sciecominfo
Nordic-Baltic Forum for Scientific Communication

PRODUCING CONNECTIONS BETWEEN RESEARCH DATA AND PUBLICATION

Iris Alfredsson, Adam Brenthel and Birger Jerlehag

Introduction

Most research processes follow a cyclic path; a study concept is formulated, data is collected and analyzed, and the results of the analysis are published. From the published reports, new research questions arise, new data will be collected, and new results will be published, and so forth. Alternatively, old data will be reanalyzed, as the primary researcher seldom has explored everything in the collected data, or researchers from other disciplines, with different research questions, can reuse the data in a new way.

The possibilities to reanalyze data are dependent on the information upon which data was used and where it was found. To be able to use data for secondary analysis one needs a lot of information about the data: information about the concept, sampling, fieldwork, etc. One also needs information about reports published on results from earlier analysis of the data. Finding published reports is rather straightforward – we have a long tradition where libraries take care of the published research reports, and develop good search systems to find them. Libraries have also the possibility to supply a good overview of everything that is published. But how does one locate the data?

The history of data archives does not extend so far back as the history of libraries. When the technique and methodology for collecting mass data developed during the middle of the twentieth century, a need arose for creating institutions taking care of data and assisting in the process of sharing data. Roper Center, the first social science data archive in the world, was founded in 1947. During the 1960´s and 1970´s a number of social science data archives emerged in Europe and USA, starting with the Zentral Archive (ZA) in Cologne in 1960. [1]

Among the Nordic countries Norway and Denmark were the pioneers. The Norwegian Social Science Data Services (NSD) and Danish Data Archive (DDA) were established in the beginning of the 1970´s. Swedish researchers had to wait until 1981, when the Swedish Social Science Data Service (SSD) was established. At the end of the 1990´s the Finnish Social Science Data Archive (FSD) and the Estonian Social Science Data Archive (ESSDA) were founded. Swedish National Data Service (SND) was re-established 2007.

As part of the Swedish Research Council's (VR) major infrastructure initiative, the Database Infrastructure Committee (DISC)[2] was founded in 2006. DISC's mission is to promote the development of an effective infrastructure for sharing research data resources in Sweden. Organisationally, DISC is subordinate to the Committee for Research Infrastructures (KFI). One of the first key issues for DISC was to transform the existing Social Science Data Service (SSD) into the Swedish National Data Service (SND). The new organisation covers a broader scope, which includes social sciences, the humanities and part of medicine, mainly epidemiology. In the autumn of 2006 there was a call for applications to host the new data service and in the autumn of 2007 an agreement was signed between VR and University of Gothenburg, establishing the university as the host for SND during the next five years.

The main purposes for SND are to mediate information on databases and other collections of digital material for research, to facilitate access to research data and to serve as a knowledge node for documenting, managing research data and adherent methodologies in several knowledge fields. Thus, a very important task for SND is to strengthen the altruistic reception of the importance of data sharing and open access among researchers. There are two key areas that serve as barriers for reaching these goals; legal barriers and possessive barriers. The legal barriers are hinders in current Swedish laws and statues but these laws and statues are also the protection against misuse of information, making it a delicate question. The possessive barriers are attitudes among researchers; many consider produced research material financed by tax money their own property. A strategy to overcome these barriers is a combination of "top-down" and "bottom-up" activities. An example of a "top-down" activity is to influence research financiers to put higher demands on future open access to data when completion of studies. Another example is to provide means and to support researchers through the whole research process, e.g. with interpretations of different legal aspects of open access. Examples of "bottom-up"

[1] Mochmann, E. (2002)International Social Science Data Service. Scope and Accessibility. Report for the International Social Science Council. Cologne: Zentralarchiv für Empirische Sozialforschung

[2] http://www.disc.vr.se/

activities include SND´s presence in different research contexts for example, at conferences and seminars propagating the benefits of sharing data.

