Sustainability through the liaison with data archive users

Michaela Kudrnáčová[1] and Ilona Trtíková[2]

Abstract

As a social science data archive, we focus on collecting and archiving research data. However, there are more responsibilities that come with data archiving: cooperation on international social surveys (ISSP, ESS), supporting secondary data analysis, and much more. A significant part of our work is to communicate with students and researchers, and to educate them about data management and data analysis. Although the relationship we have is functional and seems sufficient, we tend to ask ourselves: who are the data archive users and what do they expect from us?

We decided to employ user-centered design methods and tools to define a typical user of our services, find their motivations for using our data archive and the specific functions they do (or do not) use and appreciate, and thereby get a better understanding of their needs. Moreover, we wondered about the role of open science and its impact on the users' needs and future requirements arising from the open science environment. The information obtained is a starting point for redesigning archival services to satisfy new demands our users have regarding more data resources, new techniques for scientific work, and better interconnection between different platforms.

Keywords

CSDA, data archive, open science, social science, user, survey, user-centered design

Kudrnáčová, Michaela and Ilona Trtíková (2020) Sustainability through the liaison with data archive users, *IASSIST Quarterly* 44(4), pp. 1-13. DOI: https://doi.org/10.29173/iq976

1. Introduction

'The Czech Social Science Data Archive (CSDA) at the Institute of Sociology of the Czech Academy of Sciences accesses, processes, documents and stores data files from social science research projects and promotes their dissemination to make them widely available for secondary use in academic research and for educational purposes.' (CSDA, 2020)

In order to support these primal activities, we occasionally come up with other activities and events such as seminars and workshops. Due to the endeavor of constantly developing archive services, we decided to focus on understanding the needs of users: their perceived needs and how we, as a data archive, manage to meet these needs and what should be considered for their future needs. Unfortunately, only a handful of studies focus on data consumers (Borgman *et al.*, 2015; Late and Kekäläinen, 2020). Moreover, the cited articles only describe users' backgrounds and the characteristics of the downloaded data. Our ambition is to find out more about their interests and preferences, establish closer contacts, and involve our data consumers in the process of data archive development. We believe this topic is utterly important because the users are the key motivation that leads to building a sustainable data culture, data management and archiving, and making the overall concept work.

This paper is organized as follows: firstly, we focus on the role of the data archive, stressing the general activities in which it is currently engaged and the need to promote and support open science; we then move on to the idea of applying user-centered design within social science data archives. The core of the paper is dedicated to CSDA users: what we already know about them and the description of a user survey and its methodology. Thanks to the data we collected, we can describe 'the typical user' within a separate subchapter. Finally, we will discuss the results of our survey, debate the employment of user-centered design in the context of the digital data archive, and determine if the chosen approach can sustainably ensure the effective functioning of the data archive.

2. The Role of Data Archive

The Czech Social Science Data Archive, a national resource center for social science research, acquires, processes, documents, archives, and preserves digital datasets from Czech and international social research and makes these data publicly available for two purposes: (1) secondary analysis in academic research and (2) training purposes at higher education institutions. At the same time, CSDA serves as a Czech node within the pan-European distributed research infrastructure CESSDA ERIC (Consortium of European Social Science Data Archives European Research Infrastructure Consortium) and is the CESSDA ERIC Service Provider in the Czech Republic (CSDA, 2020).

The data archive responds to changes in scientific work and supports open science. Its data services are based on the FAIR data principles, emphasizing the re-use of data in academic research. It documents and processes data for purposes of secondary use, connects it with relevant research information and contextualizes it with other data and materials. At the same time, it is a source of research tools and procedures that have been verified in prior research, thus supporting the implementation of new surveys. Moreover, it supports the use of secondary data analysis in research by: (1) providing training courses and taking part in educational programs at universities; (2) mapping and analyzing available data sources, providing information services, and user support on data sources; and (3) connecting Czech and international data sources and research in the field of data standardization and harmonization (CSDA, 2020).

The data archive also promotes open science, which aims to open the entire research cycle, encouraging sharing and collaboration. It is based on collaboration and new ways of disseminating knowledge using digital technologies and collaboration tools (INITIATIVE, 2014). The data archive can be considered an essential open science service. The archive role in open science is to create a trusted and sustainable environment for social science data. The goal is, therefore, to remove any obstacles preventing the sharing of all types of outputs at any stage of the research process. It turns out that data sharing is highly dependent on the behavior of scientists (Koltay, 2017). This is evidenced by various research, such as the Data Literacy Multinational Study questionnaire at Charles University in the Czech Republic. The survey showed that social scientists are less willing to share their data and are more concerned about potential barriers (Jarolimkova and Drobikova, 2019). These findings correlate with other research around the world (Tenopir *et al.*, 2011; Kim and Adler, 2015) and also, to some extent, with our experience.

