10th Conference on Grey Literature and Repositories: proceedings [online]. Prague: National Library of Technology, 2017. ISSN 2336-5021. Available from: **http://nrgl.techlib.cz/conference/conference-proceedings/**.

DATA DEPOSIT INTO THE ASEP REPOSITORY

Zdeňka Chmelařová

chmelarova@lib.cas.cz

The Czech Academy of Sciences, Library

Jana Doleželová

dolezelova@knav.cz

The Czech Academy of Sciences, Library

Abstract

The ASEP repository has provided bibliographic records deposit for the results of scientific research at the Czech Academy of Sciences since 1993. In 2012, the database was expanded to include a repository in which full text documents are stored. Since 2017, ASEP also supports data records and data file storage. The bibliographic records of the results can be linked to metadata records so that the user gets not only the full text of a result, but also the data on which the result is based. Each dataset has its own description and metadata follows international standards. In the paper, we will introduce the repository workflow, describe dataset deposit and international standards used as well as different types of user interfaces.

Keywords

Data Repository; ASEP; Czech Academy of Sciences; Library of the Czech Academy of Sciences

Introduction

The Library of the Czech Academy of Sciences (the "Library")[1] has, since 1993, been the administrator of the ASEP[2] (Automatizovaný systém evidence publikací – (Register of Publication Activity of the CAS) bibliographical database, in which bibliographical records and the full texts of documents of all fundamental results of basic research produced at the institutions of the Czech Academy of Sciences (CAS) are stored. Bibliographical records are the fundamental data pillar for the internal evaluation of institutions conducted by the management of CAS and for international evaluation of the results of science and research produces with financial support from the budget of the Czech Republic[3].

The function of the Library is not passive. Its range of duties include the development of the entire system, from the structure of data to the user environment and, last but not least, passing on information to all users that use the system in a way which is understandable. The development and modification of the ASEP system mainly focus on users from institutions of the CAS – authors, managers of individual institutions and the management of CAS itself. The Library monitors international development in the sphere of science and research and subsequently develops and modifies the system. A superstructure to the ASEP database was created in 2012 in the form of a repository of complete texts, meaning that each bibliographical record in ASEP can be accompanied by the full text of the document and the full text of reviews, or responses. The practical use of this function was positively received in evaluation by institutions at the CAS in 2015, when evaluators had on-line documents at their disposal for peer review. The ASEP system was further expanded in 2017 to include another superstructure – a data repository.

Storing and archiving data files

Storing data files and sharing them with the scientific public is nothing new – there are many open institutional, multi-discipline and area-specific repositories that many authors from the CAS have used for a long time now. Area-specific repositories have been established at institutions of the CAS themselves, for example the Czech Social Science Data Archive (ČSDA)[4] at the Institute of Sociology, while the Institute of the Czech Language was a partner to the creation of the Lindat/Clarin repository[5]. An internal survey was conducted at the CAS to concern the archiving of data at institutes of the CAS and the interest shown by institutes in storing data in a data repository. An analysis of this survey shows that awareness of secure archiving is not at the sort of level it would merit. Files containing scientific data are most commonly stored on local computers and servers, not the safest places for archiving. We come across similar experiences in international surveys on the approach of scientists to storing and

[1] *Library of the Academy of Sciences of the Czech Republic* [online]. Prague: Knihovna AV ČR, v. v. i., ©2017 [cit. 26.9.2017]. Available from: **https://www.lib.cas.cz**

[2] *Online catalogue of the ASEP database* [online]. Prague: Knihovna AV ČR, v. v. i., ©1993-2017 [cit. 26.9.2017]. Available from: **https://asep.lib.cas.cz/arl-cav/cs/rozsirene-vyhledavani/**

[3] More on the evaluation of research, development and innovation: R&D Council. Evaluation of research and development. *Research and development in the Czech Republic* [online]. Prague: Research and Development Council, ©2015 [cit. 26.6.2017]. Available from: **http://www.vyzkum.cz/FrontClanek.aspx?idsekce=18748**