The emergence of more and more actors on the European level involved in the process of collecting, preserving, processing and distributing research data, created a need of cooperation between organizations. At the end of the 1970´s the Council of European Social Science Data Archives (CESSDA)[3], was founded. CESSDA extends to 20 countries across Europe and SND is the Swedish node in the network. During the years the role of CESSDA has expanded and today CESSDA hosts a gateway to social science data via the CESSDA data portal[4], providing access to 25,000 data collections, and delivering over 70,000 data collections per annum. CESSDA is also one of 35 projects listed in the European Strategy Forum on Research Infrastructures (ESFRI) European Roadmap for Research Infrastructures. As consequence of this, CESSDA was funded for a two-year preparatory phase project (CESSDA PPP)[5]. This project, which commenced in January 2008, is intended to result in a major upgrade of CESSDA in order to strengthen, widen and make the existing research infrastructure more comprehensive, efficient, effective and integrated. Such an upgraded research infrastructure aims to enable researchers, not only between disciplines but also between countries, to work together developing leading-edge research methods and efficiently analysing large and complex datasets. In essence, making it possible for researchers to sit at their computer, locate, access, merge and analyse data from a number of different sources. Hence, facilitating the potential for increased cross-disciplinary and cross-national research and cooperation.

Cooperation however, needs standards. The main task for a data service is to make the actual data, the ones and zeros, available for reuse among current and future researchers. To make this possible, SND describe the data in great detail, otherwise no one in the future will be able to interpret the data. This description, or metadata, is stored in a standard used by most data archives. Within the social sciences the Data Documentation Initiative (DDI)[6] is an effort to create an international standard in XML for metadata describing social science data. In April 2008, version 3.0 of DDI was launched. DDI 3.0 represents a major advancement for DDI by fully incorporating XML schemas and moving to a data life cycle approach, meaning that the whole cyclic research process is covered. The DDI is used in the CESSDA data portal and it includes the Dublin Core[7] elements, a basic set of tags to describe a resource. Another effort to create an international standard is the Text Encoding Initiative (TEI)[8]. Its chief deliverable is a set of guidelines, which specify encoding methods for machine-readable texts, chiefly in the humanities, social sciences and linguistics.

University libraries provide facilities for making scientific papers and publications, in electronic form, accessible to academia. SwePub[9] is a joint effort to make "unified access to and reporting of Swedish scientific publications" stored in the various campus-based repositories. The SwePub initiative uses the OAI-PMH[10] protocol to harvest the local repositories. This protocol also includes Dublin Core elements. This means we have a common denominator to use to exchange information between our systems. The problem is to know when to use it.

When a publication in a repository is based upon data from a data archive, it should be linked back to the actual dataset. At the data archive the description of the dataset should contain a link to all publications based upon it. To make this possible we have to agree on how to include these links in our respective documents without violating the standards they are built upon. As SwePub only gathers information about the publications and provide a link back to the full document at the local repository, this means that SND and the individual university libraries must do the practical work.

When the connections between research data and e-publication are produced in cooperation the advantages for the end-user will be significant. The gain is foremost scientific but also economic. The scientific gain is the possibility to reanalyse research data in order to assess the interpretations made by other researcher. It can also initiate new collaborations, and perhaps counteract the possibility that two researchers are double working. The cyclic research process will accelerate as research material becomes more searchable and accessible. The economic gain is obvious; at least from a top-down perspective, however, to promote researchers to deposit data in the archives, economic incentives for the individual researcher must be incorporated into the system. There are also improved possibilities for example, sociologists of science to go upstream from the published material towards the empirical material. It will also be possible to go downstream from a research material to find what results has been produced from it. Over time a network will arise that connects publications, material and the researchers engaged in the field. The keyword to make this possible is cooperation. The possibilities

[3] http://www.cessda.org/
[4] http://www.cessda.org/accessing/catalogue/
[5] http://www.cessda.org/project/
[6] http://www.ddialliance.org/
[7] http://www.dublincore.org/
[8] http://www.tei-c.org/
[9] http://www.swepub.se/
[10] http://www.openarchives.org/

are wide-ranging and promote openness and knowledge production.

Birger Jerlehag is the IT Coordinator with responsibility for development and security, working with infrastructure standards and following technological development within SND's areas of responsibilities.

Adam Brenthel works with marketing SND and informing the research community in the construction of new databases and data collections. He is also responsible for the information strategies and creating an overview of conferences and seminars within the humanities, social sciences and epidemiology.

Iris Alfredsson Assistant Manager at SND works with collecting and documenting research data, documentation standards as well as participates as SND's representative in the European cooperative project CESSDA PPP.