As portrayed in the surveys mentioned in the previous paragraph, it is important to communicate with users to break down potential barriers. Since the researchers are not only contributors but also data consumers, it is essential to ensure they feel support and concern from the digital data archive. Therefore, it is desirable that archive services correspond to user needs. User-centered design can be the right way to design services that increase archive reliability and motivate users to collaborate through regular data sharing.

3. User-centered design

User-centered design is the process by which end-users have an impact on design services. (Abras, Maloney-Krichmar and Preece, 2004) This process is comprised of a set of methods is used to develop applications and websites. The aim is to give users a quick orientation and thereby increase the use of services and websites. It is not limited to a specific area; the methods are universal for use in any development or redesign. Several methods can be employed. Designers and service designers choose appropriate methods, according to the particular situation, to meet the goal. The advantage of user-centered design is that a future user is considered from the beginning of the design (Abras, Maloney-Krichmar and Preece, 2004).

First of all, we have to understand who our users are and what they desire (Garrett, 2010b). The field of user research is devoted to collecting the data needed to develop that understanding. This design serves as a tool to define the users and to respond better to their demands and needs. It is necessary to collect data regarding what users want and need, mapping their experience in user-designed environments and their motivations to use specific applications or websites, and explore what makes it possible for them to use these features effectively versus which obstacles stand in the way of successful use. Various methods, including analysis of web site visits, surveys, interviews, use cases, observations and more, can be used to collect data (Still and Crane, 2017).

An important step is to divide our users into smaller groups defined by key characteristics (Garrett, 2010a). By creating these groups, we can then identify those groups that represent the user population and are best to work with when designing an application or a website. On the basis of information about user groups, we define typical users or personas with more detailed personal characteristics (Garrett, 2010b). This helps to gain a better image of our users and will be useful for developers in tailoring the resulting design to the target users, in accordance with their needs, behaviors, and personal characteristics.

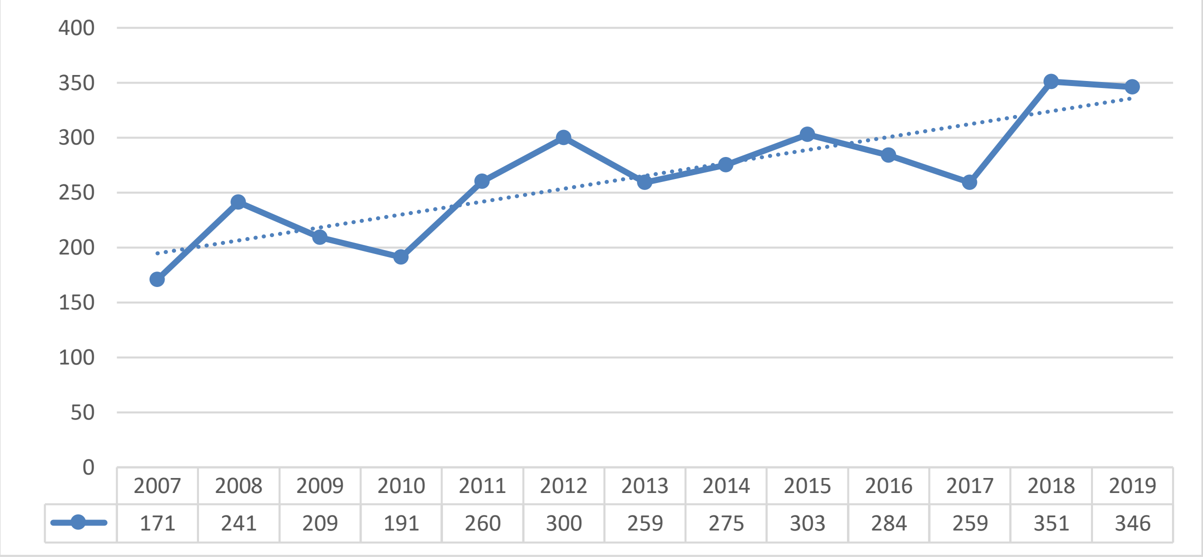
4. Who are CSDA users?

CSDA are at the beginning of the process of changing the system for managing and publishing research data. This is driven by several factors, including the currently used Nesstar system's lack of suitability for data security and, in our view, its limited user friendliness. However, beyond our personal perceptions, we believe it is necessary to involve the users in the process to get a better idea of their key issue and their perceptions of the current data archive's functionality and services. We have accordingly implemented methods of user centered design. To define and understand our users and profile typical users, we have chosen to conduct a survey among existing archive users and interviews with selected users. Moreover, we already have existing data at our disposal (see details in the chapter down below) and operational experience that have further helped us define the user.

We have two different sources from which we can draw conclusions. The first is the registration form our users need to fill in while signing up, and the other is a short survey we conducted at the beginning of 2020. Before addressing the survey, we will briefly reflect on some basic statistical measures that we are able to extract on existing data from registration forms from the year 2019.

