Research and Innovation Action

CESSDA Strengthening and Widening

Project Number: 674939 Start Date of Project: 01/08/2015 Duration: 27 months

Deliverable D4.5- Provision of development support services on the basis of identified demand

Dissemination Level	PU
Due Date of Deliverable	July 31, 2017
Actual Submission Date	July 31, 2017
Work Package	WP4 Strengthening and widening through knowledge exchange - Development of the necessary administrative, technical, and methodological support needed to establish and develop data archives
Task	T4.4
Type	Other
EC Approval Status	16 November 2017
Version	V1.0
Number of Pages	p.1 - p.74

Abstract: The objective of this task is to establish the conditions for creating new or reinforcing existing social science data services. The task will establish the needs for the provisioning of development support services for new and existing CESSDA Service Providers. Seven pilot studies of Service Provider to Service Provider support services are described that match categories of support activities and services defined. Conclusions are drawn from these studies with respect to the value of development support services.

The information in this document reflects only the author's views and the European Community is not liable for any use that may be made of the information contained therein. The information in this document is provided "as is" without guarantee or warranty of any kind, express or implied, including but not limited to the fitness of the information for a particular purpose. The user thereof uses the information at his/ her sole risk and liability.

Project funded by the EU Horizon 2020 Research and Innovation Programme under the agreement No.674939

History

Version	Date	Reason	Revised by
0.1	29.04.2017	Create	Mike Priddy
0.2	10.07.2017	Integration pilot reports	Mike Priddy / Marion Wittenberg / Vyascheslav Tykhonov
0.3	15.07.2017	Peer review	Maja Dolinar / Irena Vipavc Brvar / Trond Kvamme
0.4	18.07.2017	Document for review by MO	Marion Wittenberg

Author List

Organisation	Name	Contact Information
DANS	Mike Priddy	mike.priddy@dans.knaw.nl
DANS	Marion Wittenberg	marion.wittenberg@dans.knaw.nl
DANS	Vyascheslav Tykhonov	vyacheslav.tykhonov@dans.knaw.nl
ADP	Maja Dolinar	maja.dolinar@fdv.uni-lj.si
ADP	Irena Vipavc Brvar	Irena.vipavc@fdv.uni-lj.si
ČSDA	Jindřich Krejčí	jindrich.krejci@soc.cas.cz
FSD	Toni Sissala	toni.sissala@staff.uta.fi
IEN	Aleksandra Bradić-Martinović	abmartinovic@ien.bg.ac.rs
IEN	Goran Gicić	goran.gicic@ien.bg.ac.rs
FFZG	Alen Vodopijevec	alen@irb.hr
NSD	Trond Kvamme	trond.kvamme@nsd.no
TÁRKI	Péter Hegedűs	peter.hegedus@tarki.hu

Executive Summary

‘In the start-up and consolidation phases of new archives, established CESSDA archives will perform supportive tasks for starting data archives, especially offering technical facilities and expertise on software tools, backup services, and archival policies, for as long as trustworthy and sustainable services in the countries aspiring CESSDA membership are lacking. The supply and demand requirements will be matched and a selection of support services was to be implemented.’[1]

The objective of this task is to establish the conditions for creating new or reinforcing existing social science data services. The task will establish the needs for the provisioning of development support services for new and existing CESSDA Service Providers (SP). Seven pilot studies of Service Provider to Service Provider support services are described here. We have defined a framework of categories of support activities and services which have been used to ensure we have a range of development support services. Conclusions are drawn from these studies with respect to the value of development support services.

There were only very few existing support services, therefore ‘matchmaking’ between development support service (DSS) providers and consumers was not really a possibility. Moreover, partners present in the task did not necessarily need, or could supply, DSS suitable for a pilot study.

Three of the pilot studies were undertaken with CESSDA-SaW partners which were not in the task or even work package. We are grateful for their contribution and commitment to the success of this task and outcomes.

All of the pilot studies were bespoke activities tailored to the needs of the partners and many consisted of more than one service or activity, for example involving consultancy, software development, workflow development and technical support in installation and running of tools. It was clear that not all service providers are at the same capability, capacity or maturity level. It maybe such that packages of established and bespoke services and activities are required by SPs for rapid development to meet their obligations within CESSDA ERIC.

Pilot studies are only a starting point, and for a number of the pilots started further work is required to establish them as production solutions. The interest and involvement in this task indicate that there is a clear and pressing need for development support services for small, new, and aspiring SPs to help them meet their strategic goals of supporting their designated communities quickly and efficiently, as well as meeting the CESSDA Statutes Annex 2: Obligations of Service Providers[2].

There is considerable experience within the Service Provider community, and therefore the community is best placed to provide support services to others of the community. The

[1] See p.27 of Annex 1 (part A) of Grant Agreement 674939 - CESSDA-SaW

[2] https://www.cessda.eu/eng/content/download/316/2908/file/Annexes-to-Statutes-for-CESSDA-210213-Final-Version-brand.pdf

challenges are to understand what support is needed over time, how this SP-to-SP support can be provided, and how this can be maintained in the community of SPs and data archives. This will be further elaborated in the Deliverable D4.6: Report on sustainability model of development support services.

Abbreviations and Acronyms

ADP	Archiv družboslovnih podatkov
API	Application Programming Interface
CESSDA	Consortium of European Social Sciences Data Archives
CMM	CESSDA Metadata Model
CRIStin	Current Research Information System in Norway
CSDA	Czech Social Science Data Archive
DANS	Data Archiving and Networked Sciences
DDI	Data Documentation Initiative
DMP	Data Management Plan
DPO	Data Protection Official (Officer)
DSA	Data Seal of Approval
DSS	Development Support Services
EASY	Electronic Archiving System
EMD	EASY metadata
ERIC	European Research Infrastructure Consortium
FAIR	Findable, Accessible, Interoperable, & Reusable
FFZG	Sveuciliste u Zagrebu, Filozofski fakultet
FSD	Finnish Social Science Data Archive
GDPR	General Data Protection Regulation
HTTP	Hypertext Transfer Protocol
ICPSR	Inter-university Consortium for Political and Social Research
IEN	Institut ekonomskih nauka
IQSS	Harvard Institute for Quantitative Social Science
IRB	Institute Ruđer Bošković
ISSP	International Social Survey Programme
JSON	JavaScript Object Notation
LTP	Long-term Preservation

METS	Metadata Encoding And Transmission Standard
MODS	Metadata Object Description Schema
NORDi	Norwegian Open Research Data Infrastructure
NSD	Norsk samfunnsvitenskapelig datatjeneste AS
OAI:DC	Open Archives Initiative - Dublin Core
OAI-PMH	Open Archives Initiative - Protocol for Metadata Harvesting
ORCID	Open Researcher and Contributor ID
OS	Operating System
OSMH	Open Source Metadata Harvester
PID	Persistent Identifier
PaSC	Product and Services Catalogue
RCN	Research Council of Norway
REST	Representational State Transfer
SaW	Strengthening and Widening
SLA	Service Level Agreement
SP	Service Provider
SWORD	Simple Web-service Offering Repository Deposit
TÁRKI	TÁRKI Foundation
TDR	Trusted Digital Repository
UKDA	UK Data Archive
VRE	Virtual Research Environment
XML	eXtensible Markup Language
XSLT	eXtensible Stylesheet Language Transformation

Table of Contents

Introduction

The objective of this task is to establish the conditions for creating new or reinforcing existing social science data services. The task will establish the needs for the provisioning of development support services (DSS) for new and existing CESSDA service providers (SP). Although it was assumed that established SPs would provide development support services to new and aspiring service providers, in start-up and consolidation phases of development, this is not necessarily always the case. It is feasible that expertise, experience, and support may exist in any SP. Therefore, we took the view that any SP could be the provider as well as the recipient of development support services.

A further objective was to identify supply and demand requirements and make a selection of support services which needed to be implemented. Identification of supply and demand was achieved through two workshops.

A wide range of pilots was undertaken with both task members and other partners within CESSDA SaW, since the interest in development support services was greater than expected. These are described herein.

Purpose of this Report

Pilots by their nature are experimental and short term, however by selecting a range of pilots the aim is to provide insight for deliverable D4.6: Report on Sustainability Model of Development Support Services, and the understanding of how such pilots and future DSS can be financed and maintained for as long as they are required.

This report describes the work carried out on each pilot study by the partners involved.

Audience for this Report

The primary audience for this report is internal, and specifically the CESSDA SP Forum, individual new and aspiring service providers, CESSDA Main Office and Board of Directors, and possibly also the CESSDA General Assembly.

Others maybe interested in the categorisation of the development support services and the approach to establishing many short pilot studies.

The outcomes and conclusions of the pilot studies may be valuable for those service providers looking for examples of services they wish to implement.

Background

In the CESSDA statues[3] Annex 2: Obligations of Service Providers it states that "All Members and Observers must appoint Service Provider(s) able to meet these obligations:", one of which is: "11. provide member support for countries with immature and fragile national infrastructures to help them build up needed competence later to be able to fulfil tasks as Members." This obligation has directly lead to this specific task and deliverables in CESSDA-SaW Work Package 4.

Rather than this being *ad hoc.* and based upon a network of relationships between SPs & friendships between international colleagues, it should be possible to deliver consistent, and at least repeatable, development support as the CESSDA SP community grows and matures.

In this deliverable, we attempt to categorise development support services so that we can ensure we have a wide coverage of typical DSS in the pilot studies of this task.

Categorisation of Development Support Services

Support provided, formally or ad hoc, can be categorised as either a continuous service or a discrete activity[4]. Activities are discrete tasks, actions or events that have a defined start, action, and a completion. An activity may be repeated but each instance is discreet.

A continuous service, as the term suggests, is an open-ended service that is usually governed by a Service Level Agreement (SLA). The service can be offered to more than one SP customer at a time. The service is usually terminated by one or both parties involved, however, if it is a core service for the consumer then alternative arrangements must be arranged prior to termination of the agreement.

Both continuous services and activities can involve both human and/or machine agents as providers and consumers.

In tables 1 & 2 we categorise activities and services to ensure we cover a range of provisions of DSS so that these can be evaluated in this deliverable and deliverable D4.6.

Service	Examples
Core Technical Support Services	Persistent Identifier provision (minting), Checksum validation, virus scanning.
Content hosting	Back-up, dark archiving, off-site backup, VRE hosting

[3] At the time of writing of the proposal the 2014 Statutes for CESSDA Annexes were still in operation. https://www.cessda.eu/eng/content/download/316/2908/file/Annexes-to-Statutes-for-CESSDA-210213-Final-Version-brand.pdf

[4] See deliverable CESSDA-SaW deliverable D4.6 for a more indepth explanation. To be published.

Designated community services	Metadata validation, data quality checking, VRE services, help-desk, annotation service, tools.
Software & hardware provision and/or maintenance	Virtual test, acceptance and production machines, patching & maintaining operating system and core software,
Software support	Help-desk, support contract
Other	Other

Table 1: Types of continuous development support services

Activity	Examples
Consultancy (expertise)	Policy advice, accreditation advice, long-term preservation, metadata, workflow development, requirements analysis.
Software and tools (for local installation)	archive software, metadata publishing tools, metadata creation, thesaurus management tools, installation, and set-up.
Software development	Bespoke development of tool, or core infrastructure.
Training / Workshops / events & event management	Staff training, organising a conference
Educational resources / training materials / online courses	Online tutorial on OAIS model, resources to run own technical training course, webinars
Other	Other

Table 2: Types of development support activities

Identifying Needs for Support Services

To identify the needs and the provision of Support Services we organised a workshop as part of the overall CESSDA SaW "Training on Trust, Identifying Demand & Networking" workshop, held in The Hague, Netherlands on June 15th and 16th, 2016.

The aim of the workshop was:

1. to Identify needs from partner institutions for development support service;
2. to identify possible provider institutions for development support services;
3. to identify possible pilots of service provider to service provider development support services.

It was not the ambition of this workshop or of task 4.4 to create an exhaustive list of possible development support services or to identify all the needs of all service providers for such services, since not all service providers or aspiring service providers were present at the workshop. Therefore, the approach of aiding service providers was taken thus enabling them to identify their strengths, gaps and thus possible needs for support services.

This workshop identified both possible services to be delivered over time and specific activities provided discretely on separate occasions rather than over time. The main technical support service that was of interest was Dataverse, but OAI-PMH server, archival software (FORSBase) and back-up services were also discussed. Most of the interest in the final session centred around support activities such as expertise and in particular internships or site visits to other SPs, and perhaps with specific expertise support. Advocacy expertise was also discussed and offered.

It was agreed to pursue the following pilots:

- Dataverse pilot (lead by DANS) with possible involvement of NSD, IEN, TÁRKI Foundation, FFZG, ADP, and interest from SOHDA but they are already running Dataverse
- CSDA & TÁRKI will investigate data back-up as a service provision between institutions across borders.
- FSD will offer OAI-PMH server software
- Providing internships as well as a pilot of an expertise support activity, but the number and details are yet to be defined.

In a second workshop, organised in The Hague on 21st and 22nd of September, 2016, we further refined the needs for Dataverse into a number of smaller pilots, each with a different function for the SPs involved and with differing activity types.

Internships were not pursued due to the travel cost and staff involvement that were higher than budgeted for in the task. Expertise activities were included in a number of the Dataverse pilots. Furthermore, since training and policy advice were the subject of other tasks in work package 4 we decided not to pursue these types of support activities.

Pilots

Within this task, we have collectively undertaken seven pilot studies which are categorised as DSS types in table 3.

Pilot Study	Category of DSS	Example of
Hosting services for the geographic diversity of backups	Content hosting	Off-site reciprocal data backup
Kuha2 Metadata Server	Software Development	Bespoke development of tool
Dataverse and NSD-NORDi	Consultancy (expertise)	Requirements analysis

Remote Technical support for an installation of Dataverse	Software and tools (for local installation)	Installation and set-up
Self-archiving Tool for Researchers	Provision of test environment Consultancy (Expertise)	Requirements analysis
Dataverse for Dissemination - use case of TÁRKI Data Archive	Provision of test environment Consultancy (Expertise) Software Development	Workflow development Requirements analysis
Dataset Acquisition Utilising Dataverse in Tandem with Islandora for Long-term Preservation	Consultancy (Expertise), Software Development	Workflow development Bespoke development of core infrastructure

Table 3: Pilot study categorisation

Pilot 1: Hosting services for the geographic diversity of backups

Jindřich Krejčí (ČSDA), Péter Hegedűs (TÁRKI)

Introduction

The aim of this study is to explore the possibility of ensuring geographic diversity of backups in the digital data archives based on the reciprocal provision of hosting services in between two partner data organisations. The purpose for implementation of the geographical diversity of backups is to reduce the risks of data loss and thus increase the security of the preserved data. Such precaution is in line with recommendations and standards relevant to trustworthy digital data archives.

This study explores a case in which two independent research data organisations located in different countries provide each other their capacities to store backup copies of their digital data libraries. The solution described in this study is designed to avoid the direct financial costs and minimise the impact on operation and system changes in co-operating data organisations. It is demonstrated using an example of two national social science data archives collaborating in the network of pan-European distributed research infrastructure CESSDA. The collaborative environment within the broad CESSDA network and the authority of central structures of the CESSDA consortium facilitate proposed solution with a high level of mutual trust and organisational background.

This study explains why the geographical diversity of backups is useful, discusses the possibilities of different solutions and describes organizational, technical and legal issues of mutual provision of hosting services. The model service level agreement in between the two data organisations is in appendix 1.

Geographic distribution of backups

The risk management of data loss has to be among major issues of any digital data archiving policy. In this context, the geographic diversity of backups is inevitable data security measure for disaster recovery. The best practice is to store multiple copies in multiple locations (see, e.g., ICPSR 2012). The minimal requirement is to store three copies: (1) original, (2) external - local and (3) external - remote. The ideal characteristics for the external - remote backup copy are the geographic diversity and the independence of storage systems. That is why the geographic distribution of backups is a critical concept for data policies of trustworthy data archives[5].

The issue of backups location should not be neglected. It seems that the disasters such as the floods, fires or tornado, which put all on-site data facilities in danger are rare. However, the records from IT departments show a different picture. For example, the results of the IT Disaster Recovery & Business Continuity Survey (Evolve 2015), which is based on reports from 2,084 executive and IT professionals from U.S. companies shows following: (1) More than the one-third of responding companies claim to have suffered from at least one incident or outage that required disaster recovery. (2) The next leading cause of the incidents after hardware failure were environmental disasters such as flood, fire, ice storm, etc. (34 % of incidents).

In addition, with the geographic distribution the backups are usually stored within mutually independent hardware, software and organisational systems, thus increasing the security from fatal failures of these systems as well, including risks of hacker and virus attacks. Considering the recent events like WannaCry ransomware global attack[6], it is obvious that such risks are increasing and thus prudent backup policies are necessary.

There is a number of ways to manage the geographic distribution of backups. In a large organisation with more branches, it is often possible to use company's own servers located in different places. However, in such cases, the backups are not stored within fully independent systems. Another option is to choose one of the available commercial hosting services. There are different types of them, e.g. hosting of servers or cloud backup services. Of course, the commercial services are not for free even though the storage costs are continuously declining. Moreover, the quality and the configuration of those services should be considered also from the point of view of personal and other sensitive data protection.

In addition, an option of reciprocal provision of the hosting services between two organisations operating similar kinds of digital data libraries is worthwhile considering. CESSDA creates a suitable environment for such a collaboration. Long-term co-operation among the partners of CESSDA network, established organisational structures of the distributed research infrastructure, ongoing harmonisation processes and the common standards facilitate possible organisational and technical solutions. The CESSDA collaborative

[5] The Data Seal Of Approval (DSA 2016) certification for trustworthy digital data archives is among the basic requirement for the CESSDA ERIC Service Providers. Geographic diversity of backups is not strictly required by this certification, but applicants must provide details on their strategy on physical and logical security, failover, and business continuity including recovery procedures and prove that this strategy gives reasonable guaranties that data stored in the repository will not be lost (see DSA Guidelines 9 - Documented storage procedures and 16 – Security (DSA 2016)). Similarly also other criteria for trustworthy digital data archives (e.g., DIN 31644, ISO 16363) require prudent policies and plans for disaster preparedness, response, and recovery.

[6] http://www.bbc.com/news/technology-39901382

network is also a source of a relatively high level of trust among partners and the guarantees of a fair co-operation. Such an environment is a great advantage, which makes possible a low cost and undemanding arrangement of proposed reciprocal co-operation.

Methodology of the study

The possible settings of the system of reciprocal provision of hosting services were verified by the examination of needs, available capacities and conditions at two CESSDA partner data archives, which are national service providers. Interviews with IT specialists, archival system administrators, and the managers of the institutions hosting the data archives were organised, to identify all the technical, organisational, and legal issues regarding the planned co-operation. A lawyer was employed to help with the legal questions and contributed to the drafting of the service level agreement.

The project of this exchange was fully prepared for realisation, but remained unrealised. The idea of testing the system of reciprocal provision of hosting services in a real-life data archive operation was rejected as inefficient for two reasons. Both data archives had their own backup strategies in place before the start of this study and they did not seek alternative solutions. Moreover, any reliable test of a similar agreement would be time-consuming and would exceed the time span of the pilot study. However, the plans were evaluated as realistic, from multiple points of view, economic, technical, legal, and organizational. All relevant actors of the decision-making processes at both institutions participating in the study were involved in this evaluation.

ČSDA and TÁRKI - national service providers involved in the study

The following two national social science data archives were involved in this case study: (1) the Czech Social Science Data Archive (ČSDA)[7] and (2) the TÁRKI Data Archive (TÁRKI)[8] from Hungary.