[4] *Czech Social Science Data Archive* [online]. Prague: Sociologický ústav AV ČR, v. v. i., ©2005-2014 [cit. 26.9.2017]. Available from: **http://nesstar.soc.cas.cz/webview/**

[5] *Lindat/Clarin* [online]. Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles' University, ©2017 [cit. 26.9.2017]. Available from: **https://lindat.mff.cuni.cz/cs/**

sharing data conducted at, for example, the University of Tartu in Estonia[6], or in studies dedicated to research data, in which Charles' University[7] was also involved. The conclusion of the article, that "there is no simple, unambiguous institutional recommendation of how authors should work with their data, even though scientists suspect that this is an important area to them as well", is entirely apt. The ASEP repository would provide authors from the CAS with the opportunity to archive data securely and over the long-term. The majority of academic institutes expressed an interest in this in the internal survey. We consider a data repository to be an important superstructure to the ASEP database and are convinced that the scientific public will come to appreciate it over time. We see the role of the Library in this area as being one of a mediator that passes on information regarding why to archive and share data, provides a place for storage, advises on how to describe it and attends to long-term protection and archiving. The reasons for archiving and sharing data are described in a number of documents.[8] The opportunity to verify the validity of conclusions in published documents, the effective use of data obtained from public sources, preventing scientific errors, etc., are most commonly mentioned. New results in industry and in other projects can be created based on the use of archived data from original research. Certain scientific magazines (for example, Nature, Science, The American Naturalist) already lay down conditions for data storage, when a scientist is obliged to share data together with a publication. The Public Library of Science publishing house issues instructions for sharing data and presents a list of suitable repositories. Authors to have received a grant from the H2020 programme are obliged to store the full text of the document in open access and, from 2017 onwards, the data files on which their publications were produced[9].

When creating a data repository, we used a number of examples of good practice, as published in the international Registry of Research Data Repositories[10] and in an overview of open institutional repositories at the Technical University of Ostrava website.[11] The DataShare repository of the University of Edinburgh[12], tried and trusted for many years now, was inspirational to us in terms of international institutional repositories, as was the Zenodo[13] project initiated by the EU and CERN in terms of multi-discipline repositories and the Lindat/Clarin repository in terms of area-specific repositories.

[6] MUULI, Viktor. *Research Data in Estonia: collecting, storing, availability: some findings from questionnaire* [online]. Estonian Research Council, 2014. 23.10.2014 [cit. 26.9.2017]. Available from: **http://dspace.ut.ee/bitstream/handle/10062/44052/RD_questionnaire_eng_muuli_14.pdf?sequence=1&isAllowed=y**

[7] JAROLÍMKOVÁ, Adéla. Výzkumná data na Univerzitě Karlově. In: *INFORUM 2017: 23rd Annual Conference on Professional Information Resources, Prague, 30.-31.5.2016* [online]. Prague: AiP, 2016 [cit. 26.9.2017]. ISSN 1801–2213. Available from: **http://www.inforum.cz/pdf/2017/jarolimkova-adela.pdf**

[8] HRABAL, Jan. Repozitáře vědeckých dat. In: *Knihovna.cz* [online]. Brno: Division of Information and Library Studies, Faculty of Arts, Masaryk University, ©2013. 22. 2. 2016 [cit. 26.9.2017]. Available from: **http://ltp.knihovna.cz/?p=385**

[9] *REGULATION (EU) No 1290/2013 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 11 December 2013 laying down the rules for participation and dissemination in "Horizon 2020 - the Framework Programme for Research and Innovation (2014-2020)" and repealing Regulation (EC) No 1906/2006 [online]*.
In: Official Journal of the European Union. L 347/81, 20. 12. 2013, 23 s. [cit. 26.9.2017]. Available from: **https://www.h2020.cz/cs/storage/87c59c7c965787c1deb7a7c85ee5d5be89fbf58b?uid=87c59c7c965787c1deb7a7c85ee5d5be89fbf58b**