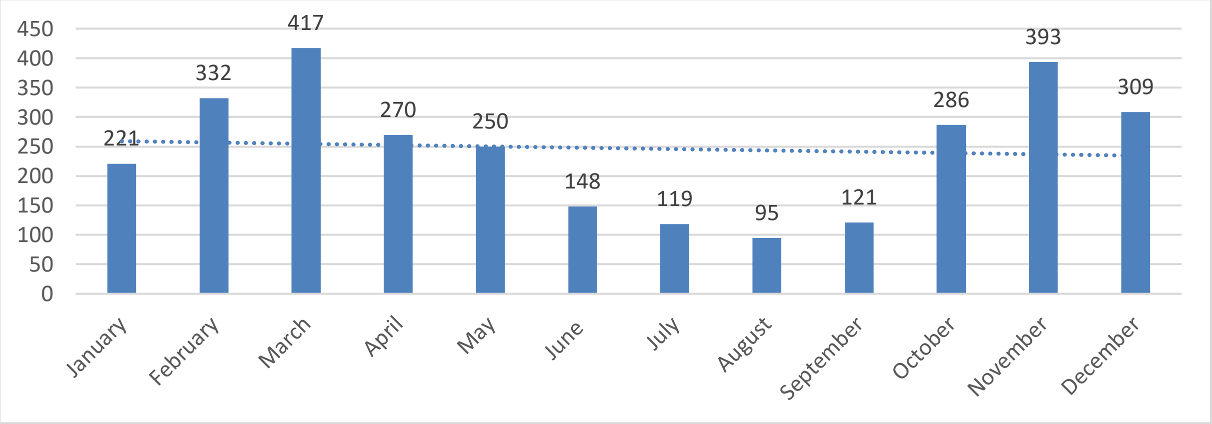
Since the existence of CSDA dates from March 2007, the first figure (Figure 1) shows a trend of growth. Allowing for annual fluctuations, the number of registered users has increased over time. Unfortunately, we can only guess the causes of the fluctuations. Since the users are mostly students, reasons may include variations over time in the effectiveness with which they share information about the archive and varying numbers of students enrolled in data management courses.

Figure 1: The number of users registered each year since its founding in 2007 until 2019



Examining the figure below (Figure 2), we can see the frequency of registrations is highest in March. This is most likely to be caused by students and their need to work with data for the purposes of assignments at the end of spring semester and due to the work on students' theses. Also, as mentioned earlier, this might be also caused by students being informed about the archive especially at the beginning of the courses in which they enroll. From June through August, there is a steady decrease in registrations, since most of the students have some time off, as is true for teachers and other academics on holidays. There is again an increase in October when the autumn semester begins. This is due to the same reasons as in the spring semester.

Figure 2: The average number of registrations since 2008 until 2019



Note: The numbers are averages from each month from 2008 through 2019. 2007 was omitted because, as it is the year in which the archive was founded, some months would be missing and the whole year would be an outlier in comparison to all the other years.

Based on the information we get from new users when registering, we are able to roughly describe the users. For that purpose, we chose the most recent year at our disposal: 2019. During this year, a total of 346 people registered into our archive. About 55% stated studies as the main reason for registering, 19% academic research and almost 16% teaching. Less common reasons included postgraduate studies, individual interest, public sector and other. As for the affiliation, two thirds were affiliated with universities (Charles University - about 30%, Masaryk University - about 21%, Palacky University Olomouc - about 11%), with students adding that they are mostly from faculties of arts or faculties of social sciences, leading us to believe it is mostly students/employees from the field of Sociology, Political studies, Social studies and Police academies. The last piece of information worth mentioning involves the country of origin: 95% of the users came from the Czech Republic, while others were from Austria, China, France, Germany, and other countries.

4.1. Survey Methodology

To find out more about the CSDA users, it was decided to conduct a short online anonymous survey. Overall, 3 398 people registered with our database were approached via their email address; 312 came back as "undeliverable", 6 were manually excluded (due to death, terminated employment, or to their request to be excluded from the survey), and an additional 174 respondents deregistered themselves because they did not wish to be contacted in connection with the survey. Over the period from the 6th of January until the 10th of February 2020, 564 individuals opened the survey, of whom 263 (46%) filled it in, representing 7.7% of the 3 398 registered users. The survey was conducted in Czech language only, since most of our users is of Czech nationality (about 90% in total); however, in the future, we aim to cover English speaking individuals as well.

