Support for Establishment of National/Regional Social Sciences Data Archives

D5.1 – Documents and Materials for Social Science Digital Data Archives

Deliverable Lead:	FORS
Related Work package:	WP5
Author(s):	Renate Kunz (FORS) Brian Kleiner (FORS) Bojana Tasić (FORS) Irena Vipavc Brvar (UL) Janez Stebe (UL) Hilary Beedham (UESSEX) Ilze Lace (UGOT) Aleksandra Bradic-Martinovic (IEN) Aleksandar Zdravkovic (IEN) Alen Vodopijevec (Rudjer Boskovic Inst) Marijana Glavica (FFZG) Teo Matkovic (FFZG) Sasa Madacki (HRC)
Dissemination level:	Public (PU)
Submission date:	31st March 2014
Project Acronym:	SERSCIDA
Website:	http://www.serscida.eu
Call:	FP7-SCIENCE-IN-SOCIETY-2011-1
Project Number:	288985
Start date of project:	1st January 2012
Duration:	30 months

Version History

Version	Date	Changes	Modified by
1.00	March 31, 2014	Released version	FORS
2.00	June 30, 2014	Revised version	FORS

Acknowledgments

This report has been developed within the "Support for Establishment of National/Regional Social Sciences Data Archives" (SERSCIDA) (www.serscida.eu) project. The participant organisations of the SERSCIDA project are:

Name	Short Name	Country
Faculty of Humanities and Social Sciences of the University of Zagreb	FFZG	HR
Slovenian Social Science Data Archives	UL	SI
Swiss Foundation for Research in Social Sciences	FORS	CH
The Human Rights Centre of the University of Sarajevo	HRC	BA
The Institute of Economic Sciences	IEN	RS
The Swedish National Data Service	UGOT	SE
United Kingdom Data Archive	UESSEX	UK

Disclaimer: The information in this report does not necessarily represent the view of the European Commission (EC). The SERSCIDA project has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 288985.

Table of Contents

Acknowledgments ***1***

Figures ***4***

Tables ***5***

Introduction ***6***

A: Charter documents ***7***

I University of Sarajevo Data Archive and Services in Social Sciences and Humanities (USDAS) ... **7**

Definition of organization 7

Scope of collection 7

Services 7

Internal structure of organization 8

Governance structure 8

Staff profiles 9

Cooperation 9

2. Serbian Social Science Data Centre (SR-SSDC) **11**

The SR-SSDC name and residence 11

SR-SSDC General activities 11

SR-SSDC Internal structure of organization and management 12

Cooperation 14

3. Croatian Data Archive for the Social Sciences (CRO-DAS) **16**

Definition of organization 16

Scope of collection 16

Services 16

Internal structure of organization 16

Governance structure 17

Staff profiles 18

Cooperation 18

B: Data Policies and Procedures ***20***

I University of Sarajevo Data Archive and Services (USDAS) **20**

1. Contextualisation and organisation of document 20

2. Conceptual models and guiding principles 20

3. Acquisition and deposit 26

4. Ingest 30

5. Archival Storage 33

6. Data Management 35

7. Preservation Planning 36

8. Access 36

9. Administration 38

II Serbian Social Sciences Data Centre – SR-SSDC **39**

1. Contextualisation and organisation of document 39

2. Conceptual models and guiding principles 39

3. Acquisition and deposit 45

4. Ingest 49
5. Archival Storage 52
6. Data Management 54
7. Preservation Planning 54
8. Access 55
9. Administration 56
III Croatian Data Archive for the Social Sciences – CRO-DAS 58
1. Contextualisation and organisation of document 58
2. Conceptual models and Guiding principles 58
3. Acquisition and deposit 64
4. Ingest 69
5. Archival Storage 71
6. Data Management 74
7. Preservation Planning 74
8. Access 74
9. Administration 76
References 78
Annex 1: Data deposit agreements 80
A1.1 Data deposit agreement USDAS 80
A1.2 Data deposit agreement SR-SSDC 82
A1.3 Data deposit agreement CRO-DAS 84
Annex 2: End user licenses 86
A2.1 End user license USDAS 86
A2.2 End user license SR-SSDC 87
A2.3 End user license CRO-DAS 88
Annex 3: Finance and budgets 89
A3.1 Finance and budget USDAS 89
A3.1 Finance and budget SR-SSDC 91
A3.3 Finance and budget CRO-DAS 93
Annex 4: Performance indicators 95
A4.1 Performance indicators, USDAS 95
A4.2 Performance indicators, SR-SSDC 97
A4.3 Performance indicators, CRO-DAS 98

Figures

Figure 1: Organizational and governance chart USDAS 9
Figure 2: SR-SSDC Organisational Chart 15
Figure 3: Organizational and governance chart CRO-DAS 17
Figure 4: Reference Model for an Open Archival Information System (OAIS); CCSDS 650.0-M-2; Consultative Committee for Space Data Systems: Washington, DC, 2002, p. 2-2. 21
Figure 5: Reference Model for an Open Archival Information System (OAIS); CCSDS 650.0-M-2; Consultative Committee for Space Data Systems: Washington, DC, 2002, p. 4-1 22
Figure 6: The DCC Curation Lifecycle Model 23
Figure 7: Reference Model for an Open Archival Information System (OAIS); CCSDS 650.0-M-2; Consultative Committee for Space Data Systems: Washington, DC, 2002, p. 2-2. 40
Figure 8: Reference Model for an Open Archival Information System (OAIS); CCSDS 650.0-M-2; Consultative Committee for Space Data Systems: Washington, DC, 2002, p. 4-1 41
Figure 9: The DCC Curation Lifecycle Model 42
Figure 10: Reference Model for an Open Archival Information System (OAIS); CCSDS 650.0-M-2; Consultative Committee for Space Data Systems: Washington, DC, 2002, p. 2-2. 59
Figure 11: Reference Model for an Open Archival Information System (OAIS); CCSDS 650.0-M-2; Consultative Committee for Space Data Systems: Washington, DC, 2002, p. 4-1 60
Figure 12: The DCC Curation Lifecycle Model 62

Tables

Table 1: Ingest Workflow according to OAIS for USDAS *33*
Table 2: Ingest Workflow according to OAIS for SR-SSDC *52*
Table 3: Ingest Workflow according to OAIS for CRO-DAS *71*
Table 4: Finance and budget USDAS *90*
Table 5: Finance and budget SR-SSDC *92*
Table 6: Finance and budget CRO-DAS *94*

Introduction

This document presents the outcomes of the work to date within WP5 of the project SERSCIDA, which is based in large part on prior work from earlier work packages. It includes two main articles concerning the future establishment of digital data archives in the countries of Bosnia and Herzegovina, Croatia, and Serbia. The first is what we are calling “Charter documents”, that is, articles that describe the future data services from an organizational and structural viewpoint. This includes information on the scope and nature of the data services, how these will be administered and staffed, as well as how these are placed within their national contexts in relation to other national and international institutions and organizations.

The second is a set of “Policy and Procedures” documents (one for each country), which lays out in some detail the future rules of the new data services: how they will operate, their purposes, their services, their workflows, their capacities, and their limitations. In a sense, these key documents together provide a nearly complete blueprint for the new data services in Bosnia and Herzegovina, Croatia, and Serbia.

Several remarks are in order, with respect to this first of three WP5 deliverables. Most notably, the intention of the participating Western Balkan countries is to establish a regional consortium of digital data archives. The aim will be to share to the greatest extent possible knowledge, tools, and experience, in order to make the most data available in the Serbo-Croatian language as possible to the greatest number of users. While the data services will certainly develop and evolve separately, according to their local contexts, they will do everything possible to harmonize their systems and practices so that data and information can easily flow across national borders. For this reason, the policies and procedures described in later in this document are quite similar across the three countries, even if differences will certainly arise in their realization.

An example of this is the focus of all three data services on the critical importance of Open access to data. All of their policies and procedures are geared toward this principle, while of course the data services will at the same time need to strike a balance between openness on the one hand and data protection and confidentiality of research respondent information on the other.

The aim of SERSCIDA is to do most of the preparatory work needed to bring to life new data services. This deliverable contributes to that end in the domain of institution-building, policy creation, and definition of workflow. The remaining SERSCIDA deliverables (due in months 28 and 30) will complete things in practice by providing the technical infrastructure and websites needed for the data services to soon go into operation.

A: Charter documents

I University of Sarajevo Data Archive and Services in Social Sciences and Humanities (USDAS)

Definition of organization

The University of Sarajevo Data Archive and Services (USDAS) for social sciences and humanities will be a national e-infrastructure public service providing long-term preservation and distribution of social sciences and humanities research data. The archive will serve as a vital data resource for researchers, teachers, and students, providing expertise in data management and promoting the secondary use of the data for scientific purposes.

Scope of collection

The holdings of the data archive will include data generated by social science and humanities researchers, or in a broader sense any data relevant to society produced by scientists, government institutions, or commercial entities. The data archive will collect both contemporary and historical data. The focus of the archive collection will be on both quantitative and qualitative data in the Social Sciences and Humanities (SSH). The collection will include quantitative data in the disciplines of anthropology, sociology, psychology, education science, information science, economics and demographics; qualitative data will encompass in the first phase oral history and life story projects, as well as other narrative studies, currently undertaken within academia and NGOs. Later on USDAS will undertake retrospective screening of empirical data produced by NGOs, IGOs, and academic institutions, as well as data from the period of the former SFR Yugoslavia. Confidential and Top Secret data will also have to be accommodated, as these data currently exist but are accessible in part only through the Information and Documentation Centre of BiH (IDC).

Services

The data archive will offer data services for national and international researchers, students, and teachers.

The following services will be available:

1) a searchable inventory of existing national projects with data;
2) preservation, documentation, and dissemination of social science and humanities data;
3) user education, user support, and consultancy;
4) regional collaboration and establishment of networks of both researchers and data repositories;
5) trainings for researchers (how to use the data service, how to prepare data for inclusion in the archive, and how to deposit prepared data); and
6) trainings for students on usage of data services.

For data providers there will be instructions available about preparing and transmitting data to the archive, as well as workshops and training sessions organized in collaboration with national and international experts in the area of data management and planning.

The data archive will also promote the use of data and the importance of long-term preservation, with special emphasis on open access to data.

The data archive will serve as coordinating body for the network of data specialists who promote and facilitate data archiving and use in their local institutions, and as the coordinating body for consortia subscriptions to commercial datasets (primarily for economic data).

Internal structure of organization

The data archive is planned to be organized as a department within a sub-organizational unit of University of Sarajevo. During the SERSCIDA project, the Human Rights Centre of the University of Sarajevo was identified as a suitable hosting institution.

Governance structure

The archive will be led by the Head of USDAS, who reports to a hosting institution head. He or she will be responsible for management of the archive, and communication with stakeholders and national, regional, and international partners.

An Oversight Board will have the responsibility of reviewing the work of the data archive. The Board will review annual plans and annual reports, and ensure that the work of the data archive is in line with goals and strategies of scientific and other institutions to which they belong. It will consist of representatives of the Steering Board of the University of Sarajevo.

A Scientific/Expert Board includes researchers, data experts/archivists, and data librarians who have an advisory role on issues related to social sciences methodology and data archiving procedures. They ensure that the needs of the scientific community and other interested communities are adequately addressed. Members will be distinguished experts in their field.

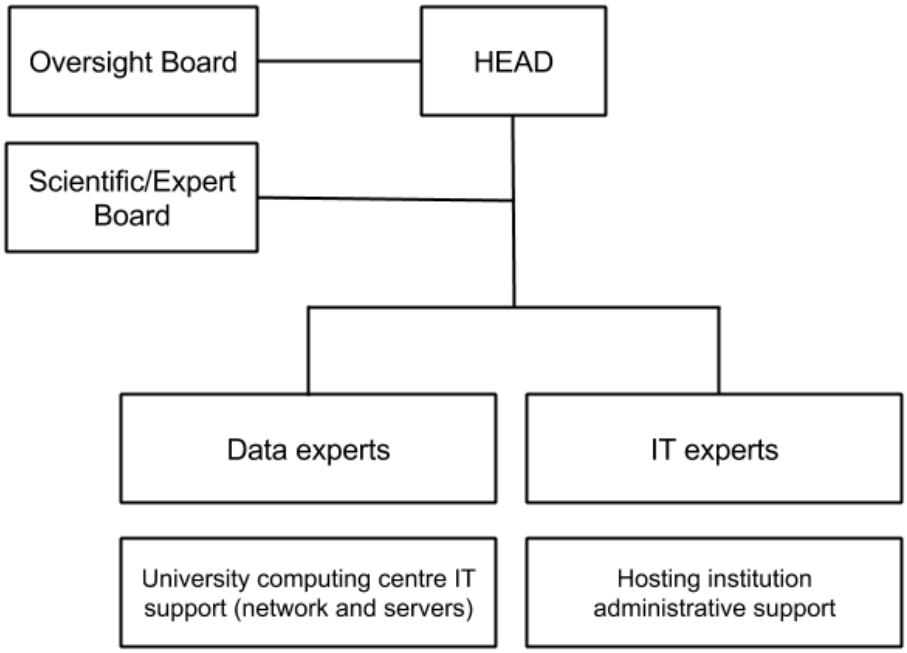


Figure 1: Organizational and governance chart USDAS

Staff profiles

The Head of the USDAS must be a person with a full understanding of the research data management workflow, as well as issues and challenges associated with these processes. He or She would preferably hold at least a Masters degree in a field of the social sciences and have experience with technical aspects of research data management.

Data expert roles can be divided into junior and senior positions based on previous work experience. Senior data experts should have demonstrated knowledge in social science methodology, statistical data processing, and data management. An MA in a field of the social sciences and an additional degree in information science are desirable. Junior-level data experts should hold at least a BA in a social science discipline. This does not require previous extensive experience in research data management, but they should have a sufficient knowledge of statistical operations and data processing.

For the IT experts, systems administration, software programming and database management skills are necessary for a position in the data archive, which may be outsourced in the initial phase. He or she has to have experience in software applications architecture, development, and maintenance. The position also requires an overall understanding of data exchange protocols in a networked environment and security issues associated with it. A BA in the field of informatics and computer science or equivalent experience is required, and previous experience with research data management and processing is desirable.

Cooperation

National and international cooperation is a crucial part of data archive operations.

On the national level it is very important for the data archive to be well recognised as a valuable national service by all relevant science institutions which are part of the science community and

higher education system in Bosnia and Herzegovina. This will be achieved by maintaining constant communication with relevant experts from higher education institutions from Bosnia and Herzegovina, and other independent institutions and individuals.

The Ministry of Science, Education and Sports of Sarajevo Canton, as the main science policy institution and financing institution for basic scientific activities of the University of Sarajevo, should be fully aware of the benefits of long-term preservation of research data, as well as the benefits connected with open access to research data. Cooperation with the National Statistical Institute, which produces national statistics, valuable and very heavily used data for the social sciences (sociology, economics), will be beneficial for both the NSI and the data archive. Also, the involvement of the Ministry of Civil Affairs of Bosnia and Herzegovina will be of crucial importance.

On the international level, cooperation with existing European data archives is crucial because of their extensive experience in research data management. One of the important goals of USDAS will be to become a member of Consortium of European Social Science Data Archives, CESSDA. CESSDA membership is conditioned by following international standards and best practice criteria, which can ensure the quality of data archive work and services. It also promotes data availability and data sharing throughout the international community.

Close cooperation with ESFRI research infrastructures as well as other international projects and initiatives dealing with managing and/or publishing research data, such as DARIAH, SHARE, ESS, OpenAIREPlus, and EUDAT, is needed to ensure development of an efficient pan-European data sharing and exchange ecosystem.

2. Serbian Social Science Data Centre (SR-SSDC)

Article 1

This Charter document regulates: the name, the residence, the activities, the scope of collection, services, management structure, job descriptions, and collaboration of the Serbian Social Science Data Center (SR-SSDC).

The SR-SSDC name and residence

Article 2

The name SR-SSDC stands for: the Serbian Social Science Data Center, The Institute of Economic Sciences, Belgrade, the Republic of Serbia, Belgrade, 12 Zmaj Jovina Street.

SR-SSDC General activities

Article 3

General activities in SR-SSDC are:

- Activities related to the collection, validation and data conversion, creation of metadata, storage, and distribution of data collections;
- Creating and improving the classification scheme of data types;
- Administration of network/system specialized in collecting, storing, and distributing data;
- Implementation of systems aimed to improve the quality and safety of using specialized infrastructure designed for data management and preservation;
- Design and implementation of new database software, and preparing reports on the use of software tools and the database management system;
- Permanent monitoring of compliance with the international standards in the field of data management and preservation;
- Mediation between the demands of data users and data providers (contracts for deposit and distribution of data, admission requirements, testing, etc.);
- Providing of services to third parties and user trainings; Customer support;
- Activities related to web portal development and maintenance;
- Production of documentation (warranties, instructions, service sheets, etc.);
- Collaboration with the wider scientific community in the field of collecting and distributing data;
- Active promotion of the SR-SSDC;
- Open Access Initiative promotion.
-

Article 4

The main goal of SR-SSDC is collecting, preserving and disseminating data generated in the process of scientific research mainly realized by institutes and faculties, but also data significant for research in social sciences and humanities generated by relevant institutions such as government agencies, the Statistical Office of the Republic of Serbia, etc.

SR-SSDC will collect, preserve, and disseminate all types of data, both quantitative and qualitative. The data scope will cover all disciplines in the social sciences and humanities. Within a regional framework SR-SSDC will develop specific knowledge and skills regarding collection and preservation of economic data.

SR-SSDC Internal structure of organization and management

Article 5

SR-SSDC is an organizational unit within the Institute of Economic Sciences in Belgrade. Its facilities are on the upper floor of the Institute of Economic Sciences building (office 307).

Article 6

SR-SSDC is organized as a separate unit managed by a Head who is report to the IES Director, the IES Scientific Council, and the SR-SSDC Scientific Board.

The SR-SSDC Head is appointed by the Scientific Council, on a proposal of IES Management, for four-year period.

Article 7

The tasks of SR-SSDC management are as follows:

- Proposing basic business policy, long-term (strategic) development plan, and the annual work and development program and taking all measures for successfully implementing policy;
- Organizing, coordinating, and monitoring the work of the Center and the fulfillment of objectives;
- Preparing and taking care of Budget and financial plans;
- Reporting on all activities to the director and Scientific Council of the Institute of Economic Sciences;
- Fulfillment all the obligations based on membership in the Scientific Council of the SR-SSDC;
- Ensuring cooperation with stakeholders and partner institutions;
- Initiating and coordinating the implementation of international cooperation and professional training of staff in the SR-SSDC;
- Taking care of promotional activities, including promotional meetings and events;

- Initiating and managing the application process for international/national projects in the field of data management and data preservation;
- Monitoring and coordinating the work of the SR-SSDC according to international standards in the field of data management and data preservation.

Required Skills

- PhD in the field of economics or other social sciences;
- Scientific or teaching position in the area of economics or other social sciences;
- Specific knowledge in the field of data management and data preservation;
- Experience of working in international projects;
- English language proficiency;
- References in the field of data management.

Article 8

Beside the Director and the IES Scientific Council, the work of SR-SSDC will be evaluated in the scope of activities connected to data servicing by the SR-SSDC Scientific Board. The Scientific Board members are mainly representatives of Belgrade University, National Library of Serbia, the Statistical Office of Republic of Serbia, institutes of social sciences, and the Institute of Economic Sciences, as well as other stakeholders.

The SR-SSDC Scientific Board has eleven representatives from the Ministry of Education, Science and Technological Development (one representative), the Institute of Economic Sciences (two), the National Library of Serbia (one), the Statistical Office of Republic of Serbia (one), institutes of social sciences (three), universities (two), and the Head of the SR-SSDC.

Article 9

The SR-SSDC Scientific Board has two regular sessions a year.

Article 10

SR-SSDC activities are realized according to the annual SR-SSDC plan.

The proposition of the annual SR-SSDC plan is adopted by the Scientific Council based on the written proposal of the Head and the SR-SSDC Scientific Board, with the previous Director`s approval.

Article 11

There are two SR-SSDC staff positions: Data Manager and IT support.

Article 12

The position of the Data Manager includes the following activities:

- Providing support to researchers (depositors) in the process of depositing the data and creating metadata,
- Data and metadata quality control;
- Preparing data for secondary use (performing data conversion into accessible formats compatible for use within statistical software);
- Management and control of the access to the data collections;
- Mediation between data users and data providers;
- The provision of services to third parties and training the users.

Required Skills

- High school education, oriented to informatics;
- Knowledge of statistics and statistical software packages and tools;
- Knowledge of server and network protocols;
- English language proficiency;
- Skills in rapid adaptation to new and different technologies.

The IT support includes network/system administration related data collection, preservation and dissemination.

Cooperation

Article 12

The cooperation with domestic and international institutions is a very important segment of SR-SSDC activities. It is essential for SR-SSDC to be the central infrastructure point for collecting, preserving, and disseminating research data in the social sciences and humanities in the Republic of Serbia.

The SR-SSDC significance in the local context is provided by the Ministry of Education, Science and Technological Development of the Republic of Serbia, the National Library of Serbia and the Statistical Office of Republic of Serbia, mainly through the membership in the SR-SSDC Scientific Board.

In the international framework SR-SSDC cooperates with a number of institutions. The first level of cooperation is with the social science digital archives of Bosnia and Herzegovina and Croatia. The second level is cooperation is with the digital archives of the members of the international organization Consortium of European Social Science Data Archives (CESSDA), and finally the cooperation with all organizations which are actively involved in digital data archiving. The cooperation with the institutions above is realized through knowledge and experience exchange and through participation in various projects.

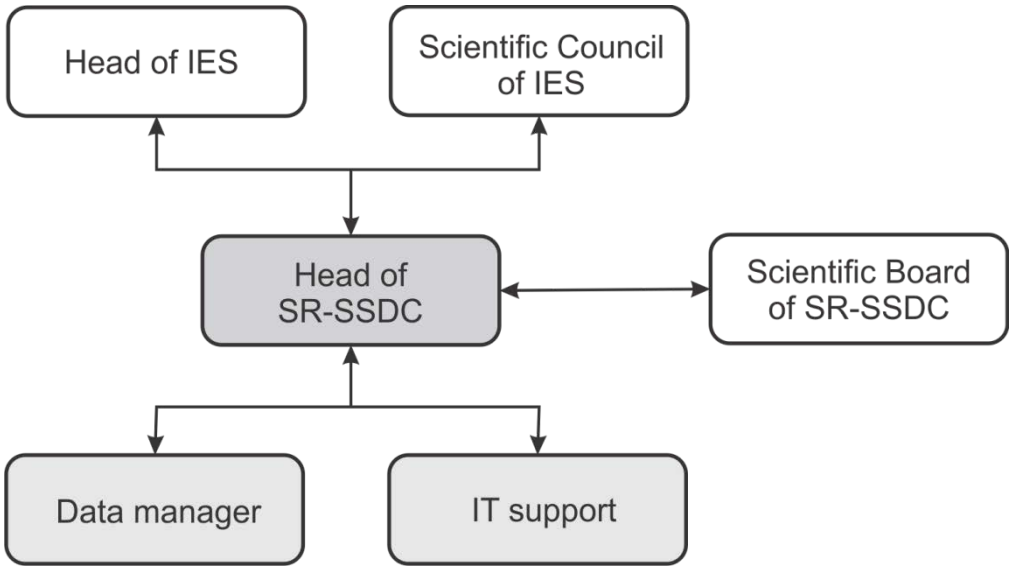


Figure 2: SR-SSDC Organisational Chart

3. Croatian Data Archive for the Social Sciences (CRO-DAS)

Definition of organization

The Croatian Data Archive for the Social Sciences (CRO-DAS) will be a national e-infrastructure public service providing long-term preservation and distribution of social science research data. The archive will serve as a vital data resource for researchers, teachers, and students, providing expertise in data management and promoting the secondary use of the data for scientific purposes.

