11th Conference on Grey Literature and Repositories: proceedings [online]. Prague: National Library of Technology, 2018. ISSN 2336-5021. Available from: **https://nusl.techlib.cz/en/conference/conference-proceedings**

ARCHIVING SOCIAL RESEARCH DATA FROM THE VIEWPOINT OF CZECH SOCIAL SCIENCE DATA ARCHIVE

Martin Vávra

martin.vavra@soc.cas.cz

Institute of Sociology, Czech Academy of Sciences

Abstract

Czech Social Science Data Archive is now an established research infrastructure within the Czech Republic and also it is a part of the European research infrastructure through CESSDA organization. The presentation will address the opportunities and constraints associated with data archiving and sharing in the social sciences. Emphasis will be placed on used standards (for metadata, keywords) and tools (on-line database solution) and on how these standards and tools help to develop a pan-European system of data services in the social sciences.

Keywords

Data archiving, data archives, open access to data, standards of data archiving

Introduction

To begin with, one thing that (I hope) is quite obvious. Scientific information includes not only texts, but research data produced as the result of research processes too. Open access to scientific information therefore covers, or at least should cover, access to research data which complies with the various recommendations articulated by, for example, OECD[1] or the European Union[2]. Moreover, it can be said that we are gradually shifting from the recommendation to make research data available for reuse to the obligation to make them available, as shown, for example, by the rules of the EU Horizon 2020 programme, where it applies in principle that open access should be ensured to the data produced within that programme, naturally with the caveat that there are legitimate reasons for being able to disengage from this obligation, such as personal data protection, or commercial obligations tied to data[3]. Development in the Czech Republic would appear to be heading in the same direction, as shown by Národní strategie otevřeného přístupu k vědeckým informacím (National Strategy for Open Access to Scientific Information) for the years 2017-2020[4].

Why is open access to data important?

There are many reasons to support open access to data to as great an extent as possible, so let us summarise the most important of these (according to Corti et al. 2014 and Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020). Open access to data:

1. makes it possible to build on previous research results and in doing so support the cumulative nature and quality of knowledge;
2. simplify the control of scientific procedures (by replicating analyses);
3. encourages collaboration and reduces the likelihood of duplication of research on the same subject-matter, this supporting the effectiveness of expenditure on science;
4. makes it easier to involve a wider range of actors in the sphere of science (see the current trend of "citizen science").

The need to store data sets and make them available is accompanied by the need to build infrastructure that makes it possible to preserve such data and ensure access to them. This is primarily a matter of constructing data repositories for storing, preserving and making accessible data files and the metadata relating to this. It is important to mention here that there are different types of data repositories and that we are interested only in the ones that are to make data accessible for further analyses as their fundamental objective (Český sociálněvědní datový archiv [Czech Social Science Data Archive] being one of them), meaning that we will leave aside repositories that are only used to preserve data for the purposes of use

---

[1] In its *Principles and Guidelines for Access to Research Data from Public Funding* (OECD 2007), OECD states that access to scientific data acquired from public sources should be simple and user-friendly (preference for online access) and without unnecessary delays (OECD 2007:150).

[2] See, for example, the updated recommendations of EC from April 2018 "On access to and preservation of scientific information", which also places emphasis on data. Available from: **https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32018H0790**

[3] More detailed information can be found in Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020. Available from: **https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf**

[4] **https://www.vyzkum.cz/FrontClanek.aspx?idsekce=851495&ad=1&attid=851502**

by a specific institution or by a specific research team. By infrastructure we do not only mean the technical side of the matter, i.e. hardware and software, but human resources, the principles of functioning and the standards and procedures used in everyday practice.

Czech Social Science Data Archive and its engagement in CESSDA

The Czech Social Science Data Archive[5] (ČSDA), which is part of the Institute of Sociology of the Czech Academy of Sciences, is such an infrastructure for social sciences in the Czech Republic. ČSDA was established in 1998 under the name of Sociologický datový archiv (Sociological Data Archive), with the name of the archive changing in 2011 to "Social Science" in order to express the broader focus of the acquisition policy. The core of the archive is its data collection, which as it stands (October 2018) numbers more than 800 data files, the overwhelming majority of which still come from sociology institutes, although we are now beginning to acquire data from other areas (historiography, psychology). Most of the data collection is available online[6], the only condition for downloading data being registration, which is possible using a web form[7]. ČSDA uses the Nesstar application[8] to make data available, the application allowing users to search metadata, analyse data online or download them to their computers for further analyses.