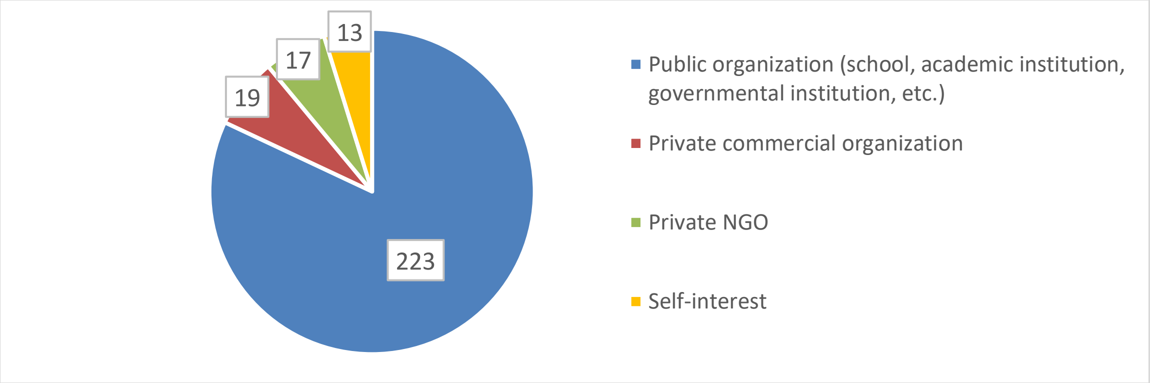
The survey consists of 10 brief questions developed to find out the usage frequency of the data archive, users' affiliations, the types of activities they perform through the archive, the Nesstar functions they use, citation practices, other data archives they use, and suggestions and complaints they have (this final question includes a space for expressing their wishes for improvements and other functionalities).

Within the survey, we focused solely on information relevant for CSDA functioning and therefore, we omitted questions regarding respondents' age, sex/gender and other sociodemographic information.

4.2. Survey Results

One question, included to help us understand respondents' backgrounds, referred to the affiliation of respondents (Figure 3). Most of the users (82%) who responded to the survey come from some sort of public organization, while the rest are comprised of people working in commercial organizations (7%) and NGOs (6.3%). Interestingly, 13 users (4.8%) stated "self-interest".

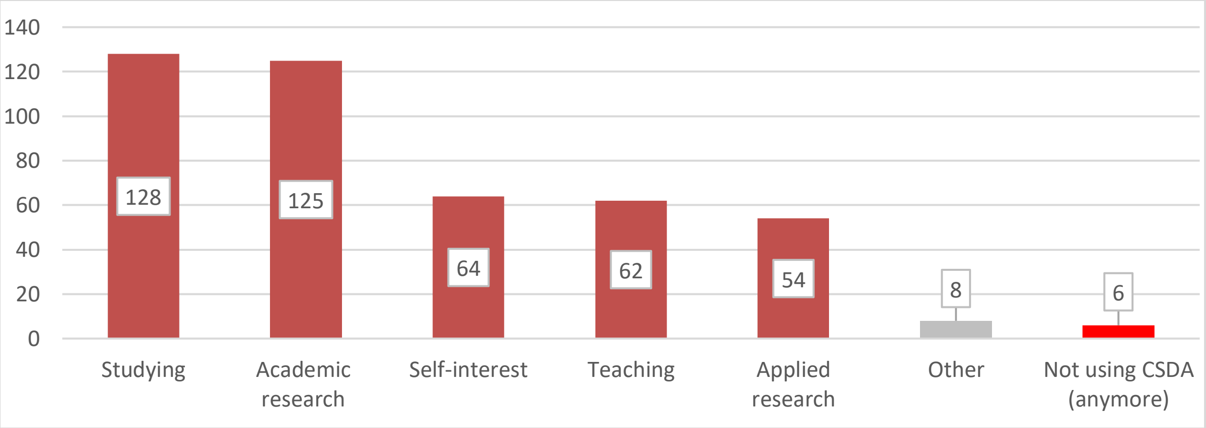
Figure 3: 'Under which affiliation do you use CSDA?'



n=263

A question regarding the activities for which respondents needed data from CSDA (Figure 4) allowed respondents to select more than one response: this revealed studying (28.6%) and academic research (28%) as the dominant activities. However, self-interest in the data (14.8%), teaching (14.3%) and applied research (12.5%) were also significant responses. The other two options were chosen by only a couple of respondents (“other” by 8 and “not using CSDA anymore” by 6). It is worth mentioning that among “other”, policy making, journalism and advocacy were mentioned.