[10] *Registry of Research data Repositories* [online]. Re3data.org Project Consortium. [Cit. 26.9.2017]. Available from: **http://www.re3data.org/**

[11] *Green Open Access* [online]. VŠB – TUO Central Library, ©1998-2016**.** Most recent update 13.3.2017 [cit. 26.9.2016]. Available from: **http://knihovna.vsb.cz/open-access/green-open-access.htm**

[12] *Datashare* [online]. University of Edinburgh. [Cit. 26.9.2017]. Available from: **http://datashare.is.ed.ac.uk/**

[13] *Zenodo* [online]. [Cit. 26.9.2017]. Available from: **https://zenodo.org/**

There are 53 institutes[14] at the CAS, divided into three areas of science: I. the area of Mathematics, Physics and Earth Sciences, II. the area of Life and Chemical Sciences, and III. the area of Humanities and Social Sciences, from which it is clear that the quantity, types and size of stored data files differ depending on individual specialisations and focus. A huge amount of data is produced during scientific research, but not all of it need be stored and archived and this is why authors should pay particular attention to file preparation. Many projects recommend, or directly demand, that the beneficiaries create a Data Management Plan, a document in which they plan and describe what data will be produced during research and how they will manage that data. A page is available to authors on the Library website that concentrates on the organisation of data, with links to guidelines and videos that might inspire them[15].

