Ahmad, N., De Backer, K., Yoon, Y. (2009/2010) An OECD perspective on microdata access: Trends, opportunities and challenges, in: Statistical Journal of the IAOS: Journal of the International Association for Official Statistics, 26, 3-4, 57 - 63

Anderson, O. (2003) From on-site to remote data access – the revolution of the Danish system for access to micro data, United Nations Statistical Commission and Economic Commission for Europe Conference of European Statisticians Working Paper No. 29.

Bender, S., Himmelreicher, R., Zühlke, S., Zwick, M. (2009) Improvement of Access to Data Set from Official Statistics in: Building on Progress – Expanding the Research Infrastructure for the Social, Economic, and Behavioral Sciences, Budrich UniPress Ltd., Opladen & Farmington Hills, MI, pp. 215 – 230

Bound, J. (2008) Michigan Center on the Demography of Aging proposal to the National Institute of Aging (P30 AG012846-16)

Borchsenius, L. (2005) New Developments in the Danish system for Access to Micro Data, Invited paper to the Joint UNECE/Eurostat work session on statistical data confidentiality (Geneva, 9-11 November 2005)

Bujnowska, A., Museux, J. (2009/2010) Release of European Union microdata, ESS projects on remote access, in: Statistical Journal of the IAOS: Journal of the International Association for Official Statistics, 26, 3-4, 89 – 94

Foster, L., Jarmin, R., Riggs, L. (2009/2010) Resolving the tension between access and confidentiality: Past experience and future plans at the U.S. Census Bureau, in: Statistical Journal of the IAOS: Journal of the International Association for Official Statistics, 26, 3-4, 119 – 128

Gadouche, K. (2011) Technological Aspects Concerned in Widening Access to Confidential Data in France, Paper presented at the New Techniques and Technologies Conference (NTTS 2011), Brussels, 23 February 2011, www.ntts2011.eu

Goldmann, G. (2009/2010) From a seed to a forest: Microdata access at Statistics Canada, in: Statistical Journal of the IAOS: Journal of the International Association for Official Statistics, 26, 3-4, 75 – 87

Grim, R., Heus, P., Mulcahy, T., Ryssevik (2009) Secure Remote Access system for an upgrated CESSDA RI, metadata technology, Cessda ppp, http://www.cessda.org/project/doc/CESSDA_RI_SRA_FINAL.pdf

Hoeve, F. (2009/2010) Microdata access in the Netherlands, in: Statistical Journal of the IAOS: Journal of the International Association for Official Statistics, 26, 3-4, 95 - 100

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Lenz, R., Longhurst, J., Schulte Norholdt, E., Seri, G., De Wolf, P (2009) Handbook on Statistical Disclosure Control – Version 1.1. A [Eurostat] Centre of Excellence for Statistical Disclosure Control, http://neon.vb.cbs.nl/cenex/CENEX-SDC_Handboo k.pdf (2009), Accessed 15 October 2009

Kommission zur Verbesserung der informellen Infrastruktur zwischen Wissenschaft und Statistik (KVI) (2000) Wege zu einer besseren informellen Infrastruktur, Nomos, Baden-Baden.

Lane J., Heus P., Mulcahy T. (2008) Data Access in a Cyber World: Making Use of Cyber-infrastructure, in: Transactions on Data Privacy, 1, 2- 16

Ritchie, F. (2009/2010) UK release practices for official microdata, in: Statistical Journal of the IAOS: Journal of the International Association for Official Statistics, 26, 3-4, 109 – 117

Rowland, S. (2003) An Examination of Monitored, Remote Microdata Access Systems, presented at the NAS Workshop on Access to Research Data: 'Assessing Risks and Opportunities', www7.nationalacademies.org/cnstat/Rowland_Paper.pdf.

Schaar, P. (2010) Data Protection and Statistics – A Dynamic and Tension-filled Relationship in: Building on Progress – Expanding the Research Infrastructure for the Social, Economic, and Behavioral Sciences, Budrich UniPress Ltd., Opladen & Farmington Hills, MI, pp. 629 - 642

Söderberg, L.-J. (2005): MONA – MICRODATA ON-LINE ACCESS AT STATISTICS SWEDEN, United Nations Statistical Commission and Economic Commission for Europe Conference of European Statisticians Working Paper No.3.