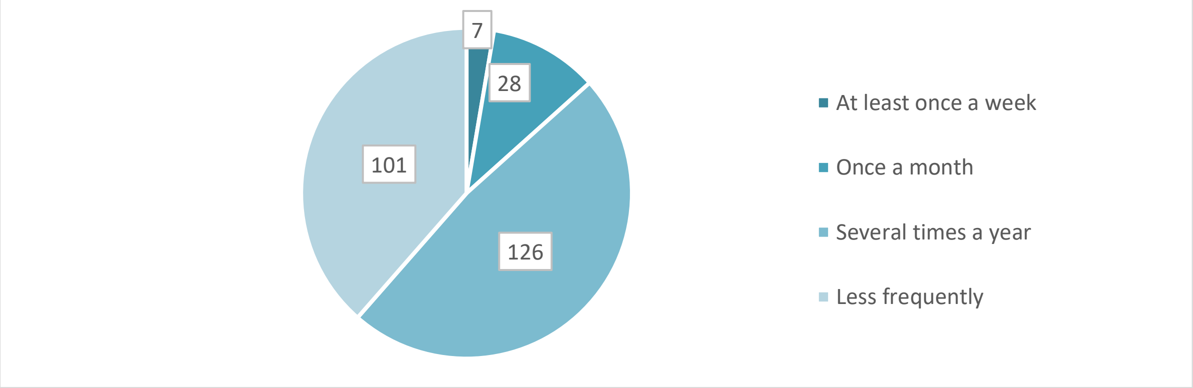
Figure 4: 'For what purpose do you use CSDA?'



n=447

The next question in the survey concerned the frequency of use of our data archive (Figure 5). There seems to be quite a variation: almost half of the respondents stated they use the data archive “several times a year” (48.1%), slightly more than one third claimed to use it “less frequently” than once a week (38.5%), 10.7% use it “once a month” and only 7 respondents (2.7%) use the archive “at least once a week”.

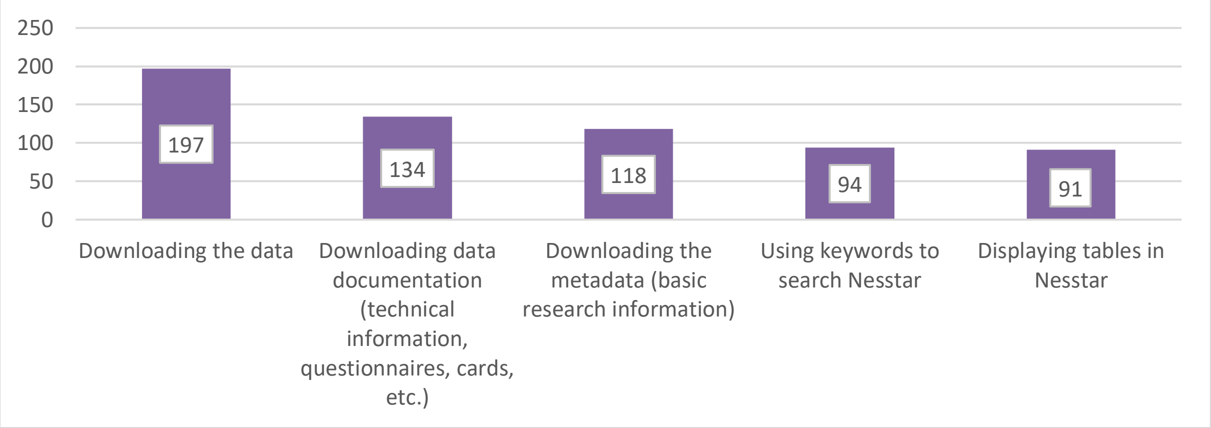
Figure 5: 'How often do you use CSDA?'



n=262

It was also crucial to extract from the users what functions they need and use (Figure 6). Almost one third of the respondents (31.3%) were "downloading the data", while 21.1% were "downloading data documentation" and 18.6% were "downloading the metadata". The last two options ("using keywords to search Nesstar" and "displaying tables in Nesstar") totaled around 15% of answers.

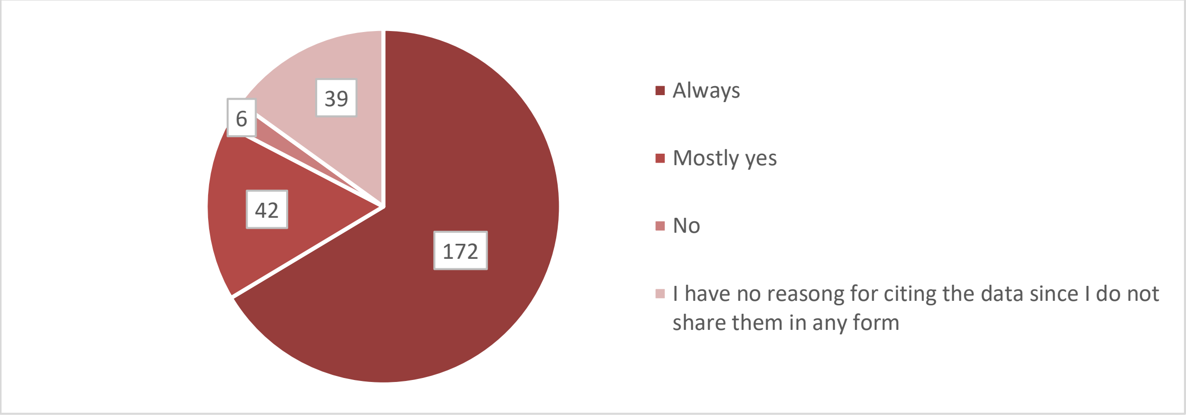
Figure 6: 'What Nesstar functions do you use?'