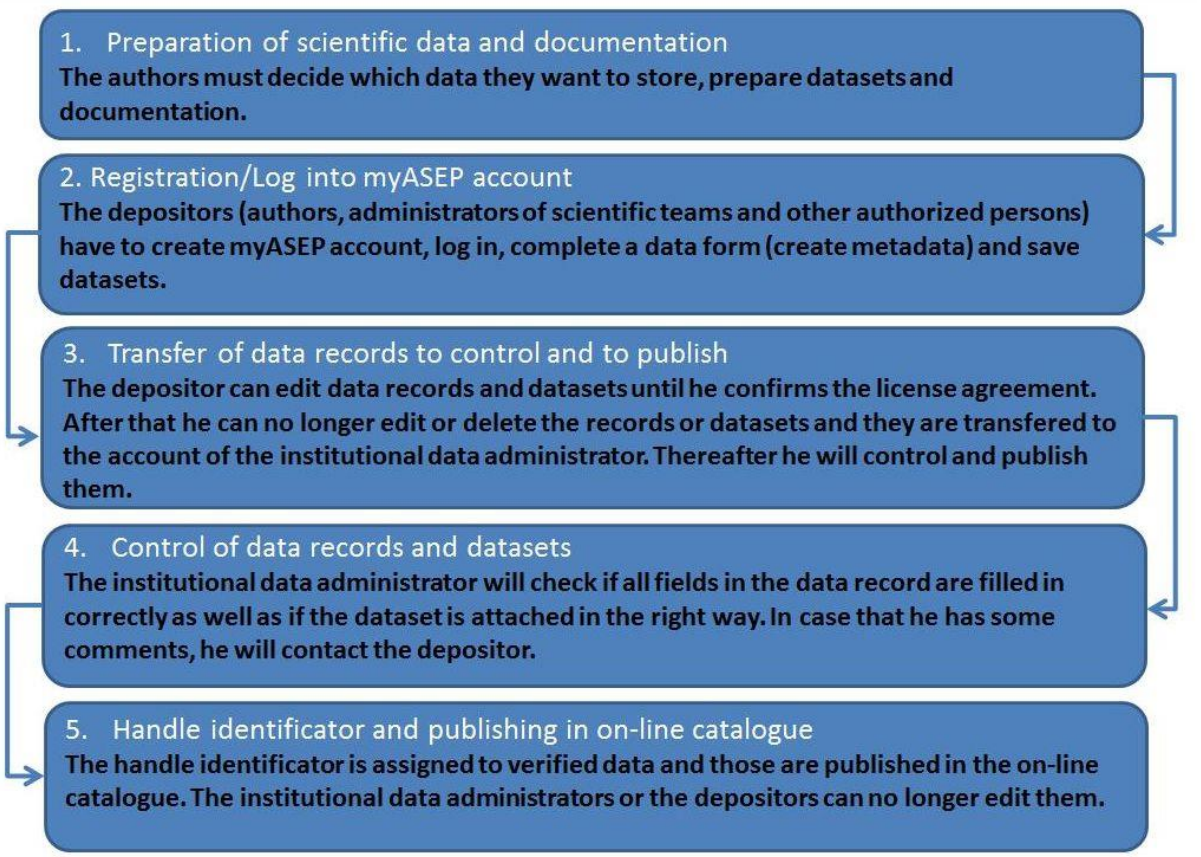
Workflow

The creation of data records and the storage of data sets in the data repository of the CAS follow on from the method of processing to date. Even though it is possible for anyone else to store data in the repository on behalf of the author, which is the common practice in storing bibliographical records and the full texts of documents, we would recommend that the authors themselves be the depositors in the case of data. Filling in metadata forms and saving data sets is a simple matter from the technical perspective. When transferring data records and data sets to data administrators for checking, the depositor confirms that he agrees with the Agreement on the Storage of Data in the ASEP repository[16]. Fundamental requirements: 1. the author must have the necessary rights to store data (the consent of joint authors); 2. sensitive information may not be published (personal numbers, names, telephone numbers, etc.); and 3. a licence for handling data sets must be submitted. The relevant data administrator undertakes a formal check of data records and of stored data sets. If everything is in order, it publishes them in the ASEP online catalogue. The workflow of storing data records with data sets in ASEP is shown in Figure 1.

[14] More about the institutes of the Czech Academy of Sciences: *CAS institutions* [online]. CAS, ©2017 [cit. 26.9.2017]. Available from: **http://www.avcr.cz/cs/o-nas/struktura/pracoviste-av/**

[15] Knihovna AV ČR, v. v. i. ASEP. *Data preparation* [online]. Knihovna AV ČR, v. v. i., ©2017 [cit. 31.10.2017]. Available from: **https://www.lib.cas.cz/asep/pro-autory/priprava-dat/**

[16] *Agreement on the Storage of Data in ASEP* [online]. Knihovna AV ČR, v. v. i., ©2017 [cit. 26.9.2017]. Available from: **https://www.lib.cas.cz/podpora/data/asep/drasep/dohoda_vkladatel.pdf**