Tam, S., Farley-Larmour, K.,Gare, M. (2009/2010) Supporting research and protecting confi-dentiality. ABS microdata: Current strategies and future directions, in: Statistical Journal of the IAOS: Journal of the International Association for Official Statistics, 26, 3-4, 65-74

Thygesen, L., Anderson, O., Schnoor, O. (2003) The Danish System for Access to Microdata; from on-site to remote access, Paper presented at Swedish Workshop on Microdata, Stockholm, www7.nationalacademies.org/cnstat/Rowland_Paper.pdf

Upfold, J., Ng, P. (2009/2010) New Zealand's approach to the provision of access to micro-data, in: Statistical Journal of the IAOS: Journal of the International Association for Official Statistics, 26, 3-4, 95 - 101

Wissenschaftsrat (German Council of Science and Humanities) (2007) Stellungnahme zum Institut für Arbeitsmarkt- und Berufsforschung (IAB), Nürnberg, Drs. 8175-07

Wright, Melanie (2009) ESRC Secure Data Service: A new vision for secure data access, Talk given by at New Services for Social Science Research: The Administrative Data Liaison Service and the Secure Data Service, Royal Statistical Society, London, 14 December 2009, http://securedata.ukda.ac.uk/news/publications.asp

NOTES

1. Stefan Bender, Institute for Employment Research (IAB), Regensburger Strasse 104, 90478 Nuernberg, Germany, Tel.: +49-911-179-3082, Fax: +49-911-179-1728, stefan.bender@iab.de; Jörg Heining (corresponding author), Institute for Employment Research (IAB), Regensburger Strasse 104, 90478 Nuernberg, Germany, Tel.: +49-911-179-1752, Fax: +49-911-179-1728, joerg.heining@iab.de. We would like to thank all of the involved project partners for collaboration and support, especially Ramona Voshage, Statistical Office of Berlin-Brandenburg, Sylvia Zühlke, Statistical Office of North Rhine-Westphalia and Maggie Levenstein, Institute for Social Research, University of Michigan. For many helpful comments and discussions we would also like to thank Peter Jacobebbinghaus, Institute for Employment Research (IAB), Karen Scott-Leuteritz as well as the participants of Statistics Canada's 2010 International Methodology Symposium and the New Techniques and Technologies in Statis-tics (NTTS 2011) conference by Eurostat. The project underlying this report was funded by the Federal Ministry for Education and Research (Bundesministerium für Bildung und Forschung) (grant number 01UW1002). The authors are responsible for the content of this publication.
2. A detailed description of these institutions is given in Bender et al. 2009.
3. Bound 2008 provides a description of the MiCDA Enclave.
4. http://www.lisproject.org/
5. https://international.ipums.org/international/
6. http://ciser.cornell.edu/CRADC/What_is_CRADC.shtml
7. http://www.norc.org/DataEnclave
8. http://nces.ed.gov/dasol/
9. http://www.dst.dk/HomeDK/TilSalg/Forskningsservice.aspx (in Danish); see also Anderson (2003), Borchsenius (2005) or Thygesen et al. (2003).
10. http://www.abs.gov.au/websitedbs/D3310114.nsf/home/CURF:+Remote+Access+Data+Laboratory+(RADL)
11. http://www.scb.se/Pages/List____257147.aspx (in Swedish); see also Söderberg (2005).
12. http://www.safe-centre.eu/
13. http://www.blue-ets.istat.it
14. An overview of the German RDCs is given on the website of the German Data Forum
15. (http://www.ratswd.de/eng/dat/fdz.html).

http://en.wikipedia.org/wiki/Thin_client

MAG

AddressingHistory:

a Web2.0 community engagement tool and API by Stuart Macdonald[1]

EDINA

Abstract

This paper will chart the development and delivery of a Web 2.0 community engagement tool and Application Programming Interface (API) developed at the EDINA in partnership with the National Library of Scotland, as part of the JISC Digitisation and e-Content Programme. Such a tool enables members of the community, both within and without academia (particularly local history groups and genealogists), to enhance and combine data from digitised historical Scottish Post Office Directories (PODs) with contemporaneous large-scale historical maps.

The paper discusses the background to Post Office Directories and the corresponding geo-referenced historic maps for Scotland, the technical platforms deployed including sustainable software components, and web applications and services. It also examines issues surrounding user generated content (UGC) created by the community such as mediation, validation and cross-checking, and the use of social media amplification for community engagement and future directions.

