Data Service Infrastructure for the Social Sciences and Humanities

EC FP7

Grant Agreement Number: 283646

Deliverable Report

Deliverable: D4.4
Deliverable Name: Comprehensive Policy-Rules for Data - Management in SSH
Deadline: 30 June 2014
Nature: R

Responsible: Norwegian Social Science Data Services (NSD)
Work Package Leader: Norwegian Social Science Data Services (NSD)

Contributing Partners and Editors: Vigdis Kvalheim (NSD), Trond Kvamme (NSD), Koenraad De Smedt (UiB), Bamba Dione (UiB), Carla Parra (UiB), Astrid Recker (GESIS), Claudia Engelhardt (UGOE), Tuomas J. Alaterä (FSD).

Abstract

A vital tool to sustain long-term preservation and accessibility of data is to provide a robust, explicit and declarative set of institutional policy-rules and requirements that can build a solid platform of trust between stakeholders involved in the creation, curation and dissemination of research outputs. A key institution in the stakeholder taxonomy of policy formation and long-term preservation is the data centre as a link between the funders and the researcher and a service centre for long-term preservation and endurable access to research data and documentation. As a key stakeholder data centres have to develop a transparent set of policy-rules and procedures that support internal data management procedures and ensure accountability and allow for external quality control. Accountability and transparency are key factors for creating *trust* in deposit service providers by funders and researchers.

A comprehensive policy framework should take into consideration both the wider strategic policies of the institution and closely connected network institutions. A preservation policy is a vital tool to establish the boundaries within which an institution operates: it supports the shorter-term management of the institutional activities while also taking into account the longer-term vision of operational activities.

The report assesses a selection of state-of-the-art guidelines and recommendations for formation of preservation policies, and describe, compare and analyse the scope of policy-rules and the requirements they set for the SSH domain. Based on the assessment of the policy rules and procedure implemented by selected data archives and services the report goes on to recommend a set of *policy-rules covering the full scope of a well-defined preservation policy.*

Table of Contents

1. Introduction, Background and Rationale

In recent years there has been an explosive growth in the amount of data generated within the various research disciplines. Data constitutes the raw material of scientific output and understanding, and the assembling, scrutinizing, organizing and disseminating of data serve an important purpose for the scientific community and the general public. In the words of the European Science Foundation (ESF): "….data sets are an important resource, which enable later verification of scientific interpretation and conclusions. They may also be the starting point for further studies. It is vital, therefore, that all primary and secondary data are stored in a secure and accessible form" [1].

As data are an important resource for verification and growth of knowledge, data should not only be stored. It has to be *shared* in an 'accessible form'. This has been further underlined by the ICSU-CODATA body (International Council for Science - Committee on Data for Science and Technology), which has stated, through a set of principles for dissemination of scientific data, that:

> *"…scientific advances rely on full and open access to data. Both science and the public are well served by a system of scholarly research and communication with minimal constraints on the availability of data for further analysis. The tradition of full and open access to data has led to breakthroughs in scientific understanding, as well as to later economic and public policy benefits. The idea that an individual or organization can control access to or claim ownership of the facts of nature is foreign to science".*
>
> *[…]Given the substantial investment in data collection and its importance to society, it is equally important that data are used to the maximum extent possible. Data that were collected for a variety of purposes may be useful to science. Legal foundations and societal attitudes should foster a balance between individual rights to data and the public good of shared data"*[2].

The combination of growth in data and information, and the heightened awareness of the importance of openness and usability of data have led to significant challenges with respect to data creation, management, curation, access and sharing, and long-term data preservation. Some of the most important challenges were summed up by the High Level Expert Group on Scientific Data, in a report to the European Commission in 2010[3]:

- How will we preserve the data? What will be the point of storing all this scientific data if, a century from now, it has degraded, been corrupted, or is simply too difficult for anyone but a well-equipped expert to use? Over time non-maintainability of essential hardware, software or support environment may make the information inaccessible and/or users may become unable to understand or use the data.

[1] European Science Foundation Policy Briefing, December 2000: *Good scientific practice in research and scholarship*: http://www.esf.org/fileadmin/Public_documents/Publications/ESPB10.pdf

[2] ICSU/CODATA (2000): ACCESS TO DATABASES - A Set of Principles for Science in the Internet Era: http://www.icsu.org/publications/icsu-position-statements/access-to-databases/389_DD_FILE_ACCESS_TO_DATABASES_Jun_00_.pdf

[3] European Union: *Riding the wave. How Europe can gain from the rising tide of scientific data. Final report of the High Level Expert Group on Scientific Data. October 2010.* http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf

- How will we protect the integrity of the data? As the 'data tide rises' higher, how will we detect unauthorised alterations? Should every researcher, and every citizen, have access to the data repositories? Should there be different levels of access allowed?

- How will we convey the context and provenance of the data? Given the emerging trend to make all publicly funded research data publicly available, just how will users from a wide range of backgrounds understand and query the data they are accessing, and recognise the special circumstances under which it was collected?

- What new funding and business models will we need, so that everyone – researchers, enterprises, citizens – has adequate incentive to contribute to the data infrastructure? What kinds of data, under what circumstances, should be free?

- How will we protect the privacy of individuals linked to the data on the one hand, while providing researchers access to vital data on the other hand?

To put it briefly, there is a need for a more robust approach to creating, maintaining and preserving the outputs of research to ensure they can be shared and reused.

There is a growing awareness among stakeholders – ranging from international organisations and interest groups, national science organisations, research funding bodies; to data centres, data archives, universities and researchers – that creation and management of research output should not be thought of as separate and independent processes but as integrated parts of a larger system of research output management.

There is also a growing realisation that to work across research disciplines and technical platforms, stakeholders are needed to produce and adopt universal rules and standards for description of data and research outputs; to define and refine the extent and content of data curation services; and to identify rules for data processing and security that are designed for use across different disciplines and technological platforms.

A crucial means to this end are *policies* and *strategies* providing guidance and orientation on the operational level. The consistency they create makes them essential to achieve long-term access to research data and to creating *trust* among stakeholders. Policy and strategy models to achieve better access to research data are being suggested and put in force on several stakeholder levels. To create synergies between these different levels, policies and policy tools are increasingly considered as parts of a larger stakeholder framework. Policies do not exist in isolation; they are a part of a wider process that involves both internal and external actors. A comprehensive policy framework should take into consideration both the wider strategic policies of the institution and closely connected network institutions. It is a vital tool to establish the boundaries within which an institution operates: it supports the shorter-term management of the institutional activities while also taking into account the longer-term vision of operational activities.

1.1. Outline, Goals and Objectives

This report was produced in the context of the DASISH project (Data Service Infrastructure for the Social Sciences and Humanities), task 4.4, "Recommendation of a set of policy-rules". The goal of this

task has been to map, describe, compare and analyse the scope of policy-rules and the requirements they set for the SSH domain, in particular in Europe and the US, and to create and establish a set of concrete policy-rules that will support the integrity of data and build trust in data preservation services.

In **Chapter 1** we look at the methodology (**1.2**) and the terminology (**1.3**) that has been applied in this report. Segment 1.3 includes an examination and discussion of the different utilisations of the term 'policy' and 'preservation policy'.

In **Chapter 2** we describe and assess the current European data preservation and data sharing context (segment **2.1** and **2.2**) by looking at a selection of initiatives that highlight the issues of *openness* on the one hand and *integrity* on the other hand. We aim to show that this duality is what drives the development of preservation policies and policy frameworks. In section **2.3** we map the wider *stakeholder context* within which a preservation policy has to be developed. This is done by examining recent data infrastructure projects that lay out 'taxonomies of stakeholders' within scientific research (i.e. the identification and categorisation of stakeholders). Further, we assess selected projects and surveys that shed light on the use of polices in data funding and in data preservation services. We look at the current situation and use of policies at the level of the data preservation centre, while drawing some parallels to the policy development within the research funder level as well.

Chapter 3 summarises the findings from our assessment of the policies and policy frameworks. First, we provide an overview of the assessed policy models and summarise their characteristics (**3.1**). Next we describe and analyse the different guidelines (**3.2**) and data services (**3.3**). In the latter we discuss the different service provider characteristics before assessing the different preservation policy characteristics.

Finally, in **Chapter 4** we provide some general considerations and recommendations (**4.1**), in addition to more concrete policy-rules recommendations (**4.2**). Case studies are presented in **Appendix 1** (CLARIN) and **Appendix 2** (EUDAT). All the individual policy models and frameworks are described and presented in detail in **Appendix 3**, **4** and **5**.

The report has the following objectives:

- Map, describe, compare and analyse the scope of policy-rules and their requirements for the SSH domain, particularly in Europe and the US, by looking at infrastructure projects, initiatives and service provider polices.
- To provide a snapshot of the current situation in the area of preservation policy development by bringing these different models, guidelines and policy examples into one source for easy comparison for potential users.
- Provide some general considerations regarding all stakeholders in the preservation strategy stakeholder taxonomy, from research funders to data curators and researchers.
- Provide a general policy model and specific policy-rules that may provide a foundation for existing and emerging data preservation initiatives that lack a coherent preservation policy.

1.2. Methodology

The principal methodology for this report has been a content analysis of existing policy and procedure templates, in addition to concrete policy examples from a selection of service providers. Each policy model was analysed and described individually before the different policy models and policy-rule elements were put into a matrix and compared to identify similarities and differences. Each policy framework has been analysed in detail before being summarised in tables (see appendices 1 and 2). In addition to the shorter assessment and summaries of service providers, we have carried out one in-depth study of one of the cases. CLARIN, an ERIC on the ESFRI roadmap, provides a special case as it is a distributed infrastructure connecting repositories in many countries. It therefore does not have a centralized preservation policy, as opposed to most of the other service providers we have looked at. Rather, it has a number of service level requirements in order to be certified as a CLARIN centre, while allowing individual centres to implement and adjust their service offers in different ways[4].

In addition, we made use of results and findings both within the DASISH project and from other recent data infrastructure projects that have policy or policy-rules as one of their main topics. In the DASISH project we have specifically drawn on the model of trust[5] and the survey of European data archive services[6].

Based on our assessments and findings we constructed a set of policy-rules in addition to some general considerations that should be taken into account when developing and refining a preservation policy. It should be noted that the selected guidelines and policy models are mostly situated in Europe and North America. One of the main reasons for this is that in these areas the development of research infrastructures and data preservation service providers has reached a certain level of maturity compared to other parts of the world. However, the selection of services, and the final set of recommendations and considerations, will only represent a snapshot of the current situation; the data infrastructure landscape is rapidly changing and new tools and service providers dissolve or emerge frequently.

1.3. Terminology and concepts

1.3.1. Data Centre

Digital preservation is a subject that interests a range of different communities, often with a distinct vocabulary and 'local' definitions for key terms. Hence, it may be helpful to draw attention to and clarify the usage of some of the key terms that are being used in this report. In general, many of the key terms have been adopted from either the OAIS Reference Model[7] and/or the ISO 16363[8]. The OAIS model aims to be "…applicable to all disciplines and organizations that do, or expect to, preserve and provide information in digital form". Hence, it uses the term 'digital archive' rather broadly as 'the organization responsible for digital preservation'.

[4] Similar measures apply to CESSDA Service Providers which have to comply with the set of service level requirement set out in the CESSDA Statutes to qualify for CESSDA membership.
[5] DASISH Deliverable 4.1: *Roadmap for Preservation and Curation in the SSH*.
http://dasish.eu/publications/projectreports/D4.1_-_Roadmap_for_Preservation_and_Curation_in_the_SSH.pdf
[6] DASISH Deliverable 4.3: *Scope and Characteristics of Data Archive Services within the DASISH Communities.*
To be published by the end of 2014.
[7] CCSDS Reference Model for an Open Archival Information System: http://public.ccsds.org/publications/archive/650x0m2.pdf
[8] Magenta draft of ISO 16363: CCSDS 652.0-M1 (2011): Audit and Certification of Trustworthy Digital Repositories:
http://public.ccsds.org/publications/archive/652x0m1.pdf

In this report, the terms 'data centre', 'data service', 'repository', 'digital repository', or 'archive' are used to convey a similar concept as OAIS definition of 'digital archive'. That is, the concept applies to all disciplines, organisations, initiatives and services that that have, or contribute to, long-term preservation responsibilities and functionality.