Scope of collection

The holdings of the data archive will include data generated by social science researchers, or in a broader sense any data relevant to society produced by scientists, government institutions, or commercial entities. The data archive will collect both contemporary and historical data. The primary focus of the archive will be on quantitative data in the social science disciplines of sociology, psychology, education science, information science, political science, and economics. The data archive will also collect and curate qualitative data.

Services

The data archive will offer data services for national and international researchers, students, and teachers.

Users of the data archive will be able to search and browse data collections using metadata about studies and variables. It will also be possible to perform simple on-line data analysis - creating tables, graphs and other analytical results. Data and related materials may be downloaded to personal computers, and converted to several file formats for use in different statistical packages (SPSS, Stata, R, etc.). The data archive will also provide materials and support for data users in the form of written instructions and advice, and will organize training sessions related to secondary use of data and data citation.

For data providers there will be instructions about preparing and transmitting data to the archive, as well as workshops and training sessions organized in collaboration with national and international experts in the area of data management and planning.

The data archive will also promote the use of data and the importance of long-term preservation, with special emphasis on open access to data.

The data archive will serve as the coordinating body for the network of data librarians who promote and facilitate data archiving and use in their local institutions, and as the coordinating body for consortia subscriptions to commercial datasets (primarily for economic data).

Internal structure of organization

The data archive will be organised as a separate unit inside an existing organisation. Preferably this should be an organisation where social scientists are concentrated. During the SERSCIDA project, and also during the active period of the Initiative for establishment of a social science data archive, the Faculty of Humanities and Social Sciences, University of Zagreb, was identified as a suitable hosting institution, in part because of its size and existing capacities. Cooperation with the University of Zagreb's University Computing Centre (SRCE) is crucial, as this institution is a recognized and

respected infrastructure with excellent capacities and skills in data and information management, and also because of its previous involvement in building basic data services for the scientific community.

Governance structure

The archive will be led by the Head of the data archive, who reports to a hosting institution head. He or she will be responsible for management of the archive and communication with stakeholders and national, regional, and international partners.

An Oversight Board will have the responsibility of reviewing the work of the data archive. The Board will review annual plans and annual reports, and will ensure that the work of the data archive is in line with goals and strategies of scientific and other institutions to which they belong. It will consist of representatives of major stakeholders: a representative of the Ministry of Science Education and Sports; representative(s) of the funders; management representatives from the hosting institution and SRCE, individuals in management positions within major research and education institutions (universities and institutes), as well as representatives of the Croatian NSI and other governmental agencies who are potential major data depositors. A Scientific/Expert Board will include researchers, data experts/archivists, and data librarians who have an advisory role on issues related to social science methodology and data archiving procedures. They will ensure that the needs of the scientific community and other interested communities are adequately addressed. Members will be distinguished experts in their field. Their work will be guided by written guidelines in cooperation with the data archive head and staff.

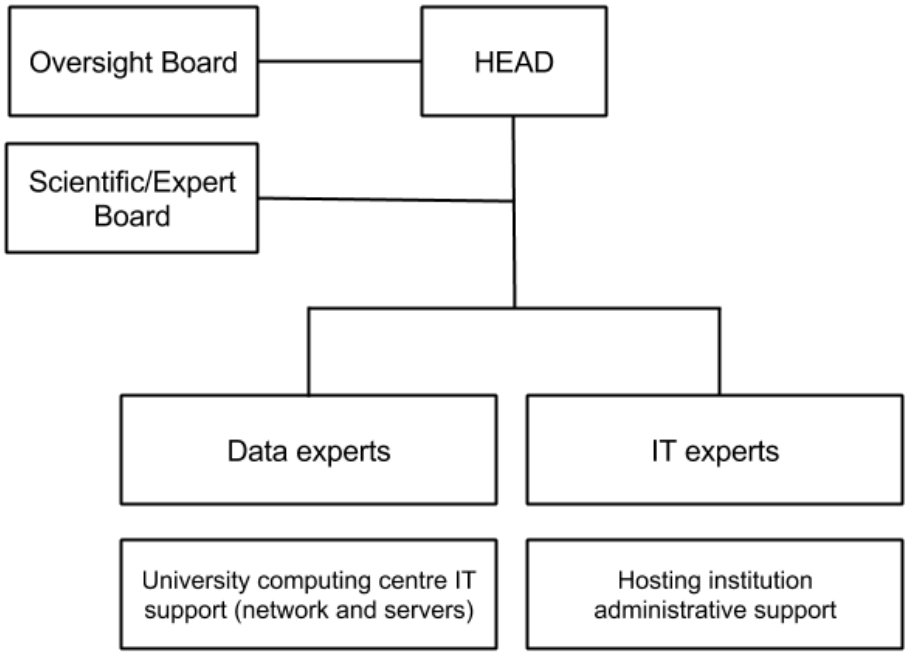


Figure 3: Organizational and governance chart CRO-DAS

Staff profiles

The Head of the data archive must be a person with a full understanding of the research data management workflow, as well as issues and challenges associated with these processes. He or She would preferably hold at least a Masters degree in a field of the social sciences and have experience with technical aspects of research data management.

Data expert roles can be divided into junior and senior positions, based on previous work experience. Senior data experts should have demonstrated knowledge in social science methodology, statistical data processing, and data management. An MA in a field of the social sciences and an additional degree in information science are desirable. Junior-level data experts should hold at least a BA in a social science field. This does not require previous extensive experience in research data management, but they should have sufficient knowledge of statistical operations and data processing.

For the IT experts, systems administration, software programming and database management skills are necessary for a position in the data archive. He or she has to have experience in software applications architecture, development, and maintenance. The position also requires an overall understanding of data exchange protocols in a networked environment and security issues associated with it. A BA in the field of informatics and computer science or equivalent experience is required, and previous experience with research data management and processing is desirable.

Cooperation

National and international cooperation is a crucial part of data archive operations.

On the national level it is very important for the data archive to be well recognised as a valuable national service by all relevant science institutions which are part of the science community and higher education system in Croatia. This will be achieved by including the relevant representatives in the Oversight Board, but also by maintaining communication with relevant institutions and individuals.

The Ministry of Science, Education and Sports, as the main science policy institution and financing institution for basic scientific activities, should be fully aware of the benefits of long- term preservation of research data, as well as the benefits connected with open access to research data. The Croatian Science Foundation, as the main funding body for competitive research projects, and also other institutions who develop criteria of scientific excellence and improvement of scientific results, may rely on existing infrastructure for data archiving when developing policies about mandating or recommending data archiving and open access to data to researchers.

Cooperation with the Croatian National Statistical Institute, which produces national statistics, valuable and very heavily used data for the social sciences (sociology, economics), is beneficial for both the NSI and the data archive. Cooperation with the National Labour Services Agency and other government agencies which produce data and statistics of value for researchers, with possibilities for data linking, is also important, as well as cooperation with the Croatian National Archive, which can play a role in long-term preservation of some kinds of data (e.g., publicly available data as national heritage).

On the national level, there are several e-infrastructure systems and services (bibliography, scientific journals, and scientists) that the future system will have to bring together if our final goal is to have a comprehensive service for the research community and the public in general. Cooperation in the field of interoperability of the data archive with those services means connecting research data with publications (integral texts and/or bibliographical entries) and scientific profiles.

On the international level, cooperation with existing European data archives is crucial because of their extensive experience in research data management. One of the important goals of the Croatian Data Archive is to become a member of Consortium of European Social Science Data Archives, CESSDA. CESSDA membership is conditioned by following international standards and best practice criteria, which can ensure the quality of data archive work and services. It also ensures data availability and data sharing throughout the international community.

Close cooperation with ESFRI research infrastructures as well as other international projects and initiatives dealing with managing and/or publishing research data, such as DARIAH, SHARE, ESS, OpenAIREPlus, and EUDAT, is needed to ensure development of an efficient pan-European data sharing and exchange ecosystem.

I University of Sarajevo Data Archive and Services (USDAS)

1. Contextualisation and organisation of document

The aim of the F7 project SERSCIDA is the establishment of a prototype for social science digital data archives[1] in the Western Balkans, which could go into operation specifically in the countries of Bosnia and Herzegovina, Croatia, and Serbia. This includes the preparation of relevant documents for the formation of new data services. The present document "Data Policies and Procedures" was created as part of Work Package 5 (WP5) of the SERSCIDA project.

The present report will follow in large part the functionalities of the international standard *Open Archival Information System (OAIS) Reference Model* (explained in chapter 2). In fact, the University of Sarajevo Data Archive and Services (USDAS) is committed to developing its digital preservation policies and strategies in accordance with the (OAIS) Reference Model. This approach aligns with policies and practices of members of the Consortium of European Social Science Data Archives CESSDA, and with the SERSCIDA deliverables from prior Work Packages, as well as the common standards of the international infrastructure CESSDA AS (Annex 2.7).[2]

The target audience of this document includes state, federal, and cantonal ministries, stakeholders of USDAS, its staff, the social science research community in Bosnia and Herzegovina, and the general public. Beyond its purpose as a deliverable for the SERSCIDA project, this policy document could be used for establishing new data archive centres in other countries.

2. Conceptual models and guiding principles

2.1. The OAIS Reference Model

A very important concept for the establishment of any digital repository is the Reference Model for an Open Archival Information System (OAIS)[3], adopted as an ISO standard in 2003 (ISO 14721:2003 OAIS). OAIS provides a framework for the understanding of digital archives and provides common terminology to be used in this context. The OAIS reference model is flexible enough to use in a wide variety of environments since it is strictly a logical model which does not depend on a particular implementation. It describes the components and services required to develop and maintain archives, in order to support long-term access to and understanding of the information in those archives. The OAIS model specifies how digital material should be preserved for a community of users (Designated Community)[4] from the moment digital material is ingested into the digital storage area, through subsequent preservation strategies, to the dissemination of digital material for the end user.

[1] We use "data archive" and "data service" interchangeably.
[2] Annexes to Statutes for CESSDA, No. 1-6 (18 June 2013).
[3] Reference Model for an Open Archival Information System (OAIS); CCSDS 650.0-M-2; Consultative Committee for Space Data Systems: Washington, DC, 2012.
[4] The Designated Community is a special class of Consumers; the subset of Consumers expected to independently understand the archived information in the form in which it is preserved and made available by the OAIS.

An OAIS-type archive carries out its preservation and access responsibilities in an environment populated by three key external stakeholders: 1) the Management (archivist, system administrator), which formulates, revises, and enforces a high-level policy framework that covers the OAIS activities; 2) the Producer (data producer, depositor, or researcher), that is, the individuals, organizations, or systems that create digital objects and/or transfer information to the OAIS for long-term preservation; and 3) the Consumer (data consumer, user, researcher) - the individuals, organizations, or systems expected to use the information preserved by the OAIS.[5]

OAIS Environment

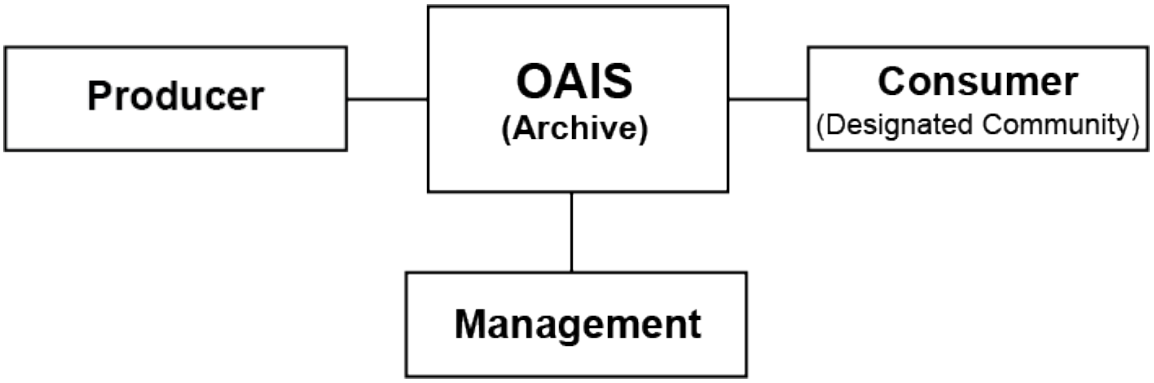


Figure 4: Reference Model for an Open Archival Information System (OAIS); CCSDS 650.0-M-2; Consultative Committee for Space Data Systems: Washington, DC, 2002, p. 2-2.

OAIS lists six functional entities that fulfil OAIS' dual role of preserving and providing access to the information: Ingest, Archival Storage, Data Management, Preservation Planning, Access, and Administration.

[5] For the sake of convenience, for the rest of this document we will generally use terminology more common to the social science archiving community in Europe (e.g., data producers, data users), rather than the OAIS terminology.[6] CCSDS 650.0-M-2, p. 3-1

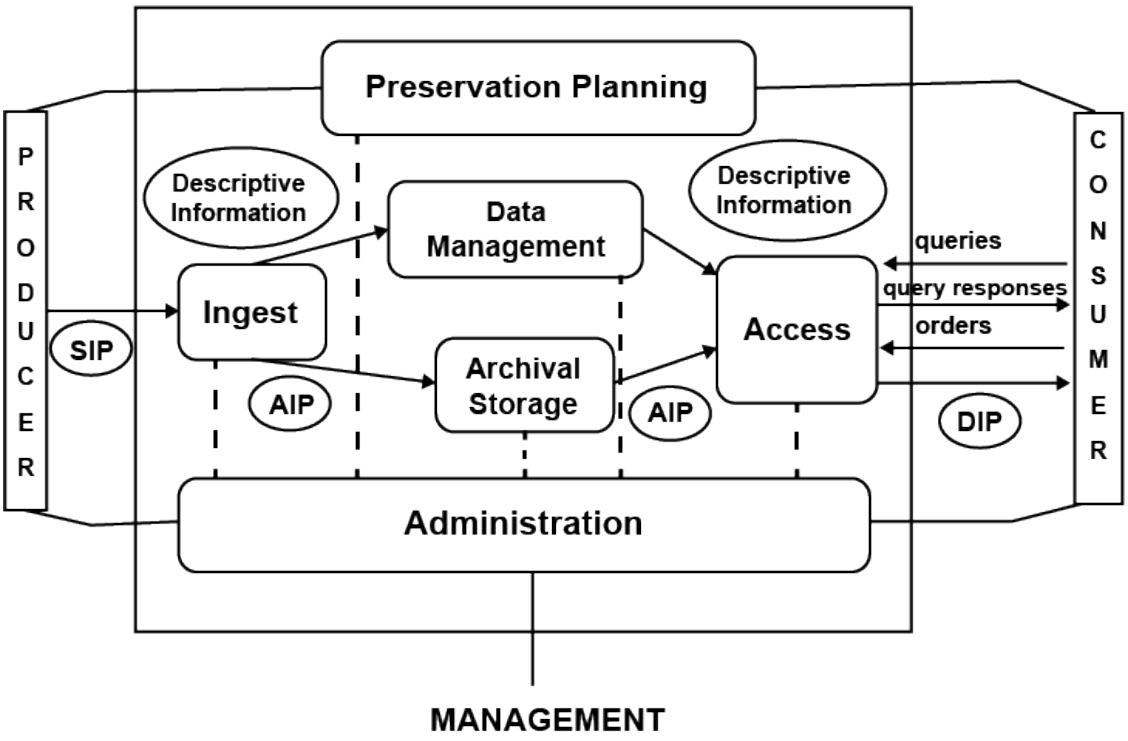


Figure 5: Reference Model for an Open Archival Information System (OAIS); CCSDS 650.0-M-2; Consultative Committee for Space Data Systems: Washington, DC, 2002, p. 4-1

An essential part of the OAIS model is the Information Package (IP), which represents the structure of information as it moves into, through, and out of the archival system. There are three types of information packages, all of them consisting of the digital object to be preserved and the metadata necessary to support its long-term preservation and access:

- SIP (Submission Information Package): information package that is transferred from the Producer to the Archive.
- AIP (Archival Information Package): information package that is stored and preserved by the Archive.
- DIP (Dissemination Information Package): information package delivered to the Consumer in response to an access request.

There are several minimum responsibilities that should be maintained by an OAIS[6]:

- Negotiate for and accept appropriate information from Producers;
- Obtain sufficient control of the information provided to the level needed to ensure long-term preservation;
- Determine, either by itself or in conjunction with other parties, which communities should become the Designated Community and, therefore, should be able to understand the information provided, thereby defining its Knowledge Base[7];
- Ensure that the information to be preserved is independently understandable to the Designated Community. In particular, the Designated Community should be able to

[6] CCSDS 650.0-M-2, p. 3-1

[7] "A person, or system, can be said to have a Knowledge Base, which allows that person or system to understand received information. For example, a person who has a Knowledge Base that includes an understanding of English will be able to read and understand an English text." (CCSDS 650.0-M-2, p. 2-3)

understand the information without needing special resources such as the assistance of the experts who produced the information;

- Follow documented policies and procedures which ensure that the information is preserved against all reasonable contingencies, including the demise of the Archive, ensuring that it is never deleted unless allowed as part of an approved strategy;
- Make the preserved information available to the Designated Community and enable the information to be disseminated as copies of, or as traceable to, the original submitted Data Objects with evidence supporting its Authenticity[8].

2.2. The DCC Curation Lifecycle Model: metadata and standards

The Data Centre will follow all the phases of the DCC **Curation Lifecycle Model**[9]. Data often have a longer lifespan than the research project that creates them. The DCC Curation Lifecycle Model can be used to plan activities within an organisation to ensure that all necessary phases are undertaken, each in the correct sequence.

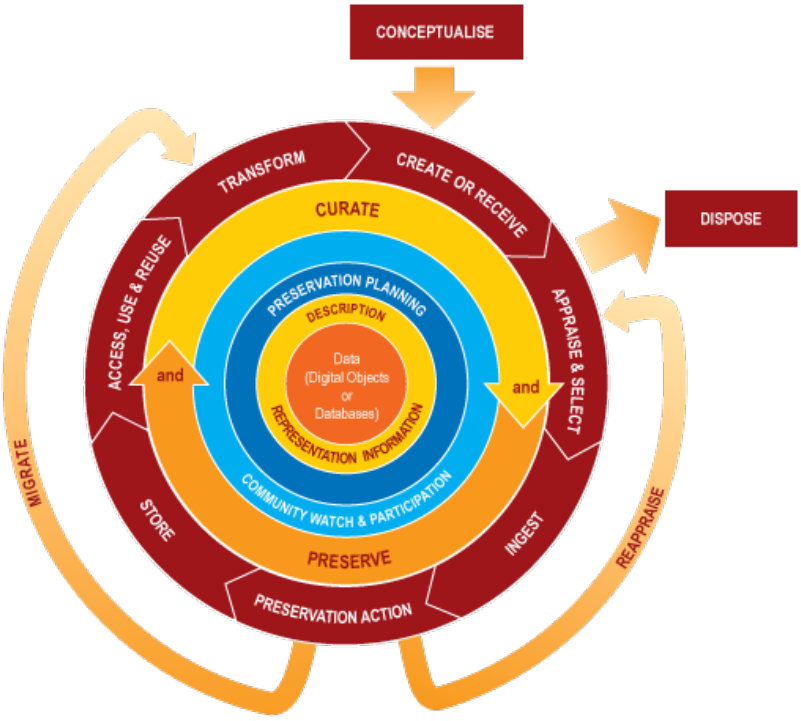


Figure 6: The DCC Curation Lifecycle Model

[8] Authenticity: relies on three provable properties: the record is what it purports to be, that it was created by whomsoever it is purported to have been created by; and that it was created at the time when it is purported to have been created. (Joint Information Systems Committee (JISC) Digital Preservation: continued access to authentic digital assets (Briefing Paper, November 2006), p.2). Other concepts designed to guide policies for information security within an organization are confidentiality (ensuring that data are not made available or disclosed to unauthorised individuals) and integrity (ensuring that data are not altered or destroyed in an unauthorized manner).

[9] DCC Curation Lifecycle Model: http://www.dcc.ac.uk/resources/curation-lifecycle-model

The collaboration between the data archive and the data producer should be a priority during the whole data curation process. It is important that the data producers are working with USDAS during the documentation process, since they know their data best and are most familiar with the study methodology. On the other hand, USDAS has the best knowledge about data documentation procedures and metadata standards. In addition, the data producers are usually responsible for the documentation, metadata description and the data themselves (data preparation, anonymisation, and cleaning). As a basic principle, it should be made clear throughout the whole archival process who bears responsibility for which tasks (the data producer or the archive).

Standardisation is a crucial element in the establishment of social science data archives. USDAS is devoted to ensuring best practice and standardised approaches in order to enable compatibility and interoperability for international data sharing.

USDAS orients itself towards international standards adopted by CESSDA AS and its service providers so that our data can be available and sharable throughout the international community. Regular reviews will be undertaken in order to guarantee best practice.

At USDAS we will be compliant with the **Data Documentation Initiative (DDI)** metadata specification, version 2.5 or higher[10]. To be more precise, USDAS will create the following types of metadata generated from accompanying documentation[11]:

- **Descriptive metadata**: Describes a resource for purposes such as discovery and identification (contains elements such as title, abstract, author, and keywords). DDI complies with Dublin Core Metadata Initiative which describes a core set of 15 elements intended to facilitate discovery of electronic resources;[12]
- **Administrative metadata:** Contains information about the use, management, and encoding processes of digital objects over time (e.g., information about data creation);
- **Technical metadata**: Provides information about the overall system environment and provides the technical information needed to use data (e.g., file format, application used and operation system)
- **Structural metadata**: Describes the logical structure of a multidimensional object. The Metadata Encoding and Transmission Standard (METS) will be used where appropriate;[13]
- **Preservation metadata**: Provides information needed to archive and preserve a resource. The PREMIS Data Dictionary for Preservation Metadata defines a core set of preservation metadata elements and describes relationships between digital preservation entities: Intellectual entity, Object, Event, Agent, and Rights.[14]

More and more digital objects require more reliable and stable references (regardless of their location and across system change). USDAS will therefore aim to assign a persistent identifier (DOI (Digital Object Identifier)/URN (Uniform Resource Name) to every study coming into the archive.

[10] Data Documentation Initiative (DDI): http://www.ddialliance.org/

[11] As a general remark, metadata standards should be machine readable (XML coded) in order to be processed automatically. XML is an open, well-supported, and widely adopted standard for encoding textual data, designed to be used regardless of the hardware and software environment.

[12] Dublin Core Metadata Initiative: http://dublincore.org/

[13] Metadata Encoding and Transmission Standard (METS) http://www.loc.gov/standards/mets/

[14] PREMIS Data Dictionary for Preservation Metadata: http://www.loc.gov/standards/premis/v2/premis-2-0.pdf

The following two international archival standards could be used as well:

- **ISAD(G)**: International Standard Archival Description (General) or ISAD(G) is an international standard which provides guidelines for creating the content of an archival description (descriptive metadata). It promotes the creation of consistent and appropriate descriptions, facilitating the retrieval and exchange of information, and the integration of descriptions into a unified information system. There is a list of elements which are considered mandatory (six elements) when writing an archival description.[15] Based on ISAD(G) is the **Encoded Archival Description (EAD)** standard;
- **ISAAR(CPF)**: is the International Standard Archival Authority Record For Corporate Bodies, Persons and Families standardised by the International Council on Archives in 2003.[16] Whereas ISAD(G) describes the archival material itself, ISAAR(CPF) is constructed to give information about the creator (authority) of the archival material with additional information on history, status, functions, structure and relations – generally speaking about the provenance of the data. Again, there has been another international standard called **Encoded Archival Context (EAC),** which is derived from ISAAR(CPF).