To conclude, the paper argues that the success of online crowdsourcing tools such as the one developed for this project will ultimately be measured by continual and extended use within the wider community.

Introduction

Post Office Directories, precursors to modern day Yellow Pages, offer a fine-grained spatial and temporal view on important social, economic and demographic circumstances. They emerged during the late seventeenth century to meet the demand for accurate information about trade and industry due to the expansion of commerce during this period. They were published more frequently than the census and generally had information about local facilities, institutions and associations, listings for private residents, traders, trades and professions, sometimes details of important people, and advertisements.

The ways in which publishers collected data varied considerably. Some obtained information by personal canvassing and combined the results with existing trade listings. Other publishers simply asked people to send in their names together with a small payment if they wanted to be included in the directory.

By the early nineteenth century methods of compilation were more organised. In part, this reflected the growing links between directories and the Post Office. Many postal officials turned their hand to directory publishing as a means of both aiding their work and augmenting their income. Information was collected by letter carriers, who circulated forms during their postal rounds, and also delivered the finished directory on commission.

For Scotland there are at least 750 Post Office Directories spanning the period 1770 – 1912. The NLS are in the process of scanning using Optical Character Recognition (OCR) techniques and publishing this historic collection in conjunction with the non-profit Internet Archive.

During the 6 month project period the AddressingHistory 'crowdsourcing' tool focussed on three volumes (1784-5; 1865; 1905-6) of the Edinburgh digitised PODs and maps from the same periods. However the specifications were such as to accommodate the full Scottish collection as and when they become available. The Web 2.0 interface and back-end storage solutions were built to be both scalable and as far as was practicable, self-standing so that multiple independent instances can be supported and customised for different audiences.

The Edinburgh Directories themselves are a unique and reliable collection of street, commercial, trades, law, court, parliamentary and postal information relating to the city of Edinburgh. They also provide a wealth of detailed information regarding residential names, occupations and addresses and include maps of both Edinburgh and Leith indicating trade and residential origins and development.

One significant deficiency of this collection, which the AddressingHistory online tool aims to redress by 'crowd sourcing', is that the addresses are not geo-referenced. Geo-referencing make possible explicit spatial search and discovery, whilst permitting a map based metaphor to be used in the exploration and visualisation of the resource. E.g. the historic distribution of shipwrights in Edinburgh can be plotted on a base map or the map itself can be used to explore the spatial distribution of selected phenomena (and their variation over time). Similarly, personalised maps illustrating family histories, maps tracking changes in local communities, and maps linking to other digitised materials such as census records and geo-referenced images, and historical addresses could all be explored through use of the Application Programming Interface (API).

The National Library of Scotland's Map Library is one of the ten largest in the world; as the Library of the Faculty of Advocates from 1689, maps of Edinburgh were actively collected; as a Copyright Library, the collections are particularly strong in the printed mapping of Scotland. Since 1998, NLS Map Library has scanned over 20,000 historical maps of Scotland, including over 500 of Edinburgh and its environs.

It is the pre-existence of large scale geo-referenced and contemporaneous maps against which the historic post office directories were contextualised that allows manual (geo)referencing down to individual house address level to be accomplished. This is achieved by simply moving a pin on the map; i.e. the map is the mechanism through which the geo-reference is allocated by the user to a particular POD entry.

To assist the geo-referencing exercise, addresses from each of the directories were parsed using Google's geocoding software[2] in order to assign a geo-reference. There were issues with the legibility of the OCR'd text (especially for the POD for the earlier period) in addition to period addresses no longer being in existence or having suffered name changes. Thus within the interface a ranking mechanism makes explicit the relative 'accuracy' of the geo-coded content.

The user interface to the tool and associated API is intuitive and easy-to-use to encourage researchers, local historians, genealogists and members of the wider community from across the age spectrum to discover, explore and contribute to rich records of social history and to create their own related maps and data sets for both academic and personal research. These were also developed to be sympathetic to tools developed by related projects including Visualising Urban Geographies (VUG), an online resource developing new insights into the spatial character and historical development of Edinburgh (http://geo.nls.uk/urbhist/).