1.3.2. Preservation policy

In the Oxford Dictionaries[9] a policy is rather broadly defined as "…a *course* or *principle of action* adopted or proposed by an organization or individual*".* In the Merriam-Webster[10] dictionary it is defined as (1) "…prudence or *wisdom* in the management of affairs", or (2) "…a definite *course or method of action* selected from among alternatives and in light of given conditions to *guide and determine present and future decisions*". Alternatively it can be defined as (3)"…a *high-level overall plan* embracing the general *goals* and acceptable *procedures* ", or (4) as "…a writing whereby a *contract of insurance* is made" (all emphasises are our own).

Based on these definitions we find that the content of a policy can range from high-level overall plan and organisational "wisdom", to the more concrete procedures and actions of the organisation including contractual arrangements ("insurance") between different actors. As we will see these are all vital elements in the formation of a preservation policy.

In the context of scientific research and the processing and storage of research output, a *preservation* policy is an important document and instrument demonstrating an organisation's commitment to the preservation of its digital collections. A well-defined policy ensures the *accountability* of the preservation organisation, and *transparency* is essential to accountability. Hence, accountability can be achieved through active, constant documentation. This is why well defined policies and transparent documentation are considered essential by several frameworks for trust. ISO 16363, DIN 31644[11] and the Data Seal of Approval[12] all underline *transparency* as an important metric to assure the organisations' trustworthiness.

The ISO 16363 defines preservation policy as a "…written statement, authorized by the repository management, that describes the approach to be taken by the repository for the preservation of objects accessioned into the repository" [13]. Further, it has a 'dualistic' approach by making a distinction between the preservation *policy* and the preservation *strategic plan* ("the preservation policy is consistent with the preservation strategic plan").

Other initiatives have sought a combined approach (i.e. merging goals and implementation) when defining and delimiting the extent of a preservation policy. Both the InterPARES[14] project and the APARSEN[15] project define a policy as "…a formal statement of direction or guidance as to how an organization will carry out its mandate, functions or activities, motivated by determined interests or programs"[16]. The definition of *preservation* is in this setting a broader term which comprises "…the whole of the principles, policies, rules and strategies aimed at prolonging the existence of an object

[9] http://www.oxforddictionaries.com/definition/english/policy
[10] http://www.merriam-webster.com/dictionary/policy
[11] Criteria for trustworthy digital archives: http://www.nabd.din.de/cmd?level=tpl-art-detailansicht&committeeid=54738855&artid=147058907&languageid=en
[12] http://datasealofapproval.org/en/
[13] CCSDS 652.0-M1 (2011): Audit and Certification of Trustworthy Digital Repositories: http://public.ccsds.org/publications/archive/652x0m1.pdf
[14] InterPARES Project: http://www.interpares.org/
[15] APARSEN: http://www.alliancepermanentaccess.org/
[16] InterPARES Terminology Database: http://www.interpares.org/ip2/ip2_terminology_db.cfm

by maintaining it in a condition suitable for use, either in its original format or in a more persistent format, while leaving intact the object's intellectual form" [17]. An earlier project, ERPANET[18] adds that among the primary aims of the preservation policy is "…to ensure the authenticity and reliability" of the digital objects and should include "…some principles and rules on specific aspects which then lay the basis of implementation"[19].

The Integrated Rule-Oriented Data System (iRODS), which is an open-source data management software, tries to position the policy in the wider context of the organisation. Here, through a 'multilevel' approach a policy is seen as part of a pyramid consisting of six levels of "functions of activities"[20]. On the highest and most abstract level is the *purpose* of an organisation, or the reason a collection is assembled. On the second level are the *properties*, or the attributes needed to ensure the purpose of the data centre. On the third level are the actual *policies*, which in the IRODS setting are defined as *controls for enforcing or implementing the desired properties* that can, but do not have to, be mapped to computer actionable rules. To enforce the desired properties the organisation needs to have a set of *procedures* in place, or functions that can assist in implementing the policies. These can also be mapped to computer executable workflows, but it is not a requirement. On the fifth level are the results of applying the procedures, or the *persistent state information,* which often is mapped to a metadata system. Finally there is the *property verification* level. That is, the validation that the state information conforms to the desired purpose (mapped to periodically executed policies). In this definition framework a policy is simply one of the parts of the broader sets of activity functions in the institution.

Figure 1: Functions of organisational activities in the iRODS framework[21]

In this model the policy is a clear and unambiguous statement of intent, a mission statement that supports the data archives, adds to their legitimacy and trust and allows the institutions to capture the purposes of their activities. Further the policy must contain a definition of how a given institutional activity should 'behave' (that is, it describes what the activity covers). Finally, it should contain a plan to guide the decisions through a precise, brief and unambiguous description of the

[17] Ibid.
[18] ERPANET (Electronic Resource Preservation and Access Network): http://www.erpanet.org/
[19] ERPANET: Digital Preservation Policy Tool: http://www.erpanet.org/guidance/docs/ERPANETPolicyTool.pdf
[20] Reagan W. Moore & Arcot Rajasekar (2012): *Policy Based Data management – iRODS*. http://irods.org/wp-content/uploads/2012/04/SC12-iRODS-overview.pdf
[21] Ibid.

processes and activities. Hence, the processes and activities start from the policies, and process descriptions should contain the detailed steps on how the policy will be implemented in the institution.

Inherent in this model there is a distinction between high-level statements (purpose and properties) on the one hand and concrete implementable activities (procedures, persistency and verification) on the other hand. High-level statements regulate fundamental constrains and strategies. They provide useful and important guidance, but they are limited to setting the framework for concrete planning rather than actually providing actionable steps for ensuring longevity of data holdings. A preservation plan (i.e. procedures, persistency and verification) on the contrary is more specific and concrete as it specifies an action plan for preserving a specific set of objects for a given purpose. The distinction between a preservation policy and a preservation plan was underlined in the PLANETS[22] project. Here, a preservation plan was defined as:

> "… a series of preservation actions to be taken by a responsible institution due to an identified risk for a given set of digital objects or records (called collection). The Preservation Plan takes into account the preservation policies, legal obligations, organisational and technical constraints, user requirements and preservation goals and describes the preservation context, the evaluated preservation strategies and the resulting decision for one strategy, including the reasoning for the decision. It also specifies a series of steps or actions (called preservation action plan) along with responsibilities and rules and conditions for execution on the collection. Provided that the actions and their deployment as well as the technical environment allow it, this action plan is an executable workflow definition."[23]

In this definition the plan is something separated from, though taking into account, the preservation policy. However, as we will show, it is possible (and common) to make the preservation plan *part of* the preservation policy through what we have called a *combined* policy approach[24].

Adding to the complexion of a preservation policy is the fact that it also needs to take into account and determine how external parties or stakeholders can interact with the services of the preservation organisation. That is, how such interactions can be formalized through contractual agreements and documents. Contractual agreements can include (but are not restricted to) User agreements; Terms of use; Legal policies; and Privacy policies.

Based on these definitions, aspects and complexities of policy content, certain reluctance to develop, adopt and commit to a full-scale policy framework is understandable. A transparent policy which can be accessed by users, partners and investors is a big commitment, and it can be quite difficult to determine the level of detail and decide on length and scope of a preservation policy. However, what organisations should keep in mind is that the content and level of detail of the policy and documentation always should be in accordance with the specific data and organisational environment and level of maturity, and the context in which the organisation operates.

---

[22] PLANETS - Preservation and Long-term Access through Networked Services: http://www.planets-project.eu/

[23] Christoph Becker, Hannes Kulovits, Mark Guttenbrunner, Stephan Strodl, Andreas Rauber, Hans Hofman (2009): "Systematic planning for digital preservation: evaluating potential strategies and building preservation plans". *International Journal on Digital Libraries*. DOI 10.1007/s00799-009-0057-1. http://www.ifs.tuwien.ac.at/~becker/pubs/becker-ijdl2009.pdf

[24] In much of the literature concerned with preservation planning a preservation *strategy* is something distinct from, or simply a part of, a preservation policy. A preservation strategy in this regard is the actual choice of preservation measures; i.e. whether to migrate or to emulate the preserved material. Throughout this report we consider the preservation strategy as an element *within* the policy, not as a policy in and by itself.

2. Preservation of Research Output in the European Context

The **OECD Principles and Guidelines for Access to Research Data from Public Funding**[25] were created to assist governments, research support and funding organisations, research institutions and researchers to overcome the barriers and challenges to international sharing of research data. OECD lists thirteen different principles, all of which are centred on two major themes, namely *openness* and *integrity*. That is, data should be made available for the international research community "at the lowest possible cost", while at the same time securing legal conformity and the quality and security of data. Several cross-national policy initiatives concerning preservation and data sharing in recent years have been concerned with this duality, or friction of interest, between openness and integrity of research output.

2.1. Preservation and openness

Some initiatives have focused mainly on openness. One of these is the **Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities**[26], which states that open access is a "comprehensive source of human knowledge and cultural heritage" and declares that open access contributions should include original scientific research results, raw data and metadata, source materials, digital representations of pictorial and graphical materials and scholarly multimedia material.

A similar initiative is the **Budapest Open Access Initiative (BOAI)**[27] which, in connection with its 10 year anniversary in 2012, issued a set of guidelines that aimed to "…usher in advances in the science, medicine and health". The guidelines cover areas such as policies, licensing and reuse, infrastructure and sustainability and advocacy and coordination. On the issue of policy it is recommended that institutions ranging from higher education to research funders should have a policy assuring that "peer-reviewed versions" of the research output are deposited in the institution's "designated" or "suitable" repository.

Another example is the **Open Access (OA) Pilot**[28], launched in 2008 and which objective is to optimise the impact of publicly funded scientific research through EU Research Framework Programmes (namely FP7 and Horizon 2020). In a report[29] from 2012 the European Commission (EC) addresses some of the major issues involved in data sharing and accessibility, namely that "…the lack of organisation and clarity about responsibilities in improving access to and use of scientific data are major barriers to change" and that the "financing models to ensure long-term access are often lacking" and that "interoperability among countries and disciplines remains an issue". In another

[25] OECD: *Principles and Guidelines for Access to Research Data from Public Funding*: http://www.oecd.org/sti/sci-tech/38500813.pdf

[26] Berlin Declaration: http://oa.mpg.de/lang/en-uk/berlin-prozess/berliner-erklarung/

[27] http://www.budapestopenaccessinitiative.org/

[28] Commission Policy Initiatives, Open Access: http://ec.europa.eu/research/science-society/index.cfm?fuseaction=public.topic&id=1294&lang=1

[29] EU Commission, COM(2012) 401: *Towards better access to scientific information*: *Boosting the benefits of public investments in research*: http://ec.europa.eu/research/science-society/document_library/pdf_06/era-communication-towards-better-access-to-scientific-information_en.pdf

report[30] released by the EC in 2012, some of the solutions to these issues are outlined. It is stated that open access should be rooted in explicit policies and that "…such policies are expected to improve conditions for conducting research by reducing duplication of efforts and by minimising the time spent searching for information and accessing it". Further, it is stated that a proper (open access) policy framework "…will speed up scientific progress and make it easier to cooperate across and beyond the EU" and respond to calls within the scientific community for "greater access to scientific information". An example of an initiative created under, and complying with, the EC OA pilot and the European Research Council Guidelines for Open Access[31], is the **OpenAIRE** project[32] which among other things aims to "…ensure localized help to researchers within their own context […], and provide a repository facility for researchers who do not have access to an institutional or discipline-specific repository".

These are just some examples of initiatives and guidelines among a wide array of bodies and movements that are aiming at promoting and promulgating the accessibility and use of research output. However, most of these initiatives are primarily focused on the immediate availability of the final research product (e.g. an article or a book) while putting less emphasis on the importance of the background material (the actual data) and the *long-term preservation* of all research output.