n=634

Since the data archives have long been struggling with students and researchers not citing the data or not citing them properly, we decided to also address this in the survey (Figure 7). About two thirds of respondents (66.4%) stated they "always" cite used research data and 16.2% admit "mostly" citing the data. As opposed to that, only 6 respondents said they do not cite the data in any form at all and, last but not least, 15.1% stated they do not cite the data simply because, since they do not use or share the data publicly or in class, they have no reason or platform for citing it. Although there is no way for us to know if the users are indeed citing the data, we are at least able to say that this outcome shows they realize the norm (i.e., they should be citing the data regardless of whether they actually reference used data).

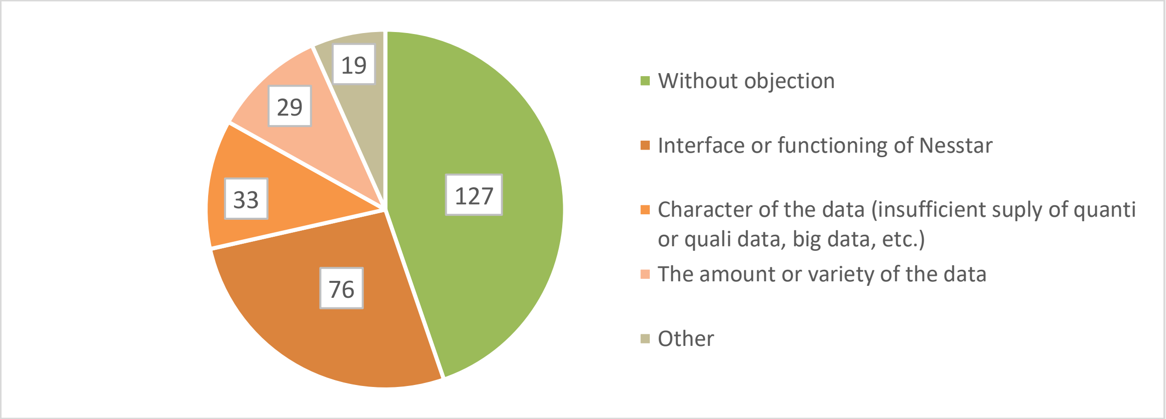
Figure 7: 'Do you cite used data? Either in written form (writing a paper) or oral form (for students during teaching classes)



n=259

As seen in Figure 8, slightly under half (44.7%) of respondents have no objections to the archive's or Nesstar's functionality. Weaknesses identified by the remaining respondents can be divided into three main groups, of which it seems the most burning issue is the "interface or Nesstar functioning" (26.8%), followed by the "character of the data" (11.6%), and 10.2% wished for the amount of data and variety of the data to be broader in general. Those citing the "character of the data" suggested that the users would like to have more quantitative data at their disposal, qualitative data and new types of data. Among "other" desires (19 responses) for the archive, examples of specific complaints and wishes included the ordering of the data, conjoint datasets, Nesstar failures, etc.

Figure 8: 'What do you find insufficient about the archive and Nesstar functioning?'