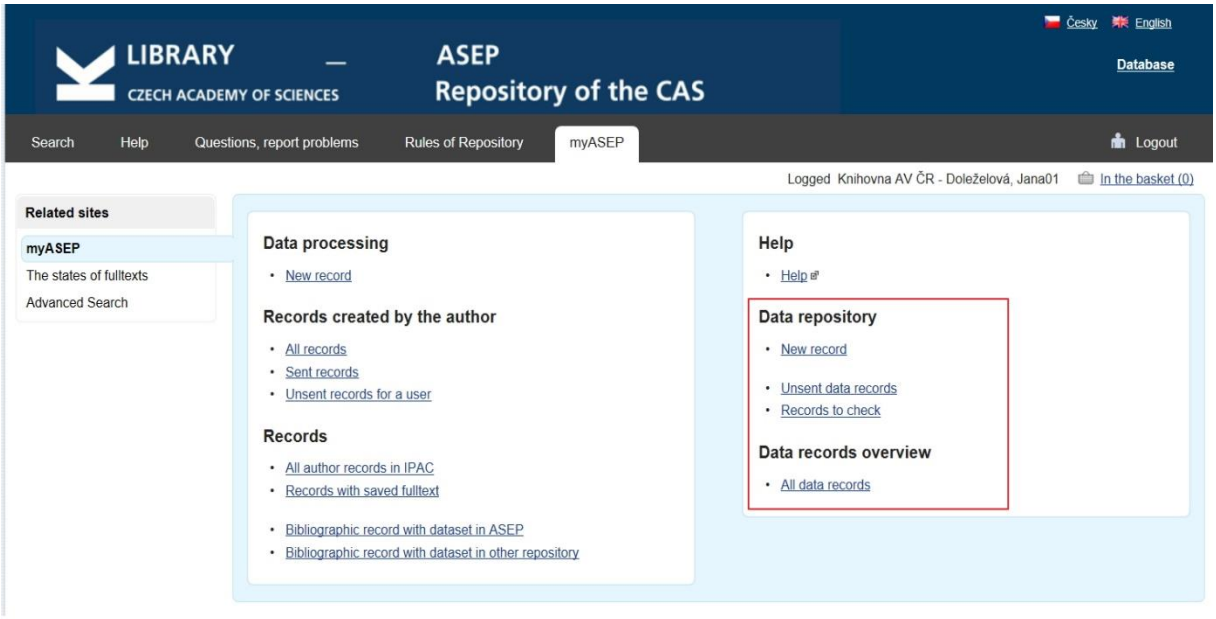


Picture 1: Dataset storage workflow into the ASEP repository

User environment - myASEP

The depositors (authors) and the administrators of the system have their own myASEP user account, from which they manage their data. Figure 2 shows the ASEP user environment for depositors. After logging in, they are able to work with their records, meaning enter new bibliographical records and citations, store the full texts of documents and reviews (the left-hand side of myASEP) and create new data records with data sets (the right-hand side of myASEP). The depositor has an overview of all records that are being processed, prepared for approval and approved and published in the online catalogue. The user account of the system administrators looks similar, but other links and functions are added, in the case of data records a link to records which depositors have submitted for checking and publication. Detailed instructions for use are available to authors and system administrators alike at the Library website[17].

[17] Knihovna AV ČR, v. v. i. ASEP. *For authors* [online]. Knihovna AV ČR, v. v. i., ©2017 [cit. 31.10.2017]. Available from: https://www.lib.cas.cz/asep/pro-autory/



Picture 2: myASEP account - author

Data records and data sets

In selecting a metadata set for the ASEP repository, we tried to ensure the maximum possible completeness of data without placing too much of a load on the scientists authors that create the metadata. We also considered individual fields according to the requirements of other systems such that cooperation would be possible in the future (for example, Data Citation Index, OpenAIRE[18]). We drew on the requirements placed on a data repository: metadata in English, information about financing – statement of projects, links to publications and other output relating to the data, description from the content and technical perspectives, statement of scientific disciplines and key words, determination of time and location. Metadata for data records is entered in an online form in which each field is provided with instructions, so that the depositor knows how to enter data in the field. Fields which are mandatory are highlighted in the form, in that it is not possible to publish a data record in the online catalogue without filling in these fields. The current metadata structure is published at the Library website.[19]

Mandatory fields include author statements, title of the dataset, stored file description, data set type, documentation language, keywords, license settings, and file access. When entering authors, we use the authority base, which enables an unambiguous identification of the author, his/her output and affiliation. Major emphasis is placed on the choice of an apt title for the data set and a description of the file/files in Czech and in English. If a longer description is required, we recommend attaching a readme.txt text file to the data set and to provide further detailed information there. The depositor determines and subsequently sets a Creative Commons licence for the item entered, or chooses his own licence, the wording of which he saves in relation to the data set. The choice of licence is entirely a matter for the author and we do not

[18] *OpenAIRE* [online]. Most recent update 22.9.2017 [cit. 26.9.2017]. Available from: **https://www.openaire.eu/**
[19] Knihovna AV ČR, v. v. i. ASEP. *Description of field – data* [online]. Knihovna AV ČR, v. v. i., ©2017 [cit. 31.10.2017]. Available from: **https://www.lib.cas.cz/asep/pro-zpracovatele/manual/popis-poli-data/**

recommend, but simply offer the choice of a CC licence, which we expect authors to use. If the depositor chooses open access or open access with time embargo, the data sets become accessible immediately after publication, or after the passing of the time embargo. If access on request is set, the user must request the data set from the author.