A detailed description of the procedure involved in archiving a data file, from the pre-acceptance stage, when negotiations are ongoing with the depositor on the provision of data for archiving, to the state of providing the data file to archive users, is provided in the ČSDA Preservation Policy[9]. Among the important moments of the whole process is the agreement on data deposition in the archive entered into between the depositor and the Institute of Sociology of the Czech Academy of Sciences[10]. The deposited data files and their formats[11] are specified in the agreement, as are any other conditions for handling the data which the depositor may impose. In light of the fact that the archiving procedures in place in ČSDA essentially refer to the OAIS system, we preserve data in original format (in OAIS terminology SIP - Submission Information Package) to maintain the authenticity of the original data, and in "archival" format (AIP - Archival Information Package), which counts on the migration of data formats such that they are readable and usable over the long term. In addition to the acquisition, preservation and disclosure of data, ČSDA workers collaborate in surveys conducted at the Institute of Sociology in which they ensure the management of research data.

The fundamental infrastructure at a trans-European level is the Consortium of European Social Science Data Archives (CESSDA) [12] . A distributed European infrastructure (European

[5] **http://archiv.soc.cas.cz/**

[6] At: **http://nesstar.soc.cas.cz/webview/**

[7] Such registration means that it is not "open access" in the strictest sense. However, the fact that there is no restriction on who can register means that it is still relatively low-threshold access. Registration serves the archive as a way of binding a specific, identifiable user to adhere to the rules of access to data and also helps us in terms of reporting our activity as an infrastructure - thanks to registration we have a better overview of the numbers and characteristics of active users.

[8] **http://www.nesstar.com/**

[9] **http://archiv.soc.cas.cz/archivacni-rad**, English version at: **http://archiv.soc.cas.cz/en/preservation-policy**

[10] A specimen, which may be modified subject to negotiation with the specific depositor, is available at: **http://archiv.soc.cas.cz/sites/default/files/dohoda_o_depozici_dat.doc**

[11] In its preservation principles, ČSDA describes preferred and acceptable formats for the transfer of data to the archive. In the same way "archival" formats in which data are preserved for the purpose of their long-term storage are specified therein.

[12] **https://www.cessda.eu/**

Research Infrastructure Consortium – ERIC) was established on the foundations of the original, informal CESSDA association in 2013. The Czech Republic became a member of this consortium and ČSDA, which had been a member of CESSDA, as an informal grouping, beforehand, became the national node for CESSDA. Eighteen European social science data archives were members of CESSDA in October 2018[13]. The institutions involved differ from each other to a fair extent. Some, such as the UK Data Service, are large organisations that make social science data available and data produced by public administration, for example. Others are "merely" departments at research organisations or at universities that employ only a few people (for example, ČSDA currently employs a total of around six full-time workers). At present, CESSDA (and through it the members' archives, including that of ČSDA) is involved in a number of projects at a European level. The objective of CESSDA for the foreseeable future is to establish a "one-stop shop", a web catalogue that will make it possible, among other, to search the data collections of all members' archives. This endeavour is also understandable with regard to the general attempt to create a genuinely trans-European integrated infrastructure for data access - the most important effort of this type is the European Open Science Cloud (EOSC), an initiative of the European Commission which aims to create a trustworthy and open space for the preservation and sharing of scientific data at a European level by the year 2020. CESSDA has signed up to the principles of open access to data and the term FAIR data is currently used in CESSDA documentation - findability, accessibility, interoperability and reusability[14]. This actually involves the updating (for more details see Wilkinson et al. 2016) and specification of the principles of open access to data that were mentioned in the introduction.

Standards of archiving

To make it possible to fulfil the principles of access to data outlined above, it is necessary to create an infrastructure in the form of "hardware" and to establish and push through standards of work and methods of evaluating their upholding in practice.

As far as the general format of the functioning of an archive is concerned, a number of the archives associated under CESSDA refer in their preservation policies and other documents to the model of the Open Archival Information System (OAIS) as a conceptual framework. As previously mentioned, the ČSDA Preservation Policy is based on OAIS and makes reference to its fundamental principles[15]. OAIS makes it possible to structure the work of archives according to basic functions, processes and positions that are responsible for the execution of functions and processes.

Entirely fundamental to the usability of data in archives are quality and comprehensible metadata, which make it possible for researchers to understand downloaded data sets and assess them from a methodological perspective. The DDI schema of description of data (the abbreviation is taken from the initiative that stands behind for this metadata standard - Data Documentation Initiative[16]) serves the archives of CESSDA, including ČSDA, for this purpose. DDI is a schema (or rather schemata - it has several variations) for data description. The standardisation of metadata is, inter alia, a prerequisite for the creation of the integrated

[13] More detailed information about members can be found at **https://www.cessda.eu/Consortium/Membership**
[14] **https://www.go-fair.org/fair-principles/**
[15] **http://archiv.soc.cas.cz/sites/default/files/csda_archivacni_rad.pdf**
[16] **https://www.ddialliance.org/**

CESSDA portal for data searching mentioned - if metadata were not to have a uniform, machine-processable format, such efforts would be unthinkable given the volume of data sets in the CESSDA members' archives.