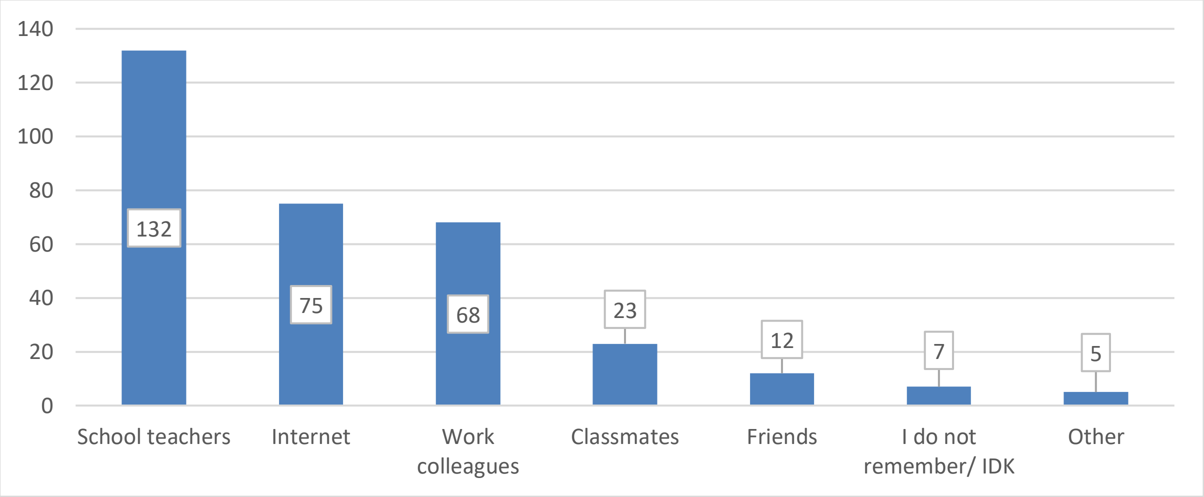
n=284

Moreover, as for the Nesstar functioning, a separate open-ended question gave respondents a chance to express precisely and more extensively their feelings about Nesstar and its areas of insufficiency. We were stunned to find that 89 open-ended responses were submitted. Most (41) referred to the interface and its chaotic arrangement, while 13 respondents commented on the functioning of the Nesstar. Interestingly, 23 responses involved wishes and interesting ideas for new features. The last 12 responses were generally positive, often mentioning similarity to other world-wide archives and stressing their "eventually finding what they need".

The survey also asked about respondents' use of other data archives ('Do you use other data archives apart from CSDA for data viewing or downloading?'). Based upon this, 53.8% of 251 respondents do not use other data archives; the rest do and mention data archives such as GESIS, Eurostat, ČSÚ and many more.

Finally, but still importantly, we were interested in who brought the users to our archive (Figure 9). Close to half of responses (41 %) suggest the users encountered CSDA thanks to their school/university teachers, almost one fourth (23.3%) learned of CSDA through the Internet, and about one fifth (21.1%) responded with the option "work colleagues". Only a small proportion of answers suggest classmates (7.1%) and friends (3.7%). Among the "other" option (five responses) were presentations, seminars and scientific literature. Seven respondents did not remember.

Figure 9: 'How did you find out about CSDA?'



n = 257

The survey concluded with an open-ended question allowing the respondent to express anything regarding the data archive that was not addressed earlier in the survey. Examples of 11 areas mentioned included the lack of localization in both Czech and English, the arrangement of surveys, large-scale downloading, and design. Moreover, we were very pleased and touched to get 20 positive comments, mostly simply thanking us for the data archive's services and its existence.

4.3. Discussion: Defining Typical User

Defining a typical user helps us to grasp our target group. The data we already had from our registration forms are more informative when combined with the user survey results. This is especially true since the survey results are not simply descriptive but were specifically designed to involve the data consumer in the data archive development and give them an opportunity to express their very own preferences in the matter.

The CSDA seems to be mostly dealing with and serving people from academia: students, academics, and teachers. It makes sense that they tend to use the data from the archive for purposes of studying and academic research, as well as occasionally for teaching. This also explains why they usually download the data at least several times a year or even once a month: academics need to write articles and books, while students need to write theses or seminar papers. Teachers might only be teaching, or their role may be broader since academics and students on a certain level of their studies have to teach as well[1].

Due to this fact, we believe there are two specific types of users that are generally similar: academic and student. Beyond their specific characteristics described at the very beginning of this subchapter, there is a bit more worth mentioning. Both academics and students state that they always or almost always cite the data. A small part of the students, though, say that they have no reason to cite the data, meaning they could be downloading the data for inspiration while conducting their own research. As for usage of Nesstar functions, it does not differ across the whole sample of respondents: academics and students mostly download the data and documentation, but generally use all offered functions. Academics and students generally concur in their objections against Nesstar, with students slightly more eager for greater quantities, variety, and types of data at their disposal. Interestingly, with regard to the usage of other data archives, two-thirds of the students within our sample (about 80 people) use no other archives, while two-thirds of academics (about 75 people) do use other archives.

These results are of a great importance to us. Our users are genuinely interested in our data archive development and clearly know their wishes for data services. Taken by themselves, the transaction log files, or registration forms would not have provided this precious information.

Figure 10: Defining a typical user of CSDA - student and researcher

STUDENT	ACADEMIC
studying at a university needs to write thesis or seminar papers one-time need due to fulfilling school duties uses Nesstar functions beyond the documentation downloads data file is citing a dataset interested in other data does not use any other archives	university employee needs to write articles and books repeated need already has ways of using Nesstar focuses on search functions needs to search questions is citing a dataset does use other archives

5. Conclusion

Within the article, we discussed the role of the data archive, specifically the role of CSDA, and our efforts to adjust to the needs of the users. We mentioned the importance of employing user-centered

[1] It is important to know that these categories were constructed based on a question regarding user activity. A total of 42 people identified as both academic and teacher, while 45 identified as both student and academic, 80 only as academics, and 83 only as students. While talking about "specific types of users", we compare academic respondents and student respondents, meaning there are duplications within the group.

design along with making use of the data we already possess about our users. A survey was conducted in order to get more user data. Finally, thanks to the information we obtained about CSDA users, we were able to describe a typical user.