2.2. Preservation and integrity

The **UNESCO** report **Policy Guidelines for the development and promotion of open access**[33] embraces open access principles and provides an extensive policy framework for open access. But it also points to some of the challenges regarding the difference between scientific *publications* and scientific *data:*

> *"Research data are increasingly covered by policies and often these policies are being implemented by smaller, niche players as well as large research funders. These policies are not usually, however, the same (Open Access) policies that cover the text-based literature. Data are exceptional because policies must take into account issues of privacy and special cases where data cannot be released for other reasons. Developing and wording Open Data policies is therefore a specialised issue that is not as straightforward as developing policies for Open Access to the literature".*

The **Royal Society** has released a report[34] on open access which highlights the need to deal with the "deluge of data created by modern technologies" in order to preserve the principle of "intelligent openness" and use/reuse of data. In a similar fashion as the UNESCO report the Royal Society points out that although most scientific research becomes publicly (but not necessarily freely) available (e.g. via academic journals), the data that underlie the research are rarely provided with the same accessibility. It is pointed out that "…ideally, all the data that underlie the research or argument presented in an article, but which is not included for reasons of space, should be accessible electronically via a link in the article". Or alternatively, that "…the publication should indicate when

[30] EU Commission, C(2012) 4890: *Recommendations on access to and preservation of scientific information*: http://ec.europa.eu/research/science-society/document_library/pdf_06/recommendation-access-and-preservation-scientific-information_en.pdf

[31] ERC, 2007: http://erc.europa.eu/sites/default/files/document/file/erc_scc_guidelines_open_access.pdf

[32] https://www.openaire.eu/

[33] UNESCO, 2012: *Policy Guidelines for the development and promotion of open access*: http://unesdoc.unesco.org/images/0021/002158/215863e.pdf

[34] Royal Society, 2012. *Science as an open enterprise*. https://royalsociety.org/~/media/policy/projects/sape/2012-06-20-saoe.pdf

and how the data will be available for others to access". They also point out (though not providing numbers to back it up) that "…an increasing number of journals have explicit policies that require data to be made available, but the rates of compliance are low". They move on by providing a set of recommendations concerning the publication of data:

> *"As a condition of publication, scientific journals should progressively enforce requirements for traceable and usable data available through an article, when they are intrinsic to the arguments in that article. This should be in line with the practical limits for that field of research. Materials should be uploaded to a repository before publication of the article, though their release may be subject to a temporary embargo. The publication should indicate when, and the conditions under which data will be available for others to access".*

Among the detailed actions provided to implement these recommendations are "actively encourage the development of standards and protocols for accessing data" and move towards "…the development of journals devoted to data publication and support the development of wider best practice and common standard".

The EC OA pilot was launched on the basis of the "lack of organisation and clarity about responsibilities"[35] and the lack of financing models to ensure long-term access. Lack of organisational clarity and financial responsibilities is also mentioned by the **Blue Ribbon Task Force on Sustainable Digital Preservation and Access (BRTF-SDPA**). In a report[36] it points out that the potential downside for such an access policy is that if there is no provision for sustaining the data over time, preservation becomes an unfunded mandate. "Open access is like any other form of access: without preservation, there will be no access, open or otherwise".

Challenges for policy-driven data sharing are also pointed out in an **ESFRI** Roadmap report[37] from 2008, where it is pointed out that one of the major obstacles for access to empirical data in Europe "…is the multitude of data access policies and regulations implemented by national governments". To overcome these barriers and make data easily available for cross national research the report finds that "…a mapping of data resources in various countries is required followed by the establishment of harmonised access regulations"[38].

However, several studies have shown that researchers still find barriers to sharing and archiving of their data. DAMVAD[39], Tenopir[40], the European Commission[41], and the Parse-Insight[42] project all

[35] EU Commission, COM(2012).
[36] BRTF-SDPA, 2010: *Sustainable Economics for a Digital Planet - Ensuring Sustainable Economics for a Digital Planet*: http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf

[37] ESFRI: *Social Sciences and Humanities*, *Roadmap Working Group Report 2008*: http://ec.europa.eu/research/infrastructures/pdf/esfri/esfri_roadmap/roadmap_2008/ssh_report_2008_en.pdf#view=fit&pagemode=none

[38] ESFRI: *Roadmap 2008 (Update)*: http://ec.europa.eu/research/infrastructures/pdf/esfri/esfri_roadmap/roadmap_2008/esfri_roadmap_update_2008.pdf#view=fit&pagemode=none

[39] DAMVAD: *Sharing and archiving of publicly funded research data - Report to the Research Council of Norway*

[40] Tenopir, et.al. (2011): *Data Sharing by Scientists: Practices and Perceptions* http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0021101

[41] European Commission (2012): Online survey on scientific information in the digital age http://ec.europa.eu/research/science-society/document_library/pdf_06/survey-on-scientific-information-digital-age_en.pdf

[42] Parse Insight (2009): *Insight into digital preservation of research output in Europe, survey report*

confirm that many researchers are still undecided on the issue of sharing data. It seems that many researchers find sharing and archiving to be a difficult and complex issue. They express concern regarding the lack of incentives for sharing. Time for preparation and lack of infrastructure are other barriers. A survey conducted by DAMVAD on commission by the Norwegian Research Council, indicate that the barriers to sharing can be divided into three main categories, namely **legal**, **sociological** and **technical**. The most important *legal* issues (for the researchers) are concerns regarding privacy, the issue of shared ownership to data (IPR and copyright issues), and the lack of knowledge on legal issues related to data. The *sociological* barriers include issues such as lack of incentives/credit to researchers; concerns about freeriding; fear of losing control over data; fear of losing 'scientific edge'; and fear that others might not understand the data. On the *technical* side there are issues such as lack of infrastructure, concerns that the sharing of data is time-consuming, lack of standards for sharing and preparing metadata, and lack of technical skills. Overall, there seems to be a perception among researchers across all scientific fields that the preparation of research data for long-term archiving and sharing is burdensome and risky.

Hence, archiving and sharing of data involves a number of technical, financial, legal, ethical, motivational and cultural obstacles. While overall legal and ethical guidelines and stakeholder policy goals on data preservation, open access and sharing are agreed upon, many questions still stand in the way of effective and successful implementation of these principles.

2.3. Preservation policy stakeholders

2.3.1. Stakeholder taxonomy

The issues discussed in the preceding segment challenge a diverse group of stakeholders within the research and data preservation community. In recent infrastructure projects attempts have been made to introduce a taxonomy of stakeholders to define and clarify the roles and interests of various actors involved in the process of producing and processing research outputs. The **PARSEinsight** project[43], which ran from 2008 to 2010 under the Seventh Framework Programme of the European Union, identified four main stakeholders in research: *funders, data managers, researchers* and *publishers*. In this taxonomy *funders* are identified as responsible for, among other things, the wider policy perspectives by developing policies, either in co-operation with other stakeholders or by themselves; to monitor and enforce policies; and to act as advocates for data curation and fund expert advisory services[44]. The funder provides resources to the *researcher* who in turn provides the necessary research output to publishers and data managers. They are also potential consumers of other researchers' findings and output as well. As creators of data the researchers are responsible for a wide variety of activities connected to data, e.g. managing data for the duration of their project, making the data available in a form that can be used by others, complying with data policies and disseminating their research work by writing articles and other publications. Additionally, as users of data they are also responsible for adhering to any license and restrictions of use, acknowledging data

http://www.parse-insight.eu/downloads/PARSE-Insight_D3-4_SurveyReport_final_hq.pdf

[43] PARSEinsight (Permanent Access to the Records of Science in Europe), project website: http://www.parse-insight.eu/
[44] Ibid.

creators and curators, give proper citation and providing feedback to the research community and data archives[45].

The *data manager* category is defined as profit and non-profit data archives, traditional memory institutions (libraries, archives and museums) as well as research and development in preservation technology itself. The *publisher* category covers publishers of academic books and journals. Both the data manager and the publisher are regarded as the most important stakeholders for disseminating the research output.

One of the tasks in the RECODE[46] project, a current (2013-2015) project also running under the Seventh Framework Programme, has been to identify the stakeholders of the 'Open Access ecosystem' with special focus on stakeholders of open research data[47]. Similar to PARSEinsight this has been done through a definition of a *functional stakeholder taxonomy* where the stakeholders have been categorised and mapped with assumed values and interests associated with each group. The ecosystem is a functional taxonomy that consists of five entities or functions (layers). These five basic functions in the Open Access ecosystem: are 1) Funders & Initiators, 2) Creators, 3) Disseminators, 4) Curators and 5) Users. These functions are represented by different performers (stakeholders) that may operate and interact in the different layers at the same time. Each identified stakeholder within each layer is then being identified with having either a primary function or a secondary function within that specific layer. The stakeholders in this taxonomy range from research councils, foundations, civil society organisations and advocacy groups, to research institutes, universities/academies, IGO's, EU funded projects and media, to mention a few.

For our purposes, the point to be drawn from these taxonomies is the illustration of the complex network of actors and stakeholder in which a preservation policy is created. A preservation policy for the individual organisation does not exist in isolation. They need to be created to a large extent in accordance with both the wider strategic polices of the organisation, and with the *wider stakeholder framework*. A policy framework should be developed and matured through the *facilitation of stakeholder understanding and cooperation*. The challenge is of course how this facilitation should take place, especially when we take into consideration the wide differences in scope and content of the various stakeholder policies. That is, if they have a policy at all.

2.3.2. The use of policies in data funding

The SHERPA/JULIET[48] service of the University of Nottingham provides a registry of open access policies from research funders worldwide. The service is community driven and funding organisations with an open access or data archiving policy can inform about their policies. One of the entries is "Data archiving is required" (that is, data archiving is a requirement in the funder policy). Statistics show that of the 139 funders in the JULIET/SHERPA registry 24 % have a data archiving requirement in their policies, 12 % encourage archiving, while as many as 65 % do not have a data archiving policy[49]. And there seems to be a discrepancy between the *data archiving* requirements on the one hand, and the *publications archiving* requirements on the other: of the 139 funders registered, 65 % require publications archiving, while the funders that require data archiving are only at 24 %. This

[45] PARSEinsight, 2010: *D3.6: Community Insight Report*: http://www.parse-insight.eu/downloads/PARSE-Insight_D3-6_InsightReport.pdf
[46] RECODE (Policy Recommendations for Open Access to Research Data in Europe) project website: http://recodeproject.eu/
[47] RECODE: *Deliverable D1: Stakeholder Values and Ecosystems*: http://recodeproject.eu/wp-content/uploads/2013/10/RECODE_D1-Stakeholder-values-and-ecosystems_Sept2013.pdf
[48] http://www.sherpa.ac.uk/juliet/
[49] JULIET/SHERPA statistics, extracted October 2014: http://www.sherpa.ac.uk/juliet/stats.php?la=en&mode=simple

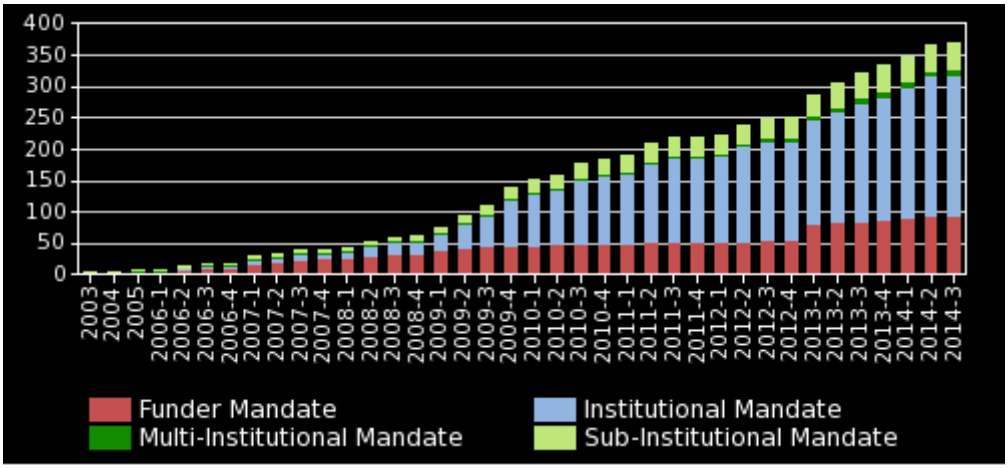
provides us with an indication of the current priorities of research funders: they seem to prioritise the preservation and sharing of articles and the final research output, but to a lesser extent seem to focus on the sharing of background data.