Within this regulatory framework, USDAS will also respect the following **national and international legal frameworks**:

- EU Copyright Directive (EU, 2001);
- Law on Higher Education (Zakon o visokom obrazovanju) ('Official Gazette of Sarajevo Canton' No:22/10);
- Framework Law on Higher Education (Okvirni zakon o visokom obrazovanju) ('Official gazette of Bosnia and Herzegovina' No:59/07);
- Law on Research (Zakon o naučno-istraživačkoj djelatnosti) ('Official Gazette of Sarajevo Canton' No:10/04);
- Law on Archival Material (Zakon o arhivskoj djelatnosti) 'Official Gazette of Sarajevo Canton' No: 2/00;
- Law on Copyrights and Related Rights (Zakon o autorskom i srodnim pravima), ('Official gazette of Bosnia and Herzegovina 63/10; and
- Law on Cultural Property (Zakon o zaštiti kulturne baštine) 'Official Gazette of Sarajevo Canton' No 2/00).

Publicly funded research data are a public good produced in the public interest, and therefore should be openly available to the maximum extent possible (OECD principle).[17] USDAS thus promotes the philosophy and practice of **Open Access,** where research data are as easily accessible as possible, while still protecting the confidentiality of respondent information and the rights of those who gathered or created the data. In addition, the basic data service of USDAS for all data producers and users will as a rule be free of charge. At the same time, USDAS will not pay for the data it acquires.

[15] General International Standard Archival Description, ISAD (G): http://www.ica.org/10207/standards/isadg-general-international-standard-archival-description-second-edition.html

[16] International Standard Archival Authority Record For Corporate Bodies, Persons and Families, ISAAR (CPF) http://www.ica.org/10203/standards/isaar-cpf-international-standard-archival-authority-record-for-corporate-bodies-persons-and-families-2nd-edition.html

[17] OECD (2007) OECD principles and guidelines for access to research data from public funding. Organisation for Economic Cooperation and Development: http://www.oecd.org/science/sci-tech/38500813.pdf

Audit and Certification: As a future goal, we will seek to become compliant with at least one of the following standards of trust: Data Seal of Approval (DSA), ISO 16363 or DIN 31644 for the Audit and Certification of Trustworthy Digital Repositories.[18] Since one of the priorities of USDAS is to become a CESSDA member, we seek to be compliant with the DSA, which is granted to repositories that are committed to best practices in archiving and providing access to scholarly research data in a sustainable way. From the beginning, we will focus on quality control mechanisms throughout the whole archiving process.

Confidentiality and disclosure: All depositors should ensure that data meet the requirements of confidentiality and non-disclosure for data collected from human subjects. Datasets should not contain direct identifiers (names, addresses, telephone numbers, etc.) or indirect identifiers (which, when combined with other variables, may provide enough details to identify an individual). In-house confidentiality checks by USDAS will help to guarantee that confidentiality requirements are met.

3. Acquisition and deposit

This section focuses on the acquisition of data and covers data selection and appraisal strategies that will be adopted, as well as the file formats accepted by USDAS. The collection development policy will answer the questions of what, from whom, why, and how are we going to collect data at USDAS. In a next step, policies for the deposit of the data and related documentation materials are discussed. This includes the conditions for depositing data, data quality, confidentiality of data, and depositor rights. Although these elements are not explicitly treated within the OAIS Reference Model, they are worth treatment in this policy document, since they are important aspects of the proper establishment of a social science data archive.

3.1. Collection development policy[19]

Data selection and appraisal play an important role in the acquisition of data in any archival setting. The collection development policy indicates the principles and criteria by which the archive develops its data collection in order to serve the Designated Community. The collection development policy of USDAS will be flexible and respond to future developments and shifting requirements that will influence the archive's data collections (technology, scientific standards, etc.).

USDAS collects data and information across the range of social science disciplines and the humanities within Bosnia and Herzegovina (including both entities and Brčko District).[20] This includes **quantitative** and **qualitative data**.

WHAT can be accepted:

[18] Data Seal of Approval: http://www.datasealofapproval.org; Audit and certification of trustworthy digital repositories: http://www.iso16363.org/; Nestor Seal for Trustworthy Digital Archives: http://www.langzeitarchivierung.de/Subsites/nestor/EN/nestor-Siegel/siegel_node.html.

[19] The content of this policy draws on the collection development policy of the UKDA: http://www.data-archive.ac.uk/media/54773/ukda067-rms-collectionsdevelopmentpolicy.pdf, and the experience of setting up new Data Archives in Finland and Slovenia (Laaksonen, H. et al, 2006): http://www.iassistdata.org/downloads/iqvol301laaksonen_0.pdf

[20] The following disciplines are embraced: Economics; Education; Employment and Labour; Environment, Conservation and Land Use; Health; History; Industry and Management; Law, Crime and Legal Systems; Media, Communication and Language; Political Science; Psychology; Sociology, Society and Culture; Social Welfare Policy and Systems, Anthropology, Ethnography.

- Research that gives evidence and greater understanding of Bosnian and Herzegovinian society or parts of it (including international surveys);
- Research data where Bosnian and Herzegovinian researchers are involved;
- Comparative or continuous research (panel data, longitudinal surveys, and time series that allow examination of trends);
- Data produced with methodological excellence;
- Works of general interest for social science research (quality and exemplary studies with analytic potential).

The deposited data will be retained for the long-term, unless otherwise specified by the depositor, funding agencies, or by government regulations.[21] USDAS will accept data and documentation in paper form only under exceptional circumstances.[22]

FROM whom are data solicited and accepted:

Eligible depositors are researchers, members of research institutions, government departments, offices of national statistics, and public and commercial sector sources (e.g., marketing research firms). Depositors may only deposit their own work and will be expected to sign a data deposit agreement (see section "data deposit" below).

WHY certain data are accepted:

There are three main reasons why data are solicited and accepted. First, because they possess potential value for secondary use and analysis for research. Second, because they may serve teaching and learning purposes. And finally, for validation reasons - to make possible the replication of research for important studies.

HOW the data will be acquired:

The acquisition process will consist of three stages: 1) identifying potential datasets (localisation), 2) negotiation with the data creators, and 3) receiving the relevant material. A very important source will be the list of potential data producers and consumers that has been established during the SERSCIDA project.

As USDAS will be in the early stages of its existence, the **acquisition strategies** will be highly proactive and persistent. This will include open calls, announcements at conferences, and direct solicitation. No national research inventory for social science research data, where staff can actively search for and proactively seek data for deposit, is currently available for this purpose in Bosnia and Herzegovina.

Similar as in some other countries, collaboration in the process of preparing and depositing data is entirely voluntary. This means that special effort will need to be invested into the process of negotiating and gaining access to the data sets:

[22] Studies in paper form will be accepted and worked on under separately funded digitisation projects, rather than as 'routine' activity. Accepting such material is fine, but they require careful storage until one has the resources to digitise and process them (i.e., special funding to digitise some of the paper deposits).
[22] Studies in paper form will be accepted and worked on under separately funded digitisation projects, rather than as 'routine' activity. Accepting such material is fine, but they require careful storage until one has the resources to digitise and process them (i.e., special funding to digitise some of the paper deposits).

- Good **working relations** with stakeholders and potential depositors (National Statistical Agency of Bosnia and Herzegovina (as well as entity agencies), research institutes and faculties in social sciences, private research companies and individual researchers) to identify and acquire new sources of data;
- **Links** with State/Entity/Historical Archives and Research Libraries;
- **Promotional activities** to raise awareness and encourage deposit of data collections (workshops, round tables, conferences, and training events);
- **Agreements** could be established with funding bodies for the disciplines in question in order to discuss deposit of material with grant-holders;
- **Data discovery** activities by staff members in publications (academic journals), news media, conferences, e-mail lists, and websites (e.g., universities' online catalogues and websites of research funding bodies).

The aim is to establish links and working relations with the stakeholders and partners mentioned above in order to establish long-term cooperation and a common framework for the deposit of data collections.

USDAS reserves the right to **refuse** to accept material under the following conditions:

- Data do not match the criteria of collection development policy and would be better dealt with at another institution (also legal conditions);
- Insufficient or poor quality documentation;
- Depositor and USDAS do not agree on access and dissemination conditions;
- There are currently no restrictions on file sizes, or number of files or studies. However, USDAS reserves the right to refuse to accept materials which are of a volume which may make it difficult or impossible to process given the resources, staff, facilities, or capacities of USDAS.

In order to make the data and documentation accessible for other users and to be able to process and use them over time, they should be in a form that best ensures their longevity. The data and documentation **file formats** that are preferred at the ingest phase by the Data Centre are non-proprietary, openly documented, unencrypted, and uncompressed that are commonly used by the research community. Several lists are available online for consultation about recommended file formats and specific requirements.[23] File format registries (e.g., PRONOM) are a way of helping to identify file formats and look up format specifications.[24]

We therefore consider the following formats as appropriate for the Submission Information Package (SIP):[25]

[23] The National Archives (2008): Selecting File Formats for Long-Term Digital Preservation: http://www.dptp.org/wp-content/uploads/2010/08/selecting-file-formats.pdf,
[24] The Technical Registry PRONOM: http://www.nationalarchives.gov.uk/PRONOM/Default.aspx
[25] The selected file formats represent a summary of different recommendations from CESSDA partners and internationally recognised institutions. FORS: Qualitative Data Archiving at FORS – Policy and Procedures: http://forscenter.ch/wp-content/uploads/2013/11/FORS_policy_and_procedures_on_qualitative_data1.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. The formats highlighted in bold are preferred over the others of the same category.

- 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 Open Office 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, such as **7-Zip** or **Winzip**.

3.2. Data deposit

In this section, the documenting of data by the researcher is explained in further detail. Data documentation explains how data were collected (context of data collection), what they mean, what are their content and structure, and specifies any manipulations that may have taken place.[26] Good documentation should be considered best practice when creating, organising, and managing data, and is equally important for data preservation.

- Data files must be clean and clear, i.e., interpretable and searchable by ways of a codebook by other researchers;
- The data should be accompanied by adequate documentation that explains the context of data collection (history, aim, and hypotheses), methodology, and any material relevant to the data. This includes not only the general information about the study itself (metadata) compatible with international standards (DDI, in order to share this metadata) but also to give as much material as possible in order to make the study as suitable as possible for secondary use (see chapter 2). This kind of accompanying material could include methodological reports, codebooks, questionnaires, coding instructions, interviewer guides, database dictionaries, bibliographies of publications related to the data, and links to online tools.

The quality of documentation can be significantly improved if its creation and collation is planned at the beginning of the data life cycle, during the project conception phase. If needed, archive staff will offer assistance in this respect.[27]

Generally speaking, the following minimal information should be made available when depositing the data at USDAS:

- Contact Information (depositor[28]): name, address, E-mail, telephone;

[26] Manipulation could involve either anonymisation, cleaning out of range values or quality assurance procedures carried out by the data producer before deposit of the data.
[27] The following manual on managing and sharing research data by the UK Data Archive is considered to be best practice: UK Data Archive (2009): Managing and Sharing Data – Best Practice for Researchers: http://www.data-archive.ac.uk/media/2894/managingsharing.pdf

- Study information: title, access level, principal investigator or data creator, data collectors, funding information;
- Accompanying documentation: study reports, questionnaires, codebook, user guides, methodology reports, publications (bibliographical reference);
- Dataset information (title, temporal coverage of collection, time dimension, method of collection, main topics, subject categories, number of variables and cases, sampling procedures, response rate, URL, etc.).

The **data deposit agreement** (see Annex A1.1: Data Deposit Agreement USDAS) describes the legal framework under which material is deposited and states the rights and responsibilities of both parties – the depositor and the archive – regarding copyright and ownership of the data and the access conditions. It is a legal agreement between the depositor and the archive that covers arrangements regarding usage rights, authenticity, data protection responsibilities, and disposal.

The following additional questions should equally be answered in advance[29]:

- How do the data come into the data archive (SIP)?[30] By means of CD-ROM, Ftp, or e-mail?
- What is the level of complexity of the study data and documentation deposited?
- Will there be an embargo or special access conditions?

The data producer can during the negotiation phase place the data under an **embargo** or special conditions, which means that data are not accessible to data users for a pre-defined period or only with certain restrictions. However, in the spirit of Open Access we highly recommended that data are made available to users as soon as possible after a research project has come to an end. The data depositor can also impose certain conditions on the use of the data (e.g., to be informed beforehand about who is going to use the data and for what purposes).

4. Ingest

Ingest is the first functional component of the OAIS reference model. It includes the set of processes responsible for accepting the information submitted by data producers (Submission Information Package, SIP) and preparing it for the archival storage. Some of these processes include:

- Validation that information received is uncorrupted and complete;
- Transformation of the submitted information into a form suitable for storage;
- Extraction and/or creation of descriptive metadata.

The tool used for ingest will be EPrints, customised by the SERSCIDA team (see SERSCIDA D5.3).

[28] Person(s) depositing the data; usually, but not always represents the licence signatory.
[29] GESIS: Archive agreement: http://www.gesis.org/fileadmin/upload/institut/wiss_arbeitsbereiche/datenarchiv_analyse/Archivierungsvertrag_GESIS_Datenarchiv_v9__englisch.pdf; Licence agreement UKDA: http://ukdataservice.ac.uk/media/28102/licenceform.pdf
[30] SIP should be a machine readable dataset, should have a codebook for raw data, and should include the survey instrument/measurement instrument (originals or copies, in a digital format if possible) and a methodology report, if available. In addition to that there should be determined beforehand how the SIP is transmitted to the archive: by e-mail, download from/upload to server, etc.

At a minimum, all data will be treated and described at study level. However, some data that are of higher analytic potential could receive more attention and processing. USDAS will use Nesstar, used by many CESSDA members to publish and disseminate variable-level metadata for high profile studies.[31]

4.1. Ingest workflow

Reception of SIP: Data arrives at USDAS

Most commonly, the data files will arrive at USDAS by e-mail (possibly encrypted) or through the ingest tool EPrints. If the SIP is submitted to USDAS by e-mail or from any other device, the archive staff subsequently needs to enter the SIP into EPrints. Ideally, the data deposit agreement should be in place before the data actually arrive at the archive.

Once reception of the SIP is approved, the following validation and quality assurance procedures are carried out:

- Review of completeness of the study data, metadata, and documentation (if needed, checked against data deposit agreement);
- Checks if the scope of documentation is sufficient (e.g., are variables described and can easily be attributed to the questions in the questionnaire);
- Comparison of data sets and data documentation to ensure that there is a match between variables, variable labels, and value labels, and that the number of observations and variables match the stated numbers in the documentation;
- Checks for anonymisation strategies if data are sensitive (proper name, postal code, profession, etc.).

Validation and quality assurance

Most of the following tasks are automatically performed by the system and corresponding tools:

- Scan and check for viruses
- Validation of the successful transfer of the SIP into EPrints
- Identify file formats
- Validate file formats (if files are what they pretend to be)
- Create checksums for each data and document file in order to guarantee data integrity during the transfer process
- Generate Unique Identifiers (UI)

Manual tasks conducted in addition by USDAS data experts include:

- Checks to see if confidentiality of data is guaranteed; and
- Manual/visual quality control assurance using user-friendly GUI interface of EPrints.

[31] The Nesstar Publisher tool is a data and metadata authoring tool that is used to prepare and publish survey data for the Nesstar system. The metadata inserted is consistent with the DDI (Data Document Initiative). The Nesstar Server tool provides the ability to publish statistical information that can be searched, browsed, analysed, and downloaded by users. (www.nesstar.com)

Normalisation : SIP is converted into a AIP

Data processing:

- The originally transmitted data files (SIP) are stored in a separate location;
- Renaming and rearranging of the data files;
- Decision of description level – study or variable;
- Metadata extraction (technical, administrational, structural, and preservation metadata);
- Integrity verification that data integrity has been fully maintained during the transfer process by checksum creation.

Documentation processing:

- Renaming and organisation of documentation files;
- Conversion of files to archival format and marking of material (e.g., bookmarking and adding headers and table of contents);
- Combination of files into one;
- Extraction of descriptive metadata, in other words, cataloguing of those sources for future dissemination (DIP)
- Creation of persistent identifiers.

The data and documentation will be registered and stored on the server located at USDAS or the University Teleinformatic Centre (UTIC), or outsourced in the initial phase of establishment.[32] The ticketing tool RequestTracker is used to record all actions and decisions made concerning the submitted material (see chapter 3.2.2 of SERSCIDA D5.3).

Pre-release procedure:

- Inform the data depositor to check the processed data files and documentation and verify with staff colleagues (quality control);
- All working copies should be deleted from the individual network area and any processing server in order to guarantee data security.

Ingest Workflow	Manually	Automatically
Receive Submission Information Package (SIP)	completeness of the study data, metadata and documentation	-
	Check scope of documentation	
	Comparison of data with documentation	
	Check anonymisation	
Quality Assurance	Check if confidentiality of data is guaranteed	Virus checking
	Perform manual/visual quality control assurance.	Validation of successful transfer of SIP into ingest tool
		Identification and validation of file formats
		Checksums creation
		Generation of Unique

[32] See „Data backup and security“ in chapter 5.

			Identifiers (UI
Generate Archival Information Package (AIP)	**Data Processing**	**Documentation processing**	-
	Rearranging of data	Reorganisation of documentation	*Extraction of metadata (technical, administrational, structural, and preservation)*
	Decide description level	-	*Conversion of files to archival format*
	-	Marking of material	
	-	Extraction of descriptive metadata	
	-	Persistent identifier (PI)	
	Extraction of metadata (technical, administrational, structural, and preservation)		
	Conversion of files to archival format		
	Separate AIP storage from SIP		
Pre-release procedure	Inform data depositor and verify with staff colleagues (quality control)		Capture correspondence (actions and decisions) between depositor and archive
	Delete working copies from individual network area and processing server		

Table 1: Ingest Workflow according to OAIS for USDAS

5. Archival Storage

The second functional component of OAIS is archival storage. The purpose is to ensure that what is added to the archive from the ingest process remains identical and accessible. More specifically, this component manages the storage and maintenance of digital materials (AIP) so that the preserved information remains complete and readable over the long-term.

The primary goal of USDAS's preservation policy is therefore to ensure the longstanding accessibility of the digital material, ensuring the highest level of authenticity of any formats. However, preservation decisions at USDAS will always be made within the context of its collections development policy, balancing the constraints of cost, research and historical value, and accessibility, alongside the requirements of levels of authenticity and legal admissibility.

5.1. Preservation policy

Storage and Migration

The aims of the preservation policy are to:

- Become a trusted digital repository in order to serve a Designated user Community[33];
- Retain authenticity, integrity and reliability of preserved datasets;
- Ensure that all data collections are protected and safely stored;
- Ensure that the relevant level of information security is applied to each data collection;

[33] See "Audit and Certification" in chapter 2.

- Apply good practice in preservation management.

USDAS will follow established best practices for managing datasets over the long-term. It assumes responsibility for the preservation and accessibility of the digital material in its collection.

Preservation in a digital archive usually means preserving the significant properties or the essential characteristics of data for these to maintain their meaning over time. Any changes (by machine/application or an individual) in the preservation process should be properly documented (either as a special document or within the archiving application).

USDAS follows international best practice in its choice of preservation formats and has listed them below. Preservation strategies are expected to consist of migration of the data files (refreshment, replication, repackaging or transformation), that is, data transferred into new data formats across several technology cycles of the data files.[34]

Preservation procedures will consist of:

- Reformatting datasets from proprietary software to software-independent formats and creating an Archival Information Package (AIP);
- Defining, timing, testing and implementing migration pathways;

When new formats are created from data files either through migration into new file formats or through the creation of new file formats for dissemination, the earlier files will be retained alongside. The creation of a metadata record for each dataset will record preservation tasks, checksums for authenticity and integrity, and logs and reports of all refreshment results.

Formats 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), Rich Text Format (RTF) (.rtf), 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).

Formats for the DIP:

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

[34] Other strategies, such as emulation (imitating hard- and software elements) or computer museum (keeping hard- and software elements) are highly time consuming and therefore not taken into account.

Data backup and security

It should be clear that data backup is a short-term data recovery solution following loss and corruption, and is fundamentally different from the preservation planning strategy of USDAS. Backing up data means making copies of files so that in case of data loss the original data can be restored from the copies. Regular backups protect against accidental data loss due to hardware, software, or media failure; failure of the power supply; human error; and fires, floods, earthquakes, and other natural disasters. Backup strategies and incident management plans are an essential part of USDAS' archival policies; however they are not sufficient to ensure survival and access to authentic digital resources over time.[35]

A reliable technical infrastructure should be in place with suitable measures to ensure the long-term usability of data collections (security and risk management). At USDAS the University Teleinformatic Centre is planned to provide this, or outsourcing for the initial phase of establishment. It is foreseen in future for USDAS to have its own infrastructure maintained by its staff. The digital preservation of the data collections must rely on a solid IT infrastructure.

USDAS will make sure that datasets are regularly backed up in accordance with established best practices. For backup and for security reasons, USDAS will aim to:

- Follow a policy of multiple copy resilience: more than one version of the complete preservation system will be maintained (e.g., main server, near-site online copy in another building, off-site online copy, disc-based offline copy=DVD, CD-R);
- Have secure computer networking systems and communications equipment;
- Ensure that all machine room computer systems are locked by a logon password system to prevent unauthorised access in case of a security breach of the rooms;

Once USDAS is officially established, a risk assessment will be conducted, in order to identify its critical functions (based on the impact of data loss or the inability of the archive to perform its duties).

6. Data Management

USDAS will follow best practice in data management. Data Management is the third functional component of an OAIS. The Data Management function involves the maintenance of all the archived information and files including the management of metadata, communications with producers and users, and access statistics. The primary functions of Data Management are carried out by archive staff and include maintaining the databases of metadata for which it is responsible, performing queries on these databases, and generating reports in response to requests from other functional components within the OAIS (Ingest, Administration, Access).

[35] http://www.jisc.ac.uk/media/documents/publications/digitalpreservationbp.pdf

7. Preservation Planning

The fourth functional component of an OAIS is Preservation Planning. This component provides the services and functions for monitoring the environment of the OAIS, providing recommendations and preservation plans to ensure that the information stored in the OAIS remains accessible to, and understandable by, the Designated Community over the long-term, even if the original computing environment becomes obsolete. Digital data are always linked to its support, which itself is subject to rapid technical change. In other words, all file formats and physical storage media will become obsolete at one time or another.

Preservation Planning functions include evaluating the contents of the archive and periodically recommending archival information updates, recommending the migration of current particular archive holdings, developing recommendations for the archive's standards and policies, providing periodic risk analysis reports, and monitoring changes in the technological environment and within the community.

USDAS will try to ensure that it is at the leading edge of technical advances by taking a strategic approach to long-term digital preservation, and by monitoring hardware and software developments and migrating its collections accordingly. The technical development of storage technology must be under constant observation and review: monitoring technological changes that will affect preservation and migration decisions, files suitable for archiving, the structure and design of the information packages, and hardware (server, storage media) and software (applications).

USDAS recognises the need for communication with other stakeholders who are active in formulating national preservation policies and standards. It also acknowledges the need to participate in national and international activities and programmes in the area of digital preservation.

8. Access

The fifth functional component of OAIS relates to access. This component provides the services and functions that support data users in determining the existence, description, location, and availability of information stored in the data archive, allowing them to request and receive data and documentation. To be more precise, as part of this functional entity, the Dissemination Information Packages (DIP) are generated through addition of descriptive metadata, and the dissemination request is processed accordingly. Access is also responsible for implementing any security or access control mechanisms associated with the archives content.

8.1. Dissemination policy

There is a tension between providing free and open access to data and at the same time protecting confidentiality of research participants and the depositor. However, anonymisation strategies on the one side, and specifications in the end user license on the other, are ways to give consideration to both aspects. The fundamental premise is that no individual should be harmed.