Contemporary science demands sharing and collaboration. Of course, this affects those systems that provide these functions. To ensure maximum support for open cooperation and the sharing of scientific data, we try to set up our service so that it best suits our users' needs. For this reason, we chose user-centered design methods that reflect the needs of our users. The survey that we conducted provided us with a lot of data. In particular, it made use of our users' willingness to cooperate in the development of our services. We have clarified and confirmed a number of data that can help us both in prioritizing the services offered and in communicating with users. Defining a typical user, the first of several phases of service design, is very important. We have acquired a great deal of information that we will use in the next stages of designing our services. We have also realized that we should put more emphasis on creating a detailed Nesstar user's manual for the students among our users, while also introducing them to other data archives around the world. To overcome existing restrictions on the use and sharing of research data, we will endeavor to remove possible obstacles in the system to publishing data and accessing services offered by our data archive. It turned out that we needed to help our users use data from other archives. The goal of our efforts is to simplify their involvement in the open science environment.

Lastly, we need to emphasize that more research is needed in this field. There is a knowledge gap regarding digital data archives and their purpose, along with knowledge of data consumers, that we intended to bridge. As is true for any business, data archives offer services, which is the reason why they should learn more about their target groups and their desires. We attempted to carry out this task differently and thereby managed to establish a greater level of contact with our users, through which we now know of their deep interest in taking part in the CSDA development. Therefore, we intend to keep in touch with our users, involving them in the next stages of user-centered design to achieve well-targeted communication, website design, data cataloging and events. To verify the continuing validity of our typical user profiles, we will need to conduct a short online survey every two years, as well as a couple of in-depth interviews. This periodic updating of the user profile will promote the long-term sustainability of our user-centered design efforts.

References

Abras, C., Maloney-Krichmar, D. and Preece, J. (2004) ‘User-Centered Design’, *Encyclopedia of Human-Computer Interaction*. Sage Publications.

Borgman, C. L. *et al.* (2015) ‘Who uses the digital data archive? An exploratory study of DANS’, *Proceedings of the Association for Information Science and Technology*, 52(1), pp. 1–4. DOI: https://doi.org/10.1002/pra2.2015.145052010096.

CSDA (2020) *About the Czech Social Science Data Archive*. Available at: https://archiv.soc.cas.cz/en/about-czech-social-science-data-archive (Accessed: 17 March 2020).

Garrett, J. J. (2010a) *The Elements of User Experience: User-Centered Design for the Web and Beyond*. 2 edition. Berkeley: New Riders.

Garrett, J. J. (2010b) *The Elements of User Experience: User-Centered Design for the Web and Beyond (2nd Edition) (Voices That Matter)*, *Elements*.

INITIATIVE, O. S. A. R. (2014) *Open Science and Research: The Open Science and Research Handbook*, *Open Science and Research: The Open Science and Research Handbook*. INITIATIVE, OPEN SCIENCE AND RESEARCH. Available at: https://www.fosteropenscience.eu/sites/default/files/original/3986.pdf.

Jarolimkova, A. and Drobikova, B. (2019) ‘Data Sharing in Social Sciences: Case Study on Charles University’, in *Communications in Computer and Information Science*. DOI: https://doi.org/10.1007/978-3-030-13472-3_52.

Kim, Y. and Adler, M. (2015) ‘Social scientists’ data sharing behaviors: Investigating the roles of individual motivations, institutional pressures, and data repositories’, *International Journal of Information Management*. Elsevier Ltd, 35(4), pp. 408–418. DOI: https://doi.org/10.1016/j.ijinfomgt.2015.04.007.

Koltay, T. (2017) ‘Data literacy for researchers and data librarians’, *Journal of Librarianship and Information Science*. DOI: https://doi.org/10.1177%2F0961000615616450.

Late, E. and Kekäläinen, J. (2020) ‘Use and users of a social science research data archive’, *PloS one*, 15(8), p. e0233455. DOI: https://doi.org/10.1371/journal.pone.0233455.

Still, B. and Crane, K. (2017) *Fundamentals of User-Centered Design : a Practical Approach*. CRC Press.

Tenopir, C. *et al.* (2011) ‘Data Sharing by Scientists: Practices and Perceptions’, *PLoS ONE*. Edited by C. Neylon. Public Library of Science, 6(6), p. e21101. DOI: https://doi.org/10.1371/journal.pone.0021101.

[1] Michaela Kudrnáčová is a PhD student at the Social Science Data Archive focused on research and its methodology, and can be reached by email: michaela.kudrnacova@soc.cas.cz (2020).

[2] Ilona Trtíková works as data manager at the Social Science Data Archive, her expertise involves sharing and retrieving research information.