The structure of data in ASEP has been based on international standards since the very outset. We use library standard UNIMARC for the storage of data, ISO 369 for language coding, ISO 3166 for country coding and Unicode (UTF-8) for symbol coding. First name, surname and institution are accompanied in sets of authorities of authors with identifiers of the Web of Science (RID) and SCOPUS (AIS) systems and the ORCID identifier[20]. Authorities of projects are provided with numbers from the code list of the Central Register of Projects of the Czech Republic (CEP), the code list of European Commission projects (CORDIS) and the code lists of CAS programmes. Subject classification corresponds to the newly-created mapping of specialisations in the Information Register of R&D Results (RIV)[21]. Each data record has a unique HANDLE[22] identifier assigned to it, an OAI-PMH and the Dublin Core DCMI metadata standard are integrated.

The concept of a data set in ASEP entails a set of files that might contain research data, documentation in which there is important information for users, and perhaps the wording of a licence if the Creative Commons licence is insufficient and the user chooses a different licence. The maximum size of one saved file is 2 GB and the total maximum size of saved files for one data set is 20 GB. Larger files can also be saved subject to agreement with the repository administrator. When choosing the format of files, we recommend using the standard open formats that are supported by various systems and programmes and whose long-term protection is ensured. For text files, for example, we recommend txt, pdf, html or csv, for images jpeg, tiff or png and for media mp3, etc. We are aware that these formats might not be sufficient because different specialisations need to store data in formats that are better suited to their data and are tried and trusted in practice by a certain community.

Links in records

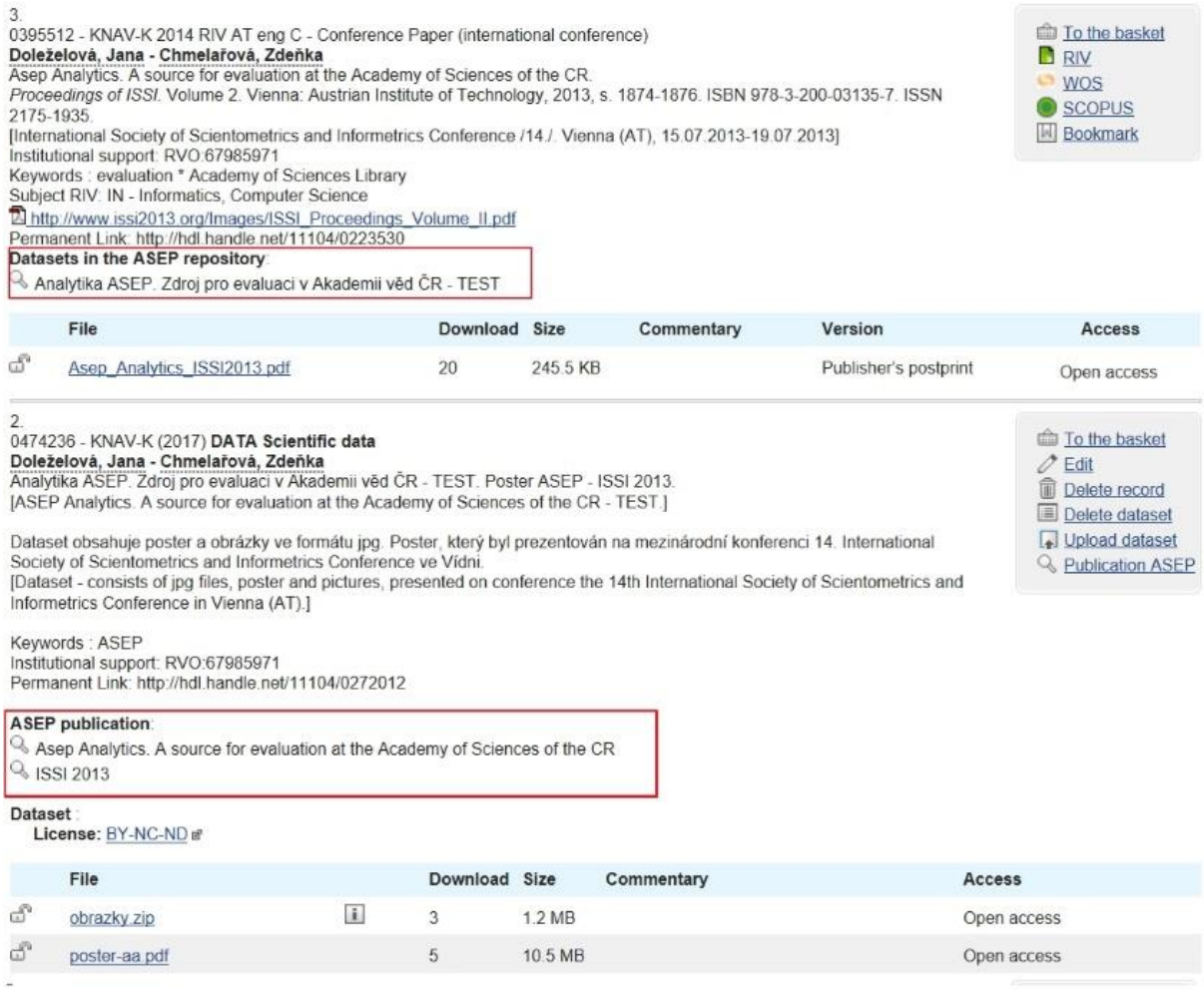
Links to publications and other scientific results (patents, applied research) that relate to the data can be entered in a data record and in the same way links to data records can be entered in bibliographical records. Figure 3 shows the interconnection of data and bibliographical records in ASEP. A data file may be attached to a data record (we favour this method), but we also make it possible for authors who have their data in, for example, an area-specific repository to create only a data record in ASEP with a link to the other repository or storage site. This might be useful in the case that such a repository does not allow the entry of metadata in the required format or to the required extent. Data is cited in much the same way as are bibliographical records, although this is not yet entirely common. Information on how the relevant data set is to be cited is available for each data record. Bibliographical citation in the ASEP database is governed by ČSN ISO 690 standard. For unspecified sources, including data files, the standard provides general rules of citation. Different citation styles and practices are used in data repositories and there is no uniform approach. We plan to introduce