Datasets and documentation held by USDAS will be publicly available within certain limitations and controlled access:

- Access to data is limited to users and individuals affiliated with a research institution such as faculty or students (for non-commercial purposes);

- Access to data could be restricted by IP address range or University network;
- If there is the need, a secure non-networked server for the storage of highly confidential datasets will be maintained;
- Preliminary registration into the system is requested to gain access to data (the registration will be done by the users or by archive staff);
- All users will have to agree to study-specific usage conditions before they access data;
- Access to data will be provided by e-mail, direct download link or ftp-server system.

Access to the collections of USDAS will be through the data catalogue (Nesstar) accessible on the institutional website: data.ba.serscida.eu. The data and metadata of USDAS are also harvested and made available in the CESSDA Catalogue, which provides seamless access to datasets from social science data archives across Europe.[36]

8.2. Access management strategy

Access to data at USDAS will be controlled depending on the level of sensitivity of the data and the specific conditions determined by the depositor. The latter are set out in the **end user license** (see Annex A2.1: End User Licence USDAS), a version of which all users must sign before being given access to any dataset. The End User Licence describes the legal framework under which the material is distributed. It specifies the conditions under which a user is granted access to data and how they may use those data (citation in scientific fashion and use following typical scientific ethical norms of conduct). According to the sensitivity of data or due to additional access conditions negotiated with the data producer, special conditions and licences might apply.

Accessibility categories

There will be three levels of access, depending on the nature of the data:

1) Open data: Data and documents that can be **accessed by anyone** by way of the archive dissemination portal; No registration is needed, but data may be subject to an open data licensing (such as Creative Commons Licence, CC);

2) Restrictive data:
a) Standard access: Data and documents that are fully anonymised and available for scientific analysis carried out in academic research and teaching, and may be **accessed without impending permission and special conditions from the data depositor,** i.e. free access via an ordering system by e-mail or online. However, registration is necessary. Users who register must have accepted the end user licence of USDAS;

b) Special conditions: Data that are released for academic research and teaching, where the user may gain access to data **subject to written approval from the depositor.** For this purpose the user must supply the archive with their details and specify their analytic intention. The Data Centre will request written permission from the depositor and will only permit access on receipt of this. In any case, USDAS informs the depositor either before or after data distribution to the user depending on their agreement. By default, the user must sign the end user licence in order to gain access to the material;

[36] The CESSDA Catalogue: http://www.cessda.net/catalogue/

c) Special license: Data that contain more detailed information and are usually not fully anonymised require users to complete a set of Special Licence forms;

3) Controlled data: USDAS reserves the right to permit access to data only through a physical or virtual secure environment or after special training. This applies to data that are considered too confidential or sensitive to be released via download.

In certain circumstances where **studies are not deposited at USDAS** but only described in the catalogue (metadata), no contract between the user and the archive is required. In this case, the research material might either be available through external websites (projects website), or the user can find addition information by directly contacting the producer.

9. Administration

Finally, USDAS will administer its digital data archive in alignment with international best practice. Administration is the sixth and final functional component of an OAIS. This component is responsible for managing the day-to-day operations of the OAIS, as well as coordinating the activities of the other five high-level OAIS services. Other responsibilities include interacting with data producers (e.g., negotiating submission agreements), data users (e.g., providing customer service support), and the management (e.g., implementing and maintaining archive policies and standards).

USDAS will use the tool RT: Request Tracker for the registration of data users (for more details and technical specifications see SERSCIDA D5.3 chapter 3.3 and 3.7).

In order to outline the financial requirements upon establishment of USDAS, a separate document has been created indicating the key variables and items of expenditure over a period of five years (see Annex A3.1: Finance and budget USDAS). In addition, as part of the administrative responsibilities of the newly established data service, an internal performance monitoring system based on a conceptual framework that links overall institutional objectives to key activities, and then to specific indicators has been created (see Annex A4.1: Performance indicators USDAS)

II Serbian Social Sciences Data Centre – SR-SSDC

1. Contextualisation and organisation of document

The aim of the F7 project SERSCIDA is the establishment of a prototype for social science digital data archives[37] in the Western Balkans, which could go into operation specifically in the countries of Bosnia and Herzegovina, Croatia, and Serbia. This includes the preparation of relevant documents for the formation of new data services. The present document "Data Policies and Procedures" was created as part of Work Package 5 (WP5) of the SERSCIDA project.

The present report will follow in large part the functionalities of the international standard *Open Archival Information System (OAIS) Reference Model* (explained in chapter 2), to which the Serbian Social Sciences Data Centre (SR-SSDC) will adhere in order to develop its digital preservation policies and strategies. This approach aligns with policies and practices of members of the Consortium of European Social Science Data Archives CESSDA, and with the SERSCIDA deliverables from prior Work Packages, as well as the common standards of the international instructure CESSDA AS (Annex 2.7).[38]

The target audience of this document includes federal ministries, stakeholders of SR-SSDC, its staff, the social science research community in Serbia, and the general public. Beyond its purpose as a deliverable for the SERSCIDA project, this policy document could be used for establishing new data archive centres in other countries.

2. Conceptual models and guiding principles

2.1. The OAIS Reference Model

A very important concept for the establishment of any digital repository is the Reference Model for an Open Archival Information System (OAIS)[39], adopted as an ISO standard in 2003 (ISO 14721:2003 OAIS). OAIS provides a framework for the understanding of digital archives and provides common terminology to be used in this context. The OAIS reference model is flexible enough to use in a wide variety of environments since it is strictly a logical model which does not depend on a particular implementation. It describes the components and services required to develop and maintain archives, in order to support long-term access to and understanding of the information in those archives. The OAIS model specifies how digital material should be preserved for a community of users (Designated Community)[40] from the moment digital material is ingested into the digital storage area, through subsequent preservation strategies, to the dissemination of digital material for the end user.

An OAIS-type archive carries out its preservation and access responsibilities in an environment populated by three key external stakeholders: 1) the Management (archivist, system administrator), which formulates, revises, and enforces a high-level policy framework that covers the OAIS activities; 2) the Producer (data producer, depositor, or researcher), that is, the individuals, organizations, or

[37] We use "data archive" and "data service" interchangeably.
[38] Annexes to Statutes for CESSDA, No. 1-6 (18 June 2013)
[39] Reference Model for an Open Archival Information System (OAIS); CCSDS 650.0-M-2; Consultative Committee for Space Data Systems: Washington, DC, 2012.
[40] The Designated Community is a special class of Consumers; the subset of Consumers expected to independently understand the archived information in the form in which it is preserved and made available by the OAIS.

systems that create digital objects and/or transfer information to the OAIS for long-term preservation; and 3) the Consumer (data consumer, user, researcher) - the individuals, organizations, or systems expected to use the information preserved by the OAIS.[41]

OAIS Environment

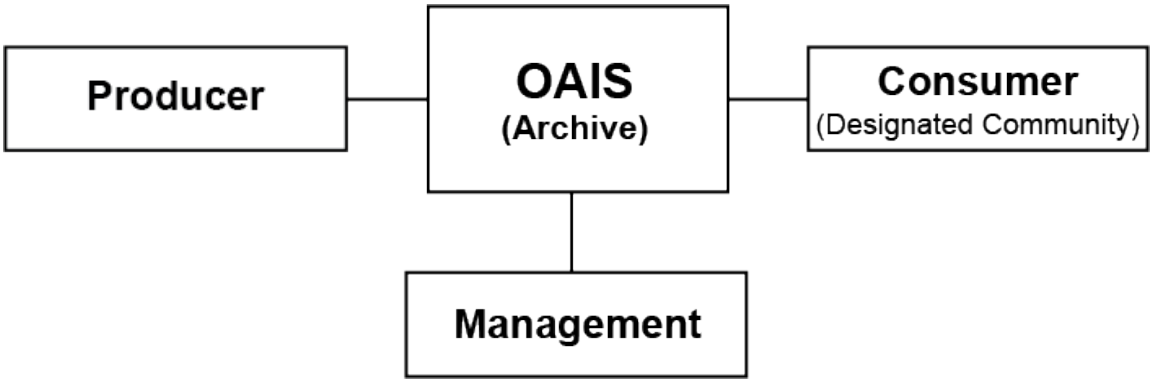
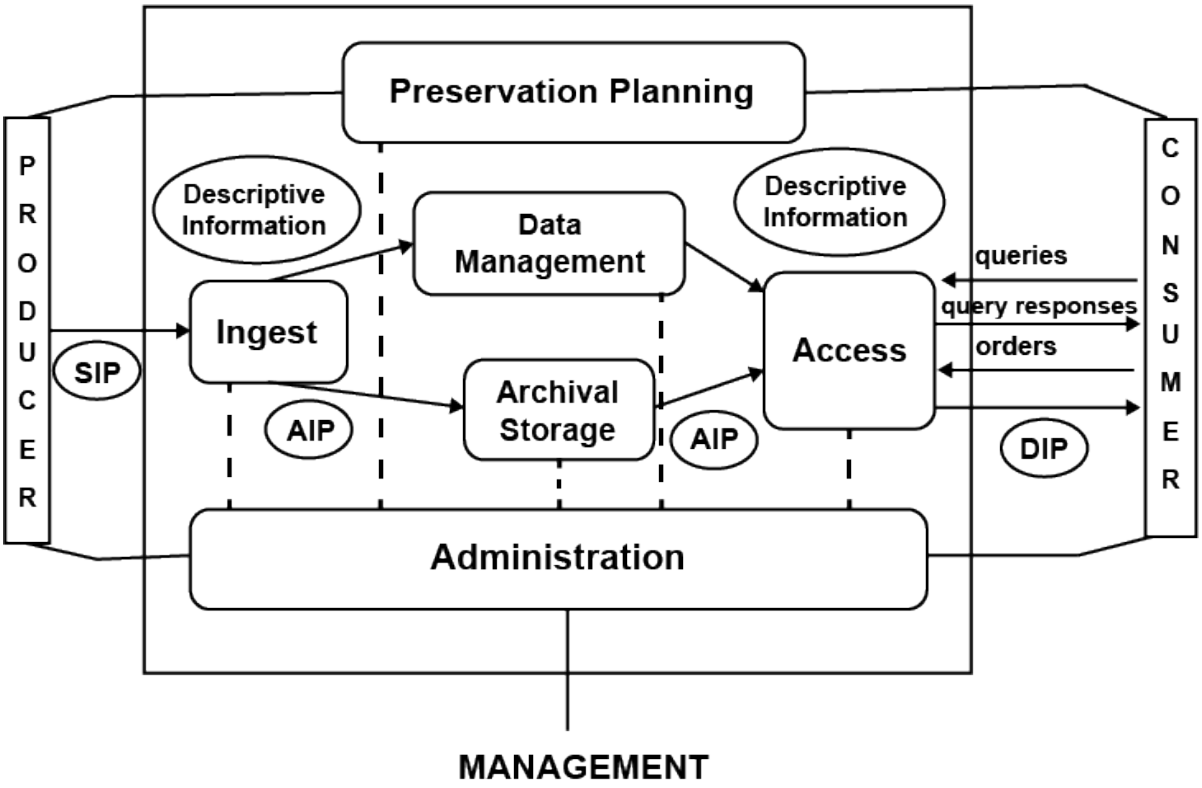


Figure 7: Reference Model for an Open Archival Information System (OAIS); CCSDS 650.0-M-2; Consultative Committee for Space Data Systems: Washington, DC, 2002, p. 2-2.

OAIS lists six functional entities that fulfil OAIS' dual role of preserving and providing access to the information: Ingest, Archival Storage, Data Management, Preservation Planning, Access, and Administration.



[41] For the sake of convenience, for the rest of this document we will generally use terminology more common to the social science archiving community in Europe (e.g., data producers, data users), rather than the OAIS terminology.

Figure 8: Reference Model for an Open Archival Information System (OAIS); CCSDS 650.0-M-2; Consultative Committee for Space Data Systems: Washington, DC, 2002, p. 4-1

An essential part of the OAIS model is the Information Package (IP), which represents the structure of information as it moves into, through, and out of the archival system. There are three types of information packages, all of them consisting of the digital object to be preserved and the metadata necessary to support its long-term preservation and access:

- SIP (Submission Information Package): information package that is transferred from the Producer to the Archive.
- AIP (Archival Information Package): information package that is stored and preserved by the Archive.
- DIP (Dissemination Information Package): information package delivered to the Consumer in response to an access request.

There are several minimum responsibilities that should be maintained by an OAIS[42]:

- Negotiate for and accept appropriate information from Producers;
- Obtain sufficient control of the information provided to the level needed to ensure long-term preservation;
- Determine, either by itself or in conjunction with other parties, which communities should become the Designated Community and, therefore, should be able to understand the information provided, thereby defining its Knowledge Base[43];
- Ensure that the information to be preserved is independently understandable to the Designated Community. In particular, the Designated Community should be able to understand the information without needing special resources such as the assistance of the experts who produced the information;
- Follow documented policies and procedures which ensure that the information is preserved against all reasonable contingencies, including the demise of the archive, ensuring that it is never deleted unless allowed as part of an approved strategy;
- Make the preserved information available to the Designated Community and enable the information to be disseminated as copies of, or as traceable to, the original submitted Data Objects with evidence supporting its Authenticity[44].

2.2. The DCC Curation Lifecycle Model: metadata and standards

[42] CCSDS 650.0-M-2, p. 3-1

[43] "A person, or system, can be said to have a Knowledge Base, which allows that person or system to understand received information. For example, a person who has a Knowledge Base that includes an understanding of English will be able to read and understand an English text." (CCSDS 650.0-M-2, p. 2-3)

[44] Authenticity: relies on three provable properties: the record is what it purports to be, that it was created by whomsoever it is purported to have been created by; and that it was created at the time when it is purported to have been created. (Joint Information Systems Committee (JISC) Digital Preservation: continued access to authentic digital assets (Briefing Paper, November 2006), p.2). Other concepts designed to guide policies for information security within an organization are confidentiality (ensuring that data are not made available or disclosed to unauthorised individuals) and integrity (ensuring that data are not altered or destroyed in an unauthorized manner).

The Data Centre will follow all the phases of the DCC **Curation Lifecycle Model**[45]. Data often have a longer lifespan than the research project that creates them. The DCC Curation Lifecycle Model can be used to plan activities within an organisation to ensure that all necessary phases are undertaken, each in the correct sequence.

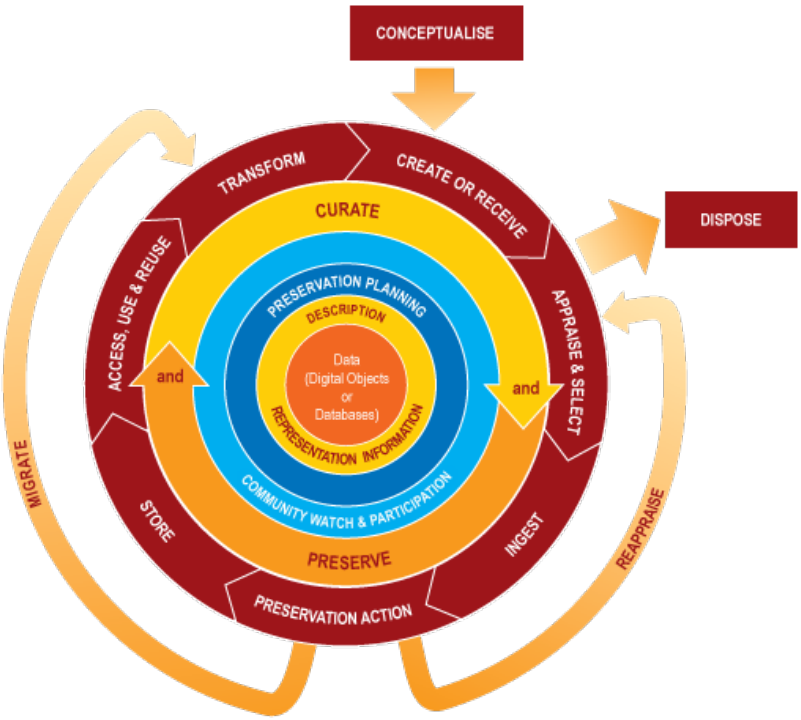


Figure 9: The DCC Curation Lifecycle Model

The collaboration between the data archive and the data producer should be a priority during the whole data curation process. It is important that the data producers are working with SR-SSDC during the documentation process, since they know their data best and are most familiar with the study methodology. On the other hand the Data Centre has the best knowledge about data documentation procedures and metadata standards. In addition, the data producers are usually responsible for the documentation, metadata description and the data themselves (data preparation, anonymisation, and cleaning). As a basic principle, it should be made clear throughout the whole archival process who bears responsibility for which tasks (the data producer or the archive).

Standardisation is a crucial element in the establishment of social science data archives. SR-SSDC is devoted to ensuring best practice and standardised approaches in order to enable compatibility and interoperability for international data sharing.

SR-SSDC orients itself towards international standards adopted by CESSDA AS and its service providers so that our data can be available and sharable throughout the international community. Regular reviews will be undertaken in order to guarantee best practice.

[45] DCC Curation Lifecycle Model: http://www.dcc.ac.uk/resources/curation-lifecycle-model

At SR-SSDC we will be compliant with the **Data Documentation Initiative (DDI)** metadata specification, version 2.5 or higher[46]. To be more precise, SR-SSDC will create the following types of metadata generated from accompanying documentation[47]:

- **Descriptive metadata**: Describes a resource for purposes such as discovery and identification (contains elements such as title, abstract, author, and keywords). DDI complies with Dublin Core Metadata Initiative which describes a core set of 15 elements intended to facilitate discovery of electronic resources;[48]
- **Administrative metadata:** Contains information about the use, management, and encoding processes of digital objects over time (e.g. information about data creation)
- **Technical metadata**: Provides information about the overall system environment and provides the technical information needed to use data (e.g. file format, application used and operation system)
- **Structural metadata**: Describes the logical structure of a multidimensional object. The Metadata Encoding and Transmission Standard (METS) will be used where appropriate;[49]
- **Preservation metadata**: Provides information needed to archive and preserve a resource. The PREMIS Data Dictionary for Preservation Metadata defines a core set of preservation metadata elements and describes relationships between digital preservation entities: Intellectual entity, Object, Event, Agent, and Rights.[50]

More and more digital objects require more reliable and stable references (regardless of their location and across system change). SR-SSDC will therefore aim to assign a persistent identifier (DOI (Digital Object Identifier)/URN (Uniform Resource Name) to every study coming into the archive.

The following two international archival standards could be used as well:

- **ISAD(G)**: International Standard Archival Description (General) or ISAD(G) is an international standard which provides guidelines for creating the content of an archival description (descriptive metadata). It promotes the creation of consistent and appropriate descriptions, facilitating the retrieval and exchange of information, and the integration of descriptions into a unified information system. There is a list of elements which are considered mandatory (six elements) when writing an archival description.[51] Based on ISAD(G) is the **Encoded Archival Description (EAD)** standard;
- **ISAAR(CPF)**: is the International Standard Archival Authority Record For Corporate Bodies, Persons and Families standardised by the International Council on Archives in 2003.[52] Whereas ISAD(G) describes the archival material itself, ISAAR(CPF) is constructed to give

[46] Data Documentation Initiative (DDI): http://www.ddialliance.org/
[47] As a general remark, metadata standards should be machine readable (XML coded) in order to be processed automatically. XML is an open, well-supported, and widely adopted standard for encoding textual data, designed to be used regardless of the hardware and software environment.
[48] Dublin Core Metadata Initiative: http://dublincore.org/
[49] Metadata Encoding and Transmission Standard (METS) http://www.loc.gov/standards/mets/
[50] PREMIS Data Dictionary for Preservation Metadata: http://www.loc.gov/standards/premis/v2/premis-2-0.pdf
[51] General International Standard Archival Description, ISAD (G): http://www.ica.org/10207/standards/isadg-general-international-standard-archival-description-second-edition.html
[52] International Standard Archival Authority Record For Corporate Bodies, Persons and Families, ISAAR (CPF) http://www.ica.org/10203/standards/isaar-cpf-international-standard-archival-authority-record-for-corporate-bodies-persons-and-families-2nd-edition.html

information about the creator (authority) of the archival material with additional information on history, status, functions, structure and relations – generally speaking about the provenance of the data. Again, there has been another international standard called **Encoded Archival Context (EAC),** which is derived from ISAAR(CPF).

Within this regulatory framework, SR-SSDC will also respect the following **national and international legal frameworks**:

- EU Copyright Directive (EU, 2001);
- Law on Scientific Research ("RS Official Gazette", no.: 110/05, 50/06) -> Law on Changes and Amendments of Law on Scientific Research ("RS Official Gazette", no: 18/10) : Strategy and programs of general interest for Serbia, criteria ensuring quality of scientific-research research activity , condition for funding and financing of the activity;
- Law on Archival Material ("FRY Official Gazette", no. 12/98, 13/98);
- Law on Copyrights and Related Rights ("RS Official Gazette", no. 104/2009);
- Law on Cultural Property ("RS Official Gazette", no. 71/94); and
- Law on Personal Data Protection ("RS Official Gazette", no. 97/2008, 104/2009, 68/2012, Decision of Constitutional Court, 107/2012.

Publicly funded research data are a public good produced in the public interest, and therefore should be openly available to the maximum extent possible (OECD principle).[53] SR-SSDC thus promotes the philosophy and practice of **Open Access,** where research data are as easily accessible as possible, while still protecting the confidentiality of respondent information and the rights of those who gathered or created the data. In addition, the basic data service of SR-SSDC for all data producers and users will as a rule be free of charge. At the same time, SR-SSDC will not pay for the data it acquires.

Audit and Certification: As a future goal, we will seek to become compliant with at least one of the following standards of trust: Data Seal of Approval (DSA), ISO 16363 or DIN 31644 for the Audit and Certification of Trustworthy Digital Repositories.[54] Since one of the priorities of SR-SSDC is to become a CESSDA member, we seek to be compliant with the DSA seal which is granted to repositories that are committed to archiving and providing access to scholarly research data in a sustainable way. From the beginning, we will focus on quality control mechanisms throughout the whole archiving process.

Confidentiality and disclosure: All depositors should ensure that data meet the requirements of confidentiality and non-disclosure for data collected from human subjects. Datasets should not contain direct identifiers (names, addresses, telephone numbers, etc.) or indirect identifiers (which, when combined with other variables, may provide enough details to identify an individual). In-house confidentiality checks by the Data Centre will help to guarantee that confidentiality requirements are met.

[53] OECD (2007) OECD principles and guidelines for access to research data from public funding. Organisation for Economic Cooperation and Development. Downloaded: http://www.oecd.org/science/sci-tech/38500813.pdf

[54] Data Seal of Approval: http://www.datasealofapproval.org; Audit and certification of trustworthy digital repositories: http://www.iso16363.org/; Nestor Seal for Trustworthy Digital Archives: http://www.langzeitarchivierung.de/Subsites/nestor/EN/nestor-Siegel/siegel_node.html.

3. Acquisition and deposit

This section focuses on the acquisition of data and covers data selection and appraisal strategies that will be adopted, as well as the file formats accepted by SR-SSDC. The collection development policy will answer the questions of what, from whom, why, and how are we going to collect data in the Data Centre. In a next step, policies for the deposit of the data and related documentation materials are discussed. This includes the conditions for depositing data, data quality, confidentiality of data, and depositor rights. Although these elements are not explicitly treated within the OAIS Reference Model, they are worth treatment in this policy document, since they are important aspects of the proper establishment of a social science data archive.

3.1. Collection development policy[55]

Data selection and appraisal play an important role in the acquisition of data in any archival setting. The collection development policy indicates the principles and criteria by which the archive develops its data collection in order to serve the Designated Community. The collection development policy of SR-SSDC will be flexible and respond to future developments and shifting requirements that will influence the archive's data collections (technology, scientific standards, etc.).