Both organisations, ČSDA and TÁRKI, are parts of larger institutions and their data services are non-profit and targeted at academic social science research community. ČSDA is a department of the Institute of Sociology of the Czech Academy of Sciences. It is a non-university public research institution obtaining the most of its budget from the state and the public grant

[7] ČSDA is the national data resource centre for social science research in the Czech Republic. It acquires, processes, documents, archives, and preserves digital datasets from Czech and international social research and makes the data publicly available for both the secondary analysis in the academic research and the training purposes at higher education. The main activities of ČSDA may be summarised as follows: (1) acquiring, archiving and providing open access to datasets from Czech social science research projects and international surveys with Czech participation; (2) providing technical and organizational support for large-scale research surveys in the Czech Republic, e.g. the Czech Household Panel Survey or Czech surveys under the International Social Survey Programme (ISSP); (3) supporting the use of secondary data analysis in research by (a) providing training courses and taking part in educational programmes in the areas of methodology and analysis of social science data; (b) mapping and analysing available data sources, providing information services and user support on data sources; and (c) connecting Czech and international data sources and research in the field of data standardization and harmonization.

[8] TÁRKI Data Archive is the national social science data archive in Hungary. The mission of this archive is to provide infrastructure service, and support for all stakeholders in social research, which includes following activities: (1) long-term preservation of digital research datasets from domestic and international studies; (2) keeping pace with technological change and participation in the development of data archiving standards; (3) providing access to data collections of empirical studies for users communities; (4) facilitating effective data use by providing access to our own and to our partners' collections.

support schemes. TÁRKI Data Archive is operated by TÁRKI Joint Research Center (TÁRKI JRC), which is a private not-for-profit association of academic and educational institutions. They are located on premises of their hosting institutions, use their administrative, technical and other organisational background and are subordinated to their managements and supervisory bodies.

Both data archives participate in the CESSDA pan-European collaborative network. CSDA is a CESSDA Service Provider in the Czech Republic, which is the CESSDA Member since the foundation of CESSDA AS in 2013. Czech membership continues also after transformation of CESSDA AS into CESSDA ERIC in June 2017. TÁRKI became the CESSDA ERIC Service Provider and Hungary became the CESSDA ERIC Member just with the launch of CESSDA ERIC during the development of this study. Before June 2017 TÁRKI was a member of the CESSDA Network of Partners[9] and participated actively in CESSDA life including participation in several CESSDA activities and projects. This extensive co-operation rooted from TÁRKI long term membership in the old CESSDA organisation during 1989 and 2013.

Both ČSDA and TÁRKI are relatively small data archives. ČSDA staff counts 8 FTE (full time equivalents). In December 2016, the data library included 774 archived studies. The majority of the collection is questionnaire-based survey data. Qualitative data and historical data in form of Nesstar cubes are included as well. The current size of backups of the data library is about 50 GB. TÁRKI Data Archive has collected and archived more than 650 empirical social research data collections that are suitable for secondary analysis. These tend to be Hungarian. Most of the collection comes from nationally representative sample-survey studies (i.e. micro datafiles). One section of the databases archived is made up of TÁRKI's own surveys, and the other section comprises surveys from other Hungarian research institutes.

In 2016 ČSDA was awarded the Data Seal of Approval certification for trustworthy digital data archives. Based on the Country report on development potential (Štebe 2017: 86-96) the digital object management and the technical infrastructure at ČSDA are well organised. Seven of twelve indicators of these measures are above average, four are slightly below average, one is missing. Preservation strategies including the backups were assessed as slightly above the CESSDA average.

The situation in TÁRKI is different. For the last few years, the data archive has been underfunded and understaffed. Based on the Country report on development potential (Štebe 2017: 166-174) TÁRKI fulfils not all of the digital object management and technical infrastructure requirements. Except for the citations, all the indicators are below CESSDA average. Preservation strategies were evaluated as in initial stage.

[9] The CESSDA Network of Partners was established within the CESSDA SaW project to formalise cooperation of CESSDA and CESSDA Service Providers with service providers and national data archiving initiatives from non-member countries. CESSDA partners take part in CESSDA life including selected events and projects. The service providers aspiring for future CESSDA membership are regularly invited guests at the CESSDA Service Providers' Forum, where they can contribute into discussions on CESSDA strategies, policies and plans on projects and activities

Major issues, findings, and proposed solutions

The requirements resulting from the aim of this study can be summarised as follows:

- Two social science data archives, Service Provider 1 and Service Provider 2, provide each other a disk space of size appropriate to host backup of a social science data library
- Each of these SPs must guarantee usual level of maintenance and security standards regarding the disk space provided
- Each of these SPs must guarantee access to the other party allowing management of the backups
- The disk space will be provided free of charge
- The impact on operation and system changes should be minimal

These requirements lead to the following issues that need to be addressed before the reciprocal provision of hosting services can be launched.

a) Legal background

A general requirement from any backup is its reliability. That is why the basic definition and characteristics of the hosting services need to be guaranteed by written and legally valid agreement in between both partners. Thus, it is recommended that the legal institutions hosting the SPs will enter into a standard Service Level Agreement (SLA). Proposal of model SLA for purposes of the reciprocal provision of hosting services in between two digital data archives is in Appendix 1 of this report. This model SLA treats all the issues discussed below and its content was accepted as appropriate at both organisations involved in the study, i.e. the Institute of Sociology of the Czech Academy of Sciences (hosts ČSDA) and the TÁRKI Joint Research Center (hosts TÁRKI Data Archive).

b) Size of the disk space available for backups

This is the question of both, needs and available capacities. In addition, the size of the disk space provided should be similar at both SPs. Otherwise it can cause disadvantage for one of the contract parties. An imbalanced contract can result into requirement of payment or tax liability resulting from the donation.

ČSDA and TÁRKI have similar goals and are currently archiving similar types of social science data. In the study, it has been found out that both organizations have free capacities, which significantly exceed current needs for hosting a complete backup of data library of the contract partner. The disk space reserved for the hosting services was set at 500 GB at both organisations.

c) Data security and maintenance

Backups should be stored under usual security standards for digital data archives. Both ČSDA and TÁRKI, have data security and maintenance policies fulfilling standards typical for organisations dealing with social science research data already in place. Requirements for such policies are part of the Data Seal of Approval (DSA) certification for trustworthy digital data archives. ČSDA has already obtained DSA, while TÁRKI's application for DSA is just under preparation. The current data policies will not be altered and their relevant measures can be applied also to the data stored under the agreement on hosting services. The following minimum requirements are set up in SLA:

- Both parties shall provide Service Maintenance. (SLA, art. 5)
- Each party must provide regular full backup of the Data Storage in other location than original Data Storage. (SLA, art. 12)
- Data Room where the Data Storage is located must conclude to general regulations and rules for such technology – fire protection, anti-theft measures, non-flooding area, emergency power supply, cooling system. On request each party must prove the concordance to such rules and regulations or enable personal inspection by a responsible employee of the other party. (SLA, art. 13)

d) Service availability

Availability of the service can be a major source of limitations of the reciprocal provision of hosting services in comparison to subscribing for professional hosting services. An agency providing this kind of services regularly, and on a professional basis, has structures and organisational background built specifically for this purpose. Based on this, it is usually able to provide 24 hours per day/7 days per week availability of the service and higher standards in response time to errors. However, building such background and hiring extra personnel in the data archive just for one case of provision hosting service is inefficient and in conflict with the objectives of low costs and low impact on operation and system changes defined in this study.

The following solution was developed based on discussions with IT staff and management at ČSDA and TÁRKI and the lawyer:

- The 24x7 availability is set up as a goal. However, it is understood as only a goal and SPs although required to use all the reasonable efforts, do not guarantee that this goal will be achieved (SLA, art. 2 and 3).
- Specific regulations are set up for the service maintenance in cases when it can cause the service unavailability. The SPs shall use reasonable efforts to limit such service maintenance to 2 hours per month and schedule it outside of business hours. If it is necessary to perform it during business hours, the other party shall be notified in advance. (SLA, art. 5 - 7)
- Business hours are defined in between 8 am and 6 pm CET/CEST. It is adjusted to the situation in ČSDA and TÁRKI and may be altered for purposes of other SPs. (SLA, art. 1-a)

e) Responsibility limitations

The service is defined for the research data (SLA, art. 8). SP providing the storage place is not responsible for the content of the data of the other SP (SLA, art. 10). Each SPs is responsible for the coding, decoding, protecting and coherence of its own data (SLA, art. 9). The disclaimer for actions caused by and/or under the control of third parties is also defined (SLA, art. 9).

f) Communication in between the partners

Each SPs is obliged to nominate its contact person (SLA, art. 19). All predefined claims should be communicated via email within seven days of the incident (SLA, art. 16).

The reciprocal hosting services are intended to be used by national SPs, thus we may assume they will come from two different countries. This fact also means that any possible disputes, controversies or claims arising from cooperation in between these two participating SPs can be settled under two different legal systems. For purposes of reciprocal hosting services organised within the CESSDA ERIC or CESSDA ERIC Network of Partners we propose to simplify this situation by the introduction of an arbitration clause into SLA and inclusion of CESSDA ERIC as an arbitrator (see SLA, art. 22).

g) Possible termination

The situation at SPs, their available capacities and needs may change over time. Participation of SP in the reciprocal provision of hosting services should not limit its development. At the same time, it is necessary to ensure the reliability of the backups, including the necessary minimum service durability. That means that contract parties need to seek a compromise between the following two requirements: (1) participation in the reciprocal provision of hosting services should not be too binding for each SPs and allow termination of its obligations in a reasonable time; (2) the length of the period in between the notice of termination and the termination itself should allow changes in the data preservation policy at each contract party and avoid threatening the security of data. At the same time the hosting party should not keep the other party's data after termination of conditions defined in the SLA. Following measures were proposed to solve these issues:

- Each SPs is authorized to terminate the cooperation. It must be done by written notice. The contract terminates more than 2 months from the delivery of such notice. (see SLA, art. 17)
- Within 7 days after termination, each SPs must confirm deletion of the data of the other SPs. (see SLA, art. 18)

Conclusions

The solution for the reciprocal provision of hosting services proposed in this study is tailored to the situation of any two social science data archives co-operating in the wide pan-European CESSDA network. It is relevant to both, SPs at CESSDA ERIC Members and Observes, as well as non-member SPs co-operating in the CESSDA ERIC Network of Partners. Specific settings of measures and values are demonstrated on the situation at ČSDA and TÁRKI. However, these settings can easily be changed according to the specific situations at any other two partners from the CESSDA network. The proposed solution stems also from existing mutual trust and a long-term collaboration in the CESSDA network. Nevertheless, most of the proposed measures can be transferred also to other areas of digital data archiving; although, mutual trust between partners will always be a prerequisite for such low-cost reciprocal provision of hosting services.

List of experts providing consultations (in addition to authors)

Petra Broskevičová, Institute of Sociology CAS, Deputy Director for Economic Development and Infrastructure
Tomáš Čížek, ČSDA, System Administrator
Krisztina, Dávid, TÁRKI, PC support, IT expert
Gergely, Kádi, TÁRKI System Administrator
Milan Paučula, Institute of Sociology CAS, IT expert
Miroslav Siva, Institute of Sociology CAS, ČSDA, IT expert
Petra Skalská, Institute of Sociology CAS, project manager
Martin Vávra, ČSDA, Acquisitions & Ingest Administrator

Martin Velík, AK-VELÍK MARTIN, law expert
Vlasta Velíková, AK-VELÍK MARTIN, law expert

Sources

DSA. 2016. Data Seal of Approval. Guidelines version 2017-2019. Hague: Data Seal of Approval. Available on-line: https://assessment.datasealofapproval.org/guidelines_54/pdf/

Evolve. 2015. The IT Disaster Recovery & Business Continuity Survey. Wayne: Evolve. Available on-line: http://www.evolveip.net/evolve_draas_survey.pdf

ICPSR. 2012. Guide to Social Science Data Preparation and Archiving. Best Practice Throughout the Data Life Cycle. 5th Edition. Ann Arbor: Inter-university Consortium for Political and Social Research (ICPSR), University of Michigan. Available on-line: https://www.icpsr.umich.edu/files/deposit/dataprep.pdf

Štebe, Janez et al. 2017. *Country Report on Development Potentials 1*. CESSDA SaW project Report D3.2. Bergen: CESSDA. Available on-line: http://cessdasaw.eu/content/uploads/2017/07/D3.2_CESSDA_SaW_v1.3.pdf

Pilot 2: Kuha2 Metadata Server

Toni Sissala (FSD)

Introduction

For this pilot study the aim is to provide development for a required (infrastructural) tool. The requirements do not come directly from a service provider, but from CESSDA ERIC in that the new metadata harvesting procedure requires SPs to deliver metadata using a new methodology.

The repositories administered by the SPs contain a wide variety of data and descriptive metadata. CESSDA aims to provide an aggregated discovery service containing the appropriate (social science related) metadata of SPs. To reach this goal the metadata provided by the SPs should somehow be made accessible to aggregating services such the Product and Services Catalogue (PaSC) of CESSDA[10].

FSD has previously developed repository handlers supporting harvesting through OAI-PMH and OSMH[11] protocols. The purpose of this pilot is to develop an open source application bundle to help developing archives to set up a metadata provisioning service that provides standards-based metadata for consuming by multiple harvesting protocols.

Description of the software

Kuha2 is a software bundle that consists of three server applications, a client application and a database. The name of the application bundle comes from a previously developed and openly sourced repository handler. Kuha2 uses DDI-C 2.5 XML-documents[12] as the source of the metadata. Harvesting protocols supported by Kuha2 include OSMH and OAI-PMH. The OSMH protocol has its own JSON-schema, while OAI-PMH can offer metadata in various XML-formats. XML-formats supported by Kuha2 via OAI-PMH are DDI-C 2.5 and OAI-DC (Dublin Core format adjusted for usage in the OAI-PMH).

The elements supported by the metadata formats in Kuha2 are the ones defined in CESSDA Metadata Model (CMM) 1.0[13] as mandatory elements that are applicable by DDI-C 2.5 and supported by the OSMH JSON-schema.

Technical Environment

Kuha2 is server software and therefore targeted to be run in server deployments. The supported operating system is Ubuntu 16.04 LTS[14]. All the components required to run Kuha2 are licensed as open-source.

[10] https://dev.cessda.net/Research-Infrastructure/Products-Services

[11] https://www.cessda.eu/eng/Projects/Work-Plans/Work-Plan-2015#osmh

[12] http://www.ddialliance.org/Specification/DDI-Codebook/2.5/

[13] https://www.cessda.eu/Consortium/Communication/News/CESSDA/CMM-1.0-CESSDA-Metadata-Management

[14] http://releases.ubuntu.com/16.04/

The software is written in Python and requires Python 3.5, which is included in vanilla Ubuntu 16.04 installations. The Python applications should be run in Python virtual environments. This provides an additional layer of confinement and separates the Python modules required by the software from the Python modules provided by the OS.

Kuha2 uses MongoDB as a persistent storage. MongoDB is a document database that stores data in JSON-like objects. It was chosen for its flexibility in the structure of the documents, its schemaless design, fast query capabilities, strong community support, good documentation and driver support, and Open-source license. MongoDB can be installed via Ubuntu repositories, however, it is recommended to use the latest MongoDB package provided by MongoDB's own repository.

Software Components

Kuha2 consists of three server applications: Document Store, OAI-PMH Repository Handler, and OSMH Repository Handler.

Document Store is responsible for ingesting and providing metadata to other components. It ingests DDI 2.5 documents and stores them to the underlying database. It provides a RESTful API for accessing and managing the metadata. It also provides a query API to access the metadata selectively.

OAI-PMH Repository Handler provides the metadata from Document Store for harvesting via OAI-PMH protocol. The OAI-PMH Repository Handler supports DDI-C 2.5 and OAI-DC metadata formats. It also supports selective harvesting and resumption tokens.

OSMH Repository Handler provides the metadata from Document Store for harvesting via OSMH protocol. It supports four OSMH Record Types: Study, Variable, Question, and StudyGroup.

Kuha2 provides command line client for submitting DDI-C 2.5 documents to the Document Store. The client is provided as an example. End users are encouraged to develop their own solution.

Kuha2 relies on a persistent storage that will be accessed via Document Store. This storage is provided by MongoDB. Kuha2 includes scripts and installation instructions to help install MongoDB.

Each software component connects to another via network sockets. Therefore all of the components may be installed on separate systems as long as they have access to each other through network ports.

Software Architecture

CESSDA Technical Framework promotes a microservice software architectural pattern and 12-factor app methodology. Kuha2 follows these recommendations as strictly as possible.

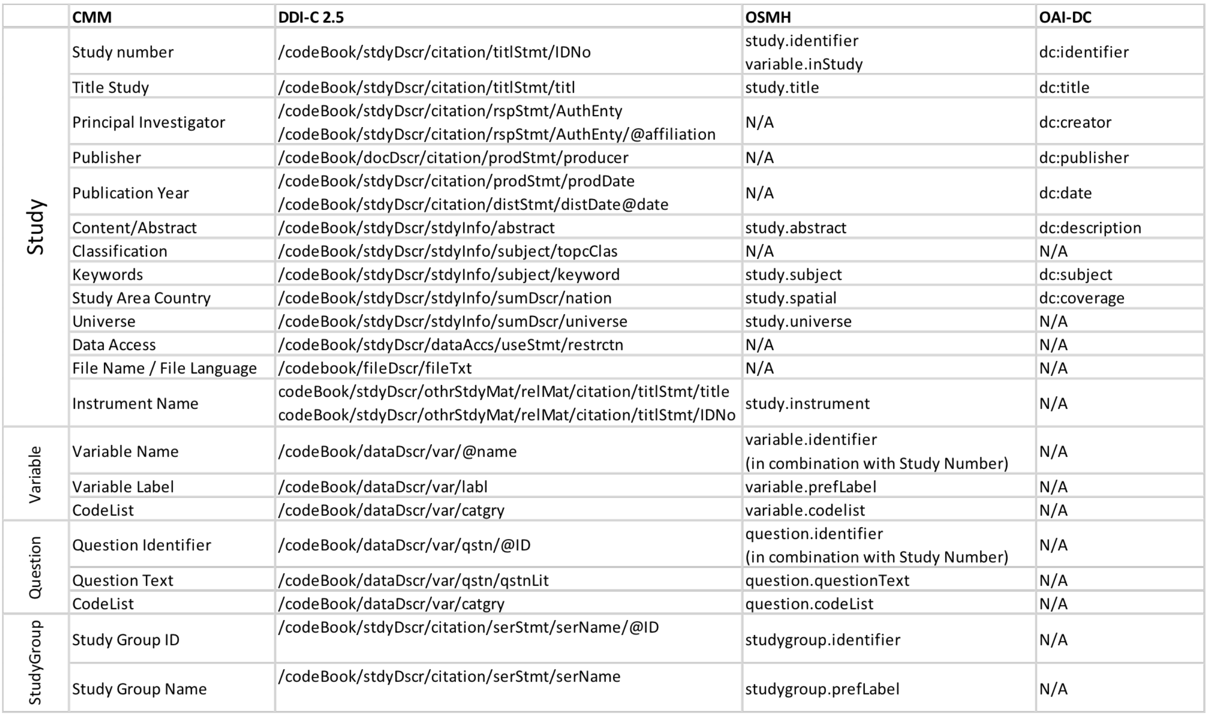
The software bundle is a distributed system, with each component having its own codebase. A small amount of shared code is factored into a library named Kuha Common, which can be installed as a dependency via Python package manager and tracked individually in version control.

The software components are built in a way that facilitate additional development. All of the server components have an API that is accessible via HTTP requests. Kuha Common library provides a ready-to-use framework to access the Document Store. Therefore it is trivial to build and develop additional software components that interact with the Document Store.

Supported Metadata Elements

The list of metadata elements supported by Kuha2 is an intersection of the elements defined in CMM 1.0 model as mandatory and the elements defined in the OSMH metadata schema (1.1.2017).

The resulting set of elements is not exhaustive, but it serves as a starting point for SPs that don't have their metadata available for harvesters. It also provides a practical example of how interoperable these metadata standards are. The following table shows the CMM elements and their corresponding counterparts in metadata formats supported by Kuha2. It also shows the source of the element in DDI-C 2.5. Note that not all the elements can be mapped one-to-one.