[20] ORCID: **https://orcid.org/**

[21] Office of the Government of the Czech Republic. *R&D Information System 2.0* [online]. Prague: Office of the Government of the Czech Republic, ©2016-2017. Most recent update 19.9.2017 [cit. 26.9.2017]. Available from: **https://www.rvvi.cz/**

[22] HANDLE: **https://www.handle.net/**

the Citace.com service in the future to provide a number of options of how to cite data (APA, Harvard, Chicago, etc.).



Picture 3: The example of link from data record to publication record

Outlook

- The bibliographical records that have a full text stored in ASEP are regularly harvested for the OpenAIRE international database using OAI-PMH and in the future we are also counting on transferring data records.
- We want to include the data repository in the Re3d register of scientific data repositories.
- Another issue we wish to concentrate on is that of large data files, their storage and long-term protection using, for example, the CESNET[23] storage site. We can take inspiration from the Lindat/Clarin repository, which is also designed for storing large data sets, and archiving large language data is ensured in cooperation with CESNET[24].

[23] ANTOŠ, David. *Způsoby využití datových úložišť CESNET aneb čekání na velká data* [online]. CESNET, 2014 [cit. 2017-9-26]. Available from: **https://www.cesnet.cz/wp-content/uploads/2014/10/CESNET_Datova-uloziste.pdf**

[24] HAJIČ, Jan. *LINDAT/CLARIN* [online]. Brno: Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles' University, 2014. 26.11.2014 [cit. 26.9.2017]. Available from: **https://www.cesnet.cz/wp-content/uploads/2014/10/LINDAT-CLARIN.pdf**

- We will endeavour to comply with the conditions of exporting data records to the Data Citation Index, the database at WOS, which monitors the citation count of research data.
- We would like to obtain certification as a reliable repository. A certificate is not simply a formal document that the repository complies with the required criteria – it is a tool with which to check the proper functioning of the repository.