SR-SSDC collects data and information across the range of social science disciplines and the humanities[56]. This includes **quantitative** and **qualitative data**.

WHAT can be accepted:

- Research that gives evidence and greater understanding of Serbian society or parts of it (including international surveys);
- Research data where Serbian researchers are involved;
- Comparative or continuous research (panel data, longitudinal surveys, and time series that allow examination of trends);
- Data produced with methodological excellence;
- Works of general interest for social science research (quality and exemplary studies with analytic potential).

Time series and panel data are of particular interest. The deposited data will be retained for the long-term, unless otherwise specified by the depositor, funding agencies, or by government regulations. SR-SSDC will accept data and documentation in paper form only under exceptional circumstances. [57]

FROM whom are data solicited and accepted:

[55] The content of this policy draws on the collection development policy of the UKDA: http://www.data-archive.ac.uk/media/54773/ukda067-rms-collectionsdevelopmentpolicy.pdf, and the experience of setting up new Data Archives in Finland and Slovenia (Laaksonen, H. et al, 2006): http://www.iassistdata.org/downloads/iqvol301laaksonen_0.pdf

[56] The following disciplines are embraced: Economics; Education; Employment and Labour; Environment, Conservation and Land Use; Health; History; Industry and Management; Law, Crime and Legal Systems; Media, Communication and Language; Political Science; Psychology; Sociology, Society and Culture; Social Welfare Policy and Systems.

[57] Studies in paper form will be accepted and worked on under separately funded digitisation projects, rather than as 'routine' activity. Accepting such material is fine, but they require careful storage until one has the resources to digitise and process them (i.e., special funding to digitise some of the paper deposits).

Eligible depositors are researchers, members of research institutions, government departments, offices of national statistics, and public and commercial sector sources (e.g., marketing research firms). Depositors may only deposit their own work and will be expected to sign a data deposit agreement (see section "data deposit" below).

WHY certain data are accepted:

There are three main reasons why data are solicited and accepted. First, because they possess potential value for secondary use and analysis for research. Second, because they may serve teaching and learning purposes. And finally, for validation reasons - to make possible the replication of research for important studies.

HOW the data will be acquired:

The acquisition process will consist of three stages: 1) identifying potential datasets (localisation), 2) negotiation with the data creators, and 3) receiving the relevant material. A very important source will be the list of potential data producers and consumers that has been established during the SERSCIDA project.

As SR-SSDC will be in the early stages of its existence, the **acquisition strategies** will be highly proactive and persistent. This will include open calls, announcements at conferences, and direct solicitation. No national research inventory for social science research data, where staff can actively search for and proactively seek data for deposit, is currently available for this purpose in Serbia.[58]

Similar as in some other countries, collaboration in the process of preparing and depositing data is entirely voluntary. This means that special effort will need to be invested in the process of negotiating and gaining access to the data sets. This will include:

- Good **working relations** with stakeholders and potential depositors (Research institutes and faculties in social sciences, National Labour Service in Serbia, private research companies and individual researchers, National Statistical Institute of the Republic of Serbia) to identify and acquire new sources of data;
- **Links** with National/Regional Archives and Research Libraries;
- **Promotional activities** to raise awareness and encourage deposit of data collections (workshops, round tables, conferences, and training events);
- **Agreements** with funding bodies for the disciplines in question in order to discuss deposit of material with grant-holders;
- **Data discovery** activities by staff members in publications (academic journals), news media, conferences, e-mail lists, and websites (e.g., universities' online catalogues and websites of research funding bodies).

The aim is to establish links and working relations with the stakeholders and partners mentioned above in order to establish long-term cooperation and a common framework for the deposit of data collections.

SR-SSDC reserves the right to **refuse** to accept material under the following conditions:

[58] There is however a list of projects financed by our Ministry for science that is accessible to SR-SSDC. This might be developed into a nation-wide research inventory in the future.

- Data do not match the criteria of collection development policy and would be better dealt with at another institution (also legal conditions);
- Insufficient or poor quality documentation;
- Depositor and Data Centre do not agree on access and dissemination conditions;
- There are currently no restrictions on file sizes, or number of files or studies. However, SR-SSDC reserves the right to refuse to accept materials which are of a volume that may make it difficult or impossible to process, given the resources, staff, facilities, or capacities of the Data Centre.

In order to make the data and documentation accessible for other users and to be able to process and use them over time, they should be in a form that best ensures their longevity. The data and documentation **file formats** that are preferred at the ingest phase by the Data Centre are non-proprietary, openly documented, unencrypted, and uncompressed that are commonly used by the research community. Several lists are available online for consultation about recommended file formats and specific requirements.[59] File format registries (e.g. PRONOM) are a way of helping to identify file formats and look up format specifications.[60]

We therefore consider the following formats as appropriate for the Submission Information Package (SIP):[61]

- 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 Open Office 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).

3.2. Data deposit

In this section, the documenting of data by the researcher is explained in further detail. Data documentation explains how data were collected (context of data collection), what they mean, what

[59] The National Archives (2008): Selecting File Formats for Long-Term Digital Preservation: http://www.dptp.org/wp-content/uploads/2010/08/selecting-file-formats.pdf,

[60] The Technical Registry PRONOM: http://www.nationalarchives.gov.uk/PRONOM/Default.aspx

[61] The selected file formats represent a summary of different recommendations from CESSDA partners and internationally recognised institutions. FORS: Qualitative Data Archiving at FORS – Policy and Procedures: http://forscenter.ch/wp-content/uploads/2013/11/FORS_policy_and_procedures_on_qualitative_data1.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. The formats highlighted in bold are preferred over the others of the same category.

are their content and structure, and specifies any manipulations that may have taken place.[62] Good documentation should be considered best practice when creating, organising, and managing data, and is equally important for data preservation.

- Data files must be clean and clear, i.e., interpretable and searchable by ways of a codebook by other researchers;
- The data should be accompanied by adequate documentation that explains the context of data collection (history, aim, and hypotheses), methodology, and any material relevant to the data. This includes not only the general information about the study itself (metadata) compatible with international standards (DDI, in order to share this metadata) but also to give as much material as possible in order to make the study as suitable as possible for secondary use (see chapter 2). This kind of accompanying material could include methodological reports, codebooks, questionnaires, coding instructions, interviewer guides, database dictionaries, bibliographies of publications related to the data, and links to online tools.

The quality of documentation can be significantly improved if its creation and collation is planned at the beginning of the data life cycle, during the project conception phase. If needed, archive staff will offer assistance in this respect. [63]

Generally speaking, the following minimal information should be made available when depositing the data at SR-SSDC:

- Contact Information (depositor[64]): name, address, E-mail, telephone;
- Study information: title, access level, principal investigator or data creator, data collectors, funding information;
- Accompanying documentation: study reports, questionnaires, codebook, user guides, methodology reports, publications (bibliographical reference);
- Dataset information (title, temporal coverage of collection, time dimension, method of collection, main topics, subject categories, number of variables and cases, sampling procedures, response rate, URL, etc.).

The **data deposit agreement** (see Annex A1.2 Data Deposit Agreement SR-SSDC) describes the legal framework under which material is deposited and states the rights and responsibilities of both parties – the depositor and the archive – regarding copyright and ownership of the data and the access conditions. It is a legal agreement between the depositor and the archive that covers arrangements regarding usage rights, authenticity, data protection responsibilities, and disposal.

The following additional questions should equally be answered in advance[65]:

[62] Manipulation could involve either anonymisation, cleaning out of range values or quality assurance procedures carried out by the data producer before deposit of the data.
[63] The following manual on managing and sharing research data by the UK Data Archive is considered to be best practice: UK Data Archive (2009): Managing and Sharing Data – Best Practice for Researchers: http://www.data-archive.ac.uk/media/2894/managingsharing.pdf
[64] Person(s) depositing the data; usually, but not always the Licence signatory.

- How do the data come into the data archive (SIP)?[66] By means of CD-ROM, Ftp, or e-mail?
- What is the level of complexity of the study data and documentation deposited?
- Will there be an embargo or special access conditions?

The data producer can during the negotiation phase place the data under an **embargo** or special conditions, which means that data are not accessible to data users for a pre-defined period or only with certain restrictions. However, in the spirit of Open Access we highly recommended that data are made available to users as soon as possible after a research project has come to an end. The data depositor can also impose certain conditions on the use of the data (e.g., to be informed beforehand about who is going to use the data and for what purposes).

4. Ingest

Ingest is the first functional component of the OAIS reference model. It includes the set of processes responsible for accepting the information submitted by data producers (Submission Information Package, SIP) and preparing it for the archival storage. Some of these processes include:

- Validation that information received is uncorrupted and complete;
- Transformation of the submitted information into a form suitable for storage;
- Extraction and/or creation of descriptive metadata.

The tool used for ingest will be EPrints, customised by the SERSCIDA team (see SERSCIDA D5.3).

At a minimum, all data will be treated and described at study level. However, some data that are of higher analytic potential could receive more attention and processing. SR-SSDC will use Nesstar, used by many CESSDA members to publish and disseminate variable-level metadata for high profile studies.[67]

4.1. Ingest workflow

Reception of SIP: Data arrives at SR-SSDC

Most commonly, the data files will arrive at SR-SSDC by e-mail (possibly encrypted) or through the ingest tool EPrints. If the SIP is submitted to SR-SSDC by e-mail or from any other device, the archive staff subsequently needs to enter the SIP into EPrints. Ideally, the data deposit agreement should be in place before the data actually arrive at the archive.

[65] Archive agreement GESIS: http://www.gesis.org/fileadmin/upload/institut/wiss_arbeitsbereiche/datenarchiv_analyse/Archivierungsvertrag_GESIS_Datenarchiv_v9__englisch.pdf; Licence agreement UKDA: http://ukdataservice.ac.uk/media/28102/licenceform.pdf

[66] SIP should be a machine readable dataset, should have a codebook for raw data, and should include the survey instrument/measurement instrument (originals or copies, in a digital format if possible) and a methodology report, if available. In addition to that there should be determined beforehand how the SIP is transmitted to the archive: by e-mail, download from/upload to server, etc.

[67] The Nesstar Publisher tool is a data and metadata authoring tool that is used to prepare and publish survey data for the Nesstar system. The metadata inserted is consistent with the DDI (Data Document Initiative). The Nesstar Server tool provides the ability to publish statistical information that can be searched, browsed, analysed, and downloaded by users. (www.nesstar.com)

Once reception of the SIP is approved, the following validation and quality assurance procedures are carried out:

- Review of completeness of the study data, metadata, and documentation (if needed, checked against data deposit agreement);
- Checks if the scope of documentation is sufficient (e.g., are variables described and can easily be attributed to the questions in the questionnaire);
- Comparison of data sets and data documentation to ensure that there is a match between variables, variable labels, and value labels, and that the number of observations and variables match the stated numbers in the documentation;
- Checks for anonymisation strategies if data are sensitive (proper name, postal code, profession, etc.).

Validation and quality assurance

Most of the following tasks are automatically performed by the system and corresponding tools:

- Scan and check for viruses
- Validation of the successful transfer of the SIP into EPrints
- Identify file formats
- Validate file formats (if files are what they pretend to be)
- Create checksums for each data and document file in order to guarantee data integrity during the transfer process
- Generate Unique Identifiers (UI)

Manual tasks conducted in addition by SR-SSDC data experts include:

- Checks to see if confidentiality of data is guaranteed; and
- Manual/visual quality control assurance using user-friendly GUI interface of EPrints.

Normalisation : SIP is converted into a AIP

Data processing:

- The originally transmitted data files (SIP) are stored in a separate location;
- Renaming and rearranging of the data files;
- Decision of description level – study or variable;
- Metadata extraction (technical, administrational, structural, and preservation metadata);
- Integrity verification that data integrity has been fully maintained during the transfer process by checksum creation.

Documentation processing:

- Renaming and organisation of documentation files;
- Conversion of files to archival format and marking of material (e.g., bookmarking and adding headers and table of contents);
- Combination of files into one;

- Extraction of descriptive metadata, in other words, cataloguing of those sources for future dissemination (DIP)
- Creation of persistent identifiers.

The data and documentation will be registered and stored on the server located at SR-SSDC or the University Teleinformatic Centre (UTIC), or outsourced in the initial phase of establishment.[68] The ticketing tool RequestTracker is used to record all actions and decisions made concerning the submitted material (see chapter 3.2.2 of SERSCIDA D5.3).

Pre-release procedure:

- Inform the data depositor to check the processed data files and documentation and verify with staff colleagues (quality control);
- All working copies should be deleted from the individual network area and any processing server in order to guarantee data security.

Ingest Workflow	**Manually**		**Automatically**
Receive Submission Information Package (SIP)	completeness of the study data, metadata and documentation		-
	Check scope of documentation		
	Comparison of data with documentation		
	Check anonymisation		
Quality Assurance	Check if confidentiality of data is guaranteed		Virus checking
	Perform manual/visual quality control assurance.		Validation of successful transfer of SIP into ingest tool
			Identification and validation of file formats
			Checksums creation
			Generation of Unique Identifiers (UI
Generate Archival Information Package (AIP)	**Data Processing**	**Documentation processing**	-
	Rearranging of data	Reorganisation of documentation	*Extraction of metadata (technical, administrational, structural, and preservation)*
	Decide description level	-	*Conversion of files to archival format*
	-	Marking of material	
	-	Extraction of descriptive metadata	
	-	Persistent identifier (PI)	
	Extraction of metadata (technical, administrational, structural, and preservation)		

[68] See „Data backup and security“ in chapter 5.

	Conversion of files to archival format	
	Separate AIP storage from SIP	
Pre-release procedure	Inform data depositor and verify with staff colleagues (quality control)	Capture correspondence (actions and decisions) between depositor and archive
	Delete working copies from individual network area and processing server	

Table 2: Ingest Workflow according to OAIS for SR-SSDC

5. Archival Storage

The second functional component of OAIS is archival storage. The purpose is to ensure that what is added to the archive from the ingest process remains identical and accessible. More specifically, this component manages the storage and maintenance of digital materials (AIP) so that the preserved information remains complete and readable over the long-term.

The primary goal of SR-SSDC's preservation policy is therefore to ensure the longstanding accessibility of the digital material, ensuring the highest level of authenticity of any formats. However, preservation decisions at SR-SSDC will always be made within the context of its collections development policy, balancing the constraints of cost, research and historical value, and accessibility, alongside the requirements of levels of authenticity and legal admissibility.

5.1. Preservation policy

Storage and Migration

The aims of the preservation policy are to:

- Become a trusted digital repository in order to serve a Designated user Community[69];
- Retain authenticity, integrity and reliability of preserved datasets;
- Ensure that all data collections are protected and safely stored;
- Ensure that the relevant level of information security is applied to each data collection;
- Apply good practice in preservation management.

SR-SSDC will follow established best practices for managing datasets over the long-term. It assumes responsibility for the preservation and accessibility of the digital material in its collection.

Preservation in a digital archive usually means preserving the significant properties or the essential characteristics of data for these to maintain their meaning over time. Any changes (by machine/application or an individual) in the preservation process should be properly documented (either as a special document or within the archiving application).

The Data Centre follows international best practice in its choice of preservation formats and has listed them below. Preservation strategies are expected to consist of migration of the data files (refreshment, replication, repackaging or transformation), that is, data transferred into new data formats across several technology cycles of the data files.[70]

[69] See "Audit and Certification" in chapter 2.

[70] Other strategies, such as emulation (imitating hard- and software elements) or computer museum (keeping hard- and software elements) are highly time consuming and therefore not taken into account.

Preservation procedures will consist of:

- Reformatting datasets from proprietary software to software-independent formats and creating an Archival Information Package (AIP);
- Defining, timing, testing and implementing migration pathways;

When new formats are created from data files either through migration into new file formats or through the creation of new file formats for dissemination, the earlier files will be retained alongside. The creation of a metadata record for each dataset will record preservation tasks, checksums for authenticity and integrity, and logs and reports of all refreshment results.

Formats 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), Rich Text Format (RTF) (.rtf), 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).

Formats for the DIP:

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

Data backup and security

It should be clear that data backup is a short-term data recovery solution following loss and corruption, and is fundamentally different from the preservation planning strategy of the Data Centre. Backing up data means making copies of files so that in case of data loss the original data can be restored from the copies. Regular backups protect against accidental data loss due to hardware, software, or media failure; failure of the power supply; human error; and fires, floods, earthquakes, and other natural disasters. Backup strategies and incident management plans are an essential part of the Data Centre's archival policies; however they are not sufficient to ensure survival and access to authentic digital resources over time.[71]

A reliable technical infrastructure should be in place with suitable measures to ensure the long-term usability of data collections (security and risk management). The digital preservation of the data collections must rely on a solid IT infrastructure.

[71] http://www.jisc.ac.uk/media/documents/publications/digitalpreservationbp.pdf

SR-SSDC will make sure that datasets are regularly backed up in accordance with established best practices. For backup and for security reasons, SR-SSDC will aim to:

- Follow a policy of multiple copy resilience: more than one version of the complete preservation system will be maintained (e.g., main server, near-site online copy in another building, off-site online copy, disc-based offline copy=DVD, CD-R);
- Have secure computer networking systems and communications equipment;
- Ensure that all machine room computer systems are locked by a logon password system to prevent unauthorised access in case of a security breach of the rooms;

Once SR-SSDC is officially established, a risk assessment will be conducted, in order to identify its critical functions (based on the impact of data loss or the inability of the Data Centre to perform its duties).

6. Data Management

SR-SSDC will follow best practice in data management. Data Management is the third functional component of an OAIS. The Data Management function involves the maintenance of all the archived information and files including the management of metadata, communications with producers and users, and access statistics. The primary functions of Data Management are carried out by archive staff and include maintaining the databases of metadata for which it is responsible, performing queries on these databases, and generating reports in response to requests from other functional components within the OAIS (Ingest, Administration, Access).

7. Preservation Planning

The fourth functional component of an OAIS is Preservation Planning. This component provides the services and functions for monitoring the environment of the OAIS, providing recommendations and preservation plans to ensure that the information stored in the OAIS remains accessible to, and understandable by, the Designated Community over the long-term, even if the original computing environment becomes obsolete. Digital data are always linked to its support which itself is subject to rapid technical change. In other words, all file formats and physical storage media will become obsolete one time or another.

Preservation Planning functions include evaluating the contents of the archive and periodically recommending archival information updates, recommending the migration of current particular archive holdings, developing recommendations for the archive's standards and policies, providing periodic risk analysis reports, and monitoring changes in the technological environment and within the community.

SR-SSDC will try to ensure that it is at the leading edge of technical advances by taking a strategic approach to long-term digital preservation, and by monitoring hardware and software developments and migrating its collections accordingly. The technical development of storage technology must be under constant observation and review: monitoring technological changes that will affect preservation and migration decisions, files suitable for archiving, the structure and design of the information packages, and hardware (server, storage media) and software (applications).

The Data Centre recognises the need for communication with other stakeholders who are active in formulating national preservation policies and standards. It also acknowledges the need to participate in national and international activities and programmes in the area of digital preservation.

8. Access

The fifth functional component of OAIS relates to access. This component provides the services and functions that support data users in determining the existence, description, location, and availability of information stored in the data archive, allowing them to request and receive data and documentation. To be more precise, as part of this functional entity, the Dissemination Information Packages (DIP) are generated through addition of descriptive metadata, and the dissemination request is processed accordingly. Access is also responsible for implementing any security or access control mechanisms associated with the archives content.

8.1. Dissemination policy

There is a tension between providing free and open access to data and at the same time protecting confidentiality of research participants and the depositor. However, anonymisation strategies on the one side, and specifications in the end user license on the other, are ways to give consideration to both aspects. The fundamental premise is that no individual should be harmed.

Datasets and documentation held by SR-SSDC will be publicly available within certain limitations and controlled access:

- Access to data is limited to users and individuals affiliated with a research institution such as faculty or students (for non-commercial purposes);
- Access to data could be restricted by IP address range or University network;
- If there is the need, a secure non-networked server for the storage of highly confidential datasets will be maintained;
- Preliminary registration into the system is requested to gain access to data (the registration will be done by the users or by archive staff);
- All users will have to agree to study-specific usage conditions before they access data;
- Access to data will be provided by e-mail, direct download link or ftp-server system.

Access to the collections of SR-SSDC will be through the data catalogue (Nesstar Server) accessible on the institutional website: data.sr.serscida.eu. The data and metadata of SR-SSDC are also harvested and made available in the CESSDA Catalogue which provides seamless access to datasets from social science data archives across Europe.[72]

8.2. Access management strategy

Access to data at SR-SSDC will be controlled depending on the level of sensitivity of the data and the specific conditions determined by the depositor. The latter are set out in the **end user license** (see Annex A1.2 End User Licence SR-SSDC), a version of which all users must sign before being given access to any dataset. The End User Licence describes the legal framework under which the material is distributed. It specifies the conditions under which a user is granted access to data and how they may use those data (citation in scientific fashion and use following typical scientific ethical norms of

[72] The CESSDA Catalogue:

conduct). According to the sensitivity of data or due to additional access conditions negotiated with the data producer special conditions and licences might apply.

Accessibility categories

There will be three levels of access, depending on the nature of the data:

1) Open data: Data and documents that can be **accessed by anyone** by way of the archive dissemination portal; No registration is needed, but data may be subject to an open data licensing (such as Creative Commons Licence, CC);

2) Restrictive data:
a) Standard access: Data and documents that are fully anonymised and available for scientific analysis carried out in academic research and teaching, and may be **accessed without impending permission and special conditions from the data depositor,** i.e. free access via an ordering system by e-mail or online. However, registration is necessary. Users who register must have accepted the end user licence of USDAS;

b) Special conditions: Data that are released for academic research and teaching, where the user may gain access to data **subject to written approval from the depositor.** For this purpose the user must supply the archive with their details and specify their analytic intention. The Data Centre will request written permission from the depositor and will only permit access on receipt of this. In any case, USDAS informs the depositor either before or after data distribution to the user depending on their agreement. By default, the user must sign the end user licence in order to gain access to the material;

c) Special license: Data that contain more detailed information and are usually not fully anonymised require users to complete a set of Special Licence forms;

3) Controlled data: USDAS reserves the right to permit access to data only through a physical or virtual secure environment or after special training. This applies to data that are considered too confidential or sensitive to be released via download.

In certain circumstances where **studies are not deposited at the Data Centre** but only described in the catalogue (metadata), no contract between the user and the Data Centre is required. In this case, the research material might either be available through external websites (projects website), or the user can find addition information by directly contacting the producer.

9. Administration

Finally, SR-SSDC will administer its digital data archive in alignment with international best practice. Administration is the sixth and final functional component of an OAIS. This component is responsible for managing the day-to-day operations of the OAIS, as well as coordinating the activities of the other five high-level OAIS services. Other responsibilities include interacting with data producers (e.g., negotiating submission agreements), data users (e.g., providing customer service support), and the management (e.g., implementing and maintaining archive policies and standards).

SR-SSDC will use the tool RT: Request Tracker for the registration of data users (for more details and technical specifications see SERSCIDA D5.3 chapter 3.3 and 3.7).

In order to outline the financial requirements upon establishment of SR-SSDC, a separate document has been created indicating the key variables and items of expenditure over a period of five years (see Annex A3.2: Finance and budget SR-SSDC). In addition, as part of the administrative responsibilities of the newly established data archive, an internal performance monitoring system based on a conceptual framework that links overall institutional objectives to key activities, and then to specific indicators has been created (see Annex A4.2: Performance indicators SR-SSDC)

III Croatian Data Archive for the Social Sciences – CRO-DAS

1. Contextualisation and organisation of document

The aim of the F7 project SERSCIDA is the establishment of a prototype for social science digital data archives[73] in the Western Balkans, which could go into operation specifically in the countries of Bosnia and Herzegovina, Croatia, and Serbia. This includes the preparation of relevant documents for the formation of new data services. The present document "Data Policies and Procedures" was created as part of Work Package 5 (WP5) of the SERSCIDA project.