Technologies

Overview

The AddressingHistory tool and API comprises several software components, each built with resilience and sustainability in mind. Open Source software was chosen in several instances, allowing for great flexibility and a feature-rich application, whilst containing costs. AddressingHistory is built as a typical 3-tier web application. For the user-facing client presentation component, JISC recommended standards such as XHTML, CSS and other relevant W3C web standards (i.e. images etc) were employed. The web interface is supported by mainstream browsers and OGC standards including the Web Map Service (WMS) Interface Standard and OpenLayers were used for web mapping components.

An API is available, allowing access to the raw data via multiple output formats. It is accessible via a RESTful web service

Development

The project followed best practice for technical development, making extensive use of a number of common and well-established libraries including the Java SDK, Spring MCV framework and the jQuery Javascript libraries. Unit testing was performed via the jUnit libraries.

Development initially began by scoping the application's requirements, designing a database structure to store the information contained in the Post Office Directories in conjunction with pre-processing and data-loading software. The structural interpretation and translation of the varied content from three eras of directories proved to be a time consuming exercise. The directory data was processed, and additional metadata such as the locations of addresses were added to the database.

The API, following JISC recommendations for API Good Practice[3] was designed to allow access to the raw data using a number of HTTP GET procedure queries including a parameter which allows web developers to specify the format (JSON, KML or TXT) they want the result returned in.

The client application was built upon the API, featuring web based mapping. To the OpenLayers mapping, we added a collection of historical maps from NLS, contemporary to the three Post Office Directories of interest. A user registration system, facilities to edit the stored data and suggest specific changes were added towards the end of the development, together with various enhancements – including a view to the original scanned directory pages.

All components of the web accessible service and API are hosted via a Solaris 10 virtual container, together with an established PostgreSQL database, hosted at EDINA. Throughout the project, the source code, tests and configuration files were stored in a Subversion version controlled repository. Software builds and releases were automated via the Apache Maven software project management tool. Documentation is stored in a shared repository.

User Generated Content

The AddressingHistory project raised a number of issues regarding user generated content (UGC) created by the community such as mediation, validation and cross-checking of UGC. At present the AddressingHistory team retain the option to check UGC and will do so on a periodic basis. As part of a sustainability plan it is envisaged that once community participation reaches a certain level an 'engaged user group' comprising active members of the user community may volunteer to conduct validation and cross-checking of UGC through a devolved mediation process.

A logging facility has been installed in order to identify inappropriate behaviour (e.g. spam) or inaccurate UGC. A registered user can be contacted in order to justify behaviour. Potentially the user or more accurately the username can be prevented from editing further content.

The AddressingHistory backend database has been designed so that the original database and that containing the UGC are maintained as separate instances thus allowing inaccurate or inappropriate user generated content to be removed from the database.

Social Media

A key element in determining the success of the project was the establishment of a mechanism whereby the 'crowd' could contribute to the creation of a fully geo-coded version of the digitised directories. In part an avenue through which such community engagement could be realised was through working with Edinburgh Beltane – a national co-ordinating centre for public engagement and with the University of Edinburgh College of Humanities and Social Sciences Knowledge Transfer Office. Social media channels were also deployed to engage the public, to develop links within the local and family history communities, and to act as a vehicle to expose the tool and API to a wider audience. The following section describes both method and mechanism used to engender public collaboration and community engagement

Building and Developing Community Connections

At the outset of the project an information page was created on the EDINA website[4] and later updated to connect to additional AddressingHistory presences. A WordPress blog[5] , was deployed as the key space for communicating and engaging with interested members of our target audiences.

Twitter was an unexpectedly useful space for the project (via the @addresshistory account) and a Facebook page[6] was also created for AddressingHistory for sharing short updates, useful links and to encourage viral sharing and recommendation.

Community Engagement

From the outset the project team encouraged blogging and discussion of the project and the Project Officer proactively sought out potential contacts, followers, and bloggers and responded to any comments on the project to ensure that mentions were complemented with links to the website and that questions were responded to.

AddressingHistory benefited from pre-existing genealogy and local history blogging and online communities, receiving regular mentions and links from a wide variety of sites and discussion boards[7] . The social media and web presences helped reach out to many interested parties[8]. However outreach activities such as events, presentations, and print communication were also instrumental in exposing the project to a wider audience.