	CMM	DDI-C 2.5	OSMH	OAI-DC
Study	Study number	/codeBook/stdyDscr/citation/titlStmt/IDNo	study.identifier variable.inStudy	dc:identifier
	Title Study	/codeBook/stdyDscr/citation/titlStmt/titl	study.title	dc:title
	Principal Investigator	/codeBook/stdyDscr/citation/rspStmt/AuthEnty /codeBook/stdyDscr/citation/rspStmt/AuthEnty/@affiliation	N/A	dc:creator
	Publisher	/codeBook/docDscr/citation/prodStmt/producer	N/A	dc:publisher
	Publication Year	/codeBook/stdyDscr/citation/prodStmt/prodDate /codeBook/stdyDscr/citation/distStmt/distDate@date	N/A	dc:date
	Content/Abstract	/codeBook/stdyDscr/stdyInfo/abstract	study.abstract	dc:description
	Classification	/codeBook/stdyDscr/stdyInfo/subject/topcClas	N/A	N/A
	Keywords	/codeBook/stdyDscr/stdyInfo/subject/keyword	study.subject	dc:subject
	Study Area Country	/codeBook/stdyDscr/stdyInfo/sumDscr/nation	study.spatial	dc:coverage
	Universe	/codeBook/stdyDscr/stdyInfo/sumDscr/universe	study.universe	N/A
	Data Access	/codeBook/stdyDscr/dataAccs/useStmt/restrctn	N/A	N/A
	File Name / File Language	/codebook/fileDscr/fileTxt	N/A	N/A
	Instrument Name	codeBook/stdyDscr/othrStdyMat/relMat/citation/titlStmt/title codeBook/stdyDscr/othrStdyMat/relMat/citation/titlStmt/IDNo	study.instrument	N/A
Variable	Variable Name	/codeBook/dataDscr/var/@name	variable.identifier (in combination with Study Number)	N/A
	Variable Label	/codeBook/dataDscr/var/labl	variable.prefLabel	N/A
	CodeList	/codeBook/dataDscr/var/catgry	variable.codelist	N/A
Question	Question Identifier	/codeBook/dataDscr/var/qstn/@ID	question.identifier (in combination with Study Number)	N/A
	Question Text	/codeBook/dataDscr/var/qstn/qstnLit	question.questionText	N/A
	CodeList	/codeBook/dataDscr/var/catgry	question.codeList	N/A
StudyGroup	Study Group ID	/codeBook/stdyDscr/citation/serStmt/serName/@ID	studygroup.identifier	N/A
	Study Group Name	/codeBook/stdyDscr/citation/serStmt/serName	studygroup.prefLabel	N/A

Table 4: Supported metadata elements

Use Case: Manage Metadata

The following use case diagram shows the available use cases for metadata management. The primary source of metadata is the DDI-C 2.5 document, but Kuha2 also provides a REST API for easy access and management of stored metadata records. When receiving a DDI-record, Kuha2 handles database inserts and updates automatically. For example, when submitting a DDI-record with a study number already known to Kuha2, it updates the study record accordingly.

The REST API facilitates further use cases and makes the software extensible. One can, for example, develop additional import functionalities by using the REST API.

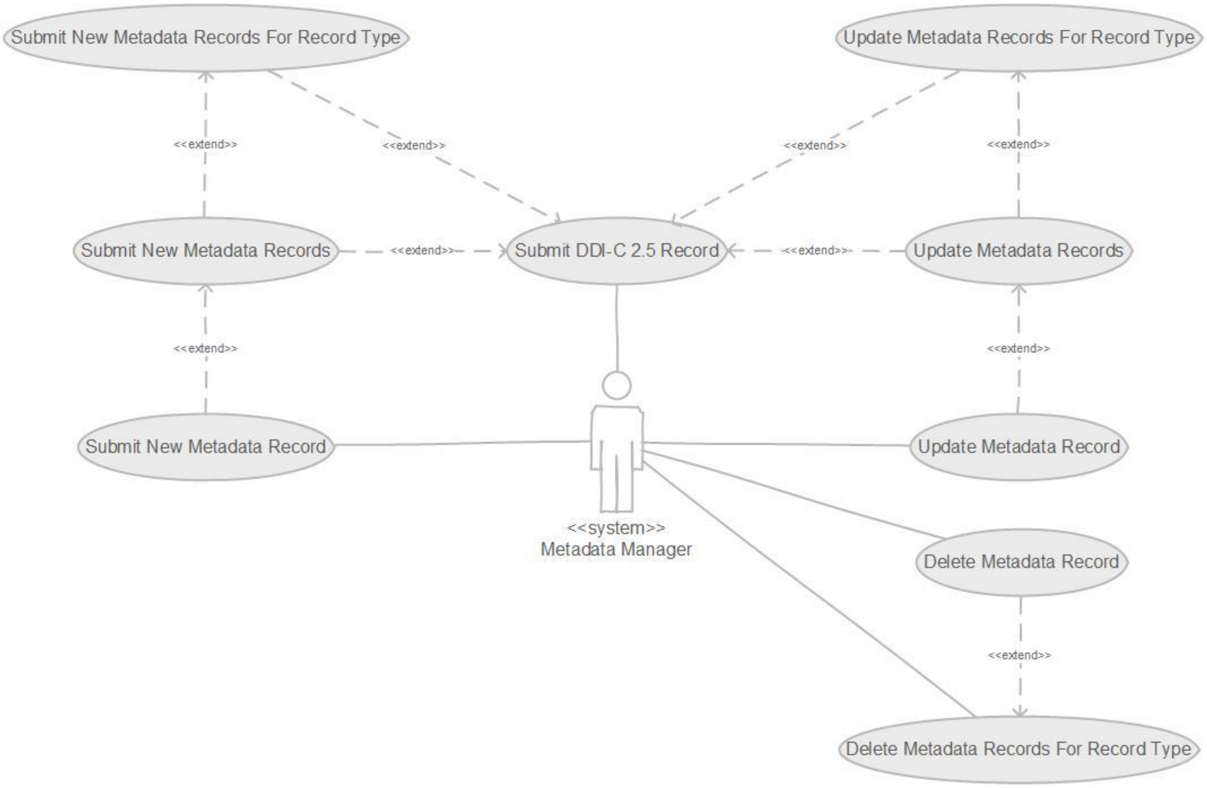


Figure 1: Manage metadata

Use Case: Harvest Metadata

Kuha2 supports metadata harvesting via OSMH and OAI-PMH protocols. The following diagram shows the available use cases for the supported protocols. It also depicts the similarities through extensions.

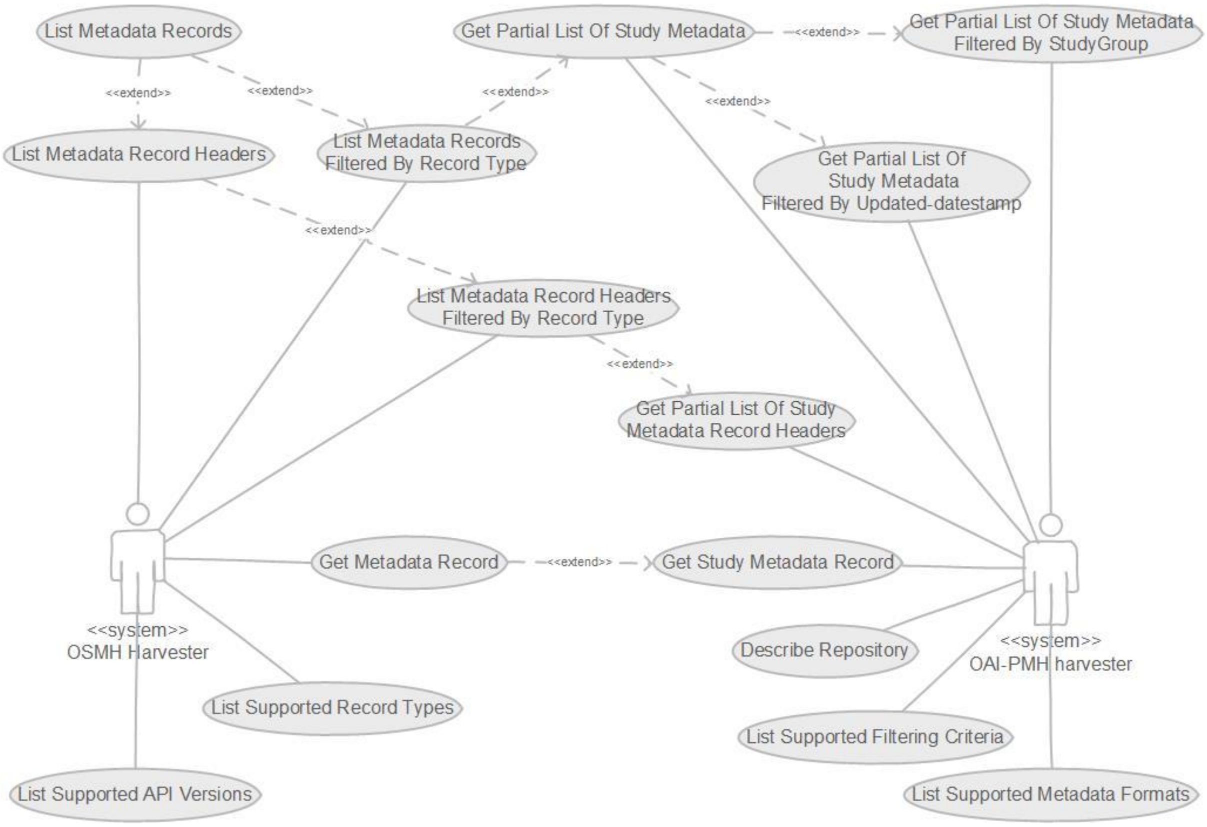

List Metadata Records
Get Partial List Of Study Metadata
<<extend>>
Get Partial List Of Study Metadata Filtered By StudyGroup
<<extend>>
<<extend>>
<<extend>>
<<extend>>
List Metadata Record Headers
List Metadata Records Filtered By Record Type
Get Partial List Of Study Metadata Filtered By Updated-datestamp
<<extend>>
List Metadata Record Headers Filtered By Record Type
<<extend>>
Get Partial List Of Study Metadata Record Headers
Get Metadata Record
<<extend>>
Get Study Metadata Record
<<system>>
OSMH Harvester
Describe Repository
<<system>>
OAI-PMH harvester
List Supported Record Types
List Supported Filtering Criteria
List Supported API Versions
List Supported Metadata Formats

Figure 2: Harvest metadata

Sequence: OSMH GetRecord request

The following diagram shows the sequence of OSMH GetRecord request. The OSMH Repository Handler and Document Store handle external requests/responses via HTTP protocol.

The initial request comes from OSMH Harvester via HTTP-GET. The request gets handled by the OSMH Repository Handler, which then submits an HTTP-POST request with a JSON-payload containing query parameters to the Document Store. The Document Store queries the database and sends the results back to the OSMH Repository Handler. After the Repository Handler has finished all of the required queries it constructs a JSON object, which is submitted as a response to the initial request.

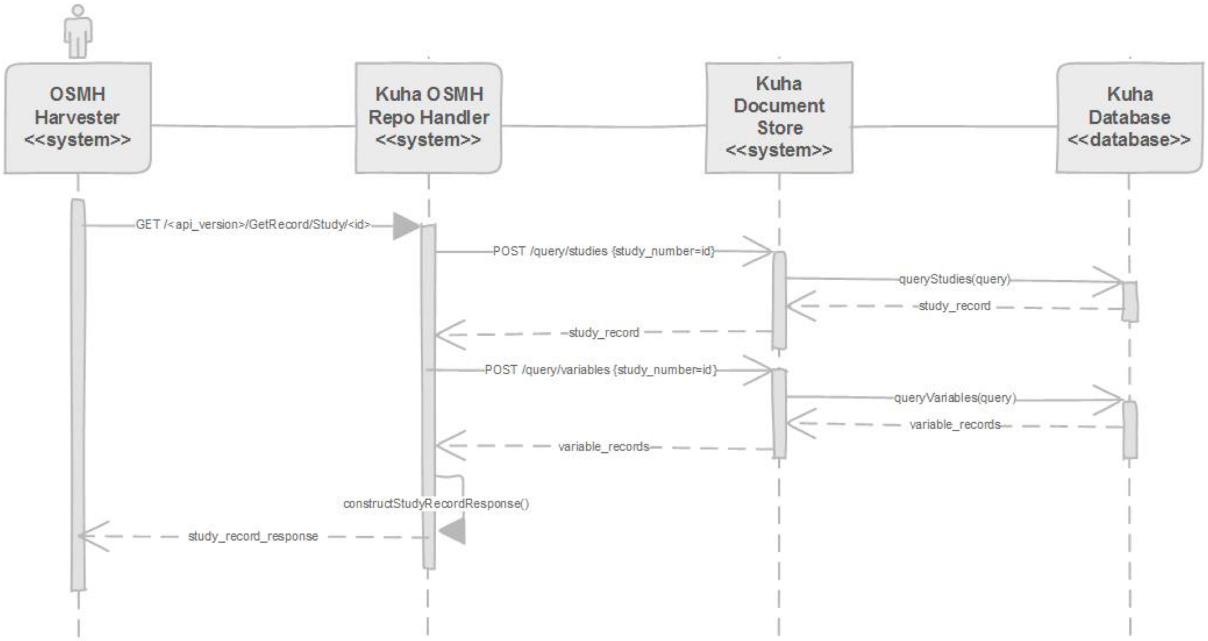

OSMH Harvester <<system>>
Kuha OSMH Repo Handler <<system>>
Kuha Document Store <<system>>
Kuha Database <<database>>
GET /<api_version>/GetRecord/Study/<id>
POST /query/studies {study_number=id}
queryStudies(query)
study_record
study_record
POST /query/variables {study_number=id}
queryVariables(query)
variable_records
variable_records
constructStudyRecordResponse()
study_record_response

Figure 3: OSMH GetRecord request

Sequence: OAI-PMH List request sequence

The OAI-PMH Repository Handler supports a list request in which the resulting list size is limited by a configurable value. This helps to limit the amount of time it takes to respond to the request and provides a way to limit the bandwidth of the responses.

The initial list request comes from OAI-PMH Harvester. If the result of the request should be larger than the configured list size, the response will contain a partial list with a resumption token. The harvester may use this token to continue the list request sequence. The OAI-PMH Repository Handler is responsible for the validity of the resumption token. If there have been changes to the repository which affect the response for the initial request, the resumption token gets marked as invalid and the harvester is advised to re-initiate the list request sequence. When the harvester receives a response with an empty resumption token, it concludes the sequence.

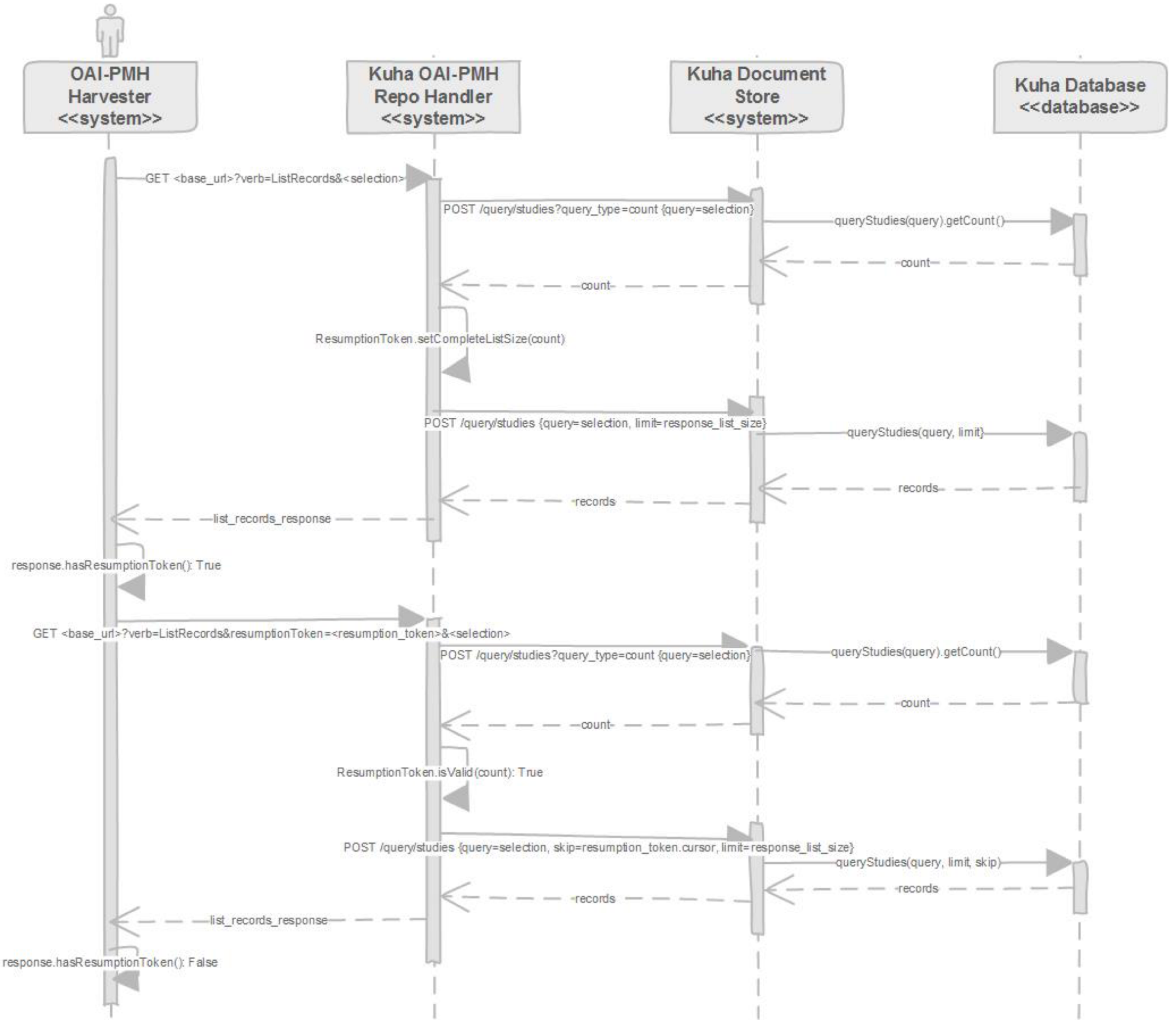

OAI-PMH Harvester <<system>>
Kuha OAI-PMH Repo Handler <<system>>
Kuha Document Store <<system>>
Kuha Database <<database>>
GET <base_url>?verb=ListRecords&<selection>
POST /query/studies?query_type=count {query=selection}
queryStudies(query).getCount()
count
count
ResumptionToken.setCompleteListSize(count)
POST /query/studies {query=selection, limit=response_list_size}
queryStudies(query, limit)
records
records
list_records_response
response.hasResumptionToken(): True
GET <base_url>?verb=ListRecords&resumptionToken=<resumption_token>&<selection>
POST /query/studies?query_type=count {query=selection}
queryStudies(query).getCount()
count
count
ResumptionToken.isValid(count): True
POST /query/studies {query=selection, skip=resumption_token.cursor, limit=response_list_size}
queryStudies(query, limit, skip)
records
records
list_records_response
response.hasResumptionToken(): False

Figure 4: OAI-PMH List request sequence

Conclusions

There is a clear requirement for the tool developed in this pilot, but that has come from CESSDA rather than a specific SP. This is most likely because many SPs are not aware of the new Technical Framework of CESSDA and that they will need to comply with the new harvesting protocol. CESSDA may need to fund and support a work programme to ensure that this new methodology is integrated, where necessary, at SPs, as some do not have the technical capacity and capabilities to do so by themselves.

We will continue to endeavour to find at least one SP, during the lifetime of SaW, to undertake testing and evaluation.

Pilot 3: Dataverse and NSD-NORDi: a national context evaluation of Dataverse

Trond Kvamme (NSD)

Pilot aim and scope

This pilot is an example of an expert requirements analysis and evaluation activity for a specific national project. Although the recipients of this expertise are not specifically another CESSDA SP the value and approach in this assessment, and frame of reference, maybe useful for other SPs in their national context of supporting social science research.

In the processes of developing and improving its services, NSD is considering the features of several of-the-shelf products and services. This pilot gives NSD the opportunity to test and assess Dataverse, and to gain insight into the advantages and disadvantages of Dataverse as a supplementary service to NSDs existing products. How does the Dataverse service align to NSD needs and the current Norwegian national research infrastructure environment? The assessment of Dataverse must take into consideration the duties, needs and commitments of NSD as a national service for the long-term preservation of, and provision of access to, research data.

The assessment is performed on a set of service provider indicators that are considered essential to NSD and the NORDi (Norwegian Open Research Data Infrastructure) project[15]. The indicators discussed are *Service integration, interoperability and information harvesting*; *Data files and formats; Metadata model; Licensing; General terms of use and trustworthiness*; and *Legal and ethical issues*.

Background and context

In 2014 the Research Council of Norway (RCN) introduced a policy on open access[16] to research data from publicly funded projects. The policy draws up a set of guidelines for the archiving, dissemination and sharing of research data, and it is stated that the guidelines will be followed up through the RCN's research funding instruments and through the National Financing Initiative for Research Infrastructure[17]. The main principles of the policy are as follows:

- Research data should be stored/archived in a safe and secure manner (the data should be *stored in secure archives*, either in a central repository at the relevant institution or in national archives).
- Research data should be made accessible for reuse.
- Research data should be made accessible at an early stage.
- Research data should be accompanied by standardised metadata.
- Research data should be provided with a license for access, reuse, and redistribution.

[15] http://www.nsd.uib.no/nordi/english.html

[16] https://www.forskningsradet.no/en/Newsarticle/Research_data_must_be_shared/1254000848864/p1177315753918ChromeHTML.BGCFCOKOXZVZWUZKTCJ4M4WR34/Shell/Open/Command

[17] https://www.forskningsradet.no/en/Funding/INFRASTRUKTUR/1232959367957

- Research data should be made accessible at the lowest possible cost.
- Research data should be provided with a long-term plan.

Through this policy, the Research Council seeks to be a driving force for the sharing and reuse of research data. Among other things they want to facilitate cooperation between the stakeholders, finance relevant activities, and provide guidance to the research communities through, e.g., implementing procedures in the application assessment process that ensure relevant grant applications include data management plans; implementing procedures in project follow-up activities that ensure the data management plans are being followed by projects granted funding, and continuing the practice in Research Council contracts that requires research data to be stored in a safe and secure manner for a minimum of 10 years.

For projects in the fields of social science, humanities, medicine and health, and environmental and development research, the Project Owner is contractually obliged to transfer copies of all research-generated data, including the necessary documentation for reuse of data/metadata, to NSD for archiving. This is to be carried out as soon as possible after receiving a request from the NSD and at the latest two years following the conclusion of the project period[18].

These policy requirements strengthen the commitments and obligations of NSD as a national service provider for the curation and preservation of research data and give some important contextual implications for the services that NSD are to perform. These considerations have to be taken into account when NSD is assessing services like Dataverse.

NSD-NORDi is a four-year project funded by the Norwegian Research Council. It will build on the existing services of NSD to develop a new comprehensive research infrastructure and system that aims to make it easier to discover, use and share research data. Support, courses and training material will be integrated into the NORDi-platform. This new infrastructure will include the following features:

- *Data Deposit Service*: a tool and service for self-deposit of data. Support for data management planning and metadata schemas.
- *Data Storage Service*: a platform for the strengthening of current storage and preservation procedures, policies and technology.
- *Data Discovery Service*: will develop a common search and discovery portal for all data holdings at NSD. It may also include data from external repositories and archives. The service will generate landing pages for data and connected metadata, which will work as endpoints for data citation.
- *Data Access Service*: will provide flexible access to all data holdings at NSD. Will handle open data as well as data with access restrictions.
- *Service for research institutions and organisations*: will provide tools and mechanisms for the tracking and documentation of research activities regarding data depositing/sharing, reuse, and citations.
- *Service for training, support, and guidance*: will be an integrated part of services where relevant.

UK Data Archive (UKDA) and the Inter-University Consortium for Political and Social Research (ICPSR) will be partners in the project. Both institutions play an important role

[18] RCN, *About project reports*: https://www.forskningsradet.no/en/Article/About_project_reports/1253979444039/p1138882213699

internationally as archives, "controlled curators" and infrastructures for research data, and will contribute as consultants.

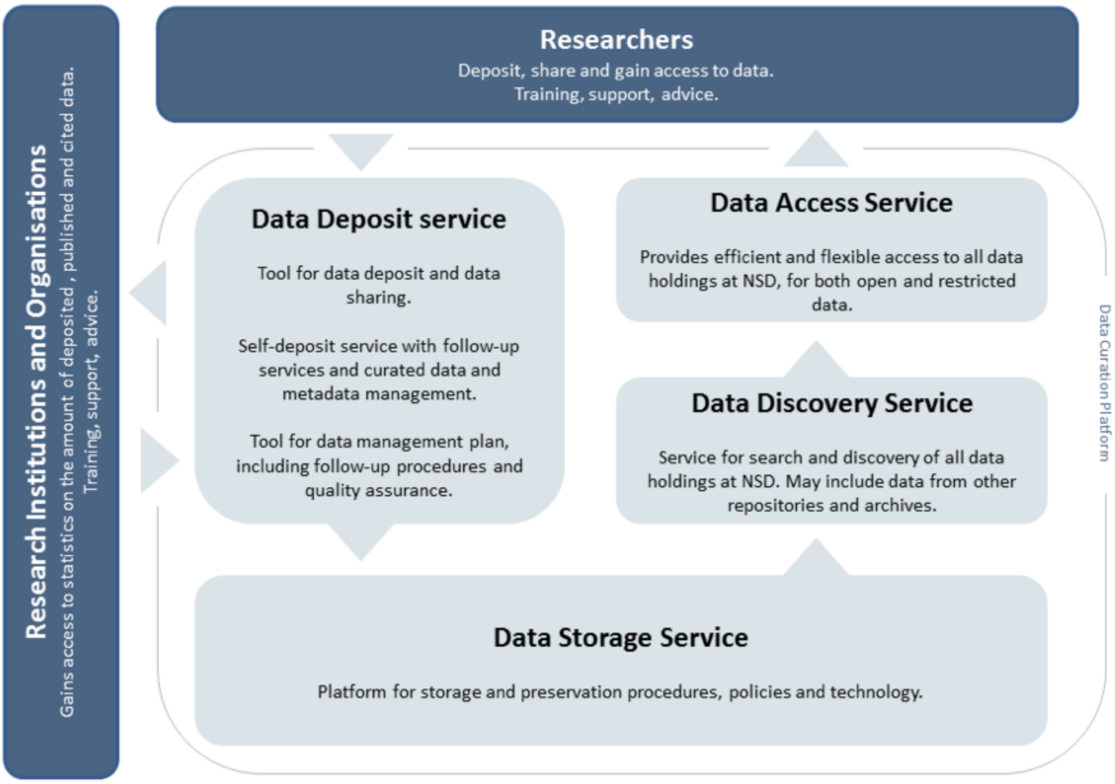

Research Institutions and Organisations
Gains access to statistics on the amount of deposited , published and cited data.
Training, support, advice.
Researchers
Deposit, share and gain access to data.
Training, support, advice.
Data Deposit service
Tool for data deposit and data sharing.
Self-deposit service with follow-up services and curated data and metadata management.
Tool for data management plan, including follow-up procedures and quality assurance.
Data Access Service
Provides efficient and flexible access to all data holdings at NSD, for both open and restricted data.
Data Discovery Service
Service for search and discovery of all data holdings at NSD. May include data from other repositories and archives.
Data Storage Service
Platform for storage and preservation procedures, policies and technology.
Data Curation Platform

Figure 5: Full model of the NSD-NORDi services

Several of the features and services inherent in NSD-NORDi are in various forms provided by numerous open, globally-scoped, general-purpose data repositories such as 4TU.Centre for Research Data[19], FigShare[20], Dryad[21], Mendeley Data[22], Zenodo[23], DataHub[24], the EUDat[25] service platform, and Dataverse[26].

Although Dataverse originates from the Harvard Institute for Quantitative Social Science (IQSS), installations of the open source software are openly available to researchers and data collectors worldwide from all disciplines. In Norway, there is currently one Dataverse installation, at the Arctic University of Norway - the UiT Open Research Data Dataverse[27]. It

[19] http://researchdata.4tu.nl/en/home/
[20] http://figshare.com
[21] http://datadryad.org/
[22] https://data.mendeley.com
[23] http://zenodo.org/
[24] https://datahub.io/
[25] https://eudat.eu/
[26] https://dataverse.org/
[27] https://opendata.uit.no/dataverse/root

consists of two sub-Dataverses, namely the Tromsø Repository of Language and Linguistics (TROLLing)[28] and the Arctic Language Technology Dataverse[29].

Service provider indicators

Service integration, interoperability and information harvesting

Dataverse provides several APIs (SWORD API, Search API, Data Access API, and Native API)[30] that can interoperate with external tools. Although these APIs are not fully tested during this pilot (we have only looked at the documentation and the guides), our impression is that they will not, without significant internal deployment costs (time, development resources), be as effective for NSD-NORDi as using self-developed tool that can interact with relevant information resources. What NSD gains from self-developed deposit and access tools is the ability to communicate with and harvest information from other national research information systems and resources like Feide/Dataporten[31] (a national login / authentication tool), CRIStin (Current Research Information System In Norway)[32], the project database[33] of the Research Council of Norway, and the Data Protection Official for Research[34].

Feide: Feide simplifies the processes for all parties involved using the concept of federated identity management. Federated identity management is based on the concept that services rely on user authentication at the user's home organization and they obtain from there some information about the user for its authorization decisions. Feide uses this federated approach to guarantee that each party remains in control of the steps relevant to it: home organizations register and authenticate their members, while service providers define their access rules. NSD-NORDi will offer both Feide-login and an NDS-account login (anyone can register an account). We aim to expand the login options in the near future, with (some of) the options that are currently available from Dataverse (e.g. ORCID, GitHub, and Google).

Cristin: CRIStin is a research information system for research institutes, universities and university colleges. One of the primary purposes of the system is to collect all the registration and reporting of research activities of institutions within the three sectors in a common system. This gives researchers a place to capture and simplify the registration of common publications. The Cristin system provides an API (the Cristin REST API[35]) which allows us (NSD) to interact with the Cristin system programmatically from our own applications. Using the Cristin API, NSD can harvest (and provide information to) Cristin information resources like Projects, Institutions, and Persons. This gives users of the NSD deposit/archive service the opportunity to automatically access or import information resources from Cristin upon deposit of data; thus avoiding entering the same information in several information platforms.

[28] https://opendata.uit.no/dataverse/trolling
[29] https://opendata.uit.no/dataverse/alt
[30] http://guides.dataverse.org/en/4.7/api/index.html
[31] https://www.uninett.no/en/service-platform-dataporten
[32] http://www.cristin.no/english/
[33] https://www.forskningsradet.no/prosjektbanken/#/
[34] http://www.nsd.uib.no/personvernombud/en/
[35] https://api.cristin.no/index.html

In addition, upon data upload in the NSD-NORDi system, data depositors will have the opportunity to reuse metadata from the Data Management Plan (if applicable; some but not all depositors will have used the NSD-DMP tool during their project period), and information from earlier deposits. They will also be able to interact with NSD's targeted training and guideline resources which will contain examples from existing metadata holdings.

RCN: The full potential of project database[36] of the Research Council of Norway is yet to be explored, as it does not currently offer any APIs. All the information stored in the project database has been transferred directly from the Research Council's internal administrative databases, with no adjustments to scientific terminology or sorting categories. The presentations of each project have not been edited in any way, and may, therefore, contain a number of abbreviations and other potentially 'incomprehensive' information.

Data Protection Official for Research: NSD is the Data Protection Official (DPO) for approximately 140 research and educational institutions, including all the Norwegian universities, university colleges, several hospitals, and a number of independent research institutions. The main task of DPO is to assist the institutions in fulfilling their statutory duties relating to internal control and quality assurance of their own research. In order to solve this task NSD/DPO offer several services, like reviewing research projects that process personal data; follow up notifications on project changes, extensions and at the end of the projects; provide guidance, training and information material for researchers, students, administration and management; provide access to tools for the institution's handling of personal information to safeguard internal control of own research; and guidance on the research subject about their rights. Many of these services overlap with the services planned in NSD-NORDi, and a fully integrated service provision between the NSD archive and NSD/DPO is crucial for the NSD-NORDi project. When it comes to data protection and data protection services, the one-size-fits-all generalist approach of services like Dataverse may impinge on the aims and goals of NSD-NORDi (see also the segments on *Licensing* and on *Legal and ethical issues* below).

Data files and formats

Generalist repositories like Dataverse are potentially exposed to a large variety of cases, e.g., an almost open ended set of dataset formats and typologies to deal with, and a large and multidisciplinary community of practices in terms of both data owners and data consumers. This potentially extreme heterogeneity of users and datasets can impose limitations on the quality, interoperability, and reusability of the data and metadata that are deposited. Hence the challenge is how to improve the specification that repositories offer, related to, e.g., the formats they natively support and the validation procedures offered[37].

A generalist repository like Dataverse does not make any assumptions on the data files and content formats that are deposited; rather there is an implicit generic approach which leads to format 'blindness'. This in turn limits/restricts the possibilities for curating and quality checking of data. Of course, Dataverse is not a curation service per se, and all quality checking and curation of data and metadata are in the hands of the depositor/researcher. This may work well for certain types of data, but in other cases, one can run into problems with

[36] https://www.forskningsradet.no/prosjektbanken/#/

[37] Assante, M. et al., (2016). *Are Scientific Data Repositories Coping with Research Data Publishing?*. Data Science Journal. 15, p.6. DOI: http://doi.org/10.5334/dsj-2016-006

interoperability. Data formats should be as much 'intelligible' and open as possible to permit cross disciplinary (re)use of data; if dataset(s) are only provided in 'narrow', specialist formats they are much less interoperable and sharable. In a more curated deposit service, the set of formats a dataset is made available in can either belong to a predefined list or be the result of a negotiation between the data depositor (and consumer) and the archive itself, aiming at identifying the format that better fits the purpose of potential consumers/re-users[38].

Metadata model

Dataverse supports several metadata schemas for citation and domain-specific metadata. Citation metadata are compliant with DDI Lite[39], DDI 2.5 Codebook[40], DataCite 3.1[41], and Dublin Core's DCMI Metadata Terms[42]. These standards also cover Geospatial metadata and Social Sciences and Humanities Metadata. In addition, Dataverse also supports metadata schemas for Astronomy and Astrophysics metadata, and Life Sciences metadata[43]. This broad spectre of metadata schemas is considered one of the main strengths of Dataverse. However, most of these schemas, even though they contain some subject-specific metadata elements, are general and limited in scope. Necessarily so, since Dataverse is meant to serve multiple unspecified target communities that may access published data for unconstrained re-uses.

The metadata model of NSD-NORDi will, although it is still in the early planning phases, also be based on DDI, Dublin Core and DataCite Metadata Schema 4.0[44]. We may also want to include elements from other schemas like the da|ra Metadata Schema[45], which is based on DataCite but provides a richer set of elements. Additionally we want to remain flexible when it comes to the discoverability and reusability of (meta)data; this might be obtained by enabling depositors to provide multiple metadata descriptions and documentation for the same datasets, each oriented to the needs of a specific target audience for a particular application (see Parsons et al., 2011). Another way to enrich data descriptions and documentation is to allow users/depositors to describe data in free-standing, free-text documents, and/or use (parts of) the data management plan as "data papers"[46] that can supplement the pre-defined metadata schemas and enrich the description of data.

The main point is that the NSD-NORDi metadata model aims for a flexibility and adaptability that the general-purpose Dataverse model currently does not support.

[38] Parsons, M.A. & Duerr, R., (2006). *Designating user communities for scientific data: challenges and solutions*. Data Science Journal. 4, pp.31–38. DOI: http://doi.org/10.2481/dsj.4.31

[39] http://www.ddialliance.org/sites/default/files/ddi-lite.html

[40] http://www.ddialliance.org/Specification/DDI-Codebook/2.5/

[41] https://schema.datacite.org/meta/kernel-3.1/

[42] http://dublincore.org/documents/dcmi-terms/

[43] Dataverse Metadata References: http://guides.dataverse.org/en/latest/user/appendix.html

[44] https://schema.datacite.org/meta/kernel-4.0/

[45] http://www.gesis.org/fileadmin/upload/forschung/publikationen/gesis_reihen/gesis_methodenberichte/2014/TechnicalReport_2014-17.pdf

[46] Candela, L., Castelli, D., Manghi, P. and Tani, A. (2015). *Data journals: A survey*. J Assn Inf Sci Tec, 66: 1747–1762. doi:10.1002/asi.23358

Licensing

In Dataverse restrictions can be set at the file level and the terms of access have to be described by the owner in free-text form. Although free-text input provides the user with flexibility, it can be argued that a set of controlled list of terms-of-access categories can be more user-friendly and can work as a guideline for depositors who are unsure of existing, common terms of access. In NSD-NORDi we also aim to provide a broader set of license options (not only CC0) through a license selector menu / dashboard licence comparison tool. This also of course implies rich descriptions and explanations through associated training material / guidelines; we would like to offer sufficient support to data owner at deposition time by clearly reporting the impact diverse licences have on the dataset usages.

The default settings in Dataverse are limited in scope with respect to dataset licensing and this can make the effective re-use of data difficult. NSD-NORDi must be able to relate, respond and adapt to the national context it operates within and will have to offer a broader set of use and control options, as "one-size-fits-all solutions" will not work. In addition, one should take into consideration that a broad application of general principles such as always practice openness, or prepare all data for sharing, may have harmful unintended consequences[47].

Managing and applying some forms of control on access and, especially, re-use conditions in heterogeneous contexts will be one of the main challenges of NSD-NORDi; we have to be more of a "controlled data collection" where staff in cooperation with user communities, make and enforce rules to control who can access data or how data can be used[48].

Trustworthiness

NSD NORDi aims to be a *curator* of research data. Data curation involves maintaining, preserving and adding value to digital research data throughout its lifecycle. It is an *active management* of research data that reduces threats to long-term research value and mitigates the risk of digital obsolescence.

It involves being a "steward" of access and providing a safe and trustworthy environment for the long-term preservation of data. Not only do we need to provide a solid platform for storage and preservation; data owners and depositors also have to trust the archive to provide access solely to authorised users, and to carry out services such as ingest processing (which includes ensuring the data are appropriately anonymised and internally consistent) and data archiving (managing the data within a secure environment) without disclosing any sensitive information. We believe that a trustworthiness built on *context awareness* and *adaptibility* can be better achived outside of the Dataverse platform.

Legal and ethical issues

Sharing of human-subjects data hinges on the development of techniques for protecting confidentiality, and new problems and issues, and ways of dealing with these issues will

[47] Sieber, J.E., (2006). *Ethics of sharing scientific and technological data: a heuristic for coping with complexity & uncertainty*. Data Science Journal. 4, pp.165–170. DOI: http://doi.org/10.2481/dsj.4.165

[48] Eschenfelder, K. R. and Johnson, A. (2014), *Managing the data commons: Controlled sharing of scholarly data*. J Assn Inf Sci Tec, 65: 1757–1774. doi:10.1002/asi.2308
See also:
Eschenfelder, K. and Johnson, A. (2011), *The Limits of sharing: Controlled data collections*. Proc. Am. Soc. Info. Sci. Tech., 48: 1–10. doi:10.1002/meet.2011.14504801062

probably always continue to arise. Data sharing policies and practices must continue to adapt to new regulations.

The new EU regulation on data protection, the General Data Protection Regulation (GDPR)[49], obtained its final legislative approval on April 14, 2016, and will be enforced in May 2018, replacing the national laws and regulations based on the 1995 EU Data Protection Directive. While the GDPR largely retains the principles and terminology of the 1995 Directive, it aims at a greater level of uniformity across Europe. Although some consider the new regulation to preserve the equilibrium between the necessity of effectively protecting data subjects' rights while allowing the processing of personal data in research http://best-practices.dataverse.org/harvard-policies/harvard-preservation-policy.html[50], including sensitive data (see for example Chassang, 2017[51]), it also adds some new principles with uncertain consequence. E.g. parts of the field remain widely regulated at national level, in particular, regarding the application of research participants' rights. There are also some uncertain consequences connected to issues like a stricter concept of consent, a requirement for data portability, and a "right to be forgotten"[52].

Well-intended phrases and concepts that are often referred to but lacks clear delimitation and instruction for implementation further occludes a clear and consistent handling of legal and ethical issues. Some examples include the FAIR data guiding principles[53], and the principle of providing research data "as open as possible, as closed as necessary" from the Data Management Guidelines of the EU H2020 Programme[54]. Although these statements and concepts are valuable as general guidelines, they may in many instances leave scientists to wonder just how such concepts are to be operationalized and what its effect on the practice of science will be.

The point is that NSD-NORDi must provide services that can capture and handle legal and ethical issues as they occur. We must also provide services that are fully integrated with training and counseling modules. Of-the-shelf-services like Dataverse may reduce our ability to act and react to occurring events and the needs of our designated community. NSD-NORDi must remain flexible and adaptable.

Conclusion

At this time, Dataverse will not be implemented as a production service at NSD. Drawbacks based on the pilot centred mainly on the lack of appropriate extensibility for the broad range of services and research data that NSD aims to curate in near future. NSD will rather refine and further develop internally developed services that can remain flexible towards the

[49] http://ec.europa.eu/justice/data-protection/

[50] http://best-practices.dataverse.org/harvard-policies/harvard-preservation-policy.html

[51] Chassang, G. (2017). *The impact of the EU general data protection regulation on scientific research*. Ecancermedicalscience, 11, 709. http://doi.org/10.3332/ecancer.2017.709

[52] InfoLawGroup: *GDPR: Getting ready for the New EU General Data Protection Regulation*: http://www.infolawgroup.com/2016/05/articles/gdpr/gdpr-getting-ready-for-the-new-eu-general-data-protection-regulation/. Retrieved 03.07.2017.

[53] https://www.force11.org/fairprinciples

[54] http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

broader user 'market', and the needs of the institutions and organisations that NSD will cooperate with.

Generalist repositories like Dataverse are potentially exposed to heterogeneity when it comes to user communities and data types/formats, which disables them to provide consistent, consolidated and shared practices and learning resources. NSD will also run into problems connected to heterogeneity among its designated communities, especially considering the expansion in scope that the archive is currently undergoing (reflected in a recent name change, from "Norwegian Social Science Data Services" to "Norwegian Centre for Research Data"). But a self-developed platform gives us the ability to act, react and adapt to user needs; different institutions and organisations that we cooperate with may have different needs and preferences. Self-developed tools give us the flexibility to address these potential differences quickly and 'smoothly'. It should be noted that this flexibility implies that sufficient resources are available; NSD is fortunate enough to be funded by the Research Council of Norway, various government ministries and other public- and private-sector grants, and, consequently houses a relatively large and experienced developer staff (e.g. Nesstar, DDI, etc.).

Users and organisations are fully aware of the benefits and value of sharing data, yet they call for support (e.g., facilities for storing and maintaining data, facilities for controlling access, facilities for getting credits, etc.). This will require processing and close communication with individual researchers and organisations (e.g. following up on content of data management plans and data papers, handle both data management planning early on in research project and capturing high-quality metadata at the end of a project, etc.), which may not be suitable for a generic service environment as Dataverse.

However, when it comes to simple operations like uploading, self-documenting and sharing uncurated project data, Dataverse can function as *a supplement* to the 'custom-made' NSD-NORDi solutions. In future service provider models, NSD may cooperate closely with some institutions and researchers, providing community training and follow up closely on DMPs and data and metadata quality, while other institutions may be using an NSD Dataverse, modeled after the Front office – Back office model implemented at DANS[55].

Providing consultancy as a DSS would appear to be an obvious and straightforward activity, however, it would take commitment for both parties involved in the process. Undertaking the analysis needs, the understanding of capability, the capabilities, goals, and direction of an SP. This may need considerable time and effort. The value of consultancy is specific and timely, as we see in the context of NSD’s analysis of Dataverse for the NORDi project.

[55] Dillo, I. and Doorn, P. (2014). *The Front Office–Back Office Model: Supporting Research Data Management in the Netherlands*. International Journal of Digital Curation, Vol 9, No 2. DOI: http://dx.doi.org/10.2218/ijdc.v9i2.333

References

Assante, M. et al., (2016). *Are Scientific Data Repositories Coping with Research Data Publishing?*. Data Science Journal. 15, p.6. DOI: http://doi.org/10.5334/dsj-2016-006

Candela, L., Castelli, D., Manghi, P. and Tani, A. (2015). *Data journals: A survey*. J Assn Inf Sci Tec, 66: 1747–1762. doi:10.1002/asi.23358

Chassang, G. (2017). *The impact of the EU general data protection regulation on scientific research*. Ecancermedicalscience, 11, 709. http://doi.org/10.3332/ecancer.2017.709

Dillo, I. and Doorn, P. (2014). The Front Office–Back Office Model: Supporting Research Data Management in the Netherlands. International Journal of Digital Curation, Vol 9, No 2. DOI: http://dx.doi.org/10.2218/ijdc.v9i2.333

Eschenfelder, K. and Johnson, A. (2011). *The Limits of sharing: Controlled data collections*. Proc. Am. Soc. Info. Sci. Tech., 48: 1–10. doi:10.1002/meet.2011.14504801062

Eschenfelder, K. R. and Johnson, A. (2014). *Managing the data commons: Controlled sharing of scholarly data*. J Assn Inf Sci Tec, 65: 1757–1774. doi:10.1002/asi.2308

Parsons, M.A. & Duerr, R., (2006). *Designating user communities for scientific data: challenges and solutions*. Data Science Journal. 4, pp.31–38. DOI: http://doi.org/10.2481/dsj.4.31

Sieber, J.E., (2006). *Ethics of sharing scientific and technological data: a heuristic for coping with complexity & uncertainty*. Data Science Journal. 4, pp.165–170. DOI: http://doi.org/10.2481/dsj.4.165

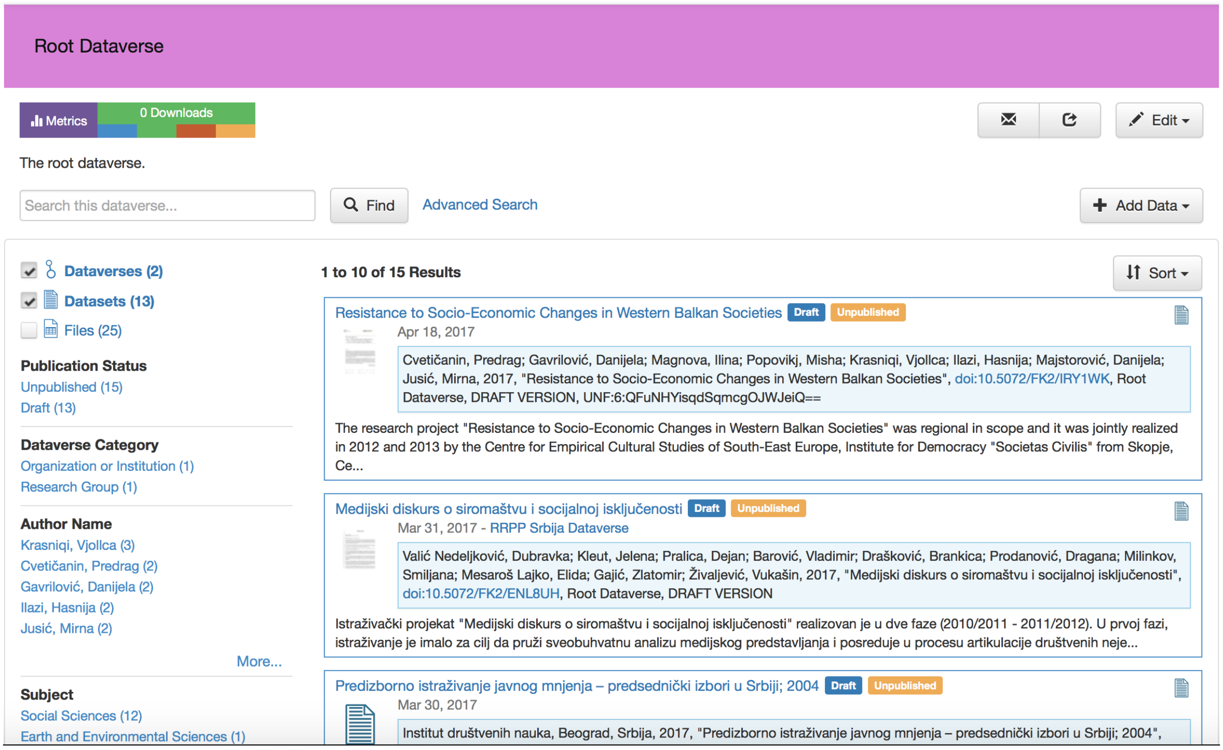
Pilot 4: Remote Technical support for an installation of Dataverse

Aleksandra Bradić-Martinović (IES), Vyacheslav Tykhonov (DANS), Goran Gicić (IES), Marion Wittenberg (DANS)

Introduction

One of the Dataverse pilots concerns a feasibility study of remote technical support. Would it be possible for one SP to give remote technical support to another SP in setting up an instance of Dataverse? The institutes involved were IEN in Serbia and DANS in the Netherlands. DANS has a lot of expertise in running Dataverse servers. IEN has no experience in Dataverse whatsoever and was lacking technical staff, so remote technical support would be very welcome.

The result of this pilot, the Serbian Dataverse instance at IEN can be found online[56]. To see the content, you need to request access.



The root dataverse.

☑ Dataverses (2)
☑ Datasets (13)
□ Files (25)

Publication Status
Unpublished (15)
Draft (13)

Dataverse Category
Organization or Institution (1)
Research Group (1)

Author Name
Krasniqi, Vjollca (3)
Cvetičanin, Predrag (2)
Gavrilović, Danijela (2)
Ilazi, Hasnija (2)
Jusić, Mirna (2)
More...

Subject
Social Sciences (12)
Earth and Environmental Sciences (1)

1 to 10 of 15 Results

Resistance to Socio-Economic Changes in Western Balkan Societies Draft Unpublished
Apr 18, 2017
Cvetičanin, Predrag; Gavrilović, Danijela; Magnova, Ilina; Popovikj, Misha; Krasniqi, Vjollca; Ilazi, Hasnija; Majstorović, Danijela; Jusić, Mirna, 2017, "Resistance to Socio-Economic Changes in Western Balkan Societies", doi:10.5072/FK2/IRY1WK, Root Dataverse, DRAFT VERSION, UNF:6:QFuNHYisqdSqmcgOJWJeiQ==
The research project "Resistance to Socio-Economic Changes in Western Balkan Societies" was regional in scope and it was jointly realized in 2012 and 2013 by the Centre for Empirical Cultural Studies of South-East Europe, Institute for Democracy "Societas Civilis" from Skopje, Ce...

Medijski diskurs o siromaštvu i socijalnoj isključenosti Draft Unpublished
Mar 31, 2017 - RRPP Srbija Dataverse
Valić Nedeljković, Dubravka; Kleut, Jelena; Pralica, Dejan; Barović, Vladimir; Drašković, Brankica; Prodanović, Dragana; Milinkov, Smiljana; Mesaroš Lajko, Elida; Gajić, Zlatomir; Živaljević, Vukašin, 2017, "Medijski diskurs o siromaštvu i socijalnoj isključenosti", doi:10.5072/FK2/ENL8UH, Root Dataverse, DRAFT VERSION
Istraživački projekat "Medijski diskurs o siromaštvu i socijalnoj isključenosti" realizovan je u dve faze (2010/2011 - 2011/2012). U prvoj fazi, istraživanje je imalo za cilj da pruži sveobuhvatnu analizu medijskog predstavljanja i posreduje u procesu artikulacije društvenih neje...

Predizborno istraživanje javnog mnjenja – predsednički izbori u Srbiji; 2004 Draft Unpublished
Mar 30, 2017
Institut društvenih nauka, Beograd, Srbija, 2017, "Predizborno istraživanje javnog mnjenja – predsednički izbori u Srbiji; 2004",

Figure 6: Serbian Dataverse

[56] http://dataverse-serbia.ien.bg.ac.rs/

Process of setting up a server

Installation & configuration of Dataverse

Operating system
A Windows Server was made available at IEN for the pilot study. Since there is currently no Windows executable version of Dataverse, the installation on a Windows Server required the setup of a virtual machine (VM) with the installation of a distribution of Linux OS and Dataverse within the VM.

Dataverse normally runs on CentOS, but at IEN only Ubuntu was available as OS[57]. This made the installation more problematic because IQSS, the institute that develops Dataverse, can't guarantee that Dataverse will run on Ubuntu without issues. Before Dataverse could be installed, IEN had to update their version of Ubuntu from version 14 to 16. Despite this update, standard installation was still not possible. Therefore DANS installed Dataverse as local DANS installation on CentOS and copied the Dataverse database to Ubuntu 16 at IEN. This made it possible to run dataverse on Ubuntu.

Proxy problems
Access to the IEN network was closed from outside. But administrative rights for DANS developers to the IEN network was required to support the installation of Dataverse. TeamViewer, computer software package for remote control was used to bypass the firewall.

Persistent Identifiers
There is no DOI or Handle service utilised by IEN so it was not possible to generate Persistent Identifiers (PID) for deposited datasets within the Serbian Dataverse. At the same time, it is not possible to publish datasets without a PID. So, for the moment there are only private URLs available for demonstration purposes. Without a PID service, it is not possible to run the Serbian Dataverse as a sustainable service.

Mail server
Dataverse sends notifications to users whenever necessary during the data management process, for example, to confirm that a user has uploaded files successfully. EIN had no mail server available in their infrastructure to support the Serbian Dataverse. This issue was solved by connecting a Google mail server to the Serbian Dataverse, to make it possible to send notifications by e-mail.

Maintenance

Backup strategy
IEN has no backup strategy at the moment for the Serbian Dataverse; it's not clear who is going to make backups and when. Within the context of this pilot study this is not essential, but must be addressed if the service is used by researchers and students.

[57] CentOS and Ubuntu are both distributions of the Linux OS

Reliability of service
The data infrastructure used by IEN does not appear to be very reliable as everything went down and crashed when electricity was disabled in Belgrad. DANS had to check the database and restore the Dataverse services. This raises the question who will be responsible if all data will be lost during disruption?

Updates and bug fixing
Updates and bug fixing require negotiations about a time when TeamViewer can be started.

Conclusions

The amount of effort taken to give remote technical support to another organisation, is dependent not only on the technical but also on the organisational capability maturity[58] of the receiving organisation. It is not always clear beforehand, which issues will be encountered. During the pilot, it became obvious that there are not only technical issues that had to be solved. A technical solution is often linked to processes that are defined in policies. For example making backups of data; this has not only to do with the technical solution how to make a backup but also or even more with policies on backup strategies. Who is responsible, when and how many times are backups being made.

This pilot was successful in the sense that DANS provided technical support to IEN and that both organisations worked together in setting up a Serbian Dataverse instance. IEN didn't have enough technical staff to do this on their own. But, on the contrary, to have a Serbian Dataverse instance up and running as a sustainable service, this will take more effort than only technical support. The later was beyond the scope of this pilot.

[58] In task 3.1 of the CESSDA SaW project a Capability Development Model has been developed which can function as a basis which upon an assessment of Social Science service provision can be made. More information about this model: https://www.cessda.eu/eng/Projects/All-projects/CESSDA-SaW/WP3/CESSDA-CDM

Pilot 5: Information Systems Support for Establishing a Localised Self-archiving Tool for Researchers

Maja Dolinar (ADP) and Irena Vipavc Brvar (ADP)

Introduction

One of the identified additional services, needed by the Slovenian Social Science Data Archive (ADP) was to offer a self-archiving tool for researchers. For this reason, ADP started examining tools that would enable such functionality. Looking at current best practices of other archives that already have similar options available, ADP identified Dataverse as one of the tools that would enable easy self-archiving services to our users. In this pilot, ADP tested Dataverse whether it fits their needs, DANS provided a sandbox for testing purposes and consultancy.

Background

At the moment ADP is offering their users only the option to deposit a study through a standard procedure of ingest, which is more thorough and time-consuming. Due to time constraints, ADP can archive only important studies and not smaller, yet perhaps also interesting, projects. In particular, ADP would like to offer the self-deposit service for archiving the data related to PhD dissertation. The process at ADP currently includes an initial decision on whether a research data is appropriate for inclusion in their Catalogue (meeting their criteria, methodological and topic vigour), thorough checking of submitted information and data files by ADP and the final preparation of different versions of descriptions and submitted files for long-term storage and dissemination.

ADP is considering usage of self-archiving tool in the next study year (2017-2018), hence they responded to the call of CESSDA SaW T4.4 team to test the Dataverse application.
Before starting the pilot, ADP identified the users of this additional feature of the ADP and the overall structure of the pilot project. The intended audience is:

1. Slovenian Ph.D. students: that are required to make data of their thesis publicly available;
2. Slovenian researchers: that produce either lower quality data (for example, data that does not reach methodological excellence) or data that are poorer in content and are likely to be used for limited range of users (for example narrower research questions, etc.);
3. Regular users of the ADP: who browse the ADP online catalogue and use the ADP research data. It is expected that this additional resource will broaden the range of topics that they have at disposal.

Based on the identified users, ADP saw that the new functionality should have the following features:

- option for easy self-depositing of a study, with the caveat that files still needs to be checked by human: ADP doesn’t want their users to submit whatever studies or files into their system, they need to enable the possibility that the ADP checks all the

submitted study descriptions and data files before publication – the depositor should not have the possibility of publishing his/her study, only the possibility of submission to the system, the final approval is on the side of the ADP,
- multilingualism: all metadata fields should be available in English and Slovenian,
- browsing of the catalogue of self-archived studies for final users: searching through the catalogue should be easy, based on metadata fields of individual studies,
- online analyses for final users: ADP would need similar functionalities for online analyses as are possible on Nesstar (desk research showed that Dataverse offers a user interface to Zelig, which is a powerful, R-based statistical computing tool, that enables our desired features and statistical analysis models)[59], and
- option for downloading the study data files on local computers.

The pilot was divided into three parts:
- Technical part: what kind of hardware and technical knowledge is needed to develop the tool and maintain it, what support could ADP receive from CESSDA MO and/or CESSDA SPs and the analysis of what is possible in the tool (available plugins and functionalities of Dataverse).
- Instructions for depositors: ADP would need to prepare clear guides for the depositors, documenting the entire workflow, especially focusing on the terms of use (recommended and accepted formats, obligatory metadata fields etc.).
- Instructions for users: ADP would need to prepare clear users guides, focusing on terms of access and instructions on browsing and analyzing research data from self-archived studies.
- For the purpose of the Deliverable 4.5, ADP focused on the technical part of the pilot, since they need to develop the Dataverse application that will fit their needs first.

Initial Analysis of Dataverse Sandbox

The first step of developing a self-archiving tool included an analysis of default functionalities of Dataverse. For this purpose, ADP was given the option to play around with different study descriptions and file uploads in the ADP Sandbox[60] that DANS created for them on the DANS Dataverse instance. Upon testing the features in Dataverse Sandbox by submitting a few studies to the system, ADP identified several issues that needed to be resolved in order to use Dataverse as a tool for self-archiving in the ADP:

Study Metadata

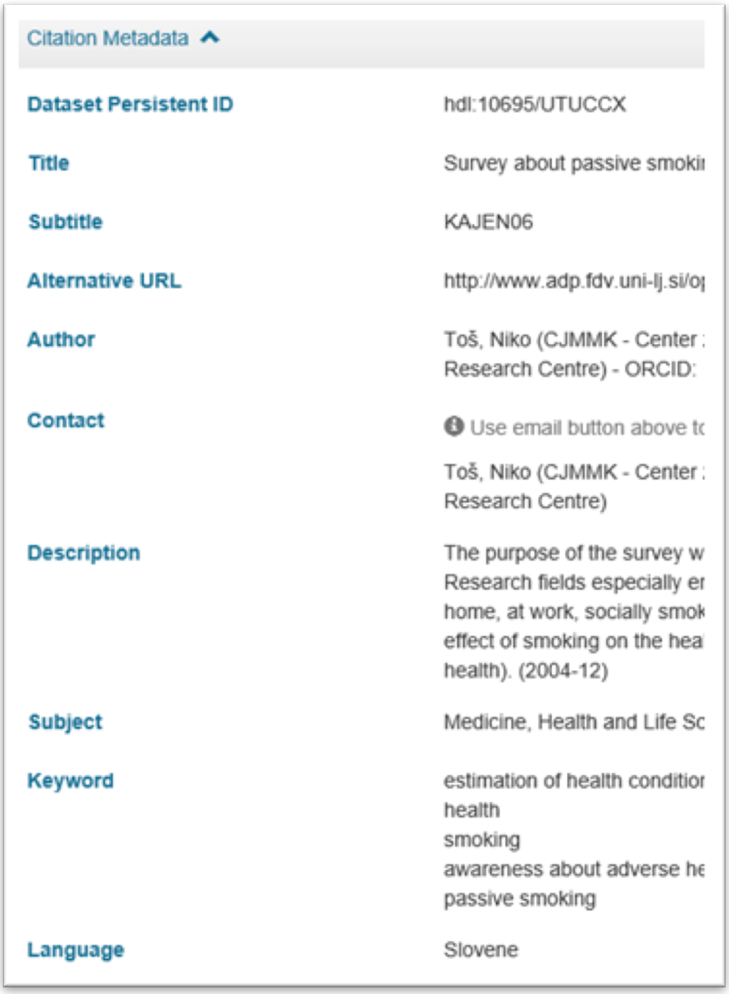
Firstly, ADP noticed that the majority of the metadata fields by default are allowing the user to write a free text. ADP would like to use a uniform list of metadata fields in order to allow browsing by category in their catalogue (as well as other international catalogues, in which the Catalogue of the ADP is also included, for example the CESSDA Catalogue). Since ADP is following the DDI Alliance controlled vocabulary in their current ingest procedure and are

[59] See Harvard Documentation http://guides.dataverse.org/en/3.6.2/dataverse-user-main.html and a tutorial of the Odum Institute https://www.youtube.com/watch?v=Zy0BvCag6ZI
[60] Available at: https://act.dataverse.nl/dataverse/adp

starting to use also the new CESSDA CV vocabulary, they need to adapt the Dataverse instance to allow for controlled vocabulary and closed list of items for metadata fields.

The other issue that was identified in the process of evaluating the default Dataverse option is the limited number of fields offered. ADP goal is that the depositor provides as much information as possible about the deposited study, as this is needed for the overall understanding of the research process and the correct re-usage of research data.

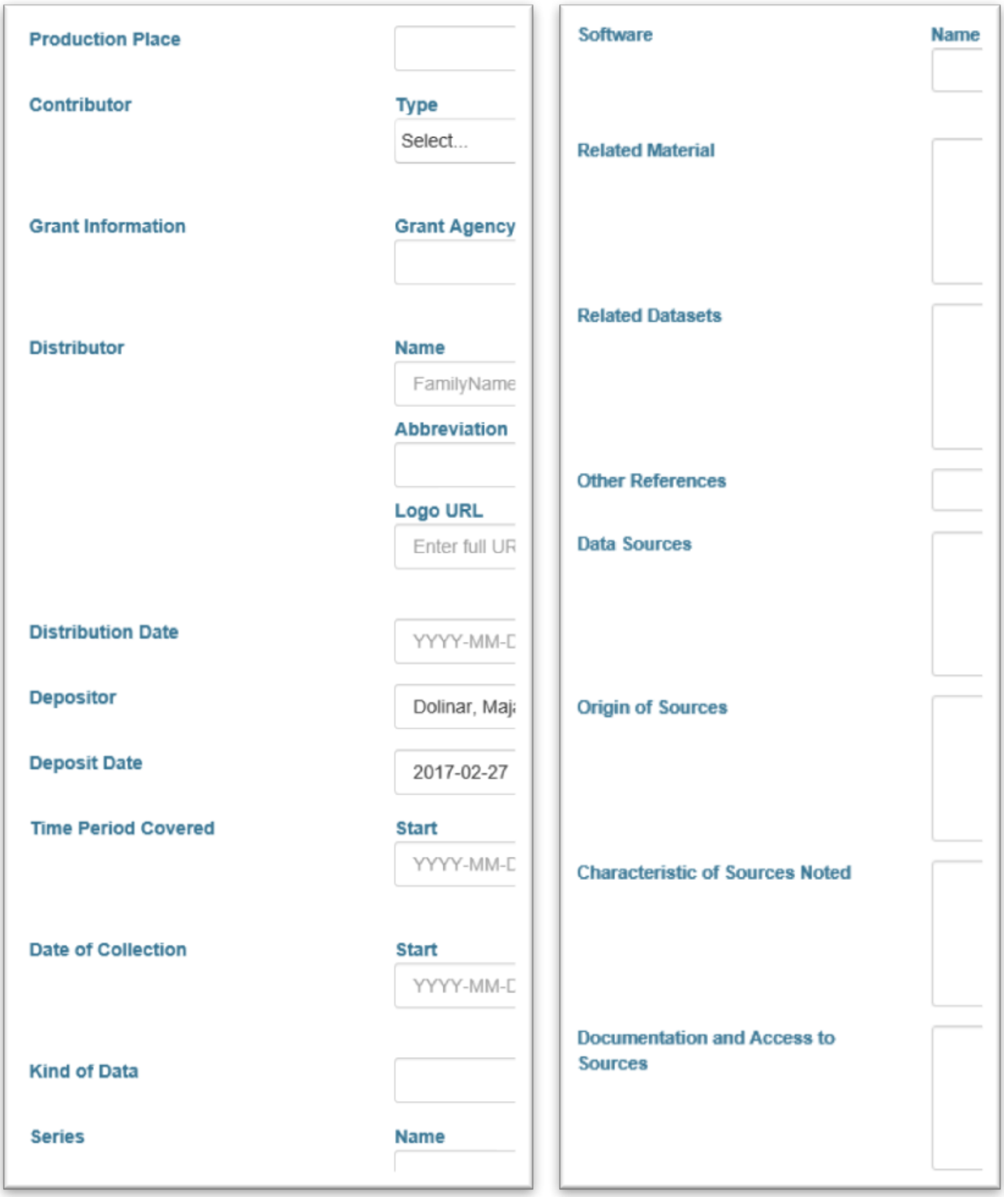
ADP's conclusion is that the Dataverse installation should have more metadata fields that are either restricted in terms of length or a closed list of items to choose from (controlled vocabularies).



Citation Metadata ^

Dataset Persistent ID	hdl:10695/UTUCCX
Title	Survey about passive smoki
Subtitle	KAJEN06
Alternative URL	http://www.adp.fdv.uni-lj.si/o
Author	Toš, Niko (CJMMK - Center Research Centre) - ORCID:
Contact	❶ Use email button above t Toš, Niko (CJMMK - Center Research Centre)
Description	The purpose of the survey w Research fields especially e home, at work, socially smok effect of smoking on the hea health). (2004-12)
Subject	Medicine, Health and Life Sc
Keyword	estimation of health conditio health smoking awareness about adverse h passive smoking
Language	Slovene

Figure 7: Snapshot of a study in the ADP Sandbox – Study metadata level



Production Place

Contributor — Type: Select...

Grant Information — Grant Agency

Distributor — Name: FamilyName — Abbreviation — Logo URL: Enter full UR

Distribution Date: YYYY-MM-D

Depositor: Dolinar, Maj

Deposit Date: 2017-02-27

Time Period Covered — Start: YYYY-MM-D

Date of Collection — Start: YYYY-MM-D

Kind of Data

Series — Name

Software — Name

Related Material

Related Datasets

Other References

Data Sources

Origin of Sources

Characteristic of Sources Noted

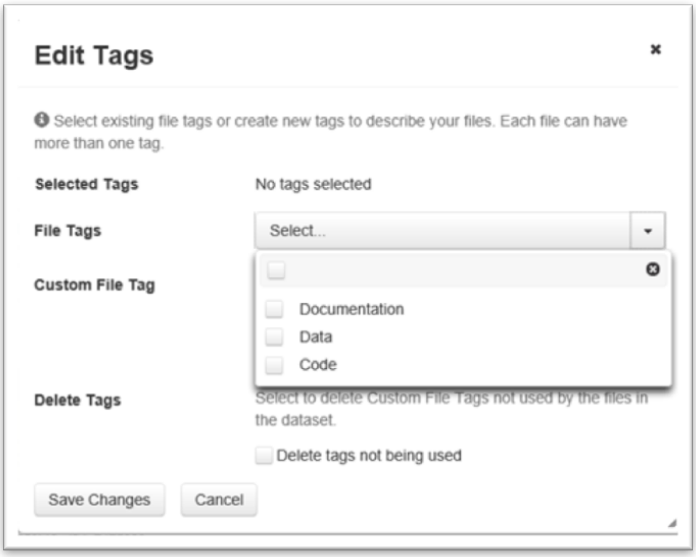
Documentation and Access to Sources

Figure 8: (Continuation) Snapshot of a study in the ADP Sandbox – Study metadata level

Dataset Metadata

On the level of data metadata, additional issues were identified. ADP would like to restrict data file formats that can be uploaded by the user, having in mind that they would like to offer support with the tools they have at hand. Dataverse, however, seems to accept nearly every data file format possible.

When ADP is going to use Dataverse, they will need to write clear instructions for depositors regarding the kind of data files they will accept and the kind of information the depositors have to provide to ADP. Dataverse provides the possibility to tag a file (*Documentation*, *Data* or *Code*). ADP would like to expand this description.



Edit Tags ×

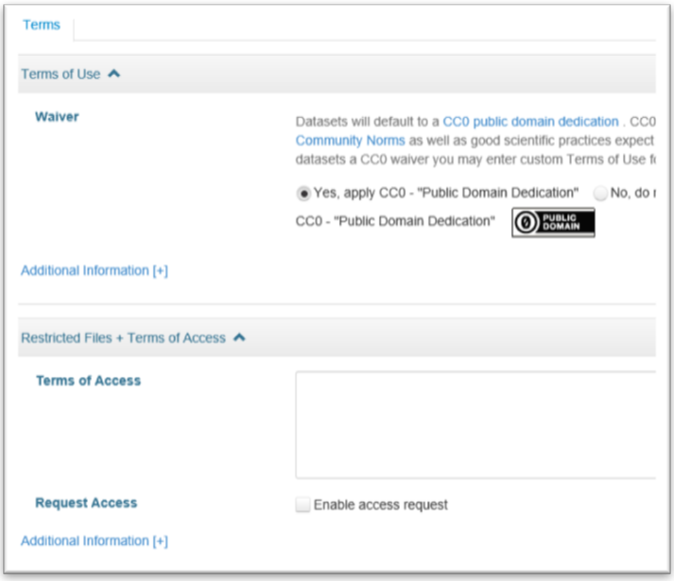
Select existing file tags or create new tags to describe your files. Each file can have more than one tag.

Selected Tags	No tags selected
File Tags	Select... ▾
Custom File Tag	□ Documentation □ Data □ Code
Delete Tags	Select to delete Custom File Tags not used by the files in the dataset. □ Delete tags not being used

Save Changes | Cancel

Figure 9: Snapshot of a study in the ADP Sandbox – File tags

Regarding of *Terms of Access* to an individual data file, Dataverse is again very liberal. The depositor may decide to offer his data file under the CC0 license or write whatever data file access restriction he/she believes it is best for the use of the study. In the ADP they would like to limit the possibilities only to their main licenses, that are CC-BY and CC-BY-NC – therefore they would need a closed list of options. Any other sort of license should be negotiated with the archive already at the deposit step.



Terms

Terms of Use

Waiver — Datasets will default to a CC0 public domain dedication . CC0 Community Norms as well as good scientific practices expect datasets a CC0 waiver you may enter custom Terms of Use f

● Yes, apply CC0 - "Public Domain Dedication" ○ No, do n

CC0 - "Public Domain Dedication" PUBLIC DOMAIN

Additional Information [+]

Restricted Files + Terms of Access

Terms of Access

Request Access — □ Enable access request

Additional Information [+]

Figure 10: Snapshot of a study in the ADP Sandbox – Terms of use

Development of Dataverse Application for Self-Archiving

Upon this initial testing in their Sandbox, ADP decided to prepare list of metadata fields (see Appendix 1) and check whether their preferences for individual fields would be possible to implement in the Dataverse.

DANS offered them help in the customization, so after the preparation of the proposed metadata fields that are following ADP’s current practices of ingest and the new CESSDA Topic Classification and CVs, DANS started working on adapting Dataverse to their needs. For now, ADP is working only on the English version of Dataverse since the question of multilingualism will be part of the follow-up project.

For the metadata fields in the User Interface (UI), Dataverse is using metadata schema saved as Tab Separated Values files stored on Harvard’s Google Drive[61]. By using native API requests, it is possible to extend Dataverse with new schemas and update fields in the existent schema. DANS experimented with adding new schema both in English and Slovenian but faced the problem with the handling of multiple fields with the same name (like title and description) as originally Dataverse was designed without considering support of multiple languages. More time should be dedicated to investigating possible solutions to this issue.

Further Developments

ADP plan to have their customized Dataverse instance installed on their server by the end of September 2017, so that they will be able to have a pilot version of their self-archiving tool ready for the new study year. To reach this point they need to have a customized version of Dataverse from DANS, which they will test, and write clear and easy instructions for their users (depositors and users).

ADP is looking forward to language specific (Slovenian) version of Dataverse in the near future, as it is necessary to have multilingual tool for their self-archiving needs (English and Slovenian). This is planned within the CESSDA’s Dataverse EU project in 2018. During 2018 ADP would like also to focus more on the online analyses possibilities within Dataverse – which is something that they were not yet able to test within this pilot project since it was not part of their sandbox installation. Based on ADP’s initial desk analysis of the possibilities, there are good prospects that Dataverse can be used also in the area of online analyses, suitable for their users.

Conclusion

This pilot proved that it is very helpful for one SP (in this instance ADP), to make use of the technical infrastructure (Sandbox) and knowledge of another SP (in this instance DANS). The cooperation between the two SPs made the process of testing for the needs of ADP more straightforward.

[61] https://docs.google.com/spreadsheets/d/13HP-jI_cwLDHBetn9UKTREPJ_F4iHdAvhjmlvmYdSSw/edit#gid=8

Pilot 6: Dataverse for Dissemination - use case of the TÁRKI Data Archive

Péter Hegedűs (TÁRKI), Vyacheslav Tykhonov (DANS), *Marion Wittenberg (DANS)*

Introduction

This pilot aimed at giving technical support in the process of investigating whether a tool, in this case Dataverse, could function as a metadata publishing system for the TÁRKI Data Archive. The metadata publishing system TÁRKI is currently using needs an upgrade and TÁRKI is searching for a more flexible system. DANS, who has much experience in the use of Dataverse, gave technical support to TÁRKI.

Background TÁRKI Data Archive

TÁRKI Data Archive, operated by TÁRKI Foundation, is the national social science archive in Hungary. Over the past three decades, TÁRKI Data Archive collected and archived more than 800 empirical social research data collections that are suitable for secondary analysis. The focus of the collection is Hungarian research. Most of the collection comprises micro data files from nationally representative sample-survey studies. Part of the collection consists of TÁRKI's own surveys, another part is coming from other Hungarian research institutes.

The research department TÁRKI takes part in international projects like International Social Survey Programme (ISSP), European Social Survey (ESS),) and (Luxembourg Income Study (LIS). More information about the services of TÁRKI on their website[62].

Metadata

The metadata schema TÁRKI is using for data documentation is based on the DDI standard. See Appendix IV for the metadata elements. The information is currently stored in an SQL database, and published on the website.

Steps of Dataverse for Dissemination Pilot

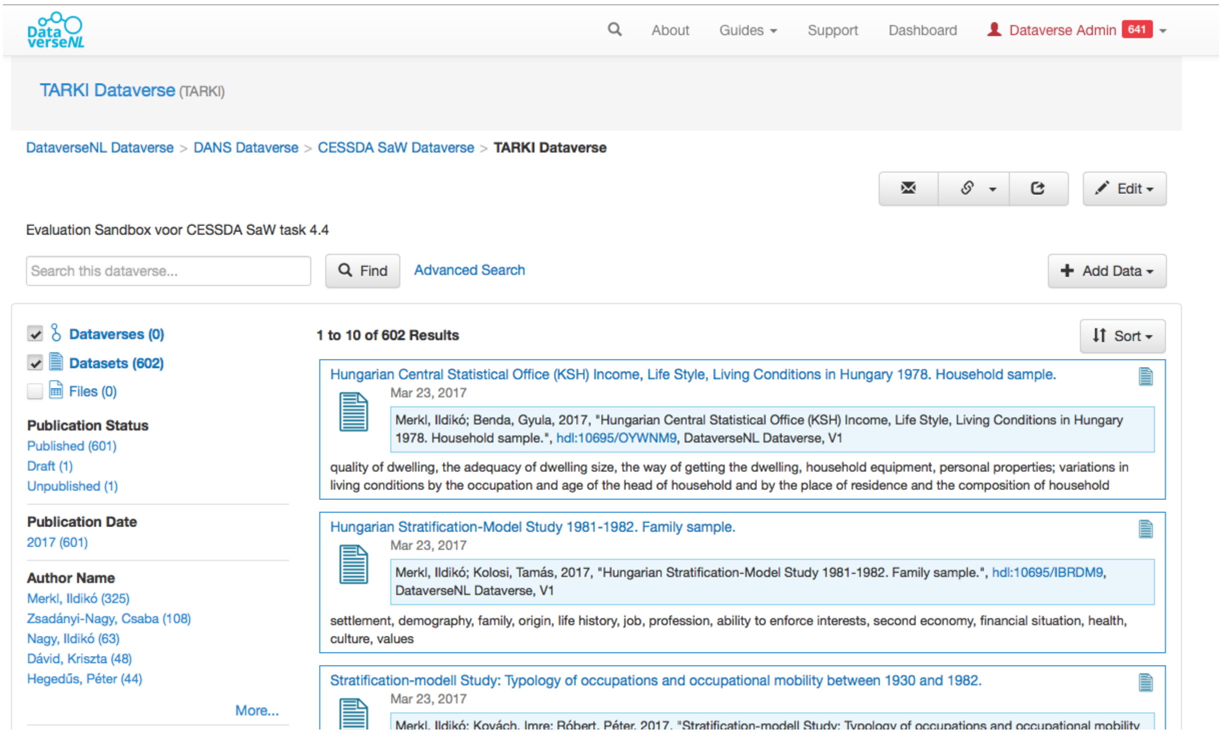
Information on TÁRKI endpoint for metadata.

TÁRKI has provided information to DANS, where to find the metadata of the catalogues of publications[63] and data collections[64]. As TÁRKI doesn't have an OAI-PMH endpoint, DANS has developed CESSDA-SaW client to harvest all their metadata directly from web pages in HTML. This client is available for the public as open source software in GitHub[65]. Mappings have been created to convert all fields to Dublin Core (DC) to be used as source to import metadata to Dataverse. The converted metadata in DC is published for review in GitHub[66]

[62] http://www.tarki.hu/en/services/da/index.html
[63] http://www.tarki.hu/cgi-bin/katalogus/biblio.pl?sorszam=TPUBL-B271
[64] http://www.tarki.hu/cgi-bin/katalogus/tarkimain_en.pl?sorszam=TDATA-D76
[65] https://github.com/DANS-KNAW/cessda-saw
[66] https://github.com/DANS-KNAW/cessda-saw/tree/master/data

Metadata transfer from TÁRKI repository to Dataverse

In order to get all TÁRKI metadata in one place, DANS used Dataverse sandbox to create separate TÁRKI Dataverse. All datasets metadata delivered as Dublin Core were imported by CESSDA-SaW client to this TÁRKI Dataverse container[67] and available for TÁRKI for evaluation. Overall, the automatic metadata transfer worked well, but it needs refinements for some metadata elements.



TARKI Dataverse (TARKI)

DataverseNL Dataverse > DANS Dataverse > CESSDA SaW Dataverse > TARKI Dataverse

Evaluation Sandbox voor CESSDA SaW task 4.4

Search this dataverse...

☑ Dataverses (0)
☑ Datasets (602)
□ Files (0)

Publication Status
Published (601)
Draft (1)
Unpublished (1)

Publication Date
2017 (601)

Author Name
Merkl, Ildikó (325)
Zsadányi-Nagy, Csaba (108)
Nagy, Ildikó (63)
Dávid, Kriszta (48)
Hegedűs, Péter (44)

1 to 10 of 602 Results

Hungarian Central Statistical Office (KSH) Income, Life Style, Living Conditions in Hungary 1978. Household sample.
Mar 23, 2017
Merkl, Ildikó; Benda, Gyula, 2017, "Hungarian Central Statistical Office (KSH) Income, Life Style, Living Conditions in Hungary 1978. Household sample.", hdl:10695/OYWNM9, DataverseNL Dataverse, V1
quality of dwelling, the adequacy of dwelling size, the way of getting the dwelling, household equipment, personal properties; variations in living conditions by the occupation and age of the head of household and by the place of residence and the composition of household

Hungarian Stratification-Model Study 1981-1982. Family sample.
Mar 23, 2017
Merkl, Ildikó; Kolosi, Tamás, 2017, "Hungarian Stratification-Model Study 1981-1982. Family sample.", hdl:10695/IBRDM9, DataverseNL Dataverse, V1
settlement, demography, family, origin, life history, job, profession, ability to enforce interests, second economy, financial situation, health, culture, values

Stratification-modell Study: Typology of occupations and occupational mobility between 1930 and 1982.
Mar 23, 2017
Merkl, Ildikó; Kovách, Imre; Róbert, Péter, 2017, "Stratification-modell Study: Typology of occupations and occupational mobility

Figure 11: TÁRKI Dataverse container in DANS sandbox

Linking files from TÁRKI storage to appropriate metadata descriptions in Dataverse

Inside of TÁRKI infrastructure files are stored in ftp archive and separated from metadata descriptions, so there is no direct link between them. When a request is coming to get files from the specific dataset TÁRKI staff is searching for it inside of ftp archive and sending the file(s) by email to the user. It was beyond the scope of this pilot to get metadata descriptions and files linked together in Dataverse. However, a lot has to be done before this can be accomplished; TÁRKI has to structure all files in a way that will allow linking of every file to the appropriate metadata description. For example, files have to put in folders related to the number of the dataset.

Evaluation of Dataverse

After the migration of metadata into the TÁRKI Dataverse, TÁRKI evaluated Dataverse whether it would fit for their purposes.

[67] https://act.dataverse.nl/dataverse/tarki

A follow-up project should look into the following aspects:

1. What are the system requirements Dataverse needs? How can Dataverse be compatible to the TÁRKI own system?
2. Multilinguality: Currently the TÁRKI metadata system is available in Hungarian and in English. For TÁRKI it is important that Dataverse would also offer this possibility.
3. Customizing possibilities of the Dataverse to fit the TÁRKI own endpoint system.
4. Is it possible to link Dataverse other repository systems, for example for publications

Conclusions

The pilot was successful in that respect that one SP (DANS) provided technical support to another SP (TÁRKI) in achieving the goal to transfer the metadata from the system TÁRKI is currently using, to the Dataverse system.

The pilot shows that the metadata system from the TÁRKI can be migrated to the Dataverse system. So, when TÁRKI would like to use the Dataverse system in the near future, migration can be done automatically. To have a production Dataverse server up and running within TÁRKI technical infrastructure, was beyond the scope of this pilot.

Pilot 7: Dataset Acquisition Utilising Dataverse in Tandem with Islandora for Long-term Preservation

Vyacheslav Tykhonov (DANS), Alen Vodopijevec (IRB)

Aim

This pilot aims to investigate the provision of workflow development and software development remotely. FFZG (Sveuciliste U Zagrebu Filozofski Fakultet) wish to use Dataverse as an infrastructure service for the acquisition of datasets from social science researchers, in cooperation with the Croatian national data archiving infrastructure (based upon Islandora) to provide long-term preservation of the datasets. DANS has experience in developing workflows connecting services and thus, was in the position to provide this DSS pilot.

Method

IRB on behalf of FFZG installed a Dataverse instance on a server in Croatia as part of the test platform, such that testing by FFZG could be carried out entirely in Croatia.

DANS developed a Dataverse Bridge that can be used to ‘plug in’ Trusted Digital Repositories (TDR) using Bagit as hierarchical file packaging format for use with, for example, Islandora, DANS EASY, Archivematica, and possibly other systems. All source code for the software has been published as open source in GitHub[68] and contributed to Dataverse Community.

An integration workflow implemented previously in an environment of Dataverse and Archivematica (see figure 12) was reused here for this pilot study, thus allowing the adaptation work to be completed rapidly. During the development of the Bridge, Dataverse functionality was extended with new API endpoint serving conversion of metadata from DDI to Trusted Digital Repository specific format (MODS, METS or EDM) in a BagIt package.

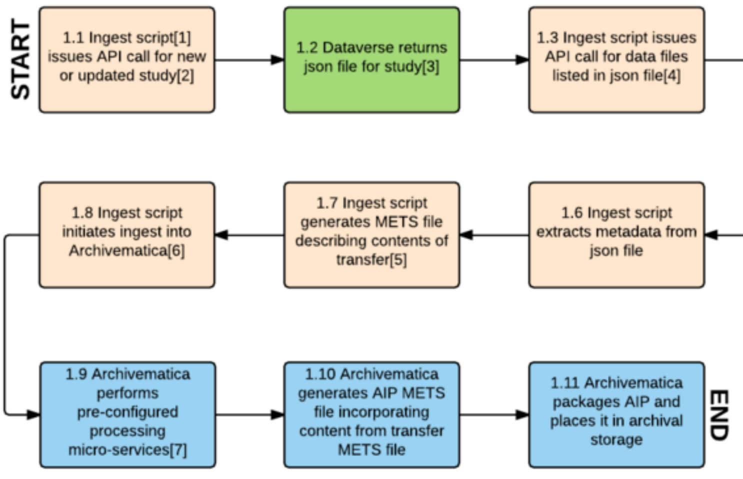

START
1.1 Ingest script[1] issues API call for new or updated study[2]
1.2 Dataverse returns json file for study[3]
1.3 Ingest script issues API call for data files listed in json file[4]
1.4 Dataverse returns all data files listed in json file
1.5 If bundled files are retrieved, ingest script unpackages the bundle(s)
1.6 Ingest script extracts metadata from json file
1.7 Ingest script generates METS file describing contents of transfer[5]
1.8 Ingest script initiates ingest into Archivematica[6]
1.9 Archivematica performs pre-configured processing micro-services[7]
1.10 Archivematica generates AIP METS file incorporating content from transfer METS file
1.11 Archivematica packages AIP and places it in archival storage
END

Figure 12: Dataverse/Archivematica integration (source: https://wiki.archivematica.org/Dataverse)

[68] https://github.com/DANS-KNAW/dataverse-bridge

SWORD (Simple Web-service Offering Repository Deposit) is used by the Dataverse Bridge to harvest all metadata for the datasets that will be archived. Usually, the process of harvesting takes a long time so we have extended the interface of Dataverse to retrieve information about the status of the archival process. After the process finishes the dataset version will get an 'ARCHIVED' status and its description will be updated with a Persistent Identifier (PID) pointing to the archived version.

Conversion process

The process of metadata conversion is very flexible and is based upon Extensible Stylesheet Language Transformations (XSLT), which facilitates the mapping of all required Dataverse metadata to any XML format, such as MODS (Islandora), METS (Archivematica) or EMD (EASY), that can be used by BagIt software to create hierarchical packages with all metadata fields and files. After the creation of a BagIt package, Dataverse Bridge transmitting it to the Trusted Digital Repository API endpoint to initiate depositing process to archive the dataset. After the archiving process is completed, the TDR returns a new PID that will be stored in Dataverse and points to the archived version of the dataset.

Testing

At the time of writing, the Dataverse Bridge has undergone testing with the test installation of the EASY system at DANS. Testing of the Islandora integration is currently taking place in Croatia and will be done by IRB/FFZG in the cooperation with Hrcak Portal of scientific journals of Croatia (Srce).

Results/Outcome

The open source code for this pilot can be found on Github[69]. Dataverse Bridge was presented during CESSDA PID workshop in May 2017 and the presentation is available online[70].

Conclusions

By extending the number of examples use cases for the workflow from Dataverse as an acquisition service with additional TDR examples, this pilot study has, as a consequence, broadened the usability of the Bridge such that may be used by other CESSDA SPs and aspiring SPs. By ensuring wider usage it is hoped that the sustainability of the Bridge will be improved as more users will have a vested interest in maintaining the software.

Collaborative software development at a distance has the additional challenge of communication and understanding of the situation, implicit requirements, development methodologies & practices, and differing prioritisation.

[69] https://github.com/DANS-KNAW/dataverse-bridge
[70] https://www.slideshare.net/vty/cessda-persistent-identifiers

Conclusions

The approach of evaluating need and supply of development support services through workshops meant that we received some immediate feedback on requirements from SaW partners. However, this was a self-selecting process and was not a complete, or in-depth, analysis across all the partners in CESSDA-SaW and CESSDA SPs of both needs and provision of DSS. It also became clear through the workshops that there were very few existing DSS on offer from established SPs[71], therefore 'matchmaking' between DSS provider and the consumer was not really a possibility[72]. Moreover, partners present in the task did not necessarily need or could supply, DSS suitable for a pilot study. By utilising the framework of long-term services and discreet activities to define the framing of the needs of the partner institutions we were able to develop a set of pilot studies that would expose the opportunities and challenges of development support services.

The self-selecting nature of the workshops did, however, develop a small community of committed institutions interested in Dataverse, for whom it was a possible test and evaluation differing forms of service and activity delivery. Many of this engaged group were not partners within the task or even work package. We are grateful for their contribution and commitment to the success of this task and outcomes. This breadth of the pilot studies undertaken will contribute to the models proposed in deliverable D4.6: Report on sustainability model of development support services.

Many of the Dataverse pilots did not consist of single types of activities, as defined in the categorisation, and were formed of a number of activity elements. For example, the pilot studies with TÁRKI, ADP, IEN and FFZG all had aspects of consultancy, requirements analysis, and technical support activities supplied by DANS. Both the TÁRKI and FFZG pilots had workflow development as an activity. Development support service pilots were bespoke for all of these partners and this may be the method expect expected in the future for new SPs, as not all SPs are at the same maturity level. It maybe such that packages of services and activities are required by SPs for rapid development to meet their obligations within CESSDA ERIC.

Software development, support, and consultancy have proved difficult at a distance. Priorities and the ability commit time and effort has not been equal between partners in the pilots. Clarity in expectations and what is to be developed in a DSS would be helped with SLAs such as for example in the Hosting services for the geographic diversity of backups pilot. However, it is clear that specialist bespoke support and expertise is needed for new and aspiring SPs. The

[71] The few identified being: GESIS DA|RA PID service, NSD NESSTAR, FORS FORSbase, and LTP by GESIS for other institutions on an *ad hoc* basis.

[72] See p.27 of Annex 1 (part A) of Grant Agreement 674939 - CESSDA-SaW:
'a. Establish the demand for development support services (also on the basis of 3.2)
b. Establish the supply of development support services
c. Pilot of delivery of development support services on the basis of a. and b.'

application of clear DSS agreements is essential when support is essentially provided ‘free’ through indirect funding (e.g. SaW and other projects).

Hosting services for the geographic diversity of backups pilot gave an interesting insight into how longterm SP-to-SP collaborative support could be undertaken. If CESSDA wishes to increase sustainable cooperation between SPs then this model of support should be investigated further.

Pilot studies are only a starting point, and for a number of pilots further work is required to establish them as production solutions. The project timeframe was too short to fully complete all pilots to a usable state as none were off-the-shelf solutions.

We did not develop pilot studies for training requirements or policy development needs as these were core aspects of other tasks in SaW, even though these were identified as needs in the first workshop. Future provision of development support services may require packaging of different activities to fulfill a need of, for example a new SP, which may require training for a core service and policy development for procedures to use the service. Whether this packaging of activities should be done *ad hoc* or as a ‘welcome package’ for new SPs is something to consider in promoting membership and as a policy for CESSDA ERIC.

The interest and involvement in this task indicate that there is a clear and pressing need for development support services for small, new, and aspiring SPs to help them meet their strategic goals of supporting their designated communities quickly and efficiently, as well as meeting the CESSDA Statutes Annex 2: Obligations of Service Providers[73]. Furthermore, as CESSDA further evolves as an ERIC, even established SPs may need DSS to develop their service provision and meet their obligations.

Each institution should regularly utilise the CESSDA-CDM (D3.1)[74] to evaluate their current situation, determine gaps, and create their own structured development plan. This will help not only identifying needs for DSS for the institution but also what services CESSDA and SPs could reasonably develop to support community maturity as a whole.

There is considerable experience within the SP community, and therefore the community is best placed to provide support services to others of the community. The challenges are to understand what support is needed over time, how this SP-to-SP support can be provided, and how this can be maintained in the community of SPs and data archives. This will be further elaborated in the Deliverable D4.6: Report on sustainability model of development support services.

[73] https://www.cessda.eu/eng/content/download/316/2908/file/Annexes-to-Statutes-for-CESSDA-210213-Final-Version-brand.pdf
[74] https://www.cessda.eu/eng/Projects/All-projects/CESSDA-SaW/WP3/CESSDA-CDM

List of Figures & Tables

Glossary and Further Information

Archivematica is a web- and standards-based, open-source application which allows your institution to preserve long-term access to trustworthy, authentic and reliable digital content. https://www.archivematica.org/en/

BagIt is a hierarchical file packaging format for the creation of standardised digital containers called 'bags,' which are used for storing and transferring digital content. A bag consists of a ‘payload’ of digital content, and ‘tags' (metadata files) to document the storage and transfer of the bag[75]. https://docs.google.com/document/d/1JqKMFn9KfeIMAAEdOGQr6LZPqNWx8Qubi12uoUXi2QU/edit?usp=sharing

CESSDA Technical Framework is concerned with building and maintaining the technical infrastructure that forms the backbone of CESSDA’s Research Infrastructure. https://www.cessda.eu/eng/Research-Infrastructure/Technical-Framework

Creative Commons Creative Commons provides free, easy-to-use copyright licenses to make a simple and standardized way to give the public permission to share and use your creative work–on conditions of your choice. https://creativecommons.org/

DSA - Data Seal of Approval is formulated in 16 quality guidelines for the application and verification of quality aspects with regard to creation, storage and (re)use of digital research data. https://www.datasealofapproval.org/en/

DataCite is a leading global non-profit organisation that provides persistent identifiers (DOIs) for research data. Our goal is to help the research community locate, identify, and cite research data with confidence. https://www.datacite.org/index.html

Dataverse is an open source web application to share, preserve, cite, explore, and analyze research data. It facilitates making data available to others, and allows you to replicate others' work more easily. https://dataverse.org/

Dataverse Bridge pushes datasets from the Dataverse temporary repository to TDR. https://github.com/DANS-KNAW/dataverse-bridge

DDI - Data Documentation Initiative is an international standard for describing the data produced by surveys and other observational methods in the social, behavioral, economic, and health sciences. https://www.ddialliance.org/

DDI-C 2.5 - DDI-Codebook 2.5 is a more lightweight version of the standard, intended primarily to document simple survey data. http://www.ddialliance.org/Specification/DDI-Codebook/2.5/

[75] http://www.dcc.ac.uk/resources/external/bagit-library

FAIR principles is a set of guiding principles to make data Findable, Accessible, Interoperable, and Re-usable. https://www.force11.org/group/fairgroup/fairprinciples

Islandora is an open-source software framework designed to help institutions and organizations and their audiences collaboratively manage, and discover digital assets using a best-practices framework. https://islandora.ca/

GitHub is a development platform where you can host and review code, manage projects, and build software alongside millions of other developers. https://github.com/

JSON Schema is a vocabulary that allows you to annotate and validate JSON documents. http://json-schema.org/

Microservice, also known as the microservice architecture, is an architectural style that structures an application as a collection of loosely coupled services, which implement business capabilities. The microservice architecture enables the continuous delivery/deployment of large, complex applications. It also enables an organization to evolve its technology stack. http://microservices.io/

MongoDB is a document database with the scalability and flexibility that you want with the querying and indexing that you need. It stores data in flexible, JSON-like documents, meaning fields can vary from document to document and data structure can be changed over time. https://www.mongodb.com/what-is-mongodb

OAI:DC is Dublin Core metadata adjusted for usage in the OAI-PMH. Schema imports the Dublin Core elements from the DCMI schema for unqualified Dublin Core.

OAI-PMH - Open Archives Initiative - Protocol for Metadata Harvesting provides an application-independent interoperability framework based on metadata harvesting. http://www.openarchives.org/OAI/openarchivesprotocol.html

ORCID - Open Researcher and Contributor ID provides an identifier for individuals to use with their name as they engage in research, scholarship, and innovation activities. We provide open tools that enable transparent and trustworthy connections between researchers, their contributions, and affiliations. We provide this service to help people find information and to simplify reporting and analysis. https://orcid.org/

Python is a programming language developed under an OSI-approved open source license, making it freely usable and distributable. https://www.python.org/about/

REST - RESTful - Representational state transfer or RESTful web services is a way of providing interoperability between computer systems on the Internet. REST-compliant Web services allow requesting systems to access and manipulate textual representations of Web resources using a uniform and predefined set of stateless operations. https://en.wikipedia.org/wiki/Representational_state_transfer

SWORD - Simple Web-service Offering Repository Deposit is a lightweight protocol for depositing content from one location to another. It stands for Simple Web-service Offering Repository Deposit and is a profile of the Atom Publishing Protocol. http://swordapp.org/about/

Twelve-factor app is a methodology for building software-as-a-service apps. It can be applied to apps written in any programming language, and which use any combination of backing services (database, queue, memory cache, etc). https://12factor.net/

Ubuntu is an open source software operating system that runs from the desktop, to the cloud, to all your internet connected things. https://www.ubuntu.com/

XML- Extensible Markup Language is a markup language that defines a set of rules for encoding documents in a format that is both human-readable and machine-readable. https://en.wikipedia.org/wiki/XML

Appendices

Appendix I - Pilot 1: Hosting services for the geographic diversity of backups

Model Service Level Agreement

SERVICE LEVEL AGREEMENT

This Service Level Agreement ("SLA") between

<<Service Provider 1>>

and

<<Service Provider 2>>

Subject

The subject of this SLA is the use of the disk space of one party by the data of another party, the access to the data, availability of the services and failures.

Definitions

1) The following are definitions of capitalized words used in this Agreement:
   a. "Business Hours" means 8:00 a.m. to 6:00 p.m. (CET/CEST), Monday through Friday, and, notwithstanding the foregoing, does not include times during Service Maintenance.
   b. "Service Maintenance" means the time when the party is maintaining the Service including software. Service Maintenance includes, without limitation, database index rebuilding, hardware upgrades, software upgrades, and network upgrades, as applicable.
   c. "Data Retention" means each contracting party shall make a full backup copy of each database and file system daily and retain each such daily backup copy for seven (7) days. The backup copy will be stored in other location than original data.
   d. "Data Storage". Each Party shall reserve the disk space of <<500 GB>> or more for data of another party.
   e. "Error" means the situation when the Contract Party cannot connect to the Data Storage or the data are inaccessible.

Service Availability

2) Service Availability - General. Both parties goal is to provide Service Availability twenty-four hours per day, seven (7) days per week (referred to as "24x7 Availability") except during times of Service Maintenance. However, the parties recognize that 24x7

Availability is only a goal, and both parties cannot represent or guarantee that such goal can be achieved.

3) Service Availability Level Goals. Both parties Services shall use reasonable efforts to achieve the target Service Availability Goal of 99.99% network uptime except during scheduled Service Maintenance ("Service Commitment"). Notwithstanding the foregoing, both parties recognize that the Internet is comprised of thousands upon thousands of autonomous systems that are beyond the control of any party. Routing anomalies, asymmetries, inconsistencies and failures of the Internet outside of the control of contract party can and will occur, and such instances shall not be considered any failure of the 99.99% network uptime. Both parties will monitor network uptime and the results of these monitoring systems shall provide the sole and exclusive determination of network uptime, response time goals cover.

4) Response Time to Error. Both parties have internal notification tools for service problems. Additionally, both parties may report problems to the other party.

Service Maintenance

5) Both parties shall provide Service Maintenance, which may cause errors or unavailability of Services and therefore both parties shall use commercially reasonable efforts to limit Service Maintenance, which causes Errors to two (2) hours per month. Each party shall notify the other party by email prior to performing any Service Maintenance which the party predicts will cause a Severity 1 Error outside of standard Service Maintenance Times.

6) Both parties shall attempt to schedule Service Maintenance outside Business Hours.

7) However, the parties agree that it may be necessary to perform Service Maintenance during Business Hours after notice to the other party.

Data Storage

8) The Data Storage is available for the research data of another party.

9) Each party is responsible for the coding, decoding, protecting and coherence of its data which are stored in the other party Data Storage.

10) Contract party providing the Data Storage is not responsible for the data content of the other party.

11) Each party must guarantee Service Availability.

12) Each party must provide regular full back-up of the Data Storage in other location than original Data Storage.

13) Data Room where the Data Storage is located must conclude to general regulations and rules for such technology – fire protection, anti-theft measures, non-flooding area, emergency power supply, cooling system. On request each party must prove the concordance to such rules and regulations or enable personal inspection by a responsible employee of the other party.

Responsibility Limitation

14) Disclaimer of Actions Caused by and/or Under the Control of Third Parties. Each party does not and cannot control the flow of data to or from Data Storage and other portions of the internet. Such flow depends on the performance of internet services provided or controlled by third parties. At times actions or inactions of such third parties can impair or disrupt other party´s connections the internet. Although each party will use reasonable efforts to take actions it deems appropriate to remedy and avoid such events, the contracting party cannot guarantee that such events will not occur. Accordingly, both parties disclaim any and all liability resulting from or related to such events.

15) Limitations. Each party cannot assume responsibility and shall not be liable for any impacts on Service Availability due to (i) any requests for non-standard environment or other party machine access; (ii) any downtime caused by other party´s produced code; or (iii) any changes to the Service by parties other than contract party. Contract party will make reasonable efforts to ensure that Service changes do not affect the other party.

Claims

16) All SLA claims should be communicated via email within seven (7) days of the incident. The notice must include all relevant information, including IP address, a full description of the incident, and any logs (if applicable).

Termination

17) Each party is authorized to terminate this SLA by written notice to the other party. The contract terminates the last day of the 3rd month from the delivery of such notice to the other party.

18) Within 7 days after termination of this contract, each party must provide written confirmation to other party stating all data including back-ups were deleted.

Closing provisions

19) Contact persons: Each contracting party is obliged to nominate its contact person responsible for communication relating to this SLA. Each party can replace such contact person by later communication.

20) This Contract may be executed in counterparts, each of which shall be deemed an original, but all of which together shall be deemed to be one and the same agreement. A signed copy of this Agreement delivered by facsimile, e-mail, or other means of electronic transmission shall be deemed to have the same legal effect as delivery of an original signed copy of this Agreement.

21) This Agreement, together with its exhibits, constitutes the complete and exclusive understanding and agreement between the parties and supersedes all prior understandings and agreements, whether written or oral, with respect to the subject matter hereof. All representations and warranties set forth herein shall survive the termination of this Agreement indefinitely. Any waiver, modification or amendment of

any provision of this Agreement will be effective only if in writing and signed by the parties hereto.

22) Any dispute, controversy or claim arising out of or in connection with this contract, or the breach, termination or invalidity thereof, shall be finally settled by arbitration. The arbitral tribunal shall be nominated by CESSDA ERIC and shall be composed of at least three arbitrators representing different CESSDA ERIC Members or CESSDA ERIC Observers, other than the countries where the contract parties are located.

Appendix II - Pilot 3: Dataverse and NSD-NORDi: a national context evaluation of Dataverse

Use-case for evaluation of Dataverse – NSD

Background / context:

The Research Council of Norway (RCN) recently adopted its first policy on open access to research data from publicly funded projects. The main principle in the new policy is clear: Research data must be shared. Research data must “in general [...] be accessible to relevant users, on equal terms, and at the lowest possible cost."[76] The policy is formulated as a set of recommendations and also points out how the Research Council is planning on implementing it. The most important instrument is the proposed requirement for data management plans for new projects, accompanied with new procedures both in project processing and follow-up at the Research Council. Further on it is emphasized that the practice with contractual requirements of data archiving should be continued as an important tool in securing research to be verified and used for new research purposes. The importance of establishing and developing good infrastructures for data storing and management is also stressed.

NSD has an important role in this respect, as we are contractually committed to make sure a copy of all data funded by the Research Council of Norway within social sciences, humanities, medicine and health, environment and development, is deposited and stored and made accessible for new research purposes.

NORD-i - Norwegian Open Research Data Infrastructure – is a four-year project funded by the Norwegian Research Council. NORD-i will build on the existing Norwegian research infrastructure to develop a comprehensive system for easy data deposit, open access and data sharing to enhance scientific transparency and trust across scientific domains and to promote the use of the existing infrastructure and deposit service.

Within the framework of NSD’s mandate and national responsibilities as a research infrastructure and service provider, the main objective of the upgraded infrastructure is to improve the ability and possibilities of researchers to manage, deposit and use their own data in an efficient, secure and cost effective way and thereby build their trust and willingness to deposit research data at NSD for data sharing purposes. The new infrastructure represents a major upgrade and to some extent a replacement for NSD's current solutions of archival storage, long-term preservation and dissemination of research data.

One of the work packages – WP1 Data deposit and Archiving Portal – aims to make it as easy as possible for institutions and researchers to prepare their research data for archiving and (re)use. Data files and packages may be uploaded via a simple drag-and-drop web-interface. The service may be used to document and share working files within projects, and ultimately to upload (ingest) specific versions into NSD’s archive. Prior to this, users will be provided with a simple, dynamically generated web-based data deposit form where necessary metadata for the data material is added. This pre-ingest or deposit documentation process will be streamlined

[76] Research Council of Norway: Research data must be shared. http://www.forskningsradet.no/en/Newsarticle/Research_data_must_be_shared/1254000848864/p1177315753918

as much as possible. Depositors will have the opportunity to reuse metadata from the Data Management Plan (which is generated in the process), information from earlier deposits and to interact with NSD's targeted guidelines with examples from existing metadata holdings in the process.

The case-study: Dataverse vs custom made solutions

Some of the issues described above may require a case-by-case processing and close communication with individual researchers (e.g. following up on content of DMPs, handle both data management planning early on in research project, and capturing high-quality metadata at the end of a project, etc.), which may not be suitable for a generic virtual research environment as Dataverse. NORD-i aims at providing tools and services that can be flexible enough to meet the different needs and requirements of NSD as a national service provider for RCN-funded data (and other data).

However, when it comes to simple operations like uploading, documenting and sharing project data, Dataverse can function as *a supplement* to the 'custom-made' NORD-i solutions. In (near) future service provider models, NSD may cooperate closely with selected institutions and researchers, providing community training and follow up closely on DMPs and metadata quality, while other institutions may be using an NSD Dataverse, modelled after the Front office – Back office model implemented at DANS.

So what we want to do in this sandbox is to 'evaluate' Dataverse and its possible utility value in the context described above. That is, as a supplemental research environment for the preservation and dissemination of research project data. More specifically we want to assess Dataverse at two 'levels', both on a 'macro' and 'micro' level.

Macro level:

- Gain more insight into DANS experience with DataverseNL (i.e. how does Dataverse solution affect institutional contact and agreements, data/metadata quality, curation efforts, etc.)
- Gain insight into how the FO-BO model works in practice (roles and responsibilities, business model, etc.)
- Estimates of costs / resources of setting up and maintaining a national Dataverse for research data
- DataverseNL and effects on data acquisition (strategy)

Micro level:

- Setup and customisation of a Dataverse: how does it align with NSD needs and the Norwegian national research environment? (I.e. metadata formats and elements applied in Dataverse compared to the needs of the institutions that NSDs will cooperate with in coming years).

Appendix III - Pilot 5: Self-archiving Tool for Researchers

ADP Metadata fields for Dataverse Customization

** Obligatory field*

LEVEL of a STUDY

Metadata field	Help text	Comments	Options for fill-in field	Repeatable
Title: Subtitle (in SI) *	Full title and subtitle by which the study is known in Slovenian.		Free text (limit: 300 characters)	NO
Title: Subtitle (in EN)*	Full title and subtitle by which the study is known in English.		Free text (limit: 300 characters)	NO
Main author*	The person(s), institution(s), corporate body(ies), or agency(ies) responsible for creating the work.	person / institution	/	YES (max 10)
- Name*	The author's Family Name, Given Name or the name of the organization responsible for this Dataset.		Free text (limit: 100 characters)	YES
- Affiliation	The organization with which the author is affiliated.		Free text (limit: 100 characters)	YES

- Identifier Scheme	Name of the identifier scheme.		Closed list [ORCID;
			SICRIS;
			Other: free text]
				YES

- Identifier	Uniquely identifies an individual author or organization, according to various schemes.		Free text (limit 100 characters)	YES

Other author	The other person(s), institution(s), corporate body(ies), or agency(ies) responsible for creating the work.	person / institution	/	YES
- Name*	The other author's Family Name, Given Name or the name of the organization responsible for this Dataset.		Free text (limit: 100 characters)	YES
- Affiliation	The organization with which the author is affiliated.		Free text (limit: 100 characters)	YES

– Identifier Scheme	Name of the identifier scheme.		Closed list [ORCID;
			SICRIS;
			Other: free text]
				YES

– Identifier	Uniquely identifies an individual author or organization, according to various schemes.		Free text (limit 100 characters)	YES
Production Year*	Year when the data collection was produced (not distributed, published or archived).		Date (YYYY)	NO
Production Place*	The location where the data collection and any other related materials were produced. Example: Ljubljana		Free text (limit 100 characters)	NO

Software	Information about the Software used to generate the Dataset. Select from the list.		Closed list [Atlas.ti;
			Microsoft Excell;
			Nvivo;
			PSPP;
			R;
			SPSS;
			Stata;
			other: free text]
				YES

Grant Information *	Grant information.		/	YES (max 5)
- Grant Agency *	Name of the granting agency. If self-funded write “self-funded”.		Free text (limit 100 characters)	YES
- Grant Number *	The grant or project number. If self-funded write “self-funded”.		Free text (limit 100 characters)	YES
Keyword - Term	Key terms that describe important aspects of the Dataset. Can be used for building keyword indexes and for classification and retrieval purposes.		Free text (limit 25 characters)	YES (max 15)
Topic Classification *	Topic or Subject term that is relevant to this Dataset. The classification field indicates the broad important topic(s) and subjects that the data cover. Select from the list.	CESSDA topic list	Closed list [https://docs.google.com/spreadsheets/d/19nXUhcylbEcdf44YmodrewLVgcrbDJskLjlTpY5awFM/edit#gid=0]	YES (max 2)
Abstract in Slovenian*	A summary describing the purpose, nature, and scope of the Dataset in the Slovenian language.		Free text (limit 1000 characters)	NO
Abstract in English*	A summary describing the purpose, nature, and scope of the Dataset in the English language.		Free text (limit 1000 characters)	NO
Time period covered	The time period to which the data refers. This item reflects the time period covered by the data, not the dates of coding or making documents machine-readable or the dates the data were collected. Also known		Start (YYYY-MM-DD) End (YYYY-MM-DD) Single (YYYY-MM-DD)	YES

	as span.			
Date of collection	Contains the date(s) when the data were collected.		Start (YYYY-MM-DD) End (YYYY-MM-DD) Single (YYYY-MM-DD)	YES
Geographic coverage	Information on the geographic coverage of the data. Includes the total geographic scope of the data.		/	/
- Country/Nation*	The country or nation that the Dataset is about. Select from the list.		Closed list [what is already given in Dataverse option] Default option: Slovenia	YES
- Region	The region that the Dataset is about. Use geographical names for correct spelling and avoid abbreviations.		Free text (limit 100 characters)	YES
- City	The name of the city that the Dataset is about. Use geographical names for correct spelling and avoid abbreviations.		Free text (limit 100 characters)	YES
- Other	Other information on the geographic coverage of the data.		Free text (limit 100 characters)	YES

Geographic unit	The lowest level of geographic aggregation covered by the Dataset, e.g., village, municipality, region. Select one of the options.		Closed list [village;
			muncipality;
			administrative unit;
			country;
			local community;
			other: free text]
				NO

Unit of Analysis	The basic unit of analysis or observation that this Dataset describes. Select one of the options.	*CESSDA CV AnalysisUnit*	Closed list [https://docs.google.com/spreadsheets/d/14XYYnjhSNs-nUrgcGOtrSbyZxGv3K-WC1EKB_maPWTU/edit#gid=0]	NO
Universe*	Description of the population covered by the data in the file; the group of people or other elements that are the object of the study and to which the study results refer. Age, nationality, and residence commonly help to delineate a given universe, but any number of other factors may be used, such as age limits, sex, marital status, race, ethnic group, nationality, income, veteran status, criminal convictions, and more. The universe may consist of elements other than persons, such as housing units, court cases, deaths, countries, and so on. In general, it should be possible to tell from the description of the universe whether a given individual or element is a member of the population under study. Also known as the universe of interest, the population of interest, and target population.		Free text (limit 1000 characters)	NO

Time method	The time method or time dimension of the data collection. Select from the list.	*CESSDA CV TimeMethod*	Closed list [https://docs.google.com/spreadsheets/d/14XYYnjhSNs-nUrgcGOtrSbyZxGv3K-WC1EKB_maPWTU/edit#gid=1942149811]	NO
Kind of Data Format*	The physical format(s) of the data documented in the dataset. Select from the list.	*CESSDA CV KindofDataFormat*	Closed list [https://docs.google.com/spreadsheets/d/14XYYnjhSNs-nUrgcGOtrSbyZxGv3K-WC1EKB_maPWTU/edit#gid=822764081]	YES
Type of Data Source	Please select whether you used any type of secondary data source. If you did not use any secondary data sources, select the option "Primary Data Collected".	*CESSDA CV DataSourceType*	Closed list [https://docs.google.com/spreadsheets/d/14XYYnjhSNs-nUrgcGOtrSbyZxGv3K-WC1EKB_maPWTU/edit#gid=1215039729]	YES
Type of Research Instrument	Type of data collection instrument used. Structured indicates an instrument in which all respondents are asked the same questions/test, possibly with precoded answers. Semi-structured indicates that the research instrument contains mainly open-ended questions. Unstructured indicates that in-depth interviews were conducted. Select from the list.	*CESSDA CV Type of Instrument*	Closed list [https://docs.google.com/spreadsheets/d/14XYYnjhSNs-nUrgcGOtrSbyZxGv3K-WC1EKB_maPWTU/edit#gid=1084871036]	YES
Sampling Procedure*	Type of sample and sample design used to select the survey respondents to represent the population. Select from the list.	*CESSDA CV SamplingProcedure*	Closed list [https://docs.google.com/spreadsheets/d/14XYYnjhSNs-nUrgcGOtrSbyZxGv3K-WC1EKB_maPWTU/edit#gid=905013970]	YES

Collection Mode*	The method used to collect the data; instrumentation characteristics. Select from the list.	*CESSDA CV ModeOfCollection*	Closed list [https://docs.google.com/spreadsheets/d/14XYYnjhSNs-nUrgcGOtrSbyZxGv3K-WC1EKB_maPWTU/edit#gid=89791466]	YES
Characteristics of Data Collection Situation	Description of noteworthy aspects of the data collection situation. Includes information on factors such as cooperativeness of respondents, duration of interviews, the number of callbacks, or similar.		Free text (limit 1000 characters)	NO
Response Rate	Percentage of sample members who provided information. When calculating the response rate, use the formula IP/(IP+R+NC+O), where IP is the number of completed interviews (realized sample), R is the number of refused interviews, NC is the number of not contacted respondents and the O is the Other (unknown or inadequate units).		Free text (50 characters)	NO
Related Publication	Publications that use the data from this Dataset		/	YES (max 4)
- Citation	The full bibliographic citation for this related publication.		Free text (limit 500 characters)	YES
- ID Type	The type of digital identifier used for this publication (e.g. Digital Object Identifier (DOI)). Select from the list.		Closed list [what is already given in Dataverse option]	YES

- ID Number	The identifier for the selected ID type.		Free text (limit 50 characters)	YES
- URL	Link to the publication web page (e.g. journal article page, archive record page, or other).		Free text (limit 100 characters)	YES
Distributor*	The organization designated by the author or producer to generate copies of the particular work including any necessary editions or revisions.	no fill-in option for users	ADP (pre-defined)	NO
Distribution Date*	The date that the work was made available for distribution/presentation.	defined by ADP	YYYY-MM-DD (automatic when published by ADP)	NO
Depositor*	Depositor (Family Name, Given Name) AND the name of the organization that deposited this Dataset to the repository.	/	/	YES (max 3)
- Name*	The depositor's Family Name, Given Name or the name of the organization.	Person / institution	Free text (limit 200 characters)	NO
- E-mail	The e-mail address of the depositor for the Dataset. This will not be displayed.		Free text (limit 200 characters)	NO
- Affiliation	The organization with which the depositor is affiliated if applicable.		Free text (limit 200 characters)	NO
Deposit date*	The date that the Dataset was deposited into the repository.	when ADP checks & publish	YYYY-MM-DD (automatic when sending for review)	NO
Contact *	The contact(s) for this Dataset.		/	YES

- Name*	The contact's Family Name, Given Name or the name of the organization.		Free text (limit 200 characters)	NO
- E-mail*	The e-mail address of the contact(s) for the Dataset. This will not be displayed.		Free text (limit 200 characters)	NO
- Affiliation	The organization with which the contact is affiliated.		Free text (limit 200 characters)	NO

LEVEL OF FILES

Metadata field	Help text	Comments	Options for filling-in field	Repeat able

Edit Tags
- File tags*
	Select existing file tags to describe your files. Each file can have only one tag.		Closed list [Questionnaire;
			Codebook;
			SPSS Syntax;
			Data;
			Related Material;
			Publication;
			Data Description (DD);
			Other Material]
				NO

TERMS

Metadata field	Help text	Comments	Options for filling-in field	Repeat able

Terms of Use*	Datasets will default to a CC0 public domain dedication. CC0 facilitates reuse and extensibility of research data. Our Community Norms, as well as good scientific practices, expect that proper credit is given via citation. If you are unable to give datasets a CC0 waiver you may choose between CC-BY and CCBY-NC licenses.		Closed list [CC0;
			CC- BY;
			CCCBY-NC]
				NO

Additional Information				
Special Permissions	Determine if any special permissions are required to access a resource (e.g. if form is needed and where to access the form)		Free text (limit 400 characters)	NO

All other options in Additional Information list or Restricted files should be at this point disabled.

Appendix IV - Pilot 6: Dataverse for Dissemination - use case of TÁRKI Data Archive

Metadata elements used by TÁRKI Data Archive[77]

Discovery metadata

1. the full title of the data collection
2. the person responsible for archiving the data sheet
3. the organization publishing the data sheet
4. the date of archiving the data sheet
5. the bibliographical reference of the data sheet

Study description

1. primary investigator
2. researchers participating in the survey
3. the producer of the data collection
4. the funding agency
5. the distributor of the data collection
6. the depositor of the data collection
7. name of series
8. the version of the data collection
9. the abstract of the data collection
10. keywords
11. the time period covered
12. the time of the data collection:
13. the geographical scope of the data collection
14. the unit of analysis
15. the kind of data: survey data
16. the time dimension of the data collection
17. sampling procedure
18. the mode of data collection
19. restrictions on data access
20. citation requirement
21. related data collection
22. related publications:
23. notes

File description

1. case count
2. variable count
3. type of file
4. other study-related materials

[77] The schema used by TÁRKI is based on DDI 2.0. More information on the TÁRKI website http://www.TÁRKI.hu/en/services/da/docs/ddi_elements_description.pdf