Thanks to the CESSDA data portal, the European Language Social Science Thesaurus[17] should be able to fully fulfil its purpose. This is a "thesaurus of key words" in social sciences, now available in 13 different languages, including Czech. As a result of this, ELSST is hierarchically arranged from top terms to specific expressions and its language versions are reciprocally transferable, meaning that it will be possible, for example, to enter a key word in the search engine of CESSDA data catalogue in Czech and the result should be all data files described using this key work in all languages of ELSST.

The final standard that I shall mention here is CoreTrustSeal [18] (originally Data Seal of Approval), which is a system of certifying digital archives. Awarding a CoreTrustSeal should mean that an archive has in place such processes and standards that ensure that the stored data will remain securely stored even over the long term. An archive that wishes to obtain a CoreTrustSeal must, for example, have clear rules in place for secure storage and back-up of data or for the updating and migration of data formats. ČSDA successfully obtained a CoreTrustSeal in 2017[19].

Challenges

There are still a number of problem areas and challenges in the archiving of social science data. It is important that we raise the willingness of researchers to share data - without this, meaning without a culture of data sharing, archiving is troublesome (especially in situations in which the principle of open access to data is not in any way enforceable).

Data should also be prepared for archiving from the very beginning of its life cycle - this means that researchers should have compiled data management plans that are ideally part of research projects, and reference would be made to them in contracts with providers of public money for research. Procedure would then follow these plans. CESSDA is also aware of the fact that archives should be more active at this stage too, i.e. in familiarising researchers with how data management should look, and for this reason compiled the Expert Tour on Data Management[20], which should allow data producers to better plan work with data and their future archiving.

It can be said that "big data" is another major challenge for archives. There has been discussion ongoing in sociology, at least since the article published by Savage and Burrows (2007), on the need to use big data in analyses. The problem is that when we look into data archives, we find practically no big data there (to be blunt, there are none at all in ČSDA). There are various reasons for this, a typical one being that in most cases there is a combination

[17] **https://elsst.ukdataservice.ac.uk/**
[18] **https://www.coretrustseal.org/**
[19] A document providing information on source documents and the results of certification is available from: **https://www.coretrustseal.org/wp-content/uploads/2018/01/Czech-Social-Science-Data-Archive.pdf**
[20] **https://www.cessda.eu/Research-Infrastructure/Training/Expert-Tour-Guide-on-Data-Management**

of problems with copyrights, personal data protection and technical problems - meaning how to tailor big data to existing technology and standards at archives.

Conclusion

The infrastructure for archiving and sharing research data in social sciences already exists: however, it obviously has its limits for potential onward development (this stands at the Czech and the EU level). CESSDA is an entirely fundamental organisation at the European level for access to social data. ČSDA is such an organisation within the Czech research environment, and is also a member of CESSDA. Thanks to the development of internet technologies and of archiving standards, this infrastructure is now relatively simple for the user to use. What is important is that the idea of a single place (in the form of a website) at which the user is able to look for data in all CESSDA members' archives is no longer merely an idea, but is at an advanced stage of development.

In order to direct future development, however, it will be necessary to determine what form of open access to scientific data we would like, how we will support it and how we will motivate scientists and their institutions to collaborate on it, such determination obviously only being possible in terms of the science policy of the state, but if possible in line with debate among the parties involved.

References

CORTI, Louise, Veerle van den EYNDEN, Libby BISHOP and Matthew WOOLLARD, 2014. *Managing and sharing research data: a guide to good practice* [online]. Los Angeles: SAGE [Accessed 19 October 2018]. ISBN 978-1-4462-6726-4. **Available from: https://data-archive.ac.uk/media/2894/managingsharing.pdf**

ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT, 2007. *OECD principles and guidelines for access to research data from public funding* [online]. Paris, France: OECD [Accessed 19 October 2018]. ISBN 9789264034020. Available from: **https://doi.org/10.1787/9789264034020-en-fr**

SAVAGE, Mike and Roger BURROWS, 2007. The Coming Crisis of Empirical Sociology. *Sociology*. **41**(5), p. 885–899. DOI: 10.1177/0038038507080443.

WILKINSON, Mark D., Michel DUMONTIER, IJsbrand Jan AALBERSBERG, et al., 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* [online]. **3** [Accessed 19 October 2018]. DOI: 10.1038/sdata.2016.18. ISSN 2052-4463. Available from: **http://www.nature.com/articles/sdata201618**