Ongoing Activity

As a longer term strategy we intend to maintain where practicable blog activity, Facebook and Twitter presences. A mailing list has been set up to ensure we can remain in contact with those interested in AddressingHistory developments and a Google group has been established aimed at users interested in using the AddressingHistory API for their own websites, projects, or mashups.

Future Directions

Features and Functionality

AddressingHistory was an ambitious project from the outset, covering a range of technologies and featuring several disparate problems. The initial processing of data extracted from the historical directories through OCR, presents a unique challenge in terms of data errors and lack of structure.

In spring 2011[9], as part of a second phase of development made possible through further funding, the AddressingHistory team will investigate and further streamline data pre-processing and loading processes with a view to providing 'cleaner' output. In conjunction with this, it will add further content (for other areas of Scotland) to broaden the user community and subsequent utility of the tool and API, as well as incorporating computer-generated metadata to POD entries such as categorising places and professions, or extracting multiple addresses.

Interesting further work may also involve investigating how best to capitalise on the social mechanics of AddressingHistory. This unique application offers opportunities for game-mechanics, awarding users for their crowd sourced information and challenging users to contribute.

Another avenue of development under consideration is the inclusion of facilities to upload and attach geo-referenced content such as images, census records, videos and sound files to AddressingHistory entries in the database. In this way, the directories would be extended to include photos of people, buildings, landmarks thus enriching the resource and broadening both utility and appeal.

Learning and Teaching

The AddressingHistory project team have met with representatives from Glow (a national intranet for education hosted by the Scottish Government[10]) with a view to using AddressingHistory as a means to create learning and teaching materials by Glow subject specialists for school pupils both within Edinburgh and beyond. Materials developed may have resonance with the recently launched *Digimap for Schools*[11], an online mapping service for use by teachers and pupils in schools hosted by EDINA

Linked Open Data

> *"Great work! but (cries) this *really* should be done with Linked Data and RDF, endless scope and endless data!"*
> - Comment received on the AddressingHistory Blog

The POD data has been processed and structured in such a manner that every person entry in the directories has a unique identifier in the AddressingHistory XML database. Each entry is accessible via the API through a URI - in much the same way that a unique place in Geonames (an open geographical database) is referenced by a unique URI. As the RDF output from GeoNames gives you the data in an XML document by using a schema of tags defined by the GeoNames ontology/vocabulary, AddressingHistory could theoretically provide output that included some place information using said GeoNames tags (i.e. you get a result for a person who lives in Leith and the result also links to information describing the spatial footprint, population, economy, demography of Leith). Or, indeed, 'celebrities' present within the PODs could be linked to DBpedia entries using the 'sameAs' tag, which declares a link between two resources that describe the same real-world thing.

Context versus Content

> *"A major feature of this project is the offer of maps, and maps which enable the user to explore and present historical information spatially. The outcome is visually attractive and exciting. There is a danger that the fun of producing the map acts as a barrier to thinking about what is happening."*
>
> - *Quote from Professor Robert Mor*ris at the Launch Event

Professor Robert Morris, Emeritus Professor of Social and Economic History at the University of Edinburgh who provided the introductory presentation at the AddressingHistory launch[12] had reservations regarding context versus content. He indicated that, where applicable, explanatory notes providing information about background, construct and content of the original directory listings should be made explicit. In addition underlying assumptions and rules about both the structure of the processed data and the translation of the structured data into a consumable and interactive format should be made clear.

This has in part been addressed by inclusion within the interface of a range of Help documents (including a Post Office Directory Guide and People, Place & Profession Search Guides), an API Guide and Frequently Asked Questions. As part of any future development, use cases and contextual essays (such those available from the Statistical Accounts for Scotland[13]) should be considered.

Two videos which help to demonstrate context were created for the launch and shared via Vimeo . One video discussed the background to the PODS[14] explaining their usefulness to researchers (amateur and professional), the second video, featured on our Facebook page on the launch day, explains the POD digitisation process[15].

Sustainability

In accordance with the project plan the AddressingHistory project partners are committed to supporting the resource for a minimum of one year whilst it gathers community traction. During this time consideration will be made to the processes necessary for ongoing dissemination, community take-up of the deliverables and their adoption by the community. AddressingHistory aim to achieve this through those social media channels established as part of the project and an on-going relationship with Edinburgh Beltane[16] and, in turn, to appropriate organizations engaged in local and family history projects. Given the broad applicability of the resource it is envisaged that a range of communities may be interested in the longer term curation and continuance of the project tools e.g. the Open Street map community has an active user base interested in both contemporary and historical addresses. It is also anticipated that the active involvement of 'engaged users' throughout the project and beyond will provide direction on longer term sustainability issues.