Conclusion

A basic data structure is defined in the ASEP data repository that is based on international standards, a system of links is in place between the data and bibliographical records stored in ASEP and between data and bibliographical records stored in other systems. Depositors are able to store data sets provided with metadata or create data sets with respect to data stored at other storage sites. The CAS has an open system that can easily be modified and enlarged as required. The base is in place and we will modify and broaden this according to the practical experiences of users and offer new functions which the Library considers important. We will, in the forthcoming period, familiarise scientists with the system of storing and describing data sets in ASEP and will also listen, so that we are able to find an intersection point between the needs of scientists and the ideas of system administrators.

References

ANTOŠ, David. *Způsoby využití datových úložišť CESNET aneb čekání na velká data* [online]. Praha: CESNET, 2014 [Accessed 26 September 2017]. Available from: **https://www.cesnet.cz/wp-content/uploads/2014/10/CESNET_Datova-uloziste.pdf**

HAJIČ, Jan. *LINDAT/CLARIN* [online]. Brno: Ústav formální a aplikované lingvistiky MFF UK, 2014. 26.11.2014 [Accessed 26 September 2017]. Available from: **https://www.cesnet.cz/wp-content/uploads/2014/10/LINDAT-CLARIN.pdf**

HRABAL, Jan. Repozitáře vědeckých dat. In: *Knihovna.cz* [online]. Brno: KISK FF MUNI, 2013. 22. 2. 2016 [Accessed 26 September 2017]. Available from: **http://ltp.knihovna.cz/?p=385**

HRABAL, Jan, HRUŠKA, Zdeněk. Úvod do problematiky dlouhodobé ochrany digitálních dokumentů – díl 2. In: *Knihovna.cz* [online]. Brno: KISK FF MUNI, 2013. [Accessed 26 September 2017]. Available from: **http://ltp.knihovna.cz/?p=249**

HRUŠKA, Zdeněk. Audit digitálních repozitářů. *Duha* [online]. 2013, **27**(4) [cit. 2017-09-26]. ISSN 1804-4255. Available from: **http://duha.mzk.cz/clanky/audit-digitalnich-repozitaru**

JAROLÍMKOVÁ, Adéla. Výzkumná data na Univerzitě Karlově. In: *INFORUM 2017: 23. ročník konference o profesionálních informačních zdrojích, Praha, 30.-31.5.2016* [online]. Praha: AiP, 2016 [Accessed 26 September 2017]. ISSN 1801-2213. Available from: **http://www.inforum.cz/pdf/2017/jarolimkova-adela.pdf**

MUULI, Viktor. *Research Data in Estonia: collecting, storing, availibility: some findings from questionnaire* [online]. Estonian Research Council, 2014. 23.10.2014 [Accessed 26 September 2017]. Available from: **http://dspace.ut.ee/bitstream/handle/10062/44052/RD_questionnaire_eng_muuli_14.pdf?sequence=1&isAllowed=y**

MYŠKA, Matěj, KYNCL, Libor, POLČÁK, Radim a ŠAVELKA, Jaromír. *Veřejné licence v České republice* [online]. Brno: Masarykova univerzita. 2012 [Accessed 26 September 2017]. ISBN: 978-80-263-0344-2. Available from: **https://is.muni.cz/www/102870/Prirucka.pdf**

ROSENTHAL, Colin, BLEKINGE-RASMUSSEN, Asger, HUTAŘ, Jan a kol. *Průvodce plánem důvěryhodného digitálního repozitáře (PLATTER)* [online]. Praha: Národní knihovna ČR, 2009 [Accessed 26 September 2017]. ISBN 978-80-7050-569-4. Available from: **http://www.ndk.cz/platter-cz/Platter.pdf**