The present report will follow in large part the functionalities of the international standard *Open Archival Information System (OAIS) Reference Model* (explained in chapter 2). The Croatian Data Archive for the Social Sciences (CRO-DAS) is committed to developing its digital preservation policies and strategies in accordance with the *(OAIS) Reference Model*. This approach aligns with policies and practices of members of the Consortium of European Social Science Data Archives CESSDA, and with the SERSCIDA deliverables from prior Work Packages, as well as the common standards of the international infrastructure CESSDA AS (Annex 2.7).[74]

The target audience of this document includes state, federal, and cantonal ministries, stakeholders of the CRO-DAS, its staff, the social science research community in Croatia, and the general public. Beyond its purpose as a deliverable for the SERSCIDA project, this policy document could be used for establishing new data archive centres in other countries.

2. Conceptual models and Guiding principles

2.1. The OAIS Reference Model

A very important concept for the establishment of any digital repository is the Reference Model for an Open Archival Information System (OAIS)[75],adopted as an ISO standard in 2003 (ISO 14721:2003 OAIS). OAIS provides a framework for the understanding of digital archives and provides common terminology to be used in this context. The OAIS reference model is flexible enough to use in a wide variety of environments since it is strictly a logical model which does not depend on a particular implementation. It describes the components and services required to develop and maintain archives, in order to support long-term access to and understanding of the information in those archives. The OAIS model specifies how digital material should be preserved for a community of users (Designated Community)[76] from the moment digital material is ingested into the digital storage area, through subsequent preservation strategies, to the dissemination of digital material for the end user.

[73] We use "data archive" and "data service" interchangeably.
[74] Annexes to Statutes for CESSDA, No. 1-6 (18 June 2013)
[75] Reference Model for an Open Archival Information System (OAIS); CCSDS 650.0-M-2; Consultative Committee for Space Data Systems: Washington, DC, 2012.
[76] The Designated Community is a special class of Consumers; the subset of Consumers expected to independently understand the archived information in the form in which it is preserved and made available by the OAIS.[77] For the sake of convenience, for the rest of this document we will generally use terminology more common to the social science archiving community in Europe (e.g., data producers, data users), rather than the OAIS terminology.[78] CCSDS 650.0-M-2, p. 3-1

An OAIS-type archive carries out its preservation and access responsibilities in an environment populated by three key external stakeholders: 1) the Management (archivist, system administrator), which formulates, revises, and enforces a high-level policy framework that covers the OAIS activities; 2) the Producer (data producer, depositor, or researcher), that is, the individuals, organizations, or systems that create digital objects and/or transfer information to the OAIS for long-term preservation; and 3) the Consumer (data consumer, user, researcher) - the individuals, organizations, or systems expected to use the information preserved by the OAIS.[77]

OAIS Environment

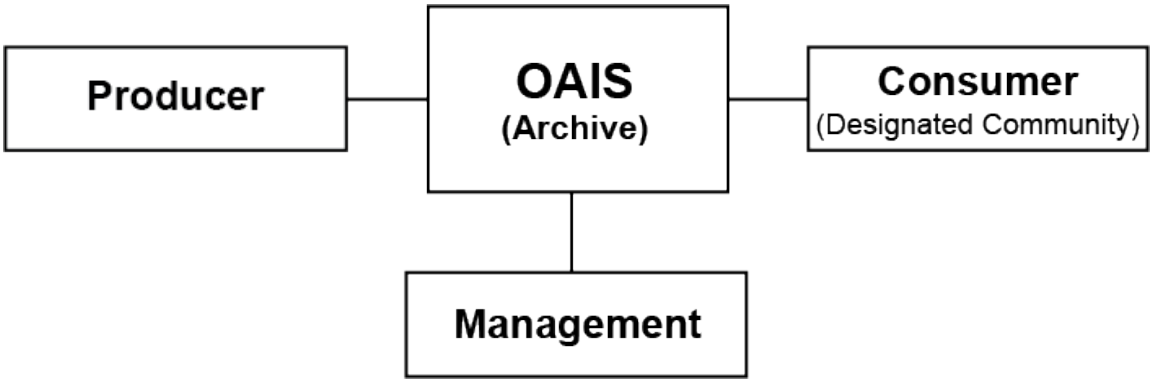


Figure 10: Reference Model for an Open Archival Information System (OAIS); CCSDS 650.0-M-2; Consultative Committee for Space Data Systems: Washington, DC, 2002, p. 2-2.

OAIS lists six functional entities that fulfil OAIS' dual role of preserving and providing access to the information: Ingest, Archival Storage, Data Management, Preservation Planning, Access, and Administration.

[77] For the sake of convenience, for the rest of this document we will generally use terminology more common to the social science archiving community in Europe (e.g., data producers, data users), rather than the OAIS terminology.[78] CCSDS 650.0-M-2, p. 3-1

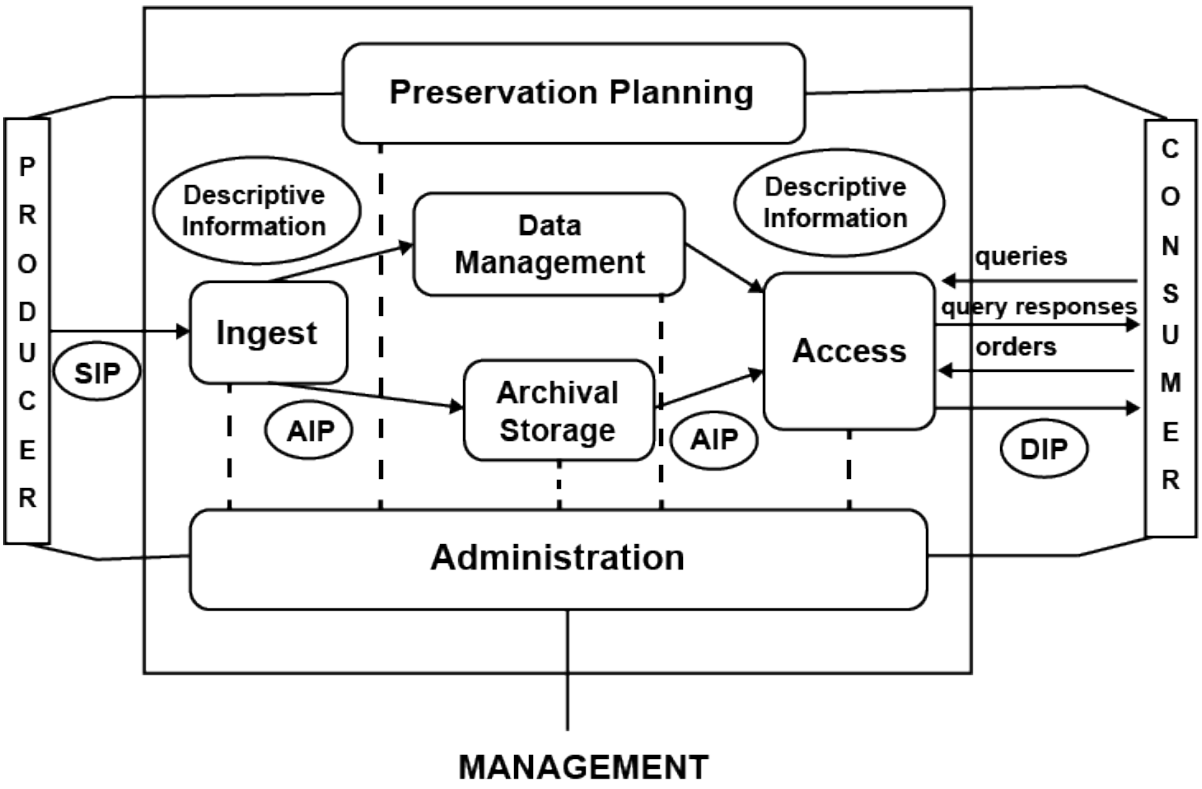


Figure 11: Reference Model for an Open Archival Information System (OAIS); CCSDS 650.0-M-2; Consultative Committee for Space Data Systems: Washington, DC, 2002, p. 4-1

An essential part of the OAIS model is the Information Package (IP), which represents the structure of information as it moves into, through, and out of the archival system. There are three types of information packages, all of them consisting of the digital object to be preserved and the metadata necessary to support its long-term preservation and access:

- SIP (Submission Information Package): information package that is transferred from the Producer to the Archive.
- AIP (Archival Information Package): information package that is stored and preserved by the Archive.
- DIP (Dissemination Information Package): information package delivered to the Consumer in response to an access request.

There are several minimum responsibilities that should be maintained by an OAIS[78]:

- Negotiate for and accept appropriate information from Producers;
- Obtain sufficient control of the information provided to the level needed to ensure long-term preservation;
- Determine, either by itself or in conjunction with other parties, which communities should become the Designated Community and, therefore, should be able to understand the information provided, thereby defining its Knowledge Base[79];

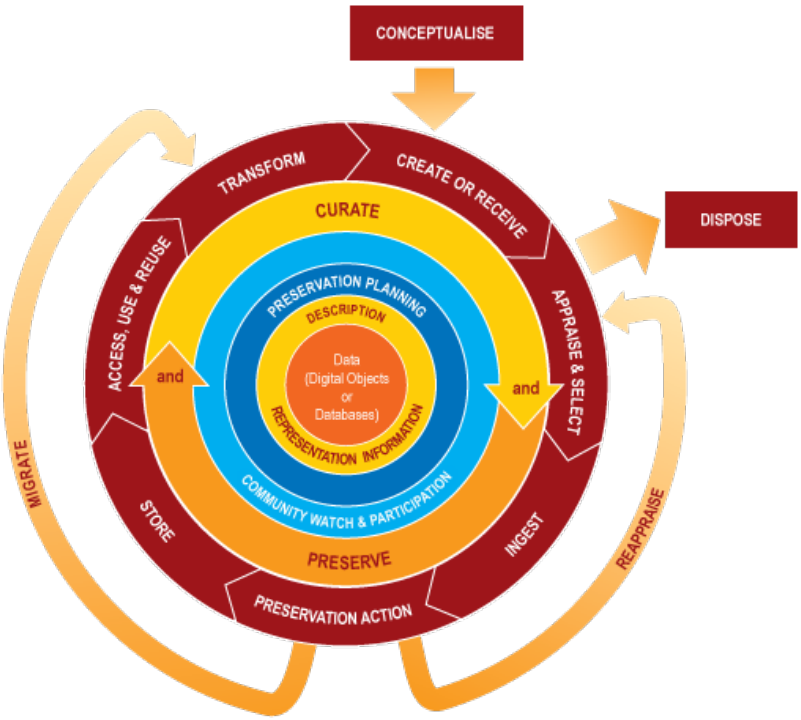
[78] CCSDS 650.0-M-2, p. 3-1

[79] "A person, or system, can be said to have a Knowledge Base, which allows that person or system to understand received information. For example, a person who has a Knowledge Base that includes an understanding of English will be able to read and understand an English text."(CCSDS 650.0-M-2, p. 2-3)

- Ensure that the information to be preserved is independently understandable to the Designated Community. In particular, the Designated Community should be able to understand the information without needing special resources such as the assistance of the experts who produced the information;
- Follow documented policies and procedures which ensure that the information is preserved against all reasonable contingencies, including the demise of the Archive, ensuring that it is never deleted unless allowed as part of an approved strategy;
- Make the preserved information available to the Designated Community and enable the information to be disseminated as copies of, or as traceable to, the original submitted Data Objects with evidence supporting its Authenticity[80].

2.2. The DCC Curation Lifecycle Model: metadata and standards

The Data Centre will follow all the phases of the DCC **Curation Lifecycle Model**[81]. Data often have a longer lifespan than the research project that creates them. The DCC Curation Lifecycle Model can be used to plan activities within an organisation to ensure that all necessary phases are undertaken, each in the correct sequence.



[80]Authenticity: relies on three provable properties: the record is what it purports to be, that it was created by whomsoever it is purported to have been created by; and that it was created at the time when it is purported to have been created Joint Information System Committee (JISC) Digital Preservation: continued access to authentic digital assets (Briefing Paper, November 2006), p.2). Other concepts designed to guide policies for information security within an organization are confidentiality (ensuring that data are not made available or disclosed to unauthorised individuals) and integrity (ensuring that data are not altered or destroyed in an unauthorized manner).

[81] DCC Curation Lifecycle Model: http://www.dcc.ac.uk/resources/curation-lifecycle-model

Figure 12: The DCC Curation Lifecycle Model

The collaboration between the data archive and the data producer should be a priority during the whole data curation process. It is important that the data producers are working with CRO-DAS during the documentation process, since they know their data best and are most familiar with the study methodology. On the other hand, CRO-DAS has the best knowledge about data documentation procedures and metadata standards. In addition, the data producers are usually responsible for the documentation, metadata description and the data themselves (data preparation, anonymisation, and cleaning). As a basic principle, it should be made clear throughout the whole archival process who bears responsibility for which tasks (the data producer or the archive).

Standardisation is a crucial element in the establishment of Social Science Data Archives. CRO-DAS is devoted to ensuring best practice and standardised approaches in order to enable compatibility and interoperability for international data sharing.

CRO-DAS orients itself towards international standards adopted by CESSDA AS and its service providers so that our data can be available and sharable throughout the international community. Regular reviews will be undertaken in order to guarantee best practice.

At CRO-DAS we will be compliant with the **Data Documentation Initiative (DDI)** metadata specification, version 2.5 or higher[82]. To be more precise, CRO-DAS will create the following types of metadata generated from accompanying documentation[83]:

- **Descriptive metadata**: Describes a resource for purposes such as discovery and identification (contains elements such as title, abstract, author, and keywords). DDI complies with Dublin Core Metadata Initiative which describes a core set of 15 elements intended to facilitate discovery of electronic resources;[84]
- **Administrative metadata:** Contains information about the use, management, and encoding processes of digital objects over time (e.g. information about data creation)
- **Technical metadata**: Provides information about the overall system environment and provides the technical information needed to use data (e.g. file format, application used and operation system)
- **Structural metadata**: Describes the logical structure of a multidimensional object. The Metadata Encoding and Transmission Standard (METS) will be used where appropriate;[85]
- **Preservation metadata**: Provides information needed to archive and preserve a resource. The PREMIS Data Dictionary for Preservation Metadata defines a core set of preservation metadata elements and describes relationships between digital preservation entities: Intellectual entity, Object, Event, Agent, and Rights.[86]

[82] Data Documentation Initiative (DDI): http://www.ddialliance.org/
[83] As a general remark, metadata standards should be machine readable (XML coded) in order to be processed automatically. XML is an open, well-supported, and widely adopted standard for encoding textual data, designed to be used regardless of the hardware and software environment.
[84] Dublin Core Metadata Initiative: http://dublincore.org/
[85] Metadata Encoding and Transmission Standard (METS) http://www.loc.gov/standards/mets/
[86] PREMIS Data Dictionary for Preservation Metadata: http://www.loc.gov/standards/premis/v2/premis-2-0.pdf

More and more digital objects require more reliable and stable references (regardless of their location and across system change). CRO-DAS will therefore aim to assign a persistent identifier (DOI (Digital Object Identifier)/URN (Uniform Resource Name) to every study coming into the archive.

The following two international archival standards could be used as well:

- **ISAD(G)**: International Standard Archival Description (General) or ISAD(G) is an international standard which provides guidelines for creating the content of an archival description (descriptive metadata). It promotes the creation of consistent and appropriate descriptions, facilitating the retrieval and exchange of information, and the integration of descriptions into a unified information system. There is a list of elements which are considered mandatory (six elements) when writing an archival description.[87] Based on ISAD(G) is the **Encoded Archival Description (EAD)** standard;
- **ISAAR(CPF)**: is the International Standard Archival Authority Record For Corporate Bodies, Persons and Families standardised by the International Council on Archives in 2003.[88] Whereas ISAD(G) describes the archival material itself, ISAAR(CPF) is constructed to give information about the creator (authority) of the archival material with additional information on history, status, functions, structure and relations – generally speaking about the provenance of the data. Again, there has been another international standard called **Encoded Archival Context (EAC),** which is derived from ISAAR(CPF).

Within this regulatory framework, the CRO-DAS will also respect the following **national and international legal frameworks**:

- EU Copyright Directive (EU, 2001);
- Scientific Activity and Higher Education Act (Official Gazette no. 123/02, 198/03, 105/04, 174/04, 02/07, 46/07, 45/09, 63/11, 94/13, 139/13);
- Official Statistics Act (Official Gazette no. 103/03, 75/09, 59/12);
- Ordinance on conditions and methods of use of confidential statistical data for scientific purposes (Official Gazette no. 137/13);
- Law on Archive Records and Archives (Official Gazette no. 105/97, 64/00, 65/09, 125/11);
- Copyright and Related Rights Act (Official Gazette no. 167/03, 79/07, 80/11, 125/11, 141/13); and
- Personal Data Protection Act (Official Gazette no. 103/03, 118/06, 41/08, 130/11, 106/12).

Publicly funded research data are a public good produced in the public interest, and therefore should be openly available to the maximum extent possible (OECD principle).[89] CRO-DAS thus promotes the philosophy and practice of **Open Access,** where research data are as easily accessible as possible, while still protecting the confidentiality of respondent information and the rights of those who gathered or created the data. In addition, the basic data service of CRO-DAS for all data producers

[87] General International Standard Archival Description, ISAD (G): http://www.ica.org/10207/standards/isadg-general-international-standard-archival-description-second-edition.html

[88] International Standard Archival Authority Record For Corporate Bodies, Persons and Families, ISAAR (CPF) http://www.ica.org/10203/standards/isaar-cpf-international-standard-archival-authority-record-for-corporate-bodies-persons-and-families-2nd-edition.html

[89] OECD (2007) OECD principles and guidelines for access to research data from public funding. Organisation for Economic Cooperation and Development. Downloaded: http://www.oecd.org/science/sci-tech/38500813.pdf

and users will as a rule be free of charge. At the same time, CRO-DAS will not pay for the data it acquires.

Audit and Certification: As a future goal, we will seek to become compliant with at least one of the following standards of trust: Data Seal of Approval (DSA), ISO 16363 or DIN 31644 for the Audit and Certification of Trustworthy Digital Repositories.[90] Since one of the priorities of CRO-DAS is to become a CESSDA member, we seek to be compliant with the DAS seal which is granted to repositories that are committed to archiving and providing access to scholarly research data in a sustainable way. From the beginning, we will focus on quality control mechanisms throughout the whole archiving process.

Confidentiality and disclosure: All depositors should ensure that data meet the requirements of confidentiality and non-disclosure for data collected from human subjects. Datasets should not contain direct identifiers (names, addresses, telephone numbers, etc.) or indirect identifiers (which, when combined with other variables, may provide enough details to identify an individual). In-house confidentiality checks by CRO-DAS help to guarantee that confidentiality requirements are met.

3. Acquisition and deposit

This section focuses on the acquisition of data and covers data selection and appraisal strategies that will be adopted, as well as the file formats accepted by CRO-DAS. The collection development policy will answer the questions of what, from whom, why, and how are we going to collect data at CRO-DAS. In a next step, policies for the deposit of the data and related documentation materials are discussed. This includes the conditions for depositing data, data quality, confidentiality of data, and depositor rights. Although these elements are not explicitly treated within the OAIS Reference Model, they are worth treatment in this policy document, since they are important aspects of the proper establishment of a social science data archive.

3.1. Collection development policy[91]

Data selection and appraisal play an important role in the acquisition of data in any archival setting. The collection development policy indicates the principles and criteria by which the Archive develops its data collection in order to serve the Designated Community. The collection development policy of the CRO-DAS will be flexible and respond to future developments and shifting requirements that will influence the Archive's data collections (technology, scientific standards, etc.).

The CRO-DAS collects data and information across the range of social science disciplines.[92] This includes **quantitative** and **qualitative data**.

[90] Data Seal of Approval: http://www.datasealofapproval.org; Audit and certification of trustworthy digital repositories: http://www.iso16363.org/; Nestor Seal for Trustworthy Digital Archives: http://www.langzeitarchivierung.de/Subsites/nestor/EN/nestor-Siegel/siegel_node.html.[91] The content of this policy draws on the collection development policy of the UKDA: http://www.data-archive.ac.uk/media/54773/ukda067-rms-collectionsdevelopmentpolicy.pdf, and the experience of setting up new Data Archives in Finland and Slovenia (Laaksonen, H. et al. 2006): http://www.iassistdata.org/downloads/iqvol301laaksonen_0.pdf

[91] The content of this policy draws on the collection development policy of the UKDA: http://www.data-archive.ac.uk/media/54773/ukda067-rms-collectionsdevelopmentpolicy.pdf, and the experience of setting up new Data Archives in Finland and Slovenia (Laaksonen, H. et al. 2006): http://www.iassistdata.org/downloads/iqvol301laaksonen_0.pdf

WHAT can be accepted:

- Research that gives evidence and greater understanding of Croatian society or parts of it (including international surveys);
- Research data where Croatian researchers are involved;
- Comparative or continuous research (panel data, longitudinal surveys, and time series that allow examination of trends);
- Data produced with methodological excellence;
- Works of general interest for social science research (quality and exemplary studies with analytic potential).

The data collected will cover the social science disciplines within Croatia. The deposited data will be retained for the long-term, unless otherwise specified by the depositor, funding agencies or by government regulations. CRO-DAS will accept data and documentation in paper form only under exceptional circumstances. [93]

FROM whom are data solicited and accepted:

Eligible depositors are researchers, members of research institutions, government departments, offices of national statistics, and public and commercial sector sources (e.g., marketing research firms). Depositors may only deposit their own work and will be expected to sign a data deposit agreement (see section "data deposit" below).

WHY certain data are accepted:

There are three main reasons why data are solicited and accepted. First, because they possess potential value for secondary use and analysis for research. Second, because they may serve teaching and learning purposes. And finally, for validation reasons - to make possible the replication of research for important studies.

HOW the data will be acquired:

The acquisition process will consist of three stages: 1) identifying potential datasets (localisation), 2) negotiation with the data creators, and 3) receiving the relevant material. A very important source will be the list of potential data producers and consumers that has been established during the SERSCIDA project.

As CRO-DAS will be in the early stages of its existence, the **acquisition strategies** will be highly proactive and persistent. This will include open calls, announcements at conferences, and direct

[92] The following disciplines are embraced: sociology, psychology, educational science, information science, political science and economics.

[93] Studies in paper form will be accepted and worked on under separately funded digitisation projects, rather than as 'routine' activity. Accepting such material is fine, but they require careful storage until one has the resources to digitise and process them (i.e., special funding to digitise some of the paper deposits).[94] There is however a publicly available list of projects financed by our Ministry for Science. In the future, there will be list of projects financed by Croatian Science Foundation, since all scientific projects are now financed by the Foundation. This might be developed into a national research inventory for the social sciences. Additionally, a small amount of financial support is coming from the University of Zagreb, and the list of these projects is also available.

solicitation. No national research inventory for social science research data, where staff can actively search for and proactively seek data for deposit, is currently available for this purpose in Croatia.[94]

Similar as in some other countries, collaboration in the process of preparing and depositing data is entirely voluntary. This means that special effort will need to be invested in the process of negotiating and gaining access to the data sets. This will include:

- **Good working** relations with stakeholders and potential depositors (Croatian Bureau of Statistics, research institutes and faculties in social sciences, Croatian Employment Service and other government institutions, private research companies and individual researchers) to identify and acquire new data;
- **Links** with State/Entity/Historical Archives and Research Libraries;
- **Promotional activities** to raise awareness and encourage deposit of data collections (workshops, round tables, conferences, and training events);
- **Agreements** with funding bodies for the disciplines in question in order to discuss deposit of material with grant-holders;
- **Data discovery** activities by staff members in publications (academic journals), news media, conferences, e-mail lists, and websites (e.g., universities' online catalogues and websites of research funding bodies).