Project partners will evaluate possible business models of sustainability based on levels of demand provided they remain consistent with the underlying open philosophy e.g. revenue generation through an online donations facility, subscription model (e.g. per annum, per month, per use), a 'freemium model' (e.g. free API download of a certain number of records with payment being required for further downloads), or academic advertising.

Conclusion

AddressingHistory was an ambitious project which combined a range of technologies from data processing and database design, to Web 2.0 and web mapping services. Much was achieved within the relatively short project in terms of public engagement and amplification through social media facilities and channels, and the delivery of a robust and scalable website and API capable of empowering the 'crowd' with the facility to search and edit geo-referenced content from the Scottish Post Office Directories and digitised historic maps from the same era.

However, gauging the success of the project goes beyond the delivery of engaging and innovative online tools. It will ultimately be measured by continual and extended use within the wider community[17].

Notes

1. This paper was shown as a poster presentation at the IASSIST conference 2010 at Cornell University. Stuart Macdonald is the AddressingHistory Project Manager at EDINA & Data Library, University of Edinburgh. Email: stuart.macdonald@ed.ac.uk
2. http://code.google.com/apis/maps/documentation/javascript/v2/services.html#Geocoding
3. http://ie-repository.jisc.ac.uk/344/
4. http://edina.ac.uk/projects/addressinghistory_summary.html
5. http://addressinghistory.blogs.edina.ac.uk/
6. http://www.facebook.com/AddressingHistory/
7. E.g. Clan MacLea/Livingstone (http://clanlivingstone.info/forum/viewtopic.php?f=5&t=1084&p=9800&hilit=addressinghistory#p9800),
Rootschat (http://www.rootschat.com/forum/index.php?topic=496764)
8. Indeed many of the social media monitoring techniques that were trialled on AddressingHistory are now successfully being used to better monitor social media mentions of other EDINA projects and services.
9. At time of publication (February 2012) phase 2 of the project is nearing completion. Work focused on streamlining the geo-parsing, and extending geographic and temporal coverage of Post Office Directories within the online tool. An AddressingHistory Augmented Reality Application will also be available shortly.
10. http://www.ltscotland.org.uk/usingglowandict/glow/
11. http://digimapforschools.edina.ac.uk
12. http://tinyurl.com/375czrb
13. http://edina.ac.uk/stat-acc-scot/reading/
14. http://vimeo.com/16902845
15. http://vimeo.com/16906333
16. The partnership cultivated between AddressingHistory and Edinburgh Research & Innovation and Edinburgh Beltane has initiated ongoing communications between EDINA and both organisations with a view to enhancing community engagement from broader service level perspectives.
17. Note: A free to access index for the Glasgow Post Office Directories from 1783-1911 is now available - http://bizdirs.from-mt.com/glasgow/

MAG

IASSIST

INTERNATIONAL ASSOCIATION FOR SOCIAL SCIENCE INFORMATION SERVICE AND TECHNOLOGY

ASSOCIATION INTERNATIONALE POUR LES SERVICES ET TECHNIQUES D'INFORMATION EN SCIENCES SOCIALES

The **International Association for Social Science Information Service and Technology** (IASSIST) is an international association of individuals who are engaged in the acquistion, processing, maintenance, and distribution of machine readable text and/or numeric social science data. The membership includes information system specialists, data base librarians or administrators, archivists, researchers, programmers, and managers. Their range of interests encompases hard copy as well as machine readable data

Paid-up members enjoy voting rights benefit from reduced fees for attendance at regional and international conferences sponsored by **IASSIST**. Join today by filling in our online application:

http://www.iaassistdata.info/

Online Application

IASSIST Member ($50.00 (USD))
Subscription period: *1 year, on: July 1st*
Automatic renewal: *no*

Please fill in the information our Online Form

The application is in USD, however, we do accept Canadian Dollars, Euro, and British Pounds as well.

The membership rates in all currencies as well as the Regional Treasurers who manage them are listed on the Treasurers page