Although most of the funders registered in the SHERPA database are based in UK (70 %) it gives an indication that data archiving seems to be a low priority for many research funders. Even in the case of those who actually have a policy on data archiving and preservation, the policy statements are encouragements (to archive), not a requirement. It should be noted that one of the reasons that the SHERPA database is scant with funders from outside of Europe and the UK[50] is that infrastructures for long-term preservation and data sharing seem to be less developed in 'non-Western' countries. This was pointed out in an IFDO-report based on the results of a survey of research funders' data policies worldwide[51] released early in 2014. On the positive side the IFDO-report concludes that "…there is a growing awareness that research contributions and returns of public investments are restricted by lack of easy and open access to high quality data and an increasing political will to use strong incentives to improve this situation". However, the report also states that there is still a gap between high level policy statements and implementation and that "…enforcement of these policies and the required infrastructures to implement them are often lacking or immature and still in the process of development". Hence, although many countries and research organizations adhere to the principles of open access and oblige to follow various international open data declarations and data preservation obligations, the implementation of these obligations seems to vary significantly, or is to a great extent restricted to article or journal output, not the background material (i.e. raw data). The IFDO reports find that this is the case also in Western Europe and the US, which is considered to be in the forefront of developing open access policies and requirements. The findings also indicate that the institutional research infrastructures and data sharing requirements across the world are more developed within the social sciences than the within the humanities and medical and health sciences. These findings are supported by the assessments and survey results in DASISH WP4.

[50] Other than the fact that it is operated by a UK-based institution.
[51] International Federation of Data Organisations, 2014: *Policies for Sharing Research Data in Social Sciences and Humanities. A survey about research funders' data policies*. http://ifdo.org/wordpress/wp-content/uploads/2014/04/ifdo_survey_report.pdf

ROARMAP[52] provides an overview of the current situation by providing the number of institutions with mandatory open access repository policies in the world.

Figure 2: (data for years 2006 to 2014, onwards shown by year-quarter. Accessed September 2014).



This shows a significant growth in policies with open archive requirements, but only 90 of those listed in the ROARMAP registry are funder mandates, and most of the mandates cover only the written output of research (articles), not the data.

Similar findings have been done by a DCC study[53] on the range of policies required for and related to digital curation, and a US study[54] on current data management requirements of major US research funding agencies. Both studies found that the necessary policy information was rarely easy to find; polices were either 'dispersed' or not available online at all. Overall, the US study found that data policies were missing "…a significant number of the elements" of the pre-defined policy element framework. They also found that "…no single policy addressed all of the elements and that eleven policies addressed fewer than half of the elements, including four of the funders that appeared to have no policy at all".

2.3.3. The use of policies in data services

Several attempts have been made to gain insight into the awareness and extent of preservation policies in data centres, repositories and data archives. A survey conducted in DASISH (D4.3, to be released by the end of 2014) found that only about half of the respondents from data services mention the existence of a preservation policy. 26 (about half of the respondents) indicated that they have such a preservation policy. In 12 cases the policy is not available online (yet), and in 6 cases it was written in languages other than English. Further analysis of the answers revealed that details were not often described in the documents respondents referred to, and the information was fragmental and scattered.

But a policy might exist only internally: hence the policies are in many cases not accessible online or available in English. However, the survey shows that the majority of the data services have

[52] ROARMAP: Registry of Open Access Repositories Mandatory Archiving Policies: http://roarmap.eprints.org/

[53] http://www.dcc.ac.uk/sites/default/files/documents/reports/DCC_Curation_Policies_Report.pdf

[54] Dianne Dietrich, Trisha Adamus, Alison Miner, and Gail Steinhart. 2012. "De-Mystifying the Data Management Requirements of Research Funders". Issues in Science and Technology Librarianship, Summer 2012. DOI:10.5062/F44M92G2. http://www.istl.org/12-summer/refereed1.html

implemented deposit and user agreements and a majority of the archives have specified a long-term preservation strategy, in most cases migration[55]. However, the survey also reveals, through its design and selection of questions (separated into themes covering organisational context, deposit and ingest, archival storage and preservation, and dissemination) that many of the services analysed actually have the necessary knowledge and information. They simply have not developed and aggregated this information into an explicitly formulated and readily available preservation policy.

A different example is the 2011 DigCurV[56] survey of training needs, where more than 400 institutions were asked if/how they were engaged in the storing of digital material. One of the questions concerned the "…training needs with regard to strategic planning and skills". It turned out that of the 445 respondents that answered this question, 43.8% saw a 'great need' of such training among digital preservation staff, 43.6% indicated a 'moderate need', while the remaining percentage stated that there was 'hardly any' or no need for such training[57].

Recently, research has been carried out focusing on preservation polices in non-digital archives, libraries and museums. In 2013 a study was conducted at the National Digital Information Infrastructure and Preservation Program (NDIIPP) of the Library of Congress[58]. The study gathered English-language digital preservation policies, strategies, and/or plans, published by cultural heritage organizations. The study assesses what types of topics various institutions include in their policies and strategies, and to what extent they cover each element. From these findings the study developed a taxonomy based on most commonly cited elements. The study did not include "how to" guidelines or best practices. It is simply a template based on existing content.

Table 1: Policy elements taxonomy of the NDIIPP study

Policy element	Definition	Freq. count
Access/Use	Statement of principle which allows continued access/use of digital content	19
Accessioning/Ingest	Process through which digital objects are added into a digital repository	7
Audit	Internal/external audits conducted for authenticity/integrity	7
Bibliography	Bibliographic information included within document	13
Collaboration	Collaboration with external organizations to share/meet digital stewardship objectives	20
Content Scope	Defines digital content accepted within repository	24
Glossary/Terminology	Definitions of terminology used within digital stewardship community	17
Mandates	Digital Stewardship commitments/responsibilities to designated community	9
Metadata/Documentation	Metadata documented for preservation throughout lifecycle	10
Policy/Strategy Review	Periodic review of policy/strategy	13
Preservation Model/ Strategy	Proposed procedures for continued preservation of digital content	31
Preservation Planning	Monitor digital steward environment for changes in technology and standards/best practices to ensure long-term preservation of digital content	7
Rights and Restriction Management	Restrictions related to intellectual property/copyright, license/donor agreements, security, and user access	8
Roles and Responsibilities	High-level roles/responsibilities of institution and/or staff	19
Security Management	Risk assessment, disaster planning, and/or security procedures	15
Selection/Appraisal	Selection/collection polices related to preservation of digital content	11
Staff Training/Education	Training/continued education encouraged and/or provided for staff or producer/submitter	10
Storage, Duplication, and Backup	Duplicate/backup digital content stored in multiple locations for long-term preservation	14
Sustainability Planning	Plans to address or maintain financial stability	13

[55] Preservation strategy in this context is the specific activity of securing the longevity of data files.
[56] Digital Curator Vocational Education Europe: http://www.digcur-education.org/
[57] Strathmann, S., & Engelhardt, C. (2012). Training needs in digital preservation–A DigCurV Survey.: *Report and analysis of the survey of Training Needs*.
[58] Sheldon, M. 2013: Analysis of Current Digital Preservation Policies – Archives, Libraries and Museums: http://www.digitalpreservation.gov/documents/Analysis%20of%20Current%20Digital%20Preservation%20Policies.pdf

During work on the SCAPE project (see separate segment, policy model in appendix 2), several "real life" policies were collected. The project provides an overview with links to all polices that were collected. Although it partly builds on policies that were studied in the NDIIPP study, it gives an indication that there are more polices published from libraries and archives, than there are from data centres. However, the list is not limited to English-language material and includes Dutch, German and Danish polices. It is by no means a *complete* list of available policies, but as the SCAPE wiki is built collaboratively and receives updates from many authors from different countries, it seems fair to assume that it represents a rather comprehensive picture of the current preservation policy situation.

A third study that is worth mentioning in this regard is the Practical Policy Working Group of the Research Data Alliance (RDA – for policy model see appendix 2), which has collected and registered a series of practical policies. This has been done by conducting a survey of production data management systems with the aim of eliciting the types of policies that are being enforced[59]. The types of data management applications included archives, digital libraries, data grids for data sharing, and processing pipelines. The survey identified the highest priority policies in the surveyed sites, and based on these results the study identified eleven generic policies that were of interest to a majority of the institutions and are common to almost all data management systems[60].

Table 2: Results of the RDA survey of 30 institutions for highest priority policies

Policy	Importance
Integrity	217
Preservation	150
Access control	126
Provenance	108
Data Management plans	99
Publication	75
Replication	66
Data staging	52
Federation	37
Metadata sharing	23
Regulatory	16
Collection properties	7
Identifiers	7
Data sharing	7
Versioning	7
Licensing	6
Format	6
Data Life Cycle	6
Arrangement	5
Processing	5

These examples illustrate the various issues that are involved in the creation of a preservation policy. In the next segment we look at the findings from our own study of selected policy models and frameworks.

[59] Two-page paper from the RDA Working Group Practical Policy, available at the RDA file depot: https://www.rd-alliance.org/filedepot?cid=104&fid=557

[60] RDA: *Outcomes Policy Templates: Practical Policy Working Group, September 2014 (version August 29, 2014)*: https://www.rd-alliance.org/filedepot?cid=104&fid=557

3. Findings and analyses

3.1. Overview and characteristics of policy models

We have looked at a selection of policy models, mainly within the SSH community, with selected cross-discipline model examples. The selected models consist of two types: (1) policy statements and documents from data preservation services and (2) best practices, or guidelines, for policy development from relevant organisations and initiatives. In some cases we draw parallels to selected data management plans (DMPs)[61]. The point of including selected DMPs is to illustrate how preservation policies are, or rather should be, connected to data management requirements from research funders and vice versa. DMPs may be a vital tool for key stakeholders to secure the necessary overlap of data elements that are required for long-term preservation of data.

DMPs may limit the distance between the different stakeholders by providing tools and resources that make convergence and standardisation across the different stakeholder levels possible. However, DMPs are not one-size-fits-all, but an appropriate data management plan should take into consideration the size and complexity of the data to be collected or assembled; the likely audience for reuse of the data; and general legal and ethical requirements. Many of these elements share characteristics with elements that are common in a preservation policy.

Table 2: Overview of preservation policy models and resources

Type	Policy provider (abbr.)	Full name and link to policy resource and/or data centre	Year of policy model
Data centre	ADS	Archaeology Data Service: Preservation Policy	2011
Data centre	CentERdata (LISS)	Longitudinal Internet Studies for the Social sciences – Data Archive: Preservation and Dissemination Policy	2014
Data centre	CLARIN	Common Language Resources and Technology Infrastructure	2014
Data centre	CSDA	Czech Social Science Data Archive: Preservation Policy	-
Data centre	DANS (EASY)	Data Archiving and Networked Services – Electronic Archiving System: Preservation Policy	2014
Data centre	Dataverse	Dataverse Network Project: Data Management Plan	2014
Data centre	Dryad	Data Dryad: Terms of Services	2013
Data centre	EUDAT	European Data Infrastructure	2014
Data centre	GESIS	GESIS - Leibniz-Institute for the Social Sciences: Digital Preservation Policy	2013
Data centre	ICPSR	Inter-university Consortium for Political and Social Research: Digital Preservation Policy Framework	2012
Data centre	LOCKSS	Lots of Copies Keep Stuff Safe: Formal statement of conformance to ISO 14721	2004
Data centre	Odum	Odum Institute for Research in Social Science – Data Archive: Digital Preservation Policies	2011
Data centre	UKDA	UK Data Archive: Preservation Policy	2014
Guideline	Beagrie	Charles Beagrie Limited: Digital Preservation Policies Study	2008
Guideline	DCC / UC3	- Data Curation Centre: Preservation Policy Template for Repositories - Checklist for a Data Management Plan. v.4.0. / DMPonline - UC3 DMPTool	2010 2013 2014
Guideline	DISC-UK/DataShare	Data Information Specialist Committee – UK: Policy-making for Research Data in Repositories: A Guide	2013
Guideline	InterPares	International Research on Permanent Authentic Records in Electronic Systems: Policy and Procedures Template	2011
Guideline	nestor	Network of Expertise in Long-Term Storage of Digital Resources: Leitfaden zur Erstellung einer institutionellen Policy zur digitalen Langzeitarchivierung	2014
Guideline	OpenDOAR	Directory of Open Access Repositories: Policies Tool	2014
Guideline	RDA	Research Data Alliance: Outcomes Policy Templates	2014
Guideline	RSP	Repositories Support Project: Policies and Legal Issues	2013
Guideline	SCAPE	Scalable Preservation Environments: Catalogue of Preservation Policy Elements	2014

[61] See Horizon 2020: http://ec.europa.eu/programmes/horizon2020/en/h2020-sections

What seems to characterise most of the policy models that we have assessed, especially the best practices/guidelines and some of the policies from archives holding the Data Seal of Approval, is their similarity to the structure of the Open Archival Information System (OAIS) reference model and the ISO 16363 standard. The OAIS model can be divided into three main parts. The first part provides *purpose*, *scope*, *applicability*, *definitions* and clarification of concepts. The second part provides the details of the *responsibilities* and details of the *functional* model (i.e. the 'Information Package' and its associated objects as they follow a *lifecycle* from the data producer to the archive, and from the archive to the data consumer). The third part covers *technical issues* like digital migration across media and across new formats, and perspectives on the issues of preserving access services to digital information using software porting, wrapping, and emulation of hardware[62].

ISO 16363, which follows from the OAIS' call for a standard and an 'accreditation of archives', provides normative metrics against which a digital repository may be judged. These metrics are grouped into three main segments: *organisational infrastructure* (i.e. organisational structure and staffing, preservation policy framework, financial sustainability and contracts, licenses and liabilities); *digital object management* (i.e. data lifecycle from ingest/acquisition to preservation to access management); and *infrastructure and security risk management*.

Several of the policy models approximate this multilevel structure. In some of the models the approach is explicitly stated (i.e. 'conforms to the OAIS model'), while in others the "kinship" is implicit in the structure and content of the policy clauses. The multilevel approach is characterised by at least two sets of policy clauses: one set of 'higher' level clauses with general abstract statements, and a second level with more details and a focus on practical implementation of policy elements. In these frameworks the 'higher' level elements focus on general statements such as coverage, roles and responsibilities, and relationship to other documents. The 'lower' level elements on the other hand deal with the 'specific triggers' for digital preservation activities. These activities include elements such as migration, ingest, type of storage, etc. These models also underline the importance of stating the relationship to other policies. That is, the need to identify which other policies to take into consideration when implementing and carrying out digital preservation activities. Some of the models also have a third level that approximates the third segments of the OAIS model and the ISO 16363, i.e. a segment that covers technical infrastructure and risk management.

Some policy models combine these elements into one single policy document, while others cover the different segments through several separate policy documents. These combined policy approaches are often a joint records management and digital preservation policy which is designed to join up the creation, selection and preservation of digital records as a single managed process to ensure that all digital records are curated in the same way.

3.2. Guidelines and best practices

The Beagrie policy model is divided into two sets of clauses. The first level is called 'policy clauses', which is set at a higher level and is less technically detailed. These clauses highlight "key points of considerations" intended for the beginning of the policy. The second level, which is more of a technical implementation, constitutes a major part of the policy model and is considered as either a

[62] The OAIS reference model document is actually arranged in six segments, but thematically they can be grouped into three main subject areas.

significant part of the digital preservation policy or a part of separate detailed procedures and guidance documents which are developed to accompany it.

On the highest level of the SCAPE model we find the *guidance* policy elements which define the general long term preservation goals of the organisation. On the second level we find the procedural elements which "…describe the approach the organisation will take in order to achieve the goals as stated on the higher level". This is the level that signifies the actual policy elements and can be considered as the *implementation* level. They are detailed enough to be input for processes and workflow design but are at the same time concerned with the collection in general. These elements are similar to the implementation level elements of the Beagrie model. In addition to the implementation level the SCAPE model also introduces a third level, namely the control policy level. On this level the policies "…formulate the requirements for a specific collection, a specific preservation action or for a specific designated community". The idea behind the third level is that it should be available in machine readable and actionable form (though not a requirement, it can also be human readable) so that it can be used in automated planning and watch tools. Policies intended for machine readability require a high level of detail and clarity in the policy-rule statements, which is reflected in the rather extensive policy model of SCAPE. The RDA model also has a machine-readability approach and the level of detail is high. Under each main element there is a subset of detailed and implementable rules for machine-readability. The RDA model is mainly focused on implementation technical issues and de-emphasises the organisational aspects.

Other guidelines, like the InterPARES, RSP, the DCC template, OpenDOAR and the nestor model, have a more general approach and combine all the policy elements into a single coherent model. However, an implicit distinction between a higher level and a more specific implementation level persists in most of these models. For example, in the DCC template the higher level elements consist of entries such as 'aim' and 'content coverage', while the implementation level, or operational details, are laid out in a separate section ('implementing the strategy'). In the nestor model the higher level elements are covered in the first part ('purpose, scope and objectives') while technical details and preservation strategy are laid out in the second part ('principles and objectives of digital preservation'). The OpenDOAR tool consists of multiple policies but the final output is presented in a combined policy covering the more practical/implementable aspects of the organisation activities.

Most of the models have statements on contextual relations and administrative responsibilities. That is, principle statements on issues such as the organisational scope, aim, mandate, purpose, mission, principles and objectives. Frameworks with a more rule-based approach (like SCAPE, OpenDOAR and RDA) leave out the broader organisational statements and are more specific, but some of the contextual relationships are partly covered through a wide array of sub-elements.

In addition to providing a policy framework the DCC has been involved in the development of services aimed at the researcher. By providing a DMP checklist along with selected templates from research funders, it aims to guide the data depositor (i.e. the researcher) in the appropriate direction when planning for data preservation. DCC has been involved in the development of the DMPonline and the DMPtool (a service run by University of California Curation Center (UC3)). Both tools build on requirements from various funders (UK and US funders, respectively, in addition to requirements

from the European Commission (Horizon 2020)). The DMP elements that are generated from these tools match the selected funder template[63].

Such tools are valuable as the attention to data quality and proper data management within all stakeholder levels seems to have increased in recent years. In Europe, current and future research may depend on funding from the Horizon 2020 research program of the EU. As Horizon 2020 will be a major driving force for significant parts of European research and in years to come, both researchers and curators of long-term preservation, through the development and refining of their data and preservation polices, should take these requirements into consideration[64].

3.3. Data services

3.3.1. Service provider characteristics

Based on our findings within the rather limited selection of services, it is difficult to draw detailed conclusions and to generalise on the structure and organisation of the different service providers. But a few characteristics can be identified.

The characteristics of the service models may reflect onto the policy models: the long-term approach may be more explicit in archives with secure long-term funding, while short-term funding projects may have no policy or a vaguely defined approach to long-term storage. So before we say something about the policy models themselves it might be valuable to look at some of the different service models under which the policy models have been produced.

A key service provider distinction is between *self-archiving* and *mediated* (i.e. curated) *data deposit* models. *Self-archiving* can be defined as a repository model where the researcher deposits their work into a repository by themselves, normally to a publicly accessible website. It usually also includes the depositing of metadata and other details on what the content is about. On the other hand there is the *mediated,* or *curated,* deposit model where researchers simply supply the research output via a repository administrator (i.e. data curator). Normally the depositor is asked to provide data and sufficient metadata information in the repository's preferred formats. Sometimes it may be sufficient for the researcher to submit the materials in its original form and leave it up to the repository to convert these into an appropriate submission format, append the correct metadata and complete the deposit.

The self-archiving model may affect the quality of information that is deposited into a repository, as metadata recording, data quality assessment and correct formatting can be viewed upon as a specialist skill - especially considering the increasing complexities of modern research data - that the data depositing may be more accurate when performed by an archivist or other preservation scholars/experts.

The mediated deposit model implies less time spent for researchers and easy maintenance of internal standards for the repository. Other advantages can include more comprehensive and detailed metadata for deposited items and increased likelihood of search engines locating the item.

[63] It should be noted that some of the larger data services we have looked at in this report, like the UKDA and ICPSR , also provide comprehensive data management checklists or plans that aim to assist the researcher in collecting, managing and depositing data. But these do not necessarily cohere with the requirements from research funders.
[64] See appendix 3 for an overview of the DMP of Horizon 2020.

However, the mediated model may have its drawbacks: it has staff resourcing implications with respect to the scalability of a repository.

Another distinction that can be drawn is between *data server service* and *long-term data archives*. Examples of data server services (e.g. an ftp server) are DataShare, Dryad and Dataverse. Other examples within the commercial service landscape are Dropbox and Figshare, which are broader file sharing and storage services but which are being used, to some extent, by researchers for sharing and collaborating on data (see DASISH D 4.2 for more considerations on these services as tools for researchers). Other types of data servers may be limited to a specific *project website*, where a smaller database and/or server are being used. Examples of *long-term data archives* include UKDA, GESIS and ICPSR. Some service providers apply a combination of these two categories.

There are of course pros and cons connected to each of these service models. For example, in a service model based on a simple ftp data server, archiving data may be very fast as data and files are simply 'dragged and dropped' or 'dumped' into a server. This service model is very cheap and easy to use and it is easy to archive relatively large data sets and other research outputs. However, in such a service model the data that is being deposited can be unstructured and lack proper metadata. Metadata are vital for maintaining the fixity, viability, renderability, understandability, and/or authenticity of digital materials in a preservation context[65]. In self-archiving services data and metadata formats may vary (if metadata is provided at all); it is very likely that the total collection available on the server will vary significantly with regard to file formats, metadata content and formats, data units, etc. Without high quality metadata, data cannot be understood. Also, it is not easy for a user to search for and discover data; when metadata is lacking or inadequate the discoverability of data is very limited. It is simply very difficult to know about the existence of data when there is no structured information to complement the data collection.

Adding to the difficulties of accessibility there is the issue of data versioning. Without contextual information there may be a wide variety of different versions of a data set. And without proper maintenance and curation of versioning it is difficult for a user to know which version is the most updated, which changes were made and why.

Also, a simple data server model is not a long-term archive – the longevity of the service is unknown and data can be easily lost due to suspension or expiration of the service. How, and for how long the service is being maintained is often unknown.

Sometimes, these servers are connected to a project website. Here, the service will normally be a smaller database or ftp server. Hence it is easier for members of the project to access data since it is easier to maintain an overview of the content, as data is available at a single site. It is an easy way to inform about the project data and achievements within the project. However, the website will only represent data coming from the projects and linking to other relevant data and projects will most likely be limited. Also, as these projects often run within a limited funding time frame, project maintenance will stop and links to project data may not work after a while – data can be lost when the project funding ceases.

[65] DCC Curation Reference Manual on metadata: http://www.dcc.ac.uk/resources/curation-reference-manual/completed-chapters/preservation-metadata

Only by depositing data to a long-term data centre or data archive can the longevity and accessibility of data be secured. Moreover, such services may provide linking of data as many data centres (at least within their research discipline) are linked to each other through means of data descriptions, metadata, contextual information and persistent identifiers. Hence data can be found through different websites or in common catalogues. Data may be curated and arranged in relational databases which enable web queries and the extraction of large amounts of data – not least due to the fact that data and metadata will be structured. Also, many archives and data centres provide data sets with a PID (e.g. DOI), which makes them citable.

Research data archives with long-term funding and strong institutional backing are increasingly expanding their role as research infrastructures. They increasingly function as support services not only for data users but also for those creating data, and as experts on methodology, data quality, various data sources, metadata, and documentation. They also provide information and guidance on the ethical and legal framework (e.g. data protection laws), and provide active help with specific queries, create training and support materials, provide the technological capacity to share data, and set up legally binding user licenses. The combination of these elements can provide a secure and trustworthy data service for different data types across many research areas and disciplines.

In recent years we have seen the convergence of research disciplines. More data are being used across disciplines and with a great heterogeneity of research methods and data types, data formats, metadata, etc. Dealing with a wide variety of different data types and/or research output types of small size may not require much space for storage, but a considerable amount of manpower. An extensive data service provider that may want to offer support, advice and curation service throughout the research cycle may find that the costs will rise quickly. Hence, providing a broad selection of services may be far more time and cost intensive than simple server-based services. Often it requires a long-term commitment from funders.

What may become common tools for sharing of research outputs in coming years are solutions based on data portals, LAS (Live Access Server) and data warehouse services. Among the benefits of data portals and data warehouses is the possibility for searching all relevant data centres for new data. That is, by searching one website you search many at once. And all metadata is available in one single entry point - the data portal. Hence, changes at the different data centres are automatically applied and the latest versions of data are always used. Scientists can use such portals like a mini search engine and get a direct link to the data in need. A data warehouse solution enables online retrieval of data archived in relational databases and queries can be limited by predefined parameters. Examples of such services are the ongoing RAIRD[66] project, WAVES[67], and MIDAS[68].

3.3.2. Policies characteristics

Several of the organisations we have assessed have acquired the Data Seal of Approval. All of these provide openly accessible preservation policies. Although the subsections and policy elements are somewhat different, much of the content is similar. ADS, GESIS, ICPSR, LISS/CentERdata, DANS/Easy, Odum, UKDA, and several of the CLARIN data centres all have received the DSA[69]. CLARIN and EUDAT are somewhat different from the others, as policy content is distributed among their different

[66] http://www.raird.no/
[67] http://cdiac3.ornl.gov/waves/discrete/
[68] http://www.mimas.ac.uk/
[69] A selection of DSA case-studies will be available at the DASISH webpage.

platforms and service providers and is not available as one single coherent policy document. For more information about CLARIN and EUDAT policies, see Appendix 1 (CLARIN case study) and Appendix 2 (EUDAT case study).

All of the preservation polices within this segment have policy elements that include statements on a higher, more general level, or what we can call the *content coverage* (with entries such as 'mission', 'scope', 'purpose', 'objective', etc.). In addition they all have a data-lifecycle approach (i.e. digital object management). Finally, they all have dedicated segments on sustainability, security, risk and technical infrastructure. As discussed above, these are all elements that are built-in in the OAIS reference model and the ISO 16363 standard. Many of the archival services have an explicit statement on their conformance to the OAIS model, which explains some of the similarities. Many of the services are partners in the same SSH Research Infrastructure (e.g. CESSDA) that have common rules and guidelines for policy formation. This may also partly explain the similarities in policy structure. In the social sciences the UKDA and the ICPSR policy models are much referred to and used as a template when developing policy. For example, DANS/Easy explicitly states the UKDA as a template for their policy. UKDA conforms to the OAIS model while the ICPSR states the TDR (OCLC)[70] as their template for policy construction. The TDR was a forerunner of ISO 16363.

Most of these archival services present their policy in one single document (i.e. as a combined policy). However, larger archive services such as the UKDA connect their preservation policy to other service documentation: some elements are set out in separate procedure and guidance documents (e.g. a collection policy).

A number of data services and repositories seem to follow the OAIS reference model, and for many it has proven to be a very useful high-level model describing functional entities and the exchange of information between them. Others have simply declared their institutions as OAIS-compliant. These frameworks and models are well known and much used, but policy statements connected to them do not always separate concerns clearly and often mix *objectives* with *functional means* to implement capabilities. Verifying this conformance objectively can be difficult and complex. The impact is not always well-understood, and operations based on the conformance are complex to implement.

Another aspect of most of the OAIS-conforming services is that they are *centralised* and *curated* through *human intervention*. A curated service demands a high level of precision to avoid mistakes, and may partly explain the high level of documentation from some of these services. As the quantity of digital information grows through the convergence of research disciplines and general growth in new data sources, human intervention may become impractical. Service providers may to a larger extent want to employ automated solutions for error detection and correction. Purely human-mediated tools may not scale sufficiently for efficient application to a heterogeneous data collection, and may prove to be a less appropriate method for processes (e.g. for obtaining technical metadata).

This is among the reasons for *distributed* or *decentralised* storage solutions. The LOCKSS Program is built on the assumption that a distributed model strengthens the safety of data. It is a library-led digital preservation system built on the principle that "lots of copies keep stuff safe." The LOCKSS system allows librarians and publishers to obtain, preserve and provide access to purchased copies of e-content through network connections. The idea is that through a LOCKSS distributed network,

[70] Trusted Digital Repositories: Attributes and Responsibilities. 2002.: http://www.oclc.org/content/dam/research/activities/trustedrep/repositories.pdf?urlm=161690

libraries cooperate with one another to ensure their preserved content remains authentic and authoritative. This distributed approach is characterised by the fact that there is no human intervention (curation); there are no "trigger events" that require human intervention. The LOCKSS system has released a statement whereby they officially conform to the OAIS model. However, as opposed to most of the archives discussed above, the LOCKSS statement lays out in detail how and which functions conform to each of the elements in the general OAIS model. This is an example of how OAIS conformance can be proven through more than just a general statement (which seems to be the norm); by aligning activities to OAIS elements it is easier for external users to get a proper understanding of the preservation processes and activities.

One of the services we looked at, Dryad, uses the LOCKSS technology by being linked to the CLOCKSS network (Controlled LOCKSS, which is a closed network as opposed to the open LOCKSS network[71]). Dryad is a repository that makes the data underlying scientific publications and peer reviewed articles accessible and reusable for other researchers. Though the underlying CLOCKSS technology is based on *automated* processes, it is stated that the actual data content are *curated* "…to ensure the validity of the files and metadata". Dryad does not have a designated preservation policy. The policy is rather embedded in their Terms of Services, which contain multiple types of policies, like submission, content, payment, usage and privacy policies. It employs a combination of a centralised curation service and a distributed preservation model[72].

Another example of a decentralised and distributed preservation service is CLARIN, the Common Language Resources and Technology Infrastructure. It is not a single data archive, but a European Research Infrastructure Consortium (ERIC) operating as a network of several centres from its member countries. Some policies of some centres can be more detailed, specialized, restricted or extended than those of the infrastructure as a whole. Each centre establishes its own policies adjusted to the kind of data that it accepts in its repository or repositories. In this policy model, each of the CLARIN centres that offer preservation services has its individual restrictions, deposit rules and archiving principles. The only common elements of the CLARIN ERIC preservation and service policies are expressed through a list of certification requirements for CLARIN Centres. These requirements mainly include compliancy with the CLARIN goals, IPR and privacy statements, external assessment, certificates, federated identity management, metadata, persistent identifiers, and optionally federated content search. However, there are no requirements for the preservation of data *over time*.

In addition to the centralised data archive services (e.g. UKDA) and distributed services (e.g. CLARIN, LOCKSS) we have also considered services that operate in an intermediate area. These do not necessarily present best practice guidelines or have explicit preservation policies in place and cannot be easily categorised as either centralised or distributed services. One of the services that we assessed, the Dataverse network, uses a combination of centralisation and distribution: it has a *centralised* software installation and data repository, but with individual *distributed* data archives with their own branding. Hence, a Dataverse Network hosts multiple 'dataverses' (i.e. the individual virtual archives) where each dataverse contains studies or collections of studies and each study

[71] The Center for Research Libraries (CRL) conducted a preservation audit of CLOCKSS between September 2013 and May 2014, and on the basis of that audit CLOCKSS was certified as a trustworthy digital repository of e-journal content. http://www.crl.edu/archiving-preservation/digital-archives/certification-and-assessment-digital-repositories/clockss-report

[72] Another example of a service that links data to articles and journals is the Earth system science Data: http://www.earth-system-science-data.net/

contains cataloguing information / data descriptions plus the actual data and complementary files. However, the service does not provide a preservation policy, but the general conditions for use of the service are laid down in the 'terms of use'. These include terms for data deposit, data use, and data backup and preservation. But much like the CLARIN network the preservation policies are found in the individual data service 'modules' (dataverses in this case) and longevity and security is the responsibility of each data service provider connected to the network[73].

Dataverse, which is mainly a self-archiving service, is dependent on the depositors' ability to follow a standardised DMP[74] to make sure that the collection and documentation is coherent. By providing a DMP checklist along with a template for NSF funded research, it aims to guide the depositor in the appropriate direction when planning for data preservation. Both the DMPtool from the UC3/CDL and DMPonline from DCC provide similar tools. Both tools build on requirements from various funders (US and UK funders, respectively) and the DMP elements that are generated match the selected funder template[75].

Such tools may be valuable as the attention to data quality and proper data management within all stakeholder levels seem to have increased in recent years. In Europe, much of the current and future research depends on funding from the Horizon 2020 research program of the EU. As Horizon 2020 will be a major driving force for significant parts of European research and in years to come, both researchers and curators of long-term preservation should, through the development and refining of their data and preservation policies, take these requirements into consideration.

4. Conclusions and Recommendations

The examples of policy approaches we have provided in this report demonstrate that although there are common elements in many of the models, the decision making and procedure implementation in the organisations are to some extent based on ad-hoc solutions due to a lack of common understanding of how to implement the standardisation models. The OAIS model and corresponding criteria catalogues for trustworthy repositories specify requirements that such preservation planning and process should fulfil, but they are often general in approach and do not provide concrete guidance.

Proper preservation policies that are coherent with funder requirements and in line with proper data management plans may be a major contributor to the accessibility and integrity of research output. However, the policies should not be dictated by funder requirements or pre-customised data management plan templates. There should rather be a dialogue between the different stakeholders in the creation of a full-scale policy model framework.

[73] One of the dataverses connected to the network is the Odum institute (see appendix 3)

[74] A data management plan (DMP) can be defined as follows: "*A DMP describes the data management life cycle for all data sets that will be collected, processed or generated by the research project. It is a document outlining how research data will be handled during a research project, and even after the project is completed, describing what data will be collected, processed or generated and following what methodology and standards, whether and how this data will be shared and/or made open, and how it will be curated and preserved. The DMP is not a fixed document; it evolves and gains more precision and substance during the lifespan of the project*."

[75] It should be noted that some of the larger data services we have looked at in this report, like the UKDA and ICPSR, provide comprehensive data management checklists or plans that aim to assist the researcher in collecting, managing and depositing data. But these do not necessarily cohere with the requirements from relevant research funders.

4.1. General recommendations

Based on our findings it seems clear that an explicit, coherent and well-defined organisational policy framework does more than simply provide an abstract "high-level plan" that supports and justifies the organisational activities. Rather, it is just one of the elements of a fully matured policy framework. In addition to a high-level plan it should also *facilitate stakeholder understanding and cooperation*. A clear-cut policy model sets out the responsibilities both within the organisation and towards cooperating stakeholders and the lines of communication are clarified.

Processes are clearly defined and standardisation and conformity are not only promoted but laid out in a coherent policy framework that is consistent throughout the stakeholder network. In this lays the necessity of considering the organisational policy framework not as an isolated phenomenon; it needs to be created in accordance with both the wider strategic policies of the institution and *with all involved stakeholders*. To ensure that the policy framework is comprehensive and consistent throughout the research process and on all stakeholder levels, each organisation should seek to integrate a significant number of its policy-rules into the *wider stakeholder framework*. Hence, there should be a coherent overlap between the funder requirements, the data centre deposit and preservation rules and requirements, and the data management plans that are provided to the researcher. The DMPTool of the University of California and the DMPonline tool provided by the DCC are steps in the right direction when it comes to stakeholder synergies. However, these tools provide DMP templates based on requirements from funders and therefore they are primarily tools for researchers seeking funding. Communication to secure the long-term preservation of data (i.e. communication towards the data centres) is still lacking. A proper tool should take into consideration the requirements of the funders, the data depositing rules of the data centres and the future user of the research output.

Reports have shown us that although many research funders still lack data deposit requirements (see chapter 2.3.2), an increasing amount of funders are considering making deposit a condition of funding. A powerful tool in a full-scale policy framework is a *mandate* that not only encourages the deposit of data, but makes it compulsory, both from the funder stakeholders and from the research institutions themselves. Even enforcement of deposit may be a part of the policy framework, with possible sanctioning tools to support the enforcement (e.g. a withdrawal of funds).

A fully integrated research stakeholder policy network should not put restrictions on the flexibility of the system; it can be argued that overlapping policy-rules and requirements may create rigid systems that enforce a "discipline-irrelevant" set of rules and requirements. The overall system should be internally coherent and in accordance with the wider stakeholder taxonomy within which they operate.

In addition to the larger policy-network any organisational preservation policy should be part of, it should also help the internal planning and decision-making. A policy that is organisation-specific assists the data centre to identify and better understand and manage the risks associated with their activities. These efforts contribute towards the ongoing and day-to-day management of the data centre. It is also a valuable tool to make sure that the implications of dealing with the exposure of different types of resources are appropriately handled (e.g. that IPR are complied with).

As seen in some of the policy models that we reviewed, it is not uncommon that the data centre (or the policy model recommendation) seeks to integrate several policies into the wider organisational

framework. The main elements of this organisational framework can be divided into two main streams: *strategic* and *operational (implementation)*. The strategic policy elements have been covered above: they may consist of high-profile "vision" statements and defined agendas, mandate and designated users. But they are not independent of the rest of the policy nodes. It informs the day-to-day operations of the repository which are defined in the implementation policies. The implementation policies may be integrated in the overall preservation policy (i.e. as a combined policy) or it may be distributed among separate policy documents. Implementation policies cover areas such as submission/ingest, collection, archive/preservation and usage/access. One of our examples, UKDA, provides a single preservation policy that sets up the bigger picture while also addressing the overall data lifecycle processing involved in their preservation activities. In addition, they provide supporting documentation in separate policies like the Collections Development Policy, Cataloguing Procedures and Guidelines, Data Ingest Processing Procedures, and Data Processing Standards[76].

Some of the policies or policy elements need to formalise interaction with external actors such as a website user or a researcher. This can be achieved through legal agreements and statements connected to the data centre activities. Specifically, this involves the definition of and setting up licenses for depositing and use of data. They should be present in one or more formalized policy documents, including at least: User agreements, Terms of use, Legal policies, and Privacy policies.

Development and practical implementations of researcher policies have shown that it is crucial to involve all aspects of the relevant research community through early consultations on best practices and recognizing the different needs of different communities. That is, containing the flexibility of the data policy without sacrificing too much of the required 'formal rigidity' that is necessary to support the quality and longevity/accessibility of data. Further, the policy should be continually evolving through effective monitoring and feedback.

4.2. Policy-rules recommendations

The policy-rules and headings below provide more of a checklist to work against when developing or refining a policy than a list of normative requirements. One may not need to use all of them. Some organisations may wish to emphasise some sections more than others. Some sections may be grouped together into sub-sections of the policy, while others may be laid out in separate policy documents.

However, it is important to note that too many combinations or the joining of very different types of work in one policy could result in none of the functions being adequately supported in a policy document.

Part one: Context and purpose

Most policy models and policy recommendations have several entries concerning the general purpose and context of the organisation. These statements are in some cases described in a separate introductory segment, as is suggested here, to make a clear distinction between aims and goals, and implementation of these goals.

1.1 Purpose, objectives, scope, mandate

[76] UKDA: Documentation: http://www.data-archive.ac.uk/about/publications

This segment should contain information on *purpose, scope and/or objectives* of the organisation. These statements normally build on the organisation's founding documents, like mission statements, visions and strategic plans. These segments can be part of an introductory segment of the policy and should describe the purpose and function of the organisation, state the rationale for the preservation policy and clearly show how the policy is grounded in the organisational context by establishing clear connections between goals and implementation.

This segment should also include a statement on the *scope of the policy,* i.e. whether the preservation policy applies to all collections of the organisation and whom the policy applies to (e.g. staff, users, etc.).

1.2 Glossary, definition of terms

For some organisations it may be appropriate to include in this segment a glossary or list of definitions utilised in the policy. As the audience for a preservation policy may be diverse, it can be useful to define key terms at the outset to ensure common understanding, especially if the policy applies many technical terms or terms specific to the organisation. Key terms may typically include (digital) preservation, curation, migration, emulation, etc. Alternatively, this segment can be put at the end of the policy, in an appendix.

1.3 Preservation standards, requirements, legal and regulatory framework

Statements on how the policy is built on or supported by accepted standards, ethics and legal requirements should be included here. A preservation policy does not exist in isolation. The policy is normally influenced by a variety of external guidelines, manuals, and standards, while also taking into account the internal aims, objectives, and strategic and operational plans of the organisation. This segment should list and specify how and under which requirements and legal and regulatory frameworks the policy works. Internal documents may include strategic plan, collections of development policy or security documentation. External documents and standards should include a list of the legal regulations under which the organisation operates (e.g. the EU Copyright Directive, national Data Protection Act, etc.) and which standards (if any) it follows (OAIS, ISO 16363, etc.).

1.4 Roles and Responsibilities, financial responsibilities, cooperation

This segment should describe key stakeholders and their respective roles in digital preservation. All organisational staff has a role to play in the implementation of a preservation policy and accountability is shared among staff. This segment should clarify and define the different key roles and responsibilities within the organisation, and make an explicit statement that digital preservation is a shared responsibility requiring participants within and beyond the organization. If relevant, cite or link to documents containing more specific descriptions. It should also contain statements on financial sustainability and how the policy sits within the organisational financial plan.

Part two: Implementation clauses

As we have seen in our assessment of the different templates, checklists and policies, many are based on the OAIS reference model or take a data lifecycle approach. This part of the policy should explicitly state the organisation's approach. One possibility is to take a life-cycle approach by going through each implementation stage in the data curation process, e.g. selection/acquisition, conversion, receive, verify, determine significant properties, ingest, metadata, storage, preservation techniques, and access. Another option is to order it according to the aforementioned OAIS

terminology. This should include Preservation Planning, Ingest, Archival Storage, Data Management, Administration, Access, Deletion, and possibly a description of the different archival packages: Archival Information Package, Submission Information Package, and Dissemination Information Package.

This model may be a valuable starting-point for most organisations involved in the preservation of digital objects, as it covers all of the key activities that occur in a preservation organisation, ranging from ingest, management and storage to data access and data sharing. It may assist in clarifying the implementation of goals. Even if the organisation does not explicitly comply with the OAIS model, they are involved in the activities of acquiring, managing, storing and providing access to data. Hence it may be valuable for a preservation policy to have a data lifecycle-approach to clearly distinguish between the different activities. Whether or not the policy is built around the OAIS functionalities or a data lifecycle approach, the following points should be considered for inclusion (as a 'minimum requirement'):

2.1 Pre-ingest, selection and acquisition

The pre-ingest function is not explicitly specified in the OAIS model. However, experience has shown that inclusion of this function within the preservation model has considerable benefits: it may ensure quality, comprehensibility and accessibility of all 'information packages' by enforcing quality assurance and minimum standards at the point of 'Producer-Archive interface'. It may also contribute to reduce costs within the ingest process[77]. This segment provides the rationale and processes for developing and retaining collections based on specific parameters (e.g., formats, types of data, geographic scope). A clear articulation is important as it ensures that the selected acquired data support the institutional mission and priorities, and that necessary resources are made available for the preservation of the material. One way of solving these issues is to implement (parts of) the PAIMAS model[78].

2.2 Ingest, communication with the depositor

Often, the object is reformatted or otherwise processed before entry into the archive. These procedures should be described in this policy element and/or linked to separate procedural documentation. This segment may also mention issues such as source version vs. new version (i.e. is the original version also deposited and processed along with a new version), legal accountability now that it is a new object, statements about unique ID/naming convention for the ingested material, how the object is ingested in to the archive/repository (e.g. raw, compressed, zipped, encrypted) and data checking routines (e.g. virus check routines).

2.3 Preservation strategy

The long-term retention of research data requires measures to protect against deterioration, or outdating of data material. Hence the preservation policy should outline how the organisation approaches the storage of its data collections. The format, structure, and size of the datasets and/or the total data collection may influence how they should be stored. In addition, this policy element should state the type(s) of preservation the archive will adhere to, e.g. bit stream preservation,

[77] http://data-archive.ac.uk/media/54776/ukda062-dps-preservationpolicy.pdf
[78] CCSDS 651.0-M-1 (2004). Producer-Archive Interface Methodology Abstract Standard: http://public.ccsds.org/publications/archive/651x0m1.pdf

transformation to an open format, rendering, emulation, migration, keeping the original, or a combination of approaches.

The organisation should support and engage with global, open and persistent researcher identification initiatives to ensure connectivity and accurate attribution of researchers and data. Hence a statement on the use of persistent identifiers (PI) should be included in this section. The PI strategy of the organisation may be built on systems and formats like Digital Object Identifiers (DOI) or ORCID.

2.4 Archival storage, security

This element specifies the organization's commitment and approach to ensuring the accuracy, completeness, authenticity, integrity, and long-term protection of the organization's data assets. It should include statements on the general IT-architecture and whether or not the repository/archive is mirrored off-site, or has other external/distributed archival strategies. It should also include the type(s) of storage media and data format(s) that has been chosen and how regularly this is processed and/or upgraded (i.e. when and how often regular back-ups are carried out). Some types of data may require greater security (e.g. personal data), controlled environment and/or extra protection.

The element should also include a statement on the responsibilities for the security of the data collections, both those related to the staff and to the users. The following aspects may be addressed, either directly or through a link to a separate security policy document:

- Physical security, such as building and perimeter security.
- Security of access (e.g. access by staff and users to storage areas).
- Security of computer systems, including authorised access to and manipulation of data.

2.5 Risk management

To ensure a secure and trustworthy technical infrastructure it is necessary to include statements that demonstrate how the organisation aims to achieve this. This can include statements on technology assessments (i.e. by employing technology monitoring and technology watch), and by stating which software it makes use of. If the service provider employs community supported software like iRODS or Fedora this should be explicitly stated and explained in the policy.

Additionally it is necessary to show how the data service ensures ongoing and uninterrupted services to its designated community. This can be achieved by referring to risk, threat and control analyses that are carried out in the repository. The repository may conduct risk assessments with tools such as DRAMBORA[79] (e.g. annual assessments) and/or by demonstrating its employment of the ISO 27000 series for information security matters[80].

2.6 Data management, curation, metadata

One of the most important aspects here includes an outline of the metadata schema in use, while also specifying how the different sections of the schema are structured (e.g. descriptive metadata, structural metadata, administrative metadata, preservation metadata, etc.).

[79] Digital Repository Audit Method Based on Risk Assessment: http://www.repositoryaudit.eu/
[80] http://www.27000.org/

Further, it should include statements on version control, quality control and change procedures to ensure that any alteration to the preserved version of any part of a data collection is accurately documented. This maintains the *authenticity* of the data collection. It should also be considered to include a clause about the possible de-selection/deletion of items and/or deletion procedures.

2.7 Access, use, re-use

This segment identifies how end users interact with the archive to find, request and receive data and metadata. It includes statements on user terms (e.g. open access, barriers and/or restrictions to use) of the digital content for which the organisation is responsible. This element is often heavily dependent on other policies that are developed to further articulate access and use requirements and restrictions (e.g. access policy, deposit agreements, digital rights management rules and practices, user and licence agreements).

2.8 Intellectual property

This clause shows awareness of copyright issues and how the institution plans to recognise and tackle these key issues. The element may be integrated into one or more of the elements mentioned above, but as it is a subject that transverse the full data lifecycle it may be valuable to dedicate a separate segment in the preservation policy to IP-issues, or to put into a separate policy document.

Issues to be addressed include agreements with authors and data owners; commitment to keeping the data secure while maintaining the intellectual property rights of the depositor; tracking of changes to the digital object; registry of object creators and owners; the possibilities of reproducing/copying the digital object; agreements with authors on rights for preservation and reproduction of the object; explanation of access levels and how different levels may be assigned to different collections; deposit agreements and methods of depositing (e.g. whether there is self-archiving routines or staff-mediated/controlled by staff).

2.9 Policy review, certification

Finally, the policy should include a statement on how often a review of the policy is carried out (e.g. annually, biannually). Additionally it should include a section on how the data centre is, or aims to become, formally a trustworthy long-term preservation service. That is, how it aims to secure and guarantee the authenticity and longevity of its digital objects, either through assessment and certification through standards like the Data Seal of Approval, DIN 31644/nestor Seal, ISO 16363, etc., or through self-assessment checklists like TRAC/TDR. The number of repositories and archival services for research data are increasing and funders are seeking criteria for pointing researchers to the most trustworthy service providers. An organisation that explicitly seeks measures to strengthen its services through standards of trust will stand stronger compared to its competing preservation service providers.

Table 3: Summary of recommended preservation policy elements

	Id	Policy Element	Description
Context and purpose	**1.1**	Purpose, objectives, scope, mandate	*Should describe the purpose and function of the organisation, state the rationale for the preservation policy and clearly show how the policy is grounded in the organisational context by establishing clear connections between goals and implementation.*
	1.2	Glossary, definition of terms	*As the audience for a preservation policy may be diverse, it can be useful to define key terms at the outset to ensure common understanding, especially if the policy applies many technical terms or terms specific to the organisation*
	1.3	Preservation standards, requirements, legal and regulatory framework	*This segment should list and specify how and under which requirements and legal and regulatory frameworks the policy works.*
	1.4	Roles and Responsibilities, financial responsibilities, cooperation	*This segment should clarify and define the different key roles and responsibilities for participants involved in the long-term preservation of data.*
Implementation clauses	**2.1**	Pre-ingest, selection and acquisition	*This segment provides the rationale and processes for developing and retaining collections based on specific parameters.*
	2.2	Ingest, communication with the depositor	*Describes the processing of the data object before it is entered into the archive.*
	2.3	Preservation strategy	*Outlines how the organisation approaches the storage of its data collections (e.g. bit stream preservation, transformation to an open format, rendering, emulation, migration, etc.).*
	2.4	Archival storage, security	*Specifies the organization's commitment and approach to ensuring the accuracy, completeness, authenticity, integrity, and long-term protection of the organization's data assets.*
	2.5	Risk management	*Describes measures on how the organisation achieves a secure and trustworthy technical infrastructure.*
	2.6	Data management, curation, metadata	*Outline of the metadata schema in use. Specifies how the different sections of the schema are structured (e.g. descriptive metadata, structural metadata, administrative metadata, preservation metadata, etc.).*
	2.7	Access, use, re-use	*Identifies how end users interact with the archive to find, request and receive data and metadata.*
	2.8	Intellectual property	*This element describes how the organisation plans to recognise and deals with copyright issues.*
	2.9	Policy review, certification	*Statement on how often a review of the policy is carried out (e.g. annually, biannually). Additionally it should include a section on how the data centre is, or aims to become, formally a trustworthy long-term preservation service.*

Appendix 1: Case study 1: CLARIN

Description

CLARIN is not a single data archive, but a European Research Infrastructure Consortium (ERIC)[81] operating as a network of several centres in several countries[82]. Some policies of some centres can be more detailed, specialized, restricted or extended than those of the infrastructure as a whole.

CLARIN, the Common Language Resources and Technology Infrastructure, "is a large-scale pan-European collaborative effort to create, coordinate and make language resources and technology available and readily useable. CLARIN offers scholars the tools to allow computer-aided language processing, addressing one or more of the multiple roles language plays (i.e. carrier of cultural content and knowledge, instrument of communication, component of identity and object of study) in the Humanities and Social Sciences"[83].

Policy Model

The CLARIN ERIC expresses its preservation and service policies through the list of certification requirements for CLARIN centres. These requirements[84] mainly include compliance with the CLARIN goals, IPR and privacy statements, external assessment, certificates, federated identity management, metadata, persistent identifiers, and optionally federated content search. Certification of CLARIN centres according to this policy is performed by a Centre Assessment Committee. The CLARIN ERIC does not impose requirements for the preservation of data over time. In the current policy model, "each of the CLARIN centres that are offering preservation services at this moment has its individual restrictions, deposit rules and archiving principles"[85]. An example of a specific CLARIN centre preservation policy can be found at AVS Leipzig[86].

The headers and themes under which the following information is arranged do not reflect a CLARIN policy model as such. Rather it must be considered as a template for information gathering. The sections do not as such refer to any specific written policy document (as is the case for the other policy models that we have analysed in this report). The thematic headers are partly based on the template that was applied in DASISH 4.2[87] and partly on some of the most regular elements in most policy guidelines and recommendations.

Content coverage

Scope

CLARIN addresses the scholarly needs of any discipline working with language data. In practice, CLARIN promotes the documentation, accessibility, searchability and reusability of all scholarly digital data expressed in language or about language, such as text and speech corpora, digital literary editions, interviews, historical databases (such as church records), lexical and terminology databases, computational grammars, and psycholinguistic experimental records, sociolinguistic survey data, etc.

[81] http://ec.europa.eu/research/infrastructures/index_en.cfm?pg=eric
[82] http://www.clarin.eu/content/overview-clarin-centres
[83] http://www.clarin.eu/
[84] http://hdl.handle.net/1839/00-DOCS.CLARIN.EU-78
[85] http://www.clarin.eu/sites/default/files/preservation-CLARIN-ShortGuide.pdf
[86] http://clarin.informatik.uni-leipzig.de/repo/files/ULei_preservation_policy_v2.pdf
[87] DASISH D4.2: *Report about Preservation Service Offers*: http://dasish.eu/publications/projectreports/D4.2_-_Report_about_Preservation_Service_Offers.pdf

Mandate

The various CLARIN centres may have their own mandates which may be more restricted or more extended (than the CLARIN ERIC, see 'description'above). For example, the CLARIN Centre of the Saarland University (UdS) has "an explicit mission to archive language resources especially multilingual corpora (parallel, comparable) and corpora including specific registers"[88].

Kinds of data (selection and appraisal)

In general, CLARIN centres include digital data expressed in language or related to language research, such as the following:

- Raw or annotated primary data: text and speech corpora, audio-visual and multimedia corpora (including sign-language recordings, subtitles, etc.);
- Secondary resources: computational grammars, word lists, thesauri, ontologies, wordnets, electronic dictionaries, term bases, etc.;
- Language tools: part of speech taggers, lemmatizers, parsers, morphological transducers, tokenizers, machine translation systems, tools for phonetic alignment, audio/video analysis, etc.
- Quantitative data related to psycholinguistics, neurolinguistics, sociolinguistics or dialectology: reaction times, ECG data, fMRI data, questionnaire survey data, geographical data, etc.

The deposit offers of the various CLARIN centres may vary.[89] Some are restricted to particular languages. Some other centres contain also datasets in the Humanities and Social Sciences which are not directly related to language. For instance, the Data Archiving and Networked Services (DANS)[90] includes datasets from history, archaeology, geospatial science, and other fields.

Form/status of data

Each CLARIN centre establishes their own policies as regards the kind of data that they accept in their repositories. The CLARIN ERIC does not restrict the data to be stored by its status. Both raw and processed data as well as applications and tools may be included in the infrastructure. The results of research carried out with data coming from the repository may subsequently be stored in the same repository for further reuse. The only requisite for data to be included in the archive is that such data is adequately described by means of metadata.

The following table provides a list of the CLARIN centres offering depositing services and the form of data accepted for deposit by each of these Centres[91].

Centre	Location	Depositing offer
CLARIN Centre Vienna	Austria	Any linguistic and/or NLP data and tools

[88] http://fedora.clarin-d.uni-saarland.de/index.en.html
[89] http://www.clarin.eu/content/depositing-services
[90] DANS is not a certified CLARIN B Center yet. See http://www.clarin.eu/content/overview-clarin-centres for more details of certified CLARIN Centres.
[91] This table has been taken from the CLARIN website: http://www.clarin.eu/content/depositing-services

LINDAT-Clarin	Czech Republic	Any linguistic and/or NLP data and tools
The CLARIN Center at the University of Copenhagen (CLARIN-DK-UCPH)	Denmark	Danish language resources, with a particular focus on written and spoken language data.
Bayerisches Archiv für Sprachsignale (BAS)	Germany	Corpora of spoken languages including acoustic signals, videos, series of measurements, series of pictures, etc.
Berlin-Brandenburg Academy of Sciences and Humanities (BBAW)	Germany	Linguistic data in German, including monolingual and parallel corpora, historical prints and manuscripts, lexical resources.
Hamburger Zentrum für Sprachkorpora (HZSK)	Germany	Resources that focus on the investigation of spoken language and in particular those about multilingualism
Institut für Deutsche Sprache (IDS)	Germany	Resources on the German language
Institut für Maschinelle Sprachverarbeitung (IMS)	Germany	Any language resources and NLP tools; special focus on domain adaptation
Universität des Saarlandes (UdS)	Germany	Multilingual corpora and corpora including specific registers
Eberhard Karls Universität Tübingen (SFS)	Germany	All language resources
The Language Archive	Netherlands	Any linguistic data, with a particular focus on data related to the languages and cultures of small and endangered speech communities

Data versions / version control

For the CLARIN ERIC, versioning is optional. The certification requirements[92] do not include a requirement for using version control. Furthermore, versioning is platform-dependent. While some CLARIN centres use platforms that allow version control, other platforms used in some other centres do not provide this option.

Data formats

The CLARIN ERIC does not have restrictions, but recommendations on formats when depositing research data. A list of standards[93] and a "Standardisation Action Plan"[94] are available on CLARIN's website. Individual centres may, however, have specific restrictions on which data formats will be accepted for deposit.

Size / volume of data

Each CLARIN centre is entitled to establish its own restrictions for its depositing services. In general, there is no restriction on the size of data to be deposited.

Pre-ingest

Guidance for researcher

The CLARIN ERIC does not have specific guidelines with respect to information and guidance to the data depositor. However, individual CLARIN centres offer information and assistance/guidance as part of their services. Instructions for depositing data are usually provided during the depositing procedure. For example, the LINDAT website[95] contains instructions relating to the following procedures.

- Registration
- Login
- Selecting the type of submission
- Description of the item
- File upload
- License selection
- Review of the submitted data
- Final submission

Similarly, the CLARIN Centre at the University of Copenhagen provides "data management consultation and support in connection with the deposit"[96]. Users who want to deposit their data, are provided with information (in Danish)[97] on the following aspects:

- Who can deposit data
- How to prepare the data to be deposited
- Which kind of access may be granted to the data

Other useful information to users is provided in their website on the following:

- The authentication and authorization process[98];

[92] The CLARIN checklist is available here: hdl:1839/00-DOCS.CLARIN.EU-78.
[93] The list is available here: http://www.clarin.eu/faq/what-standards-are-recommended-clarin.
[94] The standardisation action plan can be found at: http://www.clarin.eu/system/files/private/Standardisation%20action%20plan-v8.pdf.
[95] https://lindat.mff.cuni.cz/repository/xmlui/page/deposit
[96] https://assessment.datasealofapproval.org/assessment_105/seal/html/
[97] See: http://info.clarin.dk/kom-godt-i-gang/deponer-resurser/vejledning/

- The form and kind of data accepted for deposit[99];
- The preparation of data (the specific requirements are given in forms of a user guide[100]);
- The selection of the deposit license[101].

Data ingest

Eligible depositors

CLARIN has no restrictions on the position/status of data depositors. Thus, data depositors may be associated with CLARIN member organizations or not.

However, authorization to deposit data in any certified CLARIN centre presupposes authentication (i.e. users must disclose their identity). Typically, data depositors can make use of a Federated Identity Management (FIM) service. However, there is no requirement for a depositor to be associated with a FIM member in order to deposit data. CLARIN also provides its own Id service.

Through the FIM model, several organizations define a common set of policies, practices and protocols to manage user's identities across trusted organizations[102]. As specified in the CLARIN checklist, "Centres need to join the national identity federation where available and join the CLARIN service provider federation to support single identity and single sign-on operation based on SAML2.0 and trust declarations."[103]

Review/moderation of deposited data

The CLARIN ERIC does not impose any specific requirement or provide any assessment with respect to quality/validity/accuracy of the data/metadata.[104] However, each CLARIN centre usually has its own checking procedure. For example, the DANS archivists work according to a standard protocol in order to provide long-term preservation and accessibility of the data. They also check the data to ensure that all privacy-sensitive data have been properly anonymised. If this was not the case, they anonymise the data. According to the DANS archivist standard protocol[105], the archivist shall check:

- The completeness of the dataset (files and documentation);
- The readability/accessibility of the files;
- The file formats, options to deliver or produce preferred formats or accepted formats if other formats are deposited;
- The completeness and correctness of the metadata;
- If the files or the metadata contain privacy sensitive information;
- The clarity of the dataset structure (use of file folders).

[98] http://info.clarin.dk/overblik/hvem/
[99] http://info.clarin.dk/overblik/typer/
[100] http://info.clarin.dk/kom-godt-i-gang/deponer-resurser/vejledning/
[101] The information about the available licenses can be found here: http://info.clarin.dk/overblik/licenser/.
[102] For more information, see the DASISH training module on Authentication and Authorization Infrastructure: http://training.dasish.eu/training/2/.
[103] "Single sign-on (SSO) is an authentication process that allows a user to access multiple applications with one set of login credentials" (Source: Techopedia). The Security Assertion Markup Language (SAