D6 – Report on integration of technical system: Montenegro

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

FNSNF
Swiss National Science Foundation

Deliverable Lead:	FORS
Related Work package:	WP1
Author(s):	Ivana Vujović (CeMI) Milena Nikolić (CeMI) Bojana Tasic (FORS) Irena Vipavc Brvar (ADP) Maja Dolinar (ADP)
Dissemination level:	Public (PU)
Submission date:	30rd April 2017
Project Acronym:	SEEDS
Website:	http://www.seedsproject.ch
Call:	Scientific cooperation between Eastern Europe and Switzerland (SCOPES 2013-2016)
Start date of project:	1st May 2015
Duration:	24 months

Version History

Version	Date	Changes	Modified by
1.0	February 28, 2017	Released version	FORS
2.0	April 19, 2017	1st draft	CeMI
2.1	April 21, 2017	1st revision	UL-ADP
2.2	April 28, 2017	2nd revision	CeMI
3.0	May 1, 2017	Final	FORS

Acknowledgments

This report has been developed within the "South-Eastern European Data Services" (SEEDS) (www.seedsproject.ch) project. The participant organisations of the SEEDS project are:

Name	Short Name	Country
Centre for Monitoring and Research, Podgorica	CeMI	Montenegro
Centre for Political Courage, Pristina	CPC	Kosovo
Institute for Democracy and Mediation, Tirana	IDM	Albania
Institute of Economic Sciences, Belgrade	IES	Serbia
Saints Cyril and Methodius University, Institute for Sociological, Political and Juridical Research, Skopje	ISPJR	Macedonia
University of Zagreb, Faculty of Humanities and Social Sciences, Zagreb	FFZG	Croatia
Swiss Foundation for Research in Social Sciences, Lausanne	FORS	Switzerland
University of Ljubljana, Social Science Data Archive, Ljubljana	UL-ADP	Slovenia

Table of Contents

1 Introduction

The aim of WP1 of the SEEDS project is to implement the various features of the data service establishment plans. This includes organisational, policy, and technical developments, all geared up toward preparing for “day one” of the new data services in partner countries.

The last activity of WP1 is the integration of the archiving system (chosen in D9 - Report on technical improvements) into the technical infrastructure of the partner institutions. Besides creating a set of policy documents for the data services (see D5 - Policy and procedures document) and new individual websites (see D11), it involves the development of a technical prototype that will allow for the basic archiving functions, following the OAIS model: ingest, preservation, and dissemination. Thus, as a key result of the SEEDS project, project partner have now chosen the tools to provide the capacity to take in new data, properly document, store and distribute these data, all according to international standards.

This deliverable describes the technical prototype and its related processes. The purpose is to provide the tools and processes that will allow the new data services to begin building their data collections, to structure their data and metadata in ways to allow for discovery and reuse, to store and secure data for the long-term, and to provide the conditions and platforms for data access for their future users. In sum, the prototype supplies a basic archiving infrastructure, with all needed hardware and software.

As has been the case in all previous project outputs, the intention was to put through the whole process of conducting and maintain as much commonalities as possible across new data service, especially for establishing technical platform. Common and compatible tools have been chosen among partners and determined as tools that will allow for future data and information sharing, as well as for synergies across the national services.

1.1 OAIS Model

The rapid growth of digital material in both volume and complexity, the rising expectations of archives’ users for access services, and the emerging digital preservation strategies, have all contributed to the definition of digital archive functions. The functionalities and procedures of a digital archive have been collected into the OAIS reference model, which became an ISO standard in 2003 (ISO 14721:2003). The OAIS provides both a functional model – the specific tasks performed by the archive, such as storage or access – and a corresponding information model, which includes a model for the creation of metadata to support long-term maintenance and access (see figure 1-1).

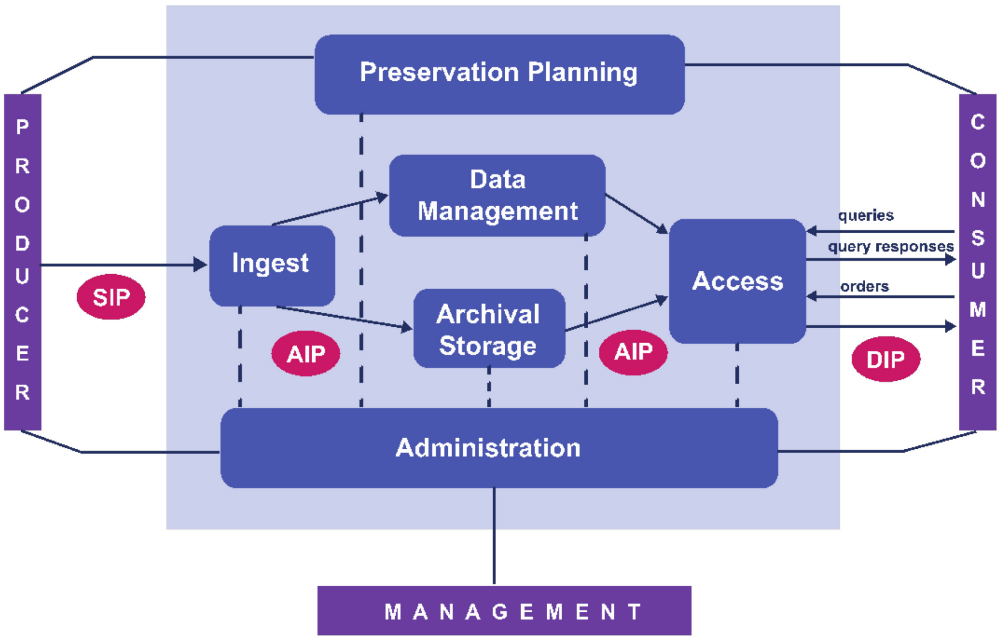

Figure 1-1: OAIS Functional Entities

The OAIS reference model is separated into six functional entities: Ingest, Data Management, Archival Storage, Preservation Planning, Administration, and Access. Outside the OAIS are the Producer (data producers, depositors, researchers), the Consumer (readers, researchers, academics, public, user community), and the Management (data managers, archivists, programmers, database managers, data centre managers). The data within the OAIS are represented as information packages (IPs). Each information package consists of metadata and physical files. There are three types of IPs: submission information package (SIP), archival information package (AIP), and dissemination information package (DIP).

2 Functional Specifications

2.1 Conceptual Model and Workflow

2.1.1 Ingest

Ingest provides the services and functions to accept SIPs from the Producer and prepare the content for Archival Storage and Data Management within the archive (see figure 2-2).

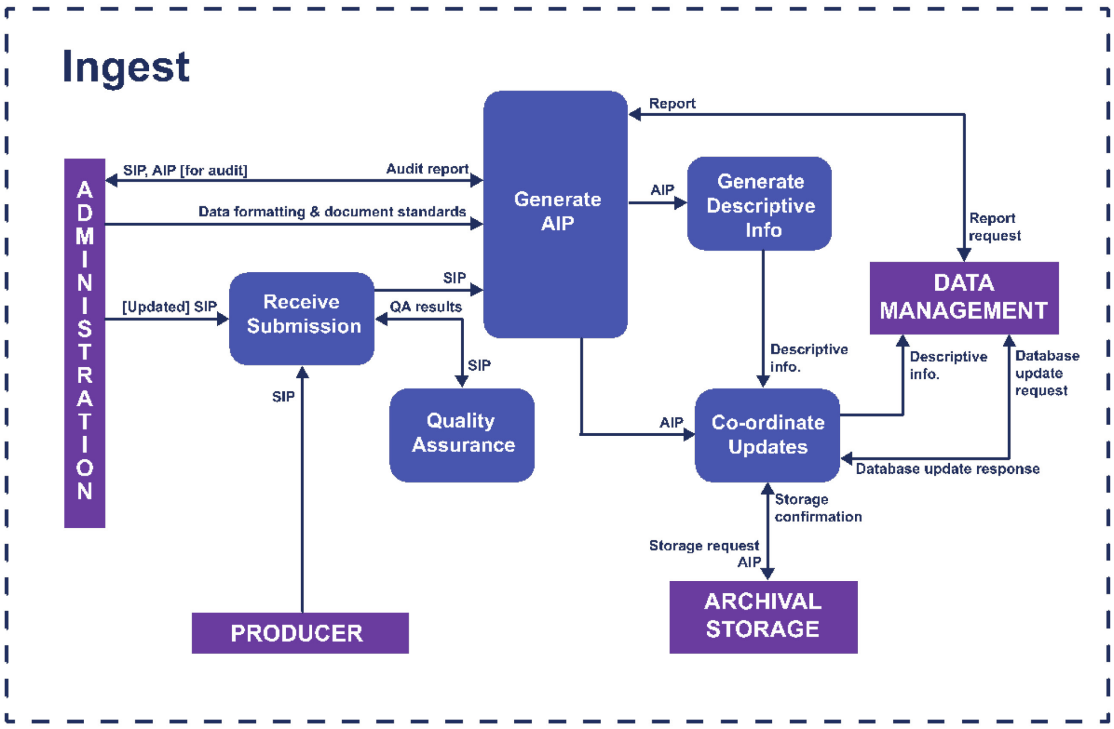

Figure 2-2: Functions of Ingest

Upon a request regarding archiving process, CeMI will provide the researcher with all necessary information on the deposit process and with related checklist. This information will also be available on the web site of the repository.

Researchers will be encouraged to independently use our online deposit system to deposit “raw” data, metadata, reports and publications, but without permission to publish their inputs publicly. Alternatively, researchers will be able to submit data and other materials via e-mail.

Upon receiving the submission, CeMI will check the inputs to ensure that data are prepared for archiving and dissemination. CeMI will check thoroughly whether or not the data are anonymised and readable, if their format is supported by the database, if all necessary codes are submitted and if codes correspond to the actual materials. CeMI will also check if all necessary metadata and related materials are provided and correctly inserted.

As data submission is received, data sets will be reviewed in details in order to determining the sensitivity and utility potentials of data and its readiness to be openly accessible after publishing. Before final archiving of datasets, the restriction level will be assessed in cooperation with depositor and archiving service team. Access to data will be determined in a contract made between the depositor and the archive, depending on data sensitivity and other specifics, as well as requests of depositors. Contract will also contain metadata and information of accessibility of all materials deposited, including datasets, codes, transcripts, syntaxes, questionnaires, reports, etc.

Researchers will be requested to submit following metadata:

Citation Metadata: Title, author name, contact e-mail, subject, keyword term, language, production date and deposit date will be required information, while other information on metadata such as subtitle, author affiliation, description, topic classification, notes and more will be optional, as they are part of Dataverse.

Social Science and Humanities Metadata: sampling procedure, target sample size, type of research instrument and estimates of sampling error will be required, while other option existing in Dataverse will be available to insert.

Journal Metadata: Information on design type, factor type and measurement type will be required.

Upon assuring the quality of a submission and before its final deposit, CeMI will send a deposit contract to the researcher or the institution that is submitting the request for deposit. The contract is to be signed between the data copyright holder and the repository. The contract will specify the data accessibility level, which could be the following:

open access, everyone will be able to see and use datasets, without registration and all related materials and inserted metadata will be available publicly. Users will be provided with the citation instructions in order to respect copyrights of the copyrights holders.

open access, upon registration to archive, certain portion of metadata will be publicly available, while, everyone will be able to see and use datasets, after registration with name, full name and affiliated institution, when all related materials and inserted metadata will become available. Copyrights holders will be provided with the information on the accesses to their data (Guestbook within Dataverse which includes name, affiliation and e-mail after download of the dataset).

restricted access, certain portion of metadata will be publicly available, while data could be used only by researchers who receive permission by the holder of the copyright related to the data, through the data archive.

For accessing restricted data, users will be required to sign a contract with the archive, in order to declare the responsibility to use the data accordingly to national and international policies and standards as well as any special requests from researchers indicated in their contract for deposition of datasets.

Policies on data sharing will be offered by CeMI and taken in consideration upon data deposition process. In this way, archived data will include the information on certain limitations or conditions of use.

2.1.2 Archival Storage

Archival Storage provides the services and functions for the storage maintenance and retrieval of AIPs (see figure 2-3). Archival Storage functions include receiving AIPs from Ingest and adding them to permanent storage, managing the storage hierarchy, refreshing the media on which archive holdings are stored, migrating files into the archival formats, performing routine and special error checking and providing disaster recovery capabilities.

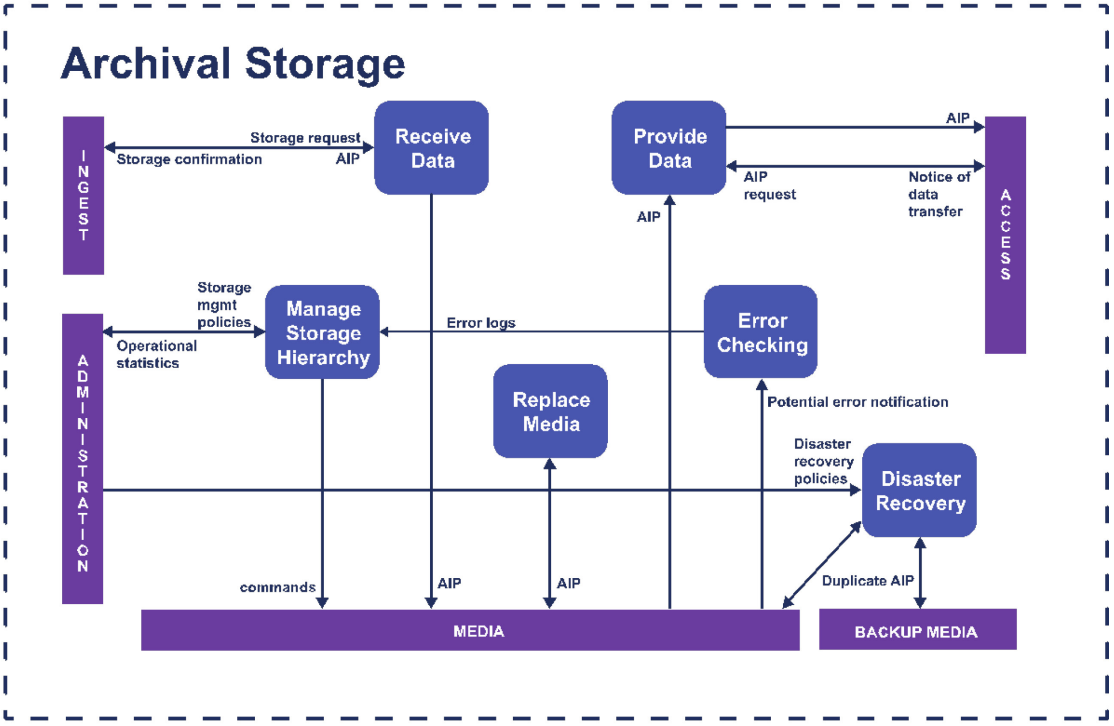

Figure 2-3: Functions of the Archival Storage

CeMI will be using Dataverse database for its archival storage and and rely on its characteristics. The template for social sciences required metadata for deposit is already provided for in the default Database installation (see paragraph on metadata in 2.1). It also supports entry of main data formats, perform automatic error checking and has a backup in case of disaster recovery. Default Dataverse metadata options will be used, and if there is need to adapt additional options in Dataverse, changes will be incorporated in the future. Depositors will be requested to submit required citation metadata and if needed optional citation metadata, social science and humanities metadata and journal metadata (see paragraph 2.1.1). Monitoring and evaluation of archiving processes will be conducted on a monthly basis, in order to determining effectiveness of the archiving processes and data availability. Therefore, analysis will show if there is need to adapt existing technical support related to Dataverse.

2.1.3 Data Management

Data Management provides the services and functions for populating, maintaining and accessing both metadata, which identify and document repository holdings, and administrative data, used to manage the repository (see figure 2-4).

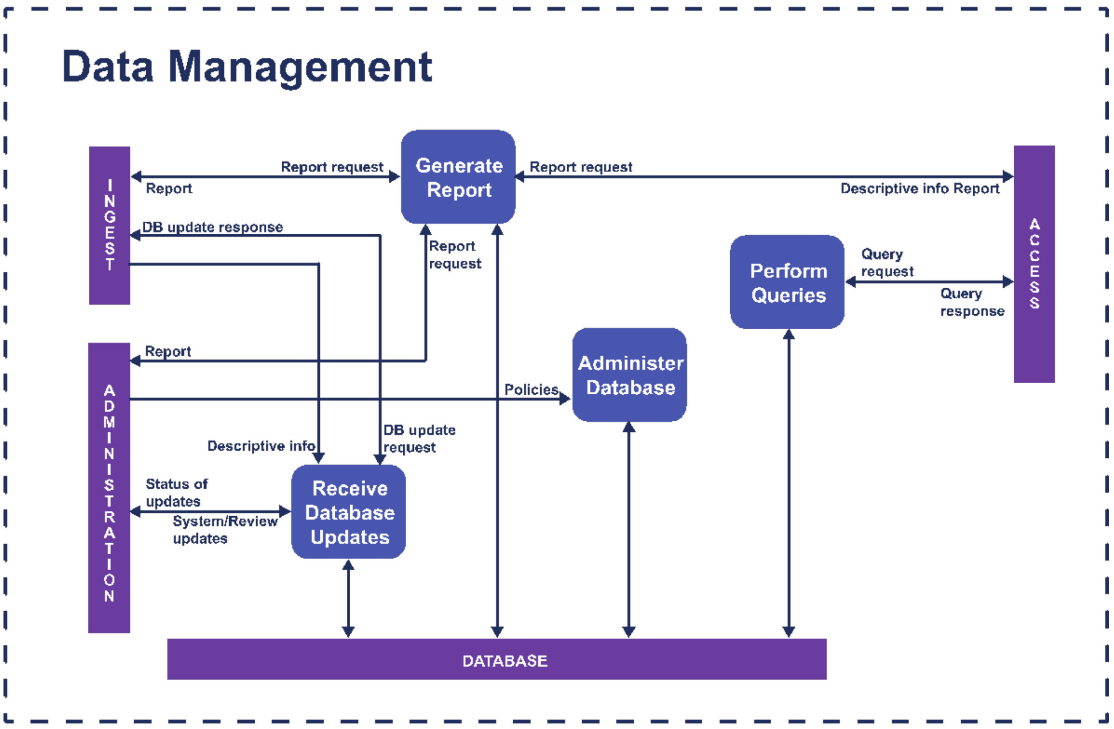

Figure 2-4: Functions of Data Management

Database will be administered by CeMI data archiving team with support from the IT staff. Database will be checked for updates at least once per month, while data will be regularly checked based on a random sample. At the beginning, we will focus on exploring possibilities for metadata and full text of reports and publications archived to be searchable. On a monthly basis, reports on database use - both related to submission and access to data - will be produced.

2.1.4 Administration

Administration provides the services and functions for the overall operation of the archive system (see figure 2-5). Administration functions include soliciting and negotiating submission agreements with the Producer, auditing submissions to ensure that they meet archive standards, and maintaining configuration management of system hardware and software. It is also responsible for establishing and maintaining archive standards and policies, providing user support, and activating stored requests.

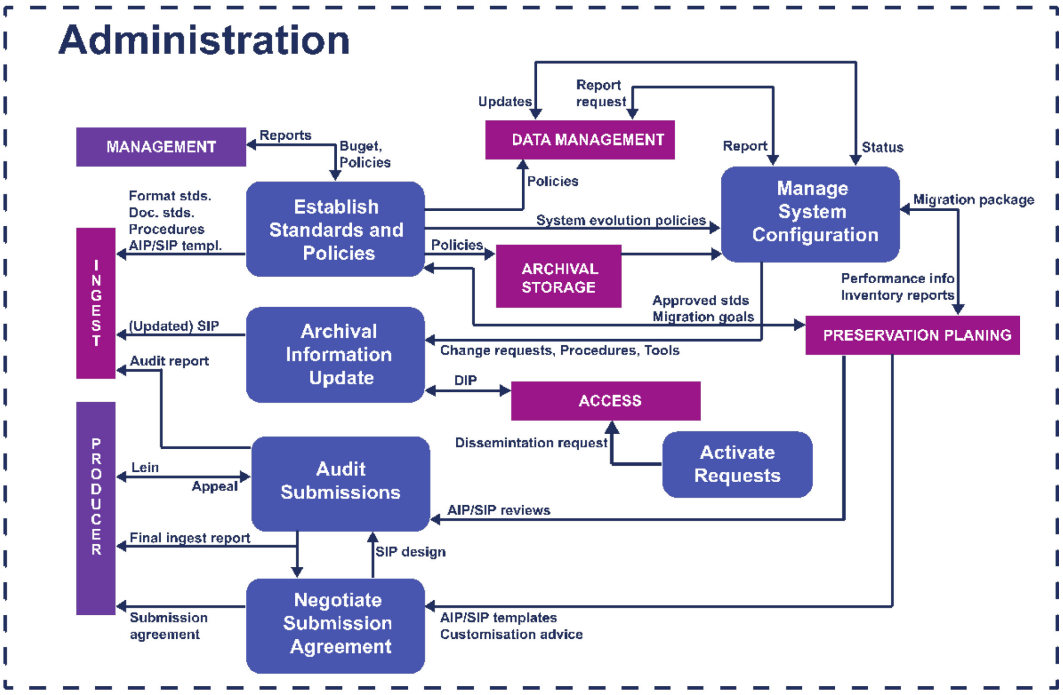

Figure 2-5: Functions of Administration

Administration process in CeMI is currently in the process of development. Once the database will be in place, information and administration packages will be fully prepared. CeMI will develop submission contracts templates, standards and policies.

Figure 2-5 shows what are the main functions of administration. Management is responsible for development and establishing if he standards and policies. Standard and policies will be developed for: data management (see 2.1.4.), archival storage, preservation planning (see 2.1.5), archival information update, system evaluation and other relevant topics.

Templates will be developed for AIP, SIP, reports and contracts.

Reports that will be created are: Performance reports, Inventory reports, Audit of submission, Ingest reports, System configuration.

2.1.5 Preservation Planning

Preservation Planning provides the services and functions for monitoring the environment of the archive and making recommendations to ensure that the information stored in the archive remain accessible over a long-term, even if the original computing environment becomes obsolete (see figure 2-6). Preservation Planning functions include evaluating the contents of the archive and periodically recommending archival information updates to migrate current archive holdings, developing recommendations for archive standards and policies, and monitoring changes in the technology environment and in the user's service requirements. Preservation Planning also develops

detailed migration plans, software prototypes, and test plans to enable implementation of Administration migration goals.

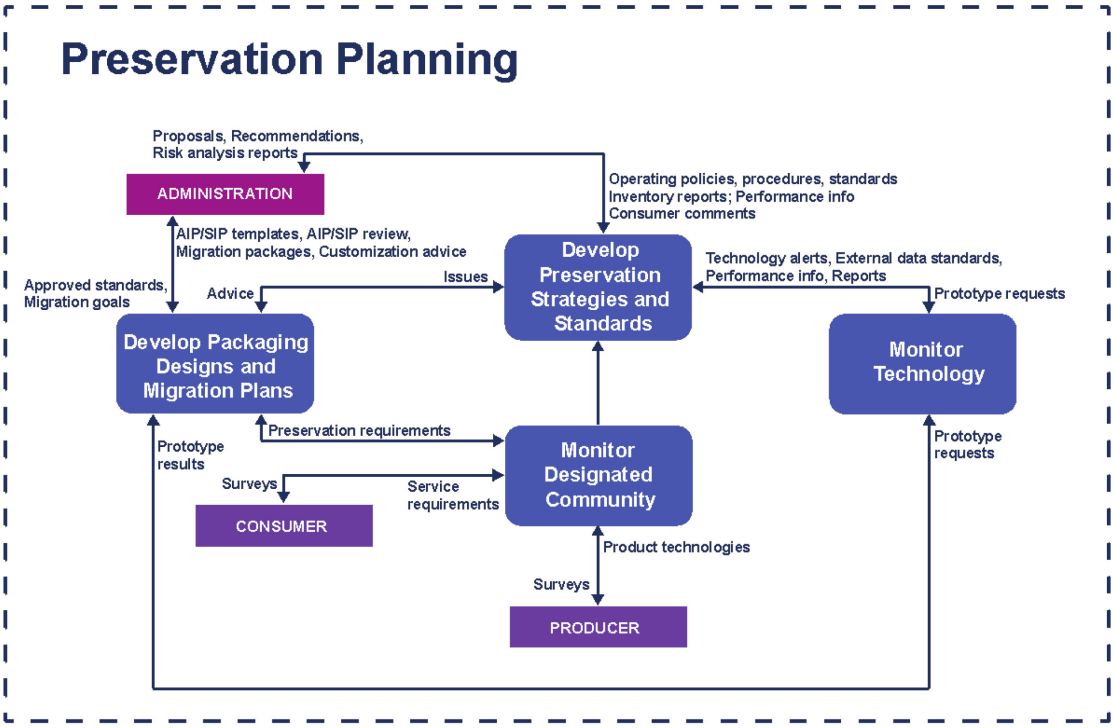

Figure 2-6: Functions of Preservation Planning

CeMI will have active role in all phases of archiving and provide continuous monitoring of the archiving process, in order to ensure that defined standards of data archiving are applied and to ensure that those standards provide sustainable data accessibility. CeMI will have access to inventory reports created once per month and will carefully review all submitted comments of archive users. Evaluation of practicality and easiness of data accessibility will be regularly carried out, in favour of improving the standards of the archiving process, especially regarding its technical development. Preservation Plan will also include recommendations on improvements that will come from our users. Monitoring of development on the level of international community, especially following CESSDA archives in partner countries, will be performed aiming to being in line with development and best practices in similar organizations. This process is strengthened with the fact that other archives in the Region have chosen the same archiving tool (Dataverse).

2.1.6 Access

Access provides the services and functions that support Consumers in determining the existence, description, location, and availability of information stored in the archive, and in allowing them to request and receive data (see figure 2-7). Access functions include communicating with Consumers

to receive requests and applying controls to limit access to specially protected information. This includes coordinating the execution of requests until its successful completion, generating responses, and delivering the responses to Consumers.

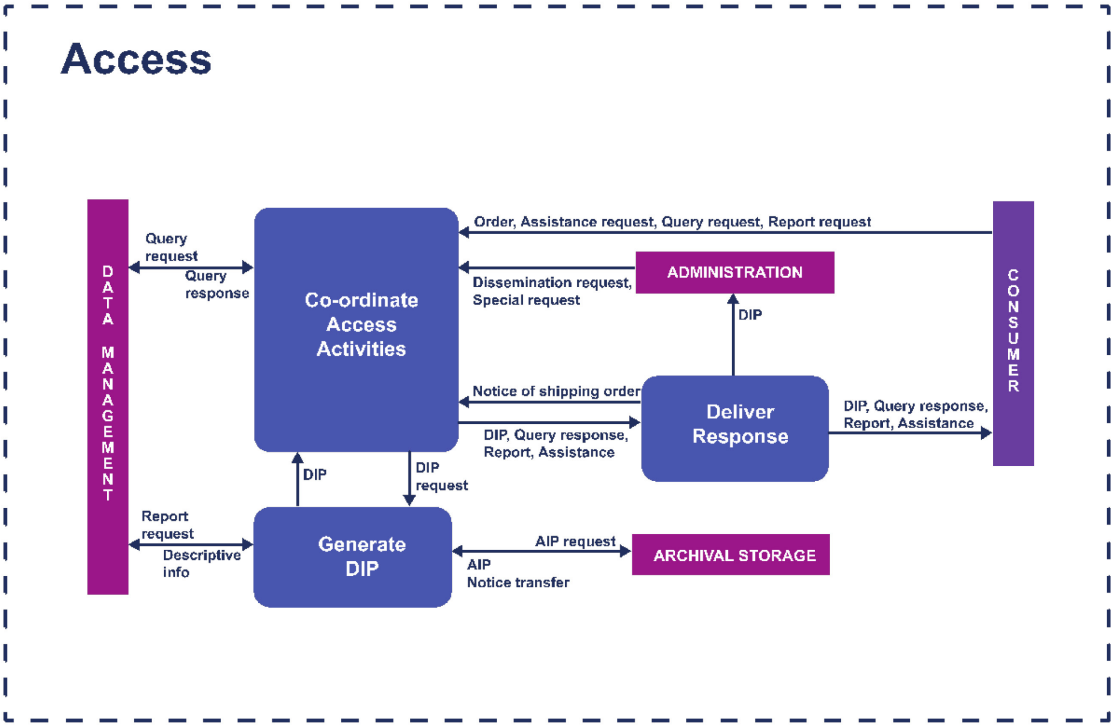

Figure 2-7: Functions of Access

Datasets deposited to archive will be available following set of policies determined for each specific dataset, such as its sensitivity and conditions determined by the depositor. Through the detailed process of data archiving, certain data sharing policies will need to be determined in cooperation with depositors. Therefore, the data will be available to users under different licences:

Open access-Datasets accessible to all users, everyone will be able to see and use datasets, without registration and all related materials and inserted metadata will be available publicly.

Open access, upon registration to archive- Certain portion of metadata will be publicly available, while, everyone will be able to see and use datasets, after registration with name, full name and affiliated institution, when all related materials and inserted metadata will become available

Restricted access- Certain portion of metadata will be publicly available, while data could be used only by researchers who receive permission by the holder of the copyright related to the data, through the data archive

Users will be able to download, or access required data in accordance with the agreement/contract

with CeMI. All accessibility protocols will be monitored, controlled and evaluated at least one per month.

2.2 Metadata Specifications

Metadata of a study will be described in Data Documentation Initiative (DDI) metadata specification, version 2.5 which is also currently supported[1] by Dataverse software.

Complete documentation is available on the DDI alliance web page[2].

The DDI is designed to be fully machine-readable and machine processable. It is defined in XML, which facilitates easy Internet access. DDI Controlled Vocabularies[3] and CESSDA topic classification are planned to be used. The use of a predefined topic classification will make possible future inclusion in the CESSDA data catalogue[4], since these are the topics that enable browsing in the catalogue.

The fields for ingest in the archive ingest tool are made using the CESSDA-recommended fields[5], relevant for study.

2.3. Files and File Formats

There are several reasons why a data archive should be concerned with file formats: they exist in big numbers, are relevant during the whole workflow of the OAIS reference model, and are largely proprietary. File formats are subject to rapid obsolescence if they are not evaluated according to crucial criteria, such as open standards, ubiquity, interoperability, and metadata support. Therefore, file formats that are well-documented, non-proprietary and usable on different hardware and software platforms are much less at risk of not being usable anymore in the future. In addition, their frequency of migration and their costs of preservation are lower.

File formats are an important issue during the entire workflow of the archive (see chapter 2.1). In the functional entity Preservation Planning, the composition and attributes of the information package are defined. This includes the selection of file formats for the SIP, the AIP and the DIP. The decisions of the archive on which file formats are acceptable as archival and distribution formats are linked to the significant properties of the files (what aspects of the digital material we want to preserve). That is why it is important that file formats are controlled and validated, according with the available specific tools, already in the Ingest phase.

There are a number of tools on the market for migrating a file format into a more reliable and sustainable file format:

- *Native Java Image library* for most image formats;
- *Imagemagick* for most image formats, esp. Raster;

[1] http://guides.dataverse.org/en/latest/user/appendix.html
[2] http://www.ddialliance.org/Specification/DDI-Codebook/2.5/
[3] http://www.ddialliance.org/controlled-vocabularies
[4] http://www.cessda.net/catalogue/
[5] https://cessda.net/content/download/709/6350/file/CESSDA%20mandatory%20and%20recommended%20metadata%20fields.pdf

- *FFMPEG* for various AV formats;
- *readpst* for email;
- *Ghostscript* for PDF;
- *LibreOffice* for Office Open XML, and word processor files – also shifts various office formats to PDF and PDF/A;
- *Inkscape* for Vector images.

When selecting target formats, the following criteria should be considered:

- Ubiquity;
- Support;
- Disclosure;
- Documentation quality;
- Stability;
- Ease of identification and validation;
- Intellectual Property Rights;
- Metadata Support;
- Complexity;
- Interoperability;
- Viability;
- Re-usability.

The selected file formats represent a summary of different recommendations from CESSDA partners and internationally recognised institutions: [6]

File formats considered as appropriate for SIPs:

- Tabular data: **SPSS portable format (.por)**, SPSS (.sav), Stata (.dta), Excel or other spreadsheet format files, which can be converted to tab- or comma-delimited text), R (.txt);
- Text: Adobe Portable Document Format (PDF/A, PDF) (.pdf), plain text data, ASCII (.txt), Rich Text Format (RTF) (.rtf), Microsoft Office and OpenOffice documents;
- Audio: **Waveform Audio Format (WAV) (.wav)** from Microsoft, Audio Interchange File Format (AIFF) (.aif) from Apple, FLAC (.flac);
- Raster (bitmap) images: **TIFF (.tif)** ideally version 6 uncompressed, JPEG (.jpeg, .jpg), PNG (.png), GIF (.gif) and BMP (.bmp) only when created in this format, Adobe Portable Document Format (PDF/A, PDF) (.pdf);
- Vector images: DFX (.dfx), SVG (.svg);
- Video: **MPEG-2 (.mpg2)**, MPEG-4 (.mpg4), motion JPEG 2000 (.mj2).

Compressed files are accepted as long as they can be uncompressed by using open and freely available software.

File formats considered as appropriate for the AIP:

- Tabular data: Microsoft Excel File Format (XLS) (.xls), ASCII, Comma Separated Values (CSV)

[6] The formats highlighted in bold are preferred over the others of the same category.
FORS: Qualitative Data Archiving at FORS – Policy and Procedures: http://www2.unil.ch/daris/IMG/pdf/Donnees_qualitatives_archivees_chez_FORS_-_Politique_et_Procedures.pdf ,
UK Data Archive: Formats table: http://www.data-archive.ac.uk/create-manage/format/formats-table, UK Data Archive: Assessment of UKDA and TNA Compliance with OAIS and METS Standards, p. 89 http://www.dptp.org/wp-content/uploads/2010/08/UKDAp90.pdf.

(.csv; .txt);

- Text: Adobe Portable Document Format (PDF/A) (.pdf), XML (.xml), Standard Generalised Markup Language (SGML) (.sgml);
- Audio: Waveform Audio File Format (.wav);
- Raster (bitmap) images: TIFF (.tif);
- Vector images: DFX (.dfx), SVG (.svg);
- Video: MPEG-2 (.mpg2).

File formats considered as appropriate for the DIP:

- Tabular data: SPSS portable format (.por), SPSS (.sav), Stata (.dta), R (.txt);
- Text: Adobe Portable Document Format (.pdf), Rich Text Format (.rtf);
- Audio: MP3 (.mp3);
- Raster (bitmap) images: JPEG (.jpg)
- Vector images: DFX (.dfx), SVG (.svg);
- Video: MPEG-4 (.mpg4).

In addition, file format registries are a way of helping to identify file formats and looking up format specifications.

3 Technical Specifications

3.1 Tools

3.1.1 DataVerse

The archive will use DataVerse software as a tool for archiving the research data. Dataverse takes in account all basic elements of the OAIS reference model (more information on the tool selection process can be found in D9).

CeMI will hire IT to provide detailed technical specification on Dataverse characteristics and its full potentials for the data archive.

3.2 Communication

3.2.1 General Communication

According to the OAIS model there are several different possibilities for how the data archive can communicate with the actors, that is the data producers and consumers. More precisely, it is the functional entities Preservation Planning and Administration that are responsible for the communication task. They include for instance the development of preservation strategies and standards of monitoring the community and technology in order to meet the needs of the producers and consumers (see figure 3-1).

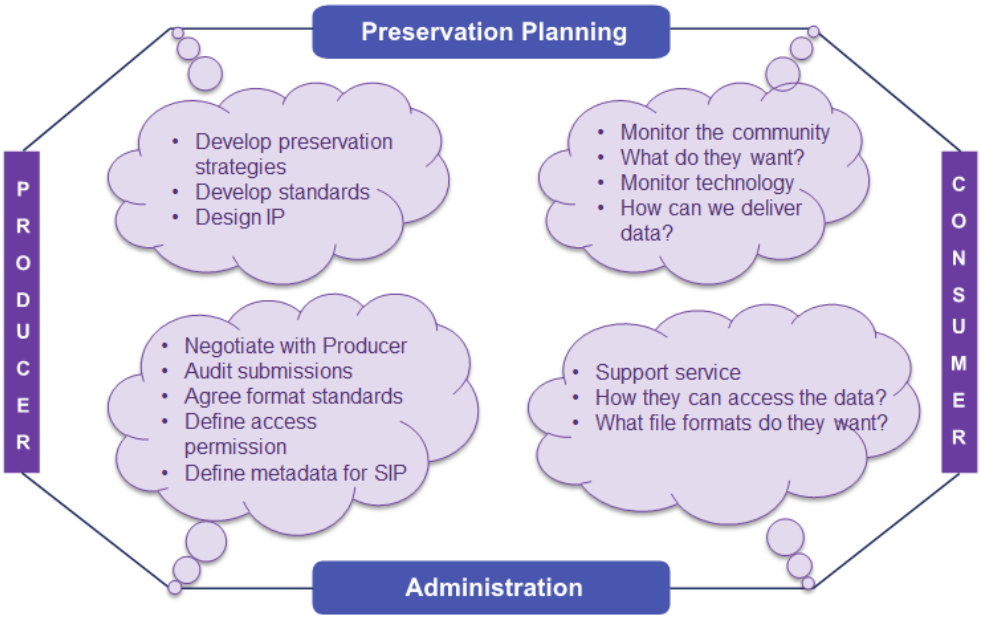

Figure 3-1: Communication

3.2.1.1 Website

The most common and wide-reaching channel to communicate with the community is by means of an institutional website. It is the showcase for interested consumers and producers of data to learn about how data can be obtained and submitted. It is the platform where the policy and procedures, reports and publications, guidelines for data preparation, description of data protection, and other training materials are made available.

The website of the repository can be visited on:

https://me.seedsproject.ffzg.hr/

3.2.1.2 Mailing Lists

A mailing list of potential users will be established by the data archive in order to inform Producers and Consumers about the latest news and upcoming events, such as training and workshops.

3.2.1.3 Direct Contact

A third way of communication is direct contact of the Producers and the Consumers or potential users of the data service through sporadic interaction on an as-needed basis (e.g., for workshops, seminars, and conferences).

3.2.2 Specific Communication

All the specific communications with the users during user registration and Ingest will be recorded and maintained via the website (https://me.seedsproject.ffzg.hr/), Dataverse, e-mails, as well as direct communication with data depositors and users.

3.3 Technical Infrastructure

Since CeMI has not yet established its entire necessary technical infrastructure, the following chapter presents only an example of how a server architecture could be specified. In the future, CeMI will determine in detail all the technical specification needed to fulfil its submission, archival administration and access policies.

3.3.1 Server Architecture (an example)

For the implementation of the SEEDS project 2 servers are needed:

- Virtual server 1 (currently in Croatia);
- Virtual server 2 (should be in each partner country);

The Virtual server 1 is used for the hosting of each national web portal. A single WordPress application with 6 website instances (one for each partner) is installed for the national web portals (see D11).

Here is the detailed Virtual server 1 configuration:

Configuration: 2 vCPU, 2GB RAM, 10GB HDD

OS: Debian GNU Linux 8.2 (Jessie)

HA: Ganneti cluster[7]

The Virtual server 2 should be used for the national catalogues and Ingest/Archival platforms.

Here is the detailed Virtual server 2 configuration example:

Configuration: 2 vCPU, 12.0 GB , 100.0 GB

OS: Debian GNU/Linux 8.2 (jessie)

HA: Not Enabled

Both Virtual server 1 and Virtual server 2 should have redundant IT infrastructure, monitoring, and backup. In addition to local (on-site) file and database backup, there should be a daily automatic offsite backup solution as well. The local servers should be used for backup purposes.

[7] http://www.ganeti.org/

The usage of virtual machines is valuable for prototype implementation and testing, but for the production system, the newly established archives should have more granular distribution of services. Sensitivity of data in the various components of the OAIS, requires us to think about different security levels of data and preservation requirements. To achieve this goal, the future architecture will be installed separately on different virtual machines, based on different platform deployment stacks:

- Website
- Dataverse

Each of these components have different deployment requirements (database, web server, runtime language stack), so it makes sense to separate components on different VMs to enable easy maintenance (migration when changing components, deploying different components for new archives in the future, firmware upgrades).

Looking at the current state of development and support probability of chosen software of the established data archives, it seems that a future change in the components will be probable. This is one of the reasons why the easy maintainability of the system is important. The staff of the archive needs to be capable of testing other available software tools (in a state accessible to them), preferably under Free/Libre/Open Source licences, by using the process described in deliverable D9-Report on technical improvements.

Since each application is installed on a separate virtual machine (and each might have its own set of issues/bugs), security issues are addressed for each virtual machine individually. This means for example that in case of security problems on the web portal, there will be no effect on the security of the archival copy of the data or any other component of the archival infrastructure.

All virtual machines should have two copies stored on different physical machines locally. Machines should be located in different buildings to ensure continuous operation in case of environmental problems in one of the buildings (fire, flooding etc.).

During the process of developing an OAIS based data archive, two distinct types of data required for keeping in the archive were identified - SIP and AIP, which require long-term preservation together with an audit log. This also requires the ability to check whether data is correctly stored on the media that requires checksums on the level of the file system (scrubbing). For this requirement, ZFS[8] storage and snapshots using LVM could be implemented to provide a long-term archival copy of current prototype on different locations (e.g. in faculty building), which should be updated daily (from computing centre location). This would enable disaster recovery in case of one location failure. It is also possible to have multiple remote copies, if needed.

The management of applications and data could be done using Ganeti[9], an open source cloud solution that enables high availability for virtual machines and provides data storage requirements outlined above.

[8] http://bit.ly/dc14-zfs
[9] http://bit.ly/dc14-ganeti

3.3.2 Network and Telecommunications

The network infrastructure and telecommunications are accessed using the host organisations' systems: Center for monitoring and research, Montenegro.

3.3.3 Hardware and Software for production systems

Based on best practices and international standards for social science data archives, the data services have determined the hardware and software they will use.

Workstation computers that will be used by future archive staff for Data Management should include the following software: office tools; conversion tools; software for statistical analysis (STATA, R, SPSS); tools for preparing metadata description of a study, etc.

If the archive wants to use a proprietary product, they will have to buy a licence or use the existing licences of their hosting institution, if available.

4 Conclusions and future development

In conclusion, the prototype described in this paper provides the technical basis for all key archiving functions, following the OAIS model. The new data services will be able on "day one" to integrate and manage new datasets, safely store and protect data, as well as disseminate data and documentation to users. Their technical systems will function according to international norms and best practices, even if some of the archiving workflow will need to be handled manually.

It should be noted, however, that while the prototype will enable certain basic services, it will not be as comprehensive or as flexible as the one used by mature social science data archives. Future work should expand the technical development to accommodate for a greater volume and variety of data, to automate more everyday practices, and to enhance communication potential and exchange with data producers and users. This work will continue for many years, and will build on experience, further training, and funding. Like any others, these new data services will have to adapt technically to the ever-changing research and policy environments.