The aim is to establish links and working relations with the stakeholders and partners mentioned above in order to establish long-term cooperation and a common framework for the deposit of data collections.

The CRO-DAS reserves the right to **refuse** to accept material under the following conditions:

- Data do not match the criteria of collection development policy and would be better dealt with at another institution (also legal conditions);
- Insufficient or poor quality documentation;
- Depositor and CRO-DAS do not agree on access and dissemination conditions;
- There are currently no restrictions on file sizes, or number of files or studies. However, the CRO-DAS reserves the right to refuse to accept materials which are of a volume which may make it difficult or impossible to process given the resources, staff, facilities, or capacities of the CRO-DAS.

In order to make the data and documentation accessible for other users and to be able to process and use them over time, they should be in a form that best ensures their longevity. The data and documentation **file formats** that are preferred at the ingest phase by the Data Centre (Submission Information Package, SIP), are non-proprietary, openly documented, unencrypted, and uncompressed that are commonly used by the research community. Several lists are available online

[94] There is however a publicly available list of projects financed by our Ministry for Science. In the future, there will be list of projects financed by Croatian Science Foundation, since all scientific projects are now financed by the Foundation. This might be developed into a national research inventory for the social sciences. Additionally, a small amount of financial support is coming from the University of Zagreb, and the list of these projects is also available.

for consultation about recommended file formats and specific requirements.[95] File format registries (e.g. PRONOM) are a way of helping to identify file formats and look up format specifications.[96]

We therefore consider the following formats as appropriate for the Submission Information Package (SIP):[97]

- 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 Open Office 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, such as **7-Zip** or **Winzip**.

3.2. Data deposit

In this section, the documenting of data by the researcher is explained in further detail. Data documentation explains how data were collected (context of data collection), what they mean, what are their content and structure, and specifies any manipulations that may have taken place.[98] Good documentation should be considered best practice when creating, organising, and managing data, and is equally important for data preservation.

- Data files must be clean and clear, i.e., interpretable and searchable by ways of a codebook by other researchers;
- The data should be accompanied by adequate documentation that explains the context of data collection (history, aim, and hypotheses), methodology, and any material relevant to the data. This includes not only the general information about the study itself (metadata) compatible with international standards (DDI, in order to share this metadata) but also to give as much material as possible in order to make the study as suitable as possible for

[95] The National Archives (2008): Selecting File Formats for Long-Term Digital Preservation: http://www.dptp.org/wp-content/uploads/2010/08/selecting-file-formats.pdf,

[96] The Technical Registry PRONOM: http://www.nationalarchives.gov.uk/PRONOM/Default.aspx

[97] The selected file formats represent a summary of different recommendations from CESSDA partners and internationally recognised institutions. FORS: Qualitative Data Archiving at FORS – Policy and Procedures: http://forscenter.ch/wp-content/uploads/2013/11/FORS_policy_and_procedures_on_qualitative_data1.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. The formats highlighted in bold are preferred over the others of the same category.

[98] Manipulation could involve either anonymisation, cleaning out of range values or quality assurance procedures carried out by the data producer before deposit of the data.

secondary use (see chapter 2). This kind of accompanying material could include methodological reports, codebooks, questionnaires, coding instructions, interviewer guides, database dictionaries, bibliographies of publications related to the data, and links to online tools.

The quality of documentation can be significantly improved if its creation and collation is planned at the beginning of the data life cycle, during the project conception phase. If needed, archive staff will offer assistance in this respect. [99]

Generally speaking, the following minimal information should be made available when depositing the data at CRO-DAS:

- Contact Information (depositor[100]): name, address, E-mail, telephone;
- Study information: title, access level, principal investigator or data creator, data collectors, funding information;
- Accompanying documentation: study reports, questionnaires, codebook, user guides, methodology reports, publications (bibliographical reference);
- Dataset information (title, temporal coverage of collection, time dimension, method of collection, main topics, subject categories, number of variables and cases, sampling procedures, response rate, URL, etc.).

The **data deposit agreement** (see Annex A1.3: Data Deposit Agreement CRO-DAS) describes the legal framework under which material is deposited and states the rights and responsibilities of both parties – the depositor and the archive – regarding copyright and ownership of the data and the access conditions. It is a legal agreement between the depositor and the archive that covers arrangements regarding usage rights, authenticity, data protection responsibilities, and disposal.

The following additional questions should equally be answered in advance[101]:

- How do the data come into the data archive (SIP)?[102] By means of CD-ROM, Ftp, or e-mail?
- What is the level of complexity of the study data and documentation deposited?
- Will there be an embargo or special access conditions?

The data producer can during the negotiation phase place the data under an **embargo** or special conditions, which means that data are not accessible to data users for a pre-defined period or only

[99] The following manual on managing and sharing research data by the UK Data Archive is considered to be best practice: UK Data Archive (2009): Managing and Sharing Data – Best Practice for Researchers: http://www.data-archive.ac.uk/media/2894/managingsharing.pdf

[100] Person(s) depositing the data; usually, but not always represents the licence signatory.

[101] Archive agreement GESIS:http://www.gesis.org/fileadmin/upload/institut/wiss_arbeitsbereiche/datenarchiv_analyse/Archivierungsvertrag_GESIS_Datenarchiv_v9__englisch.pdf; Licence agreement UKDA: http://ukdataservice.ac.uk/media/28102/licenceform.pdf

[102] SIP should be a machine readable dataset, should have a codebook for raw data, and should include the survey instrument/measurement instrument (originals or copies, in a digital format if possible) and a methodology report, if available. In addition to that there should be determined beforehand how the SIP is transmitted to the archive: by e-mail, download from/upload to server, etc.

with certain restrictions. However, in the spirit of Open Access we highly recommended that data are made available to users as soon as possible after a research project has come to an end. The data depositor can also impose certain conditions on the use of the data (e.g., to be informed beforehand about who is going to use the data and for what purposes).

4. Ingest

Ingest is the first functional component of the OAIS reference model. It includes the set of processes responsible for accepting the information submitted by data producers (Submission Information Package, SIP) and preparing it for the archival storage. Some of these processes include:

- Validation that information received is uncorrupted and complete;
- Transformation of the submitted information into a form suitable for storage;
- Extraction and/or creation of descriptive metadata.

The tool used for ingest will be EPrints, customised by the SERSCIDA team (see SERSCIDA D5.3).

At a minimum, all data will be treated and described at study level. However, some data that are of higher analytic potential could receive more attention and processing. CRO-DAS will use Nesstar, used by many CESSDA members to publish and disseminate variable-level metadata for high profile studies.[103]

4.1. Ingest workflow

Reception of SIP: Data arrives at CRO-DAS

Most commonly, the data files will arrive at CRO-DAS by e-mail (possibly encrypted) or through the ingest tool EPrints. If the SIP is submitted to CRO-DAS by e-mail or from any other device, the archive staff subsequently needs to enter the SIP into EPrints. Ideally, the data deposit agreement should be in place before the data actually arrive at the archive.

Once reception of the SIP is approved, the following validation and quality assurance procedures are carried out:

- Review of completeness of the study data, metadata, and documentation (if needed, checked against data deposit agreement);
- Checks if the scope of documentation is sufficient (e.g., are variables described and can easily be attributed to the questions in the questionnaire);
- Comparison of data sets and data documentation to ensure that there is a match between variables, variable labels, and value labels, and that the number of observations and variables match the stated numbers in the documentation;
- Checks for anonymisation strategies if data are sensitive (proper name, postal code, profession, etc.).

[103] The Nesstar Publisher tool is a data and metadata authoring tool that is used to prepare and publish survey data for the Nesstar system. The metadata inserted is consistent with the DDI (Data Document Initiative). The Nesstar Server tool provides the ability to publish statistical information that can be searched, browsed, analysed, and downloaded by users. (www.nesstar.com)

Validation and quality assurance

Most of the following tasks are automatically performed by the system and corresponding tools:

- Scan and check for viruses
- Validation of the successful transfer of the SIP into EPrints
- Identify file formats
- Validate file formats (if files are what they pretend to be)
- Create checksums for each data and document file in order to guarantee data integrity during the transfer process
- Generate Unique Identifiers (UI)

Manual tasks conducted in addition by CRO-DAS data experts include:

- Checks to see if confidentiality of data is guaranteed; and
- Manual/visual quality control assurance using user-friendly GUI interface of EPrints.

Normalisation : SIP is converted into a AIP

Data processing:

- The originally transmitted data files (SIP) are stored in a separate location;
- Renaming and rearranging of the data files;
- Decision of description level – study or variable;
- Metadata extraction (technical, administrational, structural, and preservation metadata);
- Integrity verification that data integrity has been fully maintained during the transfer process by checksum creation.

Documentation processing:

- Renaming and organisation of documentation files;
- Conversion of files to archival format and marking of material (e.g., bookmarking and adding headers and table of contents);
- Combination of files into one;
- Extraction of descriptive metadata, in other words, cataloguing of those sources for future dissemination (DIP)
- Creation of persistent identifiers.

The data and documentation will be registered and stored on the server located at CRO-DAS or the University Teleinformatic Centre (UTIC), or outsourced in the initial phase of establishment.[104] The ticketing tool RequestTracker is used to record all actions and decisions made concerning the submitted material (see chapter 3.2.2 of SERSCIDA D5.3).

Pre-release procedure:

- Inform the data depositor to check the processed data files and documentation and verify with staff colleagues (quality control);

[104] See „Data backup and security“ in chapter 5.

- All working copies should be deleted from the individual network area and any processing server in order to guarantee data security.

Ingest Workflow	**Manually**		**Automatically**
Receive Submission Information Package (SIP)	completeness of the study data, metadata and documentation		-
	Check scope of documentation		
	Comparison of data with documentation		
	Check anonymisation		
Quality Assurance	Check if confidentiality of data is guaranteed		Virus checking
	Perform manual/visual quality control assurance.		Validation of successful transfer of SIP into ingest tool
			Identification and validation of file formats
			Checksums creation
			Generation of Unique Identifiers (UI
Generate Archival Information Package (AIP)	**Data Processing**	**Documentation processing**	-
	Rearranging of data	Reorganisation of documentation	*Extraction of metadata (technical, administrational, structural, and preservation)*
	Decide description level	-	*Conversion of files to archival format*
	-	Marking of material	
	-	Extraction of descriptive metadata	
	-	Persistent identifier (PI)	
	Extraction of metadata (technical, administrational, structural, and preservation)		
	Conversion of files to archival format		
	Separate AIP storage from SIP		
Pre-release procedure	Inform data depositor and verify with staff colleagues (quality control)		Capture correspondence (actions and decisions) between depositor and archive
	Delete working copies from individual network area and processing server		

Table 3: Ingest Workflow according to OAIS for CRO-DAS

5. Archival Storage

The second functional component of OAIS is archival storage. The purpose is to ensure that what is added to the archive from the ingest process remains identical and accessible. More specifically, this component manages the storage and maintenance of digital materials (AIP) so that the preserved information remains complete and readable over the long-term.

The primary goal of CRO-DAS's preservation policy is therefore to ensure the longstanding accessibility of the digital material, ensuring the highest level of authenticity of any formats. However, preservation decisions at CRO-DAS will always be made within the context of its collections development policy, balancing the constraints of cost, research and historical value, and accessibility, alongside the requirements of levels of authenticity and legal admissibility.

5.1. Preservation policy

Storage and Migration

The aims of the preservation policy are to:

- Become a trusted digital repository in order to serve a Designated user Community[105];
- Retain authenticity, integrity and reliability of preserved datasets;
- Ensure that all data collections are protected and safely stored;
- Ensure that the relevant level of information security is applied to each data collection;
- Apply good practice in preservation management.

CRO-DAS will follow established best practices for managing datasets over the long-term. It assumes responsibility for the preservation and accessibility of the digital material in its collection.

Preservation in a digital archive usually means preserving the significant properties or the essential characteristics of data for these to maintain their meaning over time. Any changes (by machine/application or an individual) in the preservation process should be properly documented (either as a special document or within the archiving application).

CRO-DAS follows international best practice in its choice of preservation formats and has listed them below. Preservation strategies are expected to consist of migration of the data files (refreshment, replication, repackaging or transformation), that is, data transferred into new data formats across several technology cycles of the data files.[106]

Preservation procedures will consist of:

- Reformatting datasets from proprietary software to software-independent formats and creating an Archival Information Package (AIP);
- Defining, timing, testing and implementing migration pathways;

When new formats are created from data files either through migration into new file formats or through the creation of new file formats for dissemination, the earlier files will be retained alongside. The creation of a metadata record for each dataset will record preservation tasks, checksums for authenticity and integrity, and logs and reports of all refreshment results.

Formats for the AIP:

[105] See "Audit and Certification" in chapter 2.

[106] Other strategies, such as emulation (imitating hard- and software elements) or computer museum (keeping hard- and software elements) are highly time consuming and therefore not taken into account.

- 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), Rich Text Format (RTF) (.rtf), 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).

Formats for the DIP:

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

Data backup and security

It should be clear that data backup is a short-term data recovery solution following loss and corruption, and is fundamentally different from the preservation planning strategy of the CRO-DAS. Backing up data means making copies of files so that in case of data loss the original data can be restored from the copies. Regular backups protect against accidental data loss due to hardware, software, or media failure; failure of the power supply; human error; and fires, floods, earthquakes, and other natural disasters. Backup strategies and incident management plans are an essential part of its archival policies; however they are not sufficient to ensure survival and access to authentic digital resources over time.[107]

A reliable technical infrastructure should be in place with suitable measures to ensure the long-term usability of data collections (security and risk management). At CROS-DAS the University of Zagreb 's technical infrastructure department University Computing Centre (SRCE) could provide this service, which would be ensured through close formal cooperation. The digital preservation of the data collections must rely on a solid IT infrastructure.

CRO-DAS will make sure that datasets are regularly backed up in accordance with established best practices. For backup and for security reasons, the CRO-DAS will aim to:

- Follow a policy of multiple copy resilience: more than one version of the complete preservation system will be maintained (e.g., main server, near-site online copy in another building, off-site online copy, disc-based offline copy=DVD, CD-R);
- Have a secure computer networking system and communications equipment;
- Ensure that all machine room computer systems are locked by a logon password system to prevent unauthorised access in case of a security breach of the rooms;

[107] http://www.jisc.ac.uk/media/documents/publications/digitalpreservationbp.pdf

Once the CRO-DAS is officially established, a risk assessment will be conducted, in order to identify its critical functions (based on the impact of data loss or the inability of the archive to perform its duties).

6. Data Management

CRO-DAS will follow best practice in data management. Data Management is the third functional component of an OAIS. The Data Management function involves the maintenance of all the archived information and files including the management of metadata, communications with producers and users, and access statistics. The primary functions of Data Management are carried out by archive staff and include maintaining the databases of metadata for which it is responsible, performing queries on these databases, and generating reports in response to requests from other functional components within the OAIS (Ingest, Administration, Access).

7. Preservation Planning

The fourth functional component of an OAIS is Preservation Planning. This component provides the services and functions for monitoring the environment of the OAIS, providing recommendations and preservation plans to ensure that the information stored in the OAIS remains accessible to, and understandable by, the Designated Community over the long-term, even if the original computing environment becomes obsolete. Digital data are always linked to its support which itself is subject to rapid technical change. In other words, all file formats and physical storage media will become obsolete one time or another.

Preservation Planning functions include evaluating the contents of the archive and periodically recommending archival information updates, recommending the migration of current particular archive holdings, developing recommendations for the archive's standards and policies, providing periodic risk analysis reports, and monitoring changes in the technological environment and within the community.

CRO-DAS will try to ensure that it is at the leading edge of technical advances by taking a strategic approach to long-term digital preservation, and by monitoring hardware and software developments and migrating its collections accordingly. The technical development of storage technology must be under constant observation and review: monitoring technological changes that will affect preservation and migration decisions, files suitable for archiving, the structure and design of the information packages, and hardware (server, storage media) and software (applications).

CRO-DAS recognises the need for communication with other stakeholders who are active in formulating national preservation policies and standards. It also acknowledges the need to participate in national and international activities and programmes in the area of digital preservation.

8. Access

The fifth functional component of OAIS relates to access. This component provides the services and functions that support data users in determining the existence, description, location, and availability of information stored in the data archive, allowing them to request and receive data and documentation. To be more precise, as part of this functional entity, the Dissemination Information Packages (DIP) are generated through addition of descriptive metadata, and the dissemination

request is processed accordingly. Access is also responsible for implementing any security or access control mechanisms associated with the archives content.

8.1. Dissemination policy

There is a tension between providing free and open access to data and at the same time protecting confidentiality of research participants and the depositor. However, anonymisation strategies on the one side, and specifications in the end user license on the other, are ways to give consideration to both aspects. The fundamental premise is that no individual should be harmed.

Datasets and documentation held by the CRO-DAS will be publicly available within certain limitations and controlled access:

- Access to data is limited to users and individuals affiliated with a research institution such as faculty or students (for non-commercial purposes);
- If there is the need, a secure non-networked server for the storage of highly confidential datasets will be maintained;
- Preliminary registration into the system is requested to gain access to data (using an available authentication and authorisation application);
- All users will have to agree to study-specific usage conditions before they access data;
- Access to data will be provided by e-mail, direct download link or ftp-server system.

Access to the collections of CRO-DAS will be through the data catalogue (Nesstar Server) accessible on the institutional website: data.hr.serscida.eu. The data and metadata of CRO-DAS are also harvested and made available in the CESSDA Catalogue which provides seamless access to datasets from social science data archives across Europe.[108]

8.2. Access management strategy

Access to data at CRO-DAS will be controlled depending on the level of sensitivity of the data and the specific conditions determined by the depositor. The latter are set out in the **end user license** (see Annex A2.3 End User Licence CRO-DAS), a version of which all users must sign before being given access to any dataset. The End User Licence describes the legal framework under which the material is distributed. It specifies the conditions under which a user is granted access to data and how they may use those data (citation in scientific fashion and use following typical scientific ethical norms of conduct). According to the sensitivity of data or due to additional access conditions negotiated with the data producer special conditions and licences might apply.

Accessibility categories

There will be three levels of access, depending on the nature of the data:

1) Open data: Data and documents that can be **accessed by anyone** by way of the archive dissemination portal; No registration is needed, but data may be subject to an open data licensing (such as Creative Commons Licence, CC);

2) Restrictive data:

[108] The CESSDA Catalogue: http://www.cessda.net/catalogue/

a) Standard access: Data and documents that are fully anonymised and available for scientific analysis carried out in academic research and teaching, and may be **accessed without impending permission and special conditions from the data depositor,** i.e. free access via an ordering system by e-mail or online. However, registration is necessary. Users who register must have accepted the end user licence of CRO-DAS;

b) Special conditions: Data that are released for academic research and teaching, where the user may gain access to data **subject to written approval from the depositor.** For this purpose the user must supply the archive with their details and specify their analytic intention. The Data Centre will request written permission from the depositor and will only permit access on receipt of this. In any case, CRO-DAS informs the depositor either before or after data distribution to the user depending on their agreement. By default, the user must sign the end user licence in order to gain access to the material;

c) Special license: Data that contain more detailed information and are usually not fully anonymised require users to complete a set of Special Licence forms;

3) Controlled data: CRO-DAS reserves the right to permit access to data only through a physical or virtual secure environment or after special training. This applies to data that are considered too confidential or sensitive to be released via download.

In certain circumstances where **studies are not deposited at CRO-DAS** but only described in the catalogue (metadata), no contract between the user and the Data Centre is required. In this case, the research material might either be available through external websites (projects website), or the user can find addition information by directly contacting the producer.

9. Administration

Finally, CRO-DAS will administer its digital data archive in alignment with international best practice. Administration is the sixth and final functional component of an OAIS. This component is responsible for managing the day-to-day operations of the OAIS, as well as coordinating the activities of the other five high-level OAIS services. Other responsibilities include interacting with data producers (e.g., negotiating submission agreements), data users (e.g., providing customer service support), and the management (e.g., implementing and maintaining archive policies and standards).

CRO-DAS will use the tool RT: Request Tracker for the registration of data users (for more details and technical specifications see SERSCIDA D5.3 chapter 3.3 and 3.7).

In order to outline the financial requirements upon establishment of CRO-DAS, a separate document has been created indicating the key variables and items of expenditure over a period of five years (see Annex A3.3: Finance and budget CRO-DAS). In addition, as part of the administrative responsibilities of the newly established data archive, an internal performance monitoring system

based on a conceptual framework that links overall institutional objectives to key activities, and then to specific indicators has been created (see Annex A4.3: Performance indicators CRO-DAS)

References

Annexes to Statutes for CESSDA, No. 1-6 (2013): http://cessda.net/about/docs/Annexes-to-Statutes-for-CESSDA-180613-Final-Version-1.pdf

Audit and certification of trustworthy digital repositories: http://www.iso16363.org/

Beagrie, Ch. et al (2011). User Guide for Keeping Research Data Safe 2 http://www.beagrie.com/static/resource/KeepingResearchDataSafe_UserGuide_v2.pdf

DataCite: https://www.datacite.org/

Data Documentation Initiative (DDI): http://www.ddialliance.org/

Data Seal of Approval: http://www.datasealofapproval.org

DCC Curation Lifecycle Model: http://www.dcc.ac.uk/resources/curation-lifecycle-model

Dublin Core Metadata Initiative: http://dublincore.org/

EPrints: www.eprints.org

FORS: Qualitative Data Archiving at FORS – Policy and Procedures: http://forscenter.ch/wp-content/uploads/2013/11/FORS_policy_and_procedures_on_qualitative_data1.pdf

FORS: End user contract: http://forscenter.ch/wp-content/uploads/2013/11/end-user_contract_quanti_normal1.pdf.

GESIS: Archive agreement: http://www.gesis.org/fileadmin/upload/institut/wiss_arbeitsbereiche/datenarchiv_analyse/Archivierungsvertrag_GESIS_Datenarchiv_v9__englisch.pdf

GESIS: Usage regulations: http://www.gesis.org/en/services/data-analysis/data-archive-service/usage-regulations/

ISAD(G): General International Standard Archival Description http://www.ica.org/10207/standards/isadg-general-international-standard-archival-description-second-edition.html

ISAAR(CPF): International Standard Archival Authority Record For Corporate Bodies, Persons and Families http://www.ica.org/10203/standards/isaar-cpf-international-standard-archival-authority-record-for-corporate-bodies-persons-and-families-2nd-edition.html

Joint Information Systems Committee (JISC) Digital Preservation: continued access to authentic digital assets (Briefing Paper, November 2006). http://www.jisc.ac.uk/media/documents/publications/digitalpreservationbp.pdf

Laaksonen, H., Borg, S. and Stebe J. (2006) Setting up Acquisition Policies for a New Data Archive in *IASSIST Quarterly* Spring, 5-12. http://www.iassistdata.org/downloads/iqvol301laaksonen_0.pdf

Metadata Encoding and Transmission Standard (METS) http://www.loc.gov/standards/mets/

Nesstar: www.nesstar.com

Nestor Seal for Trustworthy Digital Archives: http://www.langzeitarchivierung.de/Subsites/nestor/EN/nestor-Siegel/siegel_node.html

OECD (2007) OECD principles and guidelines for access to research data from public funding. Organisation for Economic Cooperation and Development: http://www.oecd.org/science/sci-tech/38500813.pdf

PREMIS Data Dictionary for Preservation Metadata: http://www.loc.gov/standards/premis/v2/premis-2-0.pdf

Reference Model for an Open Archival Information System (OAIS); CCSDS 650.0-M-2; Consultative Committee for Space Data Systems: Washington, DC, 2012. http://public.ccsds.org/publications/archive/650x0m2.pdf

Technical Registry PRONOM: http://www.nationalarchives.gov.uk/PRONOM/Default.aspx

The CESSDA Catalogue: http://www.cessda.net/catalogue/

UK Data Archive: Assessment of UKDA and TNA Compliance with OAIS and METS Standards: http://www.dptp.org/wp-content/uploads/2010/08/UKDAp90.pdf

UK Data Archive: Development policy: http://www.data-archive.ac.uk/media/54773/ukda067-rms-collectionsdevelopmentpolicy.pdf

UK Data Archive: Formats table: http://www.data-archive.ac.uk/create-manage/format/formats-table

UK Data Archive: Licence agreement: http://ukdataservice.ac.uk/media/28102/licenceform.pdf

UK Data Archive: Terms and Conditions of Access: http://ukdataservice.ac.uk/get-data/how-to-access/conditions.aspx

UK Data Archive: Managing and Sharing Data – Best Practice for Researchers: http://www.data-archive.ac.uk/media/2894/managingsharing.pdf

Annex 1: Data deposit agreements[109]

A1.1 Data deposit agreement USDAS

1. Parties of the agreement

(Depositor) Research Institution: NAME, ADDRESS, E-MAIL, TELEPHONE, FAX

and

Data Archive: University of Sarajevo Data Archive and Services, ADDRESS, E-MAIL, TELEPHONE, FAX

2. Data to be archived

"NAME OF PROJECT"

Under the direction of "NAME OF RESEARCHER"

"REFERENCE NUMBER (if any)"

List of deposited files:

3. Depositor's rights and responsibilities

The depositor agrees to:

- Have collected and treated the data in conformity with existing laws on data protection, including the anonymisation of the data;
- Deposit them at "ADDRESS OF ARCHIVE".

The depositor has the right to:

- The copyright on the data mentioned in point 2 ;
- Be informed about any reuse of the data by third parties.

4. The Archive's rights and responsibilities

The Archive agrees to:

- Preserve the data and documentation according to international standard techniques;
- Make the data and documentation available to registered users for analytic purposes;
- Draw the following information to the attention of the user as part of the registration and data delivery process:
  1) Respect national law on data protection
  2) Cite the data source (including the institution and researchers that produced the data)

[109] GESIS: Archive agreement: http://www.gesis.org/fileadmin/upload/institut/wiss_arbeitsbereiche/datenarchiv_analyse/Archivierungsvertrag_GESIS_Datenarchiv_v9__englisch.pdf, UK Data Archive: license agreement: http://ukdataservice.ac.uk/media/28102/licenceform.pdf

3) Use the data only for their own purposes

The Archive has the right to:

- Translate or reformat datasets to ensure their preservation and future accessibility
- Make copies of deposited materials for security and backup purposes
- Provide copies of the archived data to third parties for analytic purposes
- Promote the reuse of the archived data by publishing information on the data
- To use the archived data for analytic or pedagogic purposes and to publish the results

5. Special conditions for depositing the data

Depositor (Institution/Organisation):	The Archive
Place: “PLACE “	
Date: “DATE”	Sarajevo, the “DATE”

A1.2 Data deposit agreement SR-SSDC[110]

1. Parties of the agreement

(Depositor) Research Institution: NAME, ADDRESS, E-MAIL, TELEPHONE, FAX

and

Data Archive: Serbian Social Sciences Data Centre, ADDRESS, E-MAIL, TELEPHONE, FAX

2. Data to be archived

"NAME OF PROJECT"

Under the direction of "NAME OF RESEARCHER"

"REFERENCE NUMBER (if any)"

List of deposited files:

3. Depositor's rights and responsibilities

The depositor agrees to:

- Have collected and treated the data in conformity with existing laws on data protection, including the anonymisation of the data;
- Deposit them at "ADDRESS OF ARCHIVE".

The depositor has the right to:

- The copyright on the data mentioned in point 2 ;
- Be informed about any reuse of the data by third parties.

4. The Archive's rights and responsibilities

The Archive agrees to:

- Preserve the data and documentation according to international standard techniques;
- Make the data and documentation available to registered users for analytic purposes;
- Draw the following information to the attention of the user as part of the registration and data delivery process:
  1) Respect national law on data protection
  2) Cite the data source (including the institution and researchers that produced the data)
  3) Use the data only for their own purposes

[110] GESIS: Archive agreement: http://www.gesis.org/fileadmin/upload/institut/wiss_arbeitsbereiche/datenarchiv_analyse/Archivierungsvertrag_GESIS_Datenarchiv_v9__englisch.pdf, UK Data Archive: License agreement: http://ukdataservice.ac.uk/media/28102/licenceform.pdf.

The Archive has the right to:

- Translate or reformat datasets to ensure their preservation and future accessibility
- Make copies of deposited materials for security and backup purposes
- Provide copies of the archived data to third parties for analytic purposes
- Promote the reuse of the archived data by publishing information on the data
- To use the archived data for analytic or pedagogic purposes and to publish the results

5. Special conditions for depositing the data

Depositor (Institution/Organisation):	The Archive
Place: “PLACE “	
Date: “DATE”	Belgrade, the “DATE”

A1.3 Data deposit agreement CRO-DAS

1. **Parties of the agreement**

(Depositor) Research Institution: NAME, ADDRESS, E-MAIL, TELEPHONE, FAX

and

Data Archive: Croatian Data Archive for the Social Sciences, ADDRESS, E-MAIL, TELEPHONE, FAX

2. **Data to be archived**

"NAME OF PROJECT"

Under the direction of "NAME OF RESEARCHER"

"REFERENCE NUMBER (if any)"

List of deposited files:

3. **Depositor's rights and responsibilities**

The depositor agrees to:

- Have collected and treated the data in conformity with existing laws on data protection, including the anonymisation of the data;
- Deposit them at "ADDRESS OF ARCHIVE".

The depositor has the right to:

- The copyright on the data mentioned in point 2 ;
- Be informed about any reuse of the data by third parties.

4. **The Archive's rights and responsibilities**

The Archive agrees to:

- Preserve the data and documentation according to international standard techniques;
- Make the data and documentation available to registered users for analytic purposes;
- Draw the following information to the attention of the user as part of the registration and data delivery process:

1) Respect national law on data protection
2) Cite the data source (including the institution and researchers that produced the data)
3) Use the data only for their own purposes

The Archive has the right to:

- Translate or reformat datasets to ensure their preservation and future accessibility
- Make copies of deposited materials for security and backup purposes
- Provide copies of the archived data to third parties for analytic purposes

- Promote the reuse of the archived data by publishing information on the data
- To use the archived data for analytic or pedagogic purposes and to publish the results

5. Special conditions for depositing the data

Depositor (Institution/Organisation):	The Archive
Place: “PLACE “	
Date: “DATE”	Zagreb, the “DATE”

Annex 2: End user licenses[111]

A2.1 End user license USDAS

I (the "Data User") agree to the following conditions of use

1. To use the data cited above and related instruments only for scientific research and education within an academic framework, and for no other purpose
2. To use the data with respect to standard norms for data protection and confidentiality, and not to attempt to identify, individuals, households or organisations in the data
3. To quote all used documents according to scientific standards and to inform USDAS of all publications based on this material
4. Copies of data may only be made for personal or instructional use and may not be sold or used for commercial purposes
5. To not transmit these data to third parties, whether in the same or modified form, whether paid or for free except in the framework of academic teaching (statistical or methodological courses) and after obtaining consent of USDAS for doing so
6. After completion of the project the obtained data will be deleted and the medium carrying the data will be made unreadable. In case further usage is intended, the user must apply to USDAS for new use permission
7. That any breach of the agreement will lead to immediate termination of access to the services and could result in legal action

Additional conditions for highly sensitive data:

The cited data have not been anonymised or contain information easily permitting identification of individual cases. Therefore, use of the data requires special care with respect to data protection. In addition to the points above, **I agree to the following**:

1. To respect above all individual confidentiality, and not to divulge any information that might compromise this confidentiality
2. To transmit these conditions to third parties, if need be, who are eligible to work with these data within the framework of the project for which these data were obtained

The data user:

"ADDRESS"

Place: Date:

Signature:

[111] GESIS: Usage regulations: http://www.gesis.org/en/services/data-analysis/data-archive-service/usage-regulations/, UK Data Archive: Terms and Conditions of Access: http://ukdataservice.ac.uk/get-data/how-to-access/conditions.aspx, FORS: End user contract: http://forscenter.ch/wp-content/uploads/2013/11/end-user_contract_quanti_normal1.pdf.

A2.2 End user license SR-SSDC

I (the "Data User") agree to the following conditions of use

1. To use the data cited above and related instruments only for scientific research and education within an academic framework, and for no other purpose
2. To use the data with respect to standard norms for data protection and confidentiality, and not to attempt to identify, individuals, households or organisations in the data
3. To quote all used documents according to scientific standards and to inform SR-SSDC of all publications based on this material
4. Copies of data may only be made for personal or instructional use and may not be sold or used for commercial purposes
5. To not transmit these data to third parties, whether in the same or modified form, whether paid or for free except in the framework of academic teaching (statistical or methodological courses) and after obtaining consent of SR-SSDC for doing so
6. After completion of the project the obtained data will be deleted and the medium carrying the data will be made unreadable. In case further usage is intended, the user must apply to SR-SSDC for new use permission
7. That any breach of the agreement will lead to immediate termination of access to the services and could result in legal action

Additional conditions for highly sensitive data:

The cited data have not been anonymised or contain information easily permitting identification of individual cases. Therefore, use of the data requires special care with respect to data protection. In addition to the points above, **I agree to the following**:

1. To respect above all individual confidentiality, and not to divulge any information that might compromise this confidentiality
2. To transmit these conditions to third parties, if need be, who are eligible to work with these data within the framework of the project for which these data were obtained

The data user:

“ADDRESS”

Place: Date:

Signature:

A2.3 End user license CRO-DAS

I (the "Data User") agree to the following conditions of use

1. To use the data cited above and related instruments only for scientific research and education within an academic framework, and for no other purpose
2. To use the data with respect to standard norms for data protection and confidentiality, and not to attempt to identify, individuals, households or organisations in the data
3. To quote all used documents according to scientific standards and to inform CRO-DAS of all publications based on this material
4. Copies of data may only be made for personal or instructional use and may not be sold or used for commercial purposes
5. To not transmit these data to third parties, whether in the same or modified form, whether paid or for free except in the framework of academic teaching (statistical or methodological courses) and after obtaining consent of CRO-DAS for doing so
6. After completion of the project the obtained data will be deleted and the medium carrying the data will be made unreadable. In case further usage is intended, the user must apply to CRO-DAS for new use permission
7. That any breach of the agreement will lead to immediate termination of access to the services and could result in legal action

Additional conditions for highly sensitive data:

The cited data have not been anonymised or contain information easily permitting identification of individual cases. Therefore, use of the data requires special care with respect to data protection. In addition to the points above, **I agree to the following**:

1. To respect above all individual confidentiality, and not to divulge any information that might compromise this confidentiality
2. To transmit these conditions to third parties, if need be, who are eligible to work with these data within the framework of the project for which these data were obtained

The data user:

“ADDRESS”

Place: Date:

Signature:

Annex 3: Finance and budgets

A3.1 Finance and budget USDAS

The aim of this annex is to outline the financial requirements after the establishment of digital data archives in Bosnia and Herzegovina. It describes the needs of the data archive in order to guarantee the functioning of its services and to ensure the digital preservation of its assets over a period of five years in person months (PM). The staffing figures are calculated in person months since this approach is independent of the current wage level of a particular country.

The outline is heavily based on the KRDS Activity Model put forward by Beagrie, Ch. et al. It describes the full range of activities required to support long-term preservation of research data, and supports the allocation of costs across these activities.[112]

Final calculations for costs will depend on the following key variables:

- Staffing (based on roles and responsibilities framework of the charter document of each data centre);
- Software, hardware, servers and other equipment;
- Economic adjustment such as rates of inflation

The formation phase of the data archives can be divided into two different stages: the start-up phase and the operational phase in which the cost profiles will change over a period of months or years. The start-up phase reflects the run up of activities such as the recruitment of staff and specific start-up activities such as training and further development of policies and procedures for the archive. The costs in terms of staff time in the start-up phase can be quite substantial, whereas the operational phase will reflect increasing efficiency as procedures become increasingly established and normalized.

Future development

For more detailed operations planning, there are several more in-depth analyses available to be carried out. These include, for instance, a life-cycle costing analysis by modelling a life-cycle for a specific process and then identifying cost drivers (variables that affect the costs of the activity such as data volumes and file formats), and resources (staff time, equipment etc.) in order to provide a detailed understanding of costs for that process.

[112] Beagrie, Ch. et al (2011) User Guide for Keeping Research Data Safe http://www.beagrie.com/static/resource/KeepingResearchDataSafe_UserGuide_v2.pdf. An extensive list about digital preservation and data curation costing and cost modelling of which the report by Beagrie is part of can be accessed under the following link: http://wiki.opf-labs.org/display/CDP/Home.

Function	Rate	Quantity	PM	Change	Y1	Y2	Y3	Y4	Y5	Total cost
Staff										
• Head of USDAS	50%	1	6		6	6	6	6	6	30 PM
• Data expert (junior)	60%	1	7.2		7.2	7.2	7.2	7.2	7.2	36 PM
• Data expert (senior)	50%	1	6		6	6	6	6	6	30 PM
• IT expert	30%	1	3.6		3.6	3.6	3.6	3.6	3.6	18 PM
										162 PM
Extra cost (including software, servers, hardware, additional equipment, backup and storage)										

Table 4: Finance and budget USDAS

A3.1 Finance and budget SR-SSDC

The aim of this annex is to outline the financial requirements after the establishment of digital data archives in Serbia. It describes the needs of the data archive in order to guarantee the functioning of its services and to ensure the digital preservation of its assets over a period of five years in person months (PM). The staffing figures are calculated in person months since this approach is independent of the current wage level of a particular country.

The outline is heavily based on the KRDS Activity Model put forward by Beagrie, Ch. et al. It describes the full range of activities required to support long-term preservation of research data, and supports the allocation of costs across these activities.[113]

Final calculations for costs will depend on the following key variables:

- Staffing (based on roles and responsibilities framework of the charter document of each data centre);
- Software, hardware, servers and other equipment;
- Economic adjustment such as rates of inflation

The formation phase of the data archives can be divided into two different stages: the start-up phase and the operational phase in which the cost profiles will change over a period of months or years. The start-up phase reflects the run up of activities such as the recruitment of staff and specific start-up activities such as training and further development of policies and procedures for the archive. The costs in terms of staff time in the start-up phase can be quite substantial, whereas the operational phase will reflect increasing efficiency as procedures become increasingly established and normalized.

Future development

For more detailed operations planning, there are several more in-depth analyses available to be carried out. These include, for instance, a life-cycle costing analysis by modelling a life-cycle for a specific process and then identifying cost drivers (variables that affect the costs of the activity such as data volumes and file formats), and resources (staff time, equipment etc.) in order to provide a detailed understanding of costs for that process.

[113] Beagrie, Ch. et al (2011) User Guide for Keeping Research Data Safe http://www.beagrie.com/static/resource/KeepingResearchDataSafe_UserGuide_v2.pdf. An extensive list about digital preservation and data curation costing and cost modelling of which the report by Beagrie is part of can be accessed under the following link: http://wiki.opf-labs.org/display/CDP/Home.

Function	Rate	Quantity	PM	Change	Y1	Y2	Y3	Y4	Y5	Total cost
Staff	130%	3	15.6		15.6	15.6	15.6	15.6	15.6	78 PM
• Head of SR-SSDC	50%	1	6							30 PM
• Data expert	50%	1	6		6	6	6	6	6	30 PM
• IT support	30%	1	3.6		3.6	3.6	3.6	3.6	3.6	18 PM
										156 PM
Extra cost (including software, servers, hardware, additional equipment, backup and storage)										

Table 5: Finance and budget SR-SSDC

A3.3 Finance and budget CRO-DAS

The aim of this annex is to outline the financial requirements after the establishment of digital data archives in Bosnia and Herzegovina. It describes the needs of the data archive in order to guarantee the functioning of its services and to ensure the digital preservation of its assets over a period of five years in person months (PM). The staffing figures are calculated in person months since this approach is independent of the current wage level of a particular country.

The outline is heavily based on the KRDS Activity Model put forward by Beagrie, Ch. et al. It describes the full range of activities required to support long-term preservation of research data, and supports the allocation of costs across these activities.[114]

Final calculations for costs will depend on the following key variables:

- Staffing (based on roles and responsibilities framework of the charter document of each data centre);
- Software, hardware, servers and other equipment;
- Economic adjustment such as rates of inflation

The formation phase of the data archives can be divided into two different stages: the start-up phase and the operational phase in which the cost profiles will change over a period of months or years. The start-up phase reflects the run up of activities such as the recruitment of staff and specific start-up activities such as training and further development of policies and procedures for the archive. The costs in terms of staff time in the start-up phase can be quite substantial, whereas the operational phase will reflect increasing efficiency as procedures become increasingly established and normalized.

Future development

For more detailed operations planning, there are several more in-depth analyses available to be carried out. These include, for instance, a life-cycle costing analysis by modelling a life-cycle for a specific process and then identifying cost drivers (variables that affect the costs of the activity such as data volumes and file formats), and resources (staff time, equipment etc.) in order to provide a detailed understanding of costs for that process.

[114] Beagrie, Ch. et al (2011) User Guide for Keeping Research Data Safe http://www.beagrie.com/static/resource/KeepingResearchDataSafe_UserGuide_v2.pdf. An extensive list about digital preservation and data curation costing and cost modelling of which the report by Beagrie is part of can be accessed under the following link: http://wiki.opf-labs.org/display/CDP/Home.

Function	Rate	Quantity	PM	Change	Y1	Y2	Y3	Y4	Y5	Total cost
Staff										
• Head of CRO-DAS	50%	1	6		6	6	6	6	6	30 PM
• Data expert (junior)	60%	1	7.2		7.2	7.2	7.2	7.2	7.2	36 PM
• Data expert (senior)	50%	1	6		6	6	6	6	6	30 PM
• IT expert	30%	1	3.6		3.6	3.6	3.6	3.6	3.6	18 PM
										162 PM
Extra cost (including software, servers, hardware, additional equipment, backup and storage)										

Table 6: Finance and budget CRO-DAS

Annex 4: Performance indicators

A4.1 Performance indicators, USDAS

The USDAS internal performance monitoring system is based on a conceptual framework that links overall institutional **objectives** to key **activities**, and then to specific **indicators**. It can be used over time to assess progress toward objectives.

- Objective 1: Support, enhance, and contribute to social science research in Bosnia and Herzegovina
- Objective 2: Strengthen the social sciences in Bosnia and Herzegovina by creating linkages and facilitating synergies between researchers and institutions within the country and internationally

The activities within the monitoring system represent the range of USDAS projects and services. The first two correspond to objective 1, while the third corresponds to the second objective. Indicators were selected based on their capacity to measure the extent to which USDAS is carrying out its activities toward achieving its objectives over time.

Activity 1: Prepare, document, and provide data

1. *Number of datasets distributed to or downloaded by researchers*

This indicator assesses the intensity of use of secondary data acquired by way of the USDAS. The measure is a composite of datasets distributed either by the Nesstar server or FTP.

2. *Number of visits to USDAS webpages*

This indicator is an indirect measure of researcher interest in and use of USDAS services.

3. *Number of teachers using USDAS data for teaching purposes*

This indicator allows evaluation of the use of data from USDAS for education in the social sciences. It is limited to university instruction in Bosnia and Herzegovina.

Activity 2: Advise researchers

1. *Number of cases of help and advice provided to users or to others requesting information or expertise.*

USDAS regularly provides consultation, information, and advice in response to calls and contacts from researchers and other external parties seeking help or expertise. This indicator measures the overall level of this activity (which has been documented throughout the year, and provides details according to caller characteristics.

Activity 3: Establish collaborative links and stimulate scientific exchange

1. *Workshops, seminars, and conferences organised and conducted by USDAS*

The indicator gives a sense of USDAS efforts in establishing, maintaining, and developing scientific networks nationally and internationally.

A4.2 Performance indicators, SR-SSDC

The SR-SSDC internal performance monitoring system is based on a conceptual framework that links overall institutional **objectives** to key **activities**, and then to specific **indicators**. It can be used over time to assess progress toward objectives.

- Objective 1: Support, enhance, and contribute to social science research in Serbia
- Objective 2: Strengthen the social sciences in Serbia by creating linkages and facilitating synergies between researchers and institutions within the country and internationally

The activities within the monitoring system represent the range of SR-SSDC projects and services. The first two correspond to objective 1, while the third corresponds to the second objective. Indicators were selected based on their capacity to measure the extent to which SR-SSDC is carrying out its activities toward achieving its objectives over time.

Activity 1: Prepare, document, and provide data

1. *Number of datasets distributed to or downloaded by researchers*

This indicator assesses the intensity of use of secondary data acquired by way of the SR-SSDC. The measure is a composite of datasets distributed either by the Nesstar server or FTP.

2. *Number of visits to SR-SSDC webpages*

This indicator is an indirect measure of researcher interest in and use of SR-SSDC services.

3. *Number of teachers using SR-SSDC data for teaching purposes*

This indicator allows evaluation of the use of data from SR-SSDC for education in the social sciences. It is limited to university instruction in Serbia.

Activity 2: Advise researchers

1. *Number of cases of help and advice provided to users or to others requesting information or expertise.*

SR-SSDC regularly provides consultation, information, and advice in response to calls and contacts from researchers and other external parties seeking help or expertise. This indicator measures the overall level of this activity (which has been documented throughout the year, and provides details according to caller characteristics.

Activity 3: Establish collaborative links and stimulate scientific exchange

1. *Workshops, seminars, and conferences organised and conducted by SR-SSDC*

The indicator gives a sense of SR-SSDC efforts in establishing, maintaining, and developing scientific networks nationally and internationally.

A4.3 Performance indicators, CRO-DAS

The CRO-DAS internal performance monitoring system is based on a conceptual framework that links overall institutional **objectives** to key **activities**, and then to specific **indicators**. It can be used over time to assess progress toward objectives.

- Objective 1: Support, enhance, and contribute to social science research in Croatia
- Objective 2: Strengthen the social sciences in Croatia by creating linkages and facilitating synergies between researchers and institutions within the country and internationally

The activities within the monitoring system represent the range of CRO-DAS projects and services. The first two correspond to objective 1, while the third corresponds to the second objective. Indicators were selected based on their capacity to measure the extent to which CRO-DAS is carrying out its activities toward achieving its objectives over time.

Activity 1: Prepare, document, and provide data

1. *Number of datasets distributed to or downloaded by researchers*

This indicator assesses the intensity of use of secondary data acquired by way of the CRO-DAS. The measure is a composite of datasets distributed either by the Nesstar server or FTP.

2. *Number of visits to CRO-DAS webpages*

This indicator is an indirect measure of researcher interest in and use of CRO-DAS services.

3. *Number of teachers using CRO-DAS data for teaching purposes*

This indicator allows evaluation of the use of data from CRO-DAS for education in the social sciences. It is limited to university instruction in Croatia.

Activity 2: Advise researchers

1. *Number of cases of help and advice provided to users or to others requesting information or expertise.*

CRO-DAS regularly provides consultation, information, and advice in response to calls and contacts from researchers and other external parties seeking help or expertise. This indicator measures the overall level of this activity (which has been documented throughout the year, and provides details according to caller characteristics.

Activity 3: Establish collaborative links and stimulate scientific exchange

1. *Workshops, seminars, and conferences organised and conducted by CRO-DAS*

The indicator gives a sense of CRO-DAS efforts in establishing, maintaining, and developing scientific networks nationally and internationally.