seeds

South-Eastern European Data Services

D6 – Report on integration of technical system: Croatia

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

FNSNF
SWISS NATIONAL SCIENCE FOUNDATION

Deliverable Lead:	FORS
Related Work package:	WP1

Author(s):	Alen Vodopijevec (FFZG) 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 14, 2017	Draft version	FFZG
2.1	April 24, 2017	Revision	UL- ADP
3.0	April 28, 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

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, the project partners have now chosen the tools and have the capacity to take in new social science data, and then to 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 future 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 maintain as much commonalities as possible across the six new data services, and this is especially true for the established technical platform. Common and compatible tools 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).

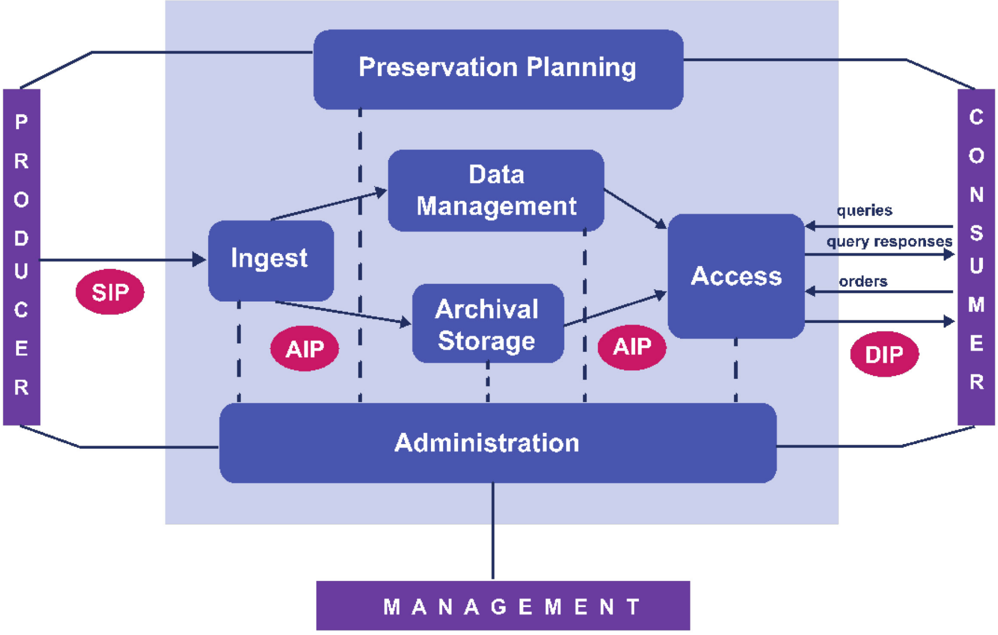

Figure 1-1: OAIS Functional Entities

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).

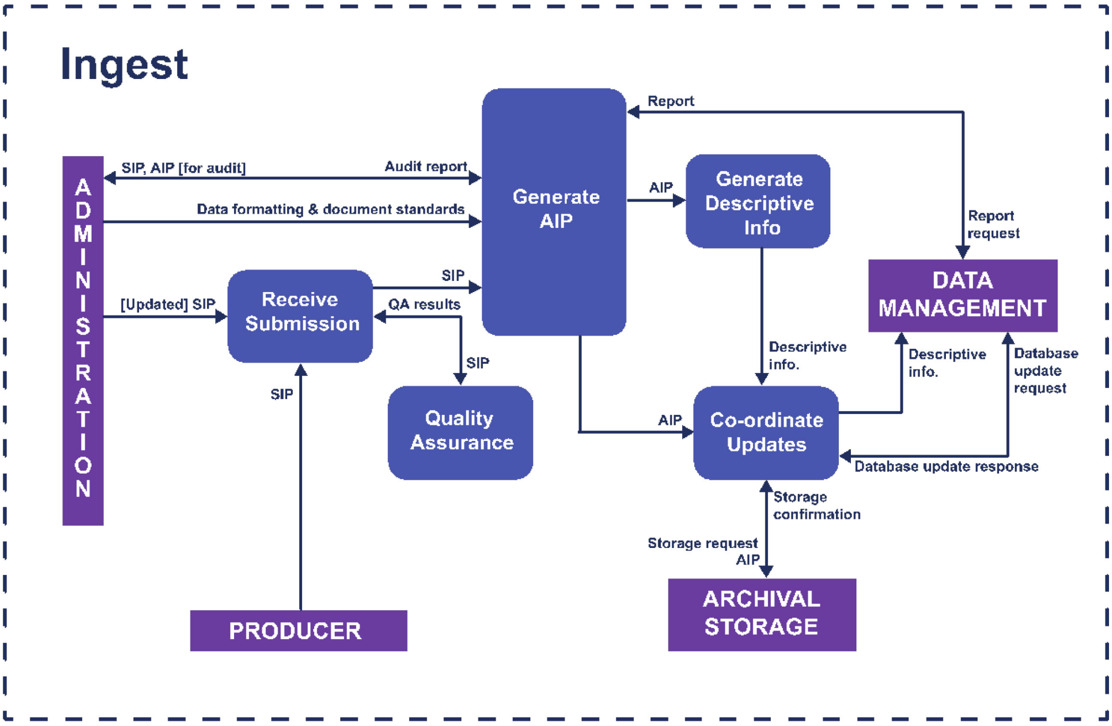

Figure 2-2: Functions of Ingest

2.1.1.1 Submission

Submission is received via an ingest tool, which supports secure file uploads and metadata creation. Archive staff will offer support to Producers in the process of transferring materials to the archive by providing advice on the content and quality of the materials and accepted file formats. All communication with the Producer concerning submitted materials will be recorded. Before creating SIP, submitted materials and metadata will be checked to ensure completeness and compliance with data archiving standards. Files will be converted to appropriate formats if necessary. All actions and decisions concerning submitted materials will be recorded.

2.1.1.2 Quality assurance

Quality assurance will include virus scanning and validation of the successful transfer of the SIP to the ingest tool. File transfer or media read/write errors will be recorded in system log. Checksums will be associated with each data file.

2.1.1.3 AIP

Transforming the SIP into the AIP will comply with the formatting and documentation standards. This process may include file format conversions, reorganization, repackaging and other transformations of the content information in the SIP. All transformations have to be recorded.

2.1.1.4 Metadata

This step includes extracting Descriptive Information from the AIP for inclusion in the discovery catalogue and coordinating updates to Archival Storage and Data Management.

2.1.1.5 Deposit contract

Registered Data Service users will be able to self-archive their datasets and to use the system as a part of the VRE (virtual research environment), keeping their data safe and sharing it with other team members. Accepting the terms of use and data service policies during registration is enough in most cases. In the case of some special conditions required by depositor or the Data Service, we will require a physically signed deposit contract.

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.

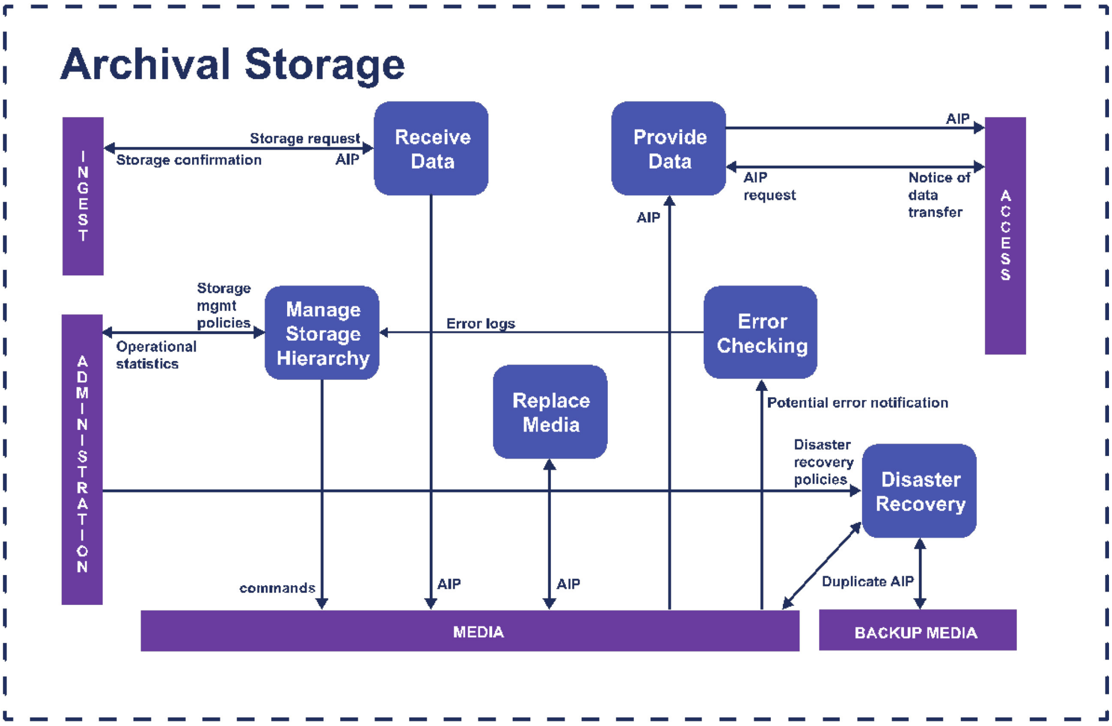

Figure 2-3: Functions of the Archival Storage

Central parts of the Data Service IT infrastructure will be hosted at the Faculty of Humanities and

Social Sciences, at the University of Zagreb in Zagreb where Basic parts of infrastructure are already installed. Virtual machine with file storage is installed in a local cluster with implemented high availability (HA) functions. Databases and files are backed up on daily basis locally with a planned offsite backup in the University computing centre's cloud.

Archival storage for Data Service will be implemented in a redundant way. Upon creating the final AIP, this package will be pushed to Dabar service at the University Computing Centre Islandora instance. Islandora uses the enterprise storage solution backed up locally (Zagreb) with a planned offsite copy on the remote locations (Osijek, Rijeka). Additionally Data Service will be keeping one copy of AIP locally at the Faculty of Humanities and Social Sciences.

2.1.2.1 Receive Data

The connection between Ingest and Archival storage will be done through Bagit[1] REST API. Once the AIP is ready for archiving, the Receive Data function will receive a storage request and an AIP from Ingest and will move the AIP to a permanent storage location within the archive. Upon completion of the transfer, this function will send a storage confirmation message to Ingest (Coordinate Updates), including the storage identification of the AIPs.

2.1.2.2 Replace Media

The Replace Media function provides the capability to reproduce the AIPs over time. Within the Replace Media function the Content Information and Preservation Description Information (PDI) must not be altered. The migration strategy must select a storage medium, taking into consideration the expected and actual rates of errors encountered in various media types, their performance, and their costs of ownership.

2.1.2.3 Error Checking and Disaster Recovery

Error Checking provides statistically accepted assurance to ensure that none of an AIP's components are corrupted via any internal storage function. The Data Service performs routine and special data integrity checking such as checksums for each individual file, and generates error reports. It also provides disaster recovery capabilities, including data backup, offsite data storage and data recovery. High availability functions and procedures as well as data recovery from backup will be checked on a regular basis.

2.1.2.4 Provide Data

The Provide Data function provides copies of stored AIPs to Access. This function receives an AIP request that identifies the requested DIP(s), and then provides them on the requested media type or transfers them to a staging area.

2.1.3 Data Management

[1] https://en.wikipedia.org/wiki/BagIt

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).

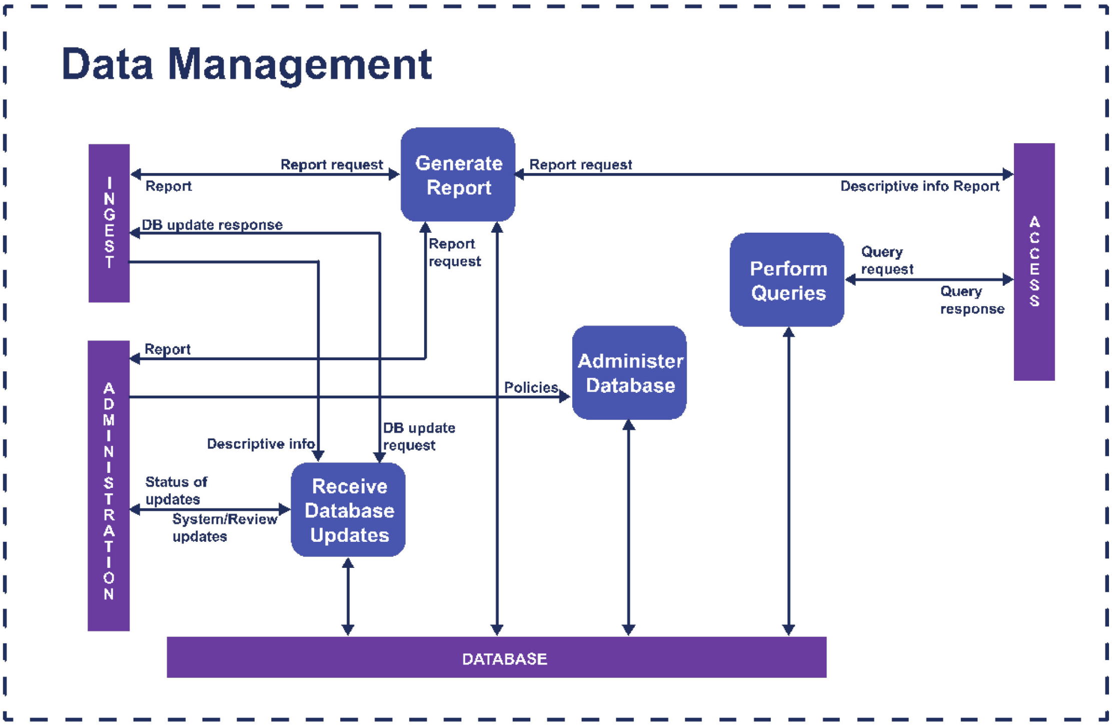

Figure 2-4: Functions of Data Management

2.1.3.1 Administer Database

The Administration Database function is responsible for creating any schema or table definitions required to support Data Management functions; for providing the capability to create, maintain and access customized user views of the contents of this storage; and for providing internal validation (e.g., referential integrity) of the contents of the database. Our Administer Database function is carried out in accordance with policies received from Administration.

2.1.3.2 Perform Queries

The Perform Queries function receives a query request from Access and executes the query to generate a result set that is transmitted to the user All queries will be verified against access rights restrictions.

2.1.3.3 Generate Report

The Generate Report function receives a report request from Ingest, Access or Administration and executes any queries or other processes necessary to generate the report that it supplies to the user. Typical reports might include summaries of archive holdings by category, or usage statistics for numbers of access to archive holdings.

2.1.3.4 Receive Database Updates

The Receive Database Updates function adds, modifies, or deletes information in the Data Management persistent storage. The main sources of updates are Ingest, which provides Descriptive Information for the new AIPs, and Administration, which provides system updates and review updates. Ingest transactions consist of Descriptive Information which identifies new AIPs stored in the archive. System updates include all system-related information (operational statistics, Consumer information, and requests status). Review updates are generated by periodic reviewing and updating of information values.

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.

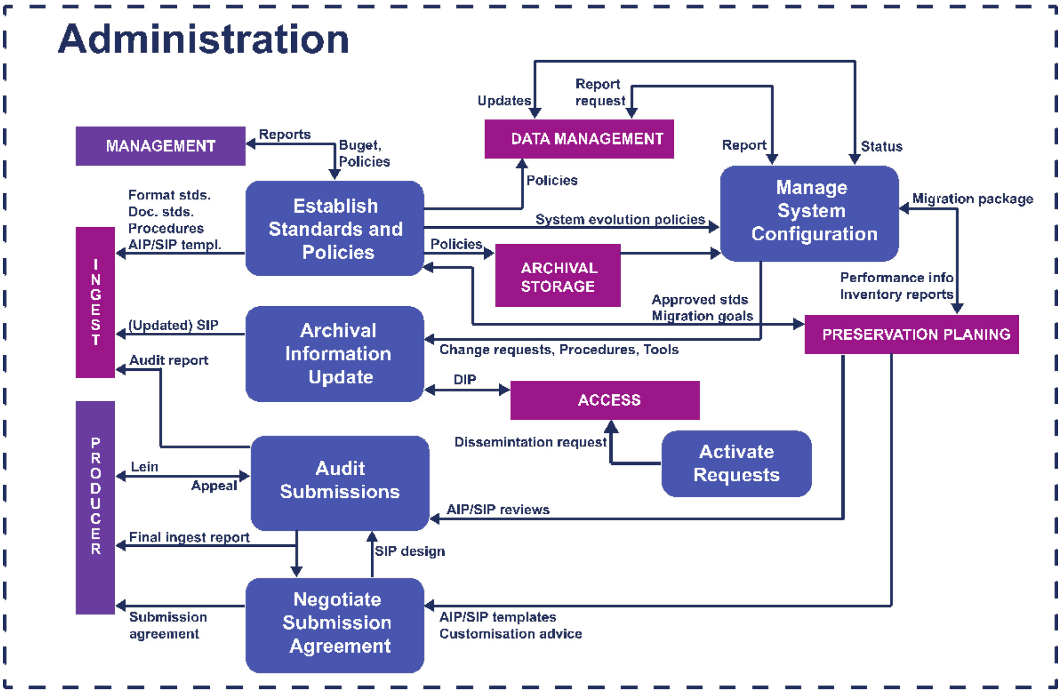

Figure 2-5: Functions of Administration

2.1.4.1 Negotiate Submission Agreement

The Negotiate Submission Agreement function negotiates submission agreements with the Producer. The submission agreement is a formal written agreement between the Producer and the archive defining the terms of the content, standards, metadata creation of the SIP, and future use by third parties. The data submission formats and procedures are clearly documented in the archive's data submission policies document.

2.1.4.2 Manage System Configuration

The Manage System Configuration function maintains integrity and tractability of the configuration during all phases of the system life cycle. It also audits system operations, system performance, and system usage. It sends report requests for system information to Data Management and receives reports; it receives operational statistics from Archival Storage.

2.1.4.3 Archival Information Update

The Archival Information Update provides a mechanism for updating Digital Objects (files and metadata) within the repository. This type of action may require the removal or update of the digital object and/or its associated metadata.

2.1.4.4 Establish Standards and Policies/Procedures

The Establish Standards and Policies function is responsible for establishing and maintaining the archive system's standards and policies. It receives recommendations for archive system enhancement, as well as proposals for new archive data standards. It also receives performance information and archive holding inventories from the Manage System Configuration. Based on these inputs, archive standards and policies are established and sent to other Administration functions and the other Functional Entities for implementation. The standards include format standards, documentation standards and the procedures to be followed during the Ingest process. It provides approved standards and migration goals to Preservation Planning. This function will also develop storage management policies (for the Archival Storage hierarchy), including migration policies to ensure that archive storage formats do not become obsolete, and database administration policies. It will develop disaster recovery policies. It will also determine security policies for the contents of the archive, including those affecting Physical Access Control and the application of error control techniques throughout the archive.

2.1.4.5 Audit Submissions

This function receives AIP/SIP reviews from Preservation Planning and may also involve an outside committee (e.g., science and technical review). The audit process must verify that the quality of the data, metadata and documentation meets the requirements of the archive and the review committee. The Audit process may determine that some portions of the SIP are not appropriate for inclusion in the archive and must be resubmitted or excluded. Producer will be informed about decision. In some cases the Producer may be requested to modify the material, or provide alternate material. All the communication and decision-making will be recorded in the ticketing system.

2.1.4.6 Activate Requests

The Activate Requests function maintains a record of event-driven requests and periodically compares it to the contents of the archive to determine if all needed data are available. This function will be performed through a regular checksum procedure.

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.

Preservation planning includes the following:

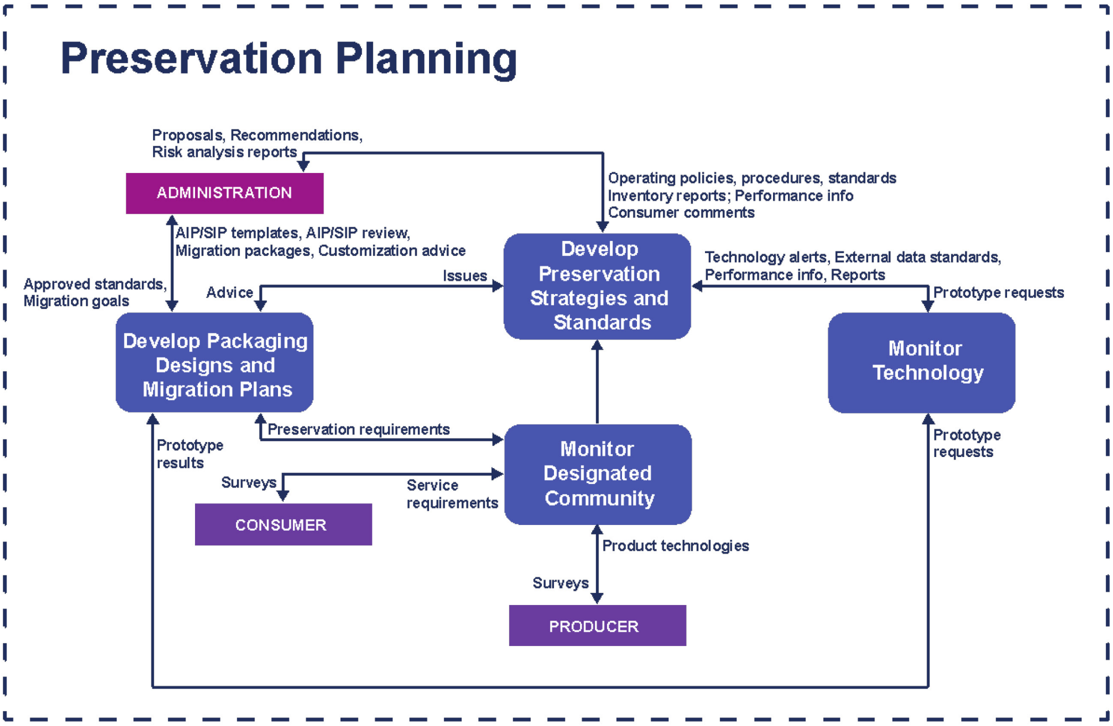
Preservation Planning
Proposals, Recommendations, Risk analysis reports
ADMINISTRATION
Operating policies, procedures, standards
Inventory reports; Performance info
Consumer comments
AIP/SIP templates, AIP/SIP review, Migration packages, Customization advice
Approved standards, Migration goals
Advice
Issues
Develop Preservation Strategies and Standards
Technology alerts, External data standards, Performance info, Reports
Prototype requests
Develop Packaging Designs and Migration Plans
Monitor Technology
Preservation requirements
Prototype results
Surveys
Service requirements
Monitor Designated Community
Prototype requests
CONSUMER
Product technologies
Surveys
PRODUCER

Figure 2-6: Functions of Preservation Planning

2.1.5.1. Monitoring of the designated community and the technology

Contact with data producers as well as data users and trying to identify changes in data acquisition processes, methodologies, data formats used etc. to be able to customize Data Service processes and procedures accordingly.

2.1.5.2 Preservation strategies and migration plans

Preservation strategies and migration plans are used to define file formats used for archiving and long term preservation as well as for monitoring technology changes and planning possible file formats migration.

2.1.5.3. Information package design

This functions develops new information package designs and detailed migration plans and prototypes, and implements Administration policies and directives; it also receives archive approved standards and migration goals from Administration. The standards include format standards, metadata standards, and documentation standards. It applies these standards to preservation requirements and provides AIP and SIP template designs to Administration.

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.

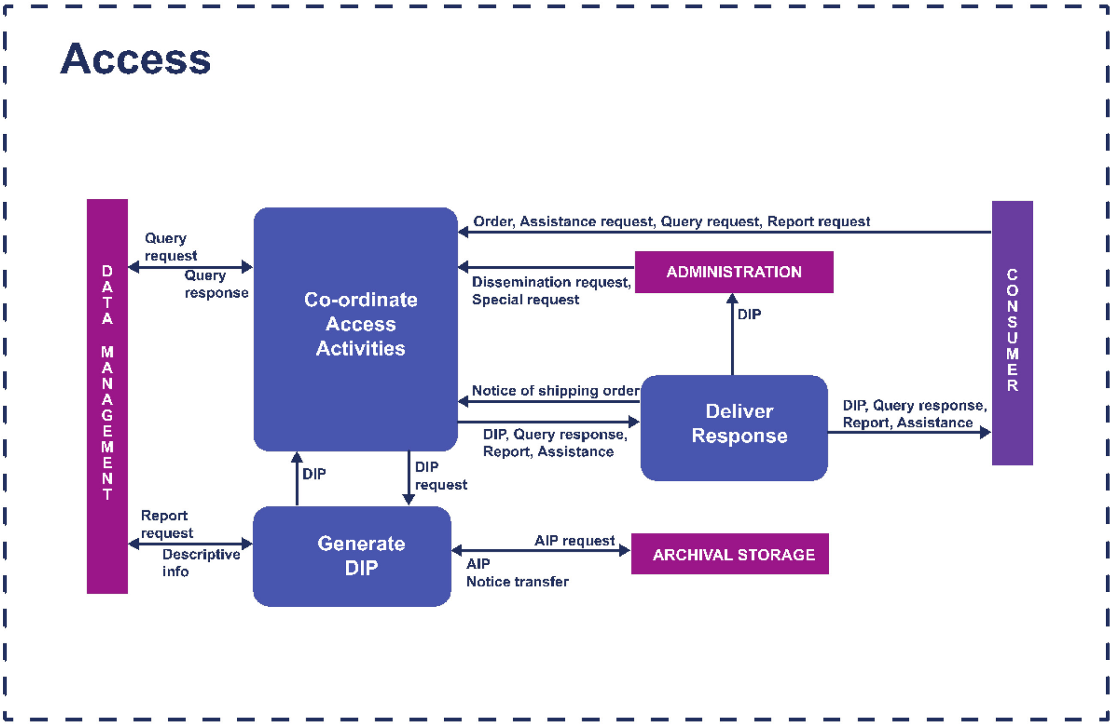

Figure 2-7: Functions of Access

Dissemination information package (DIP) will be created from the AIP. Depending on the sensitivity of data, conditions and terms of use, the DIP can be the same version as AIP but it can also consist only of the selected parts of documentation and/or subset of source data. Terms of use and access control can be applied on the whole dataset level as well as on the particular deposited file. Basic metadata will be freely available.

There will be several levels of access to datasets and documentation:

- open access
  - dataset DIPs available without restrictions
- restricted access
  - access to data is restricted to specific users (research team members, institutional users, individually authorized users)
- embargoed access
  - access to data is closed but only for a defined period of time
- closed access
  - datasets available only to data creator/depositor

Based on their permissions users will be then able to access and/or download the elements of DIP (Metadata, Documentation, Files). The network access and data delivery will be provided via secure network protocols.

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[2] by Dataverse software.

Complete documentation is available on the DDI alliance web page[3].

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[4] 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[5], 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[6],

[2] http://guides.dataverse.org/en/latest/user/appendix.html

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

[4] http://www.ddialliance.org/controlled-vocabularies

[5] http://www.cessda.net/catalogue/

[6] https://cessda.net/content/download/709/6350/file/CESSDA%20mandatory%20and%20recommended%20metadata%20fields.pdf

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;
- *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: [7]

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) (.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).

[7] 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.

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

The following software tools are used for the Ingest, Archival and Dissemination processes:

3.1.1 Dataverse / NextCloud

Dataverse will be used for receiving SIP. Alternative option is to receive SIP via NextCloud software. Both tools are using secure encrypted protocols (https) for data transfer. Receiving files through non-encrypted channels (e.g. non-encrypted email or website) is not allowed.

3.1.2 Redmine

Redmine is an opensource project management and ticketing system. It is used within data service as a tool for planning and monitoring the data curation process.

3.1.3 Islandora (DABAR)

Dabar is a national digital repository hosting service based on Islandora. It will be used as an Archival solution and a long-term preservation platform.

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).

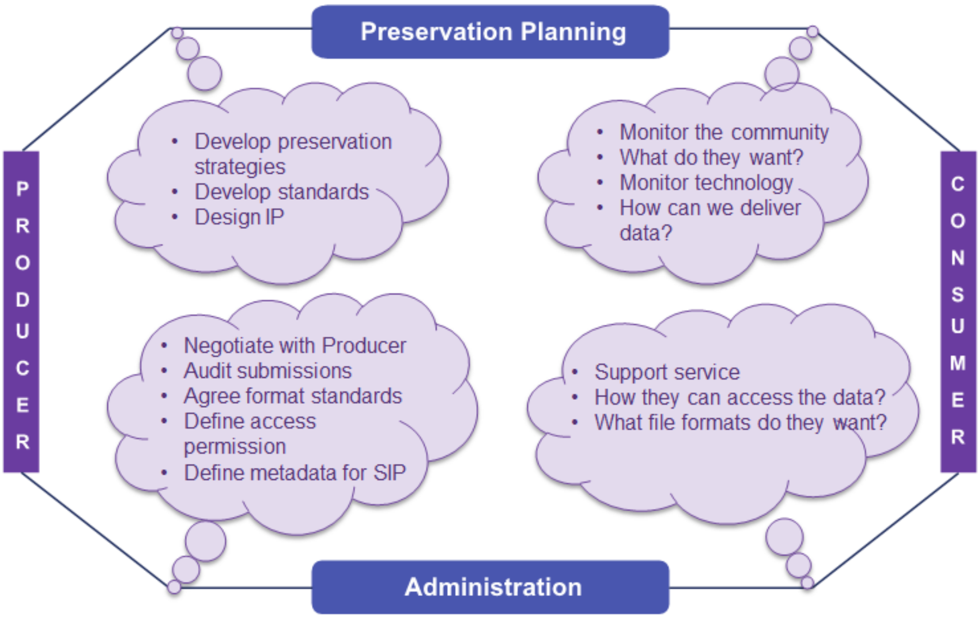

Figure 3-1: Communication

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 can be visited on:

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

3.2.1.2 Mailing Lists

A mailing list of potential users is 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 our internal project management and ticketing system based on the open-source tool Redmine[8]. We have developed default project template, which follows our processes and procedures with custom statuses and the possibility to add new subtasks if needed. Administrators are able to easily track the progress of each project and identify possible problems.

3.3 Technical Infrastructure

3.3.1 Server Architecture (an example)

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

- Virtual server 1 (Website, Cloud file storage, Communication platform);
- Virtual server 2 (Dataverse);

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[9]

[8] http://www.redmine.org/

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

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.

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:

- Centre’s website
- Virtual research environment and self-archiving tool (Dataverse)
- File storage and sharing platform (Nextcloud)
- Long-term preservation archival system (Islandora)

Each of these components have different deployment requirements (database, web server, runtime language stack), so it makes sense to separate components on different servers 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 tool evaluation and selection.

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[10] 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[11], an open source cloud solution that enables high availability for virtual machines and provides data storage requirements outlined above.

3.3.2 Network and Telecommunications

The network infrastructure and telecommunications are accessed using the Faculty of Humanities and Social Sciences facilities, University of Zagreb, backed up by the University Computing Centre's systems.

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.

[10] http://bit.ly/dc14-zfs

[11] http://bit.ly/dc14-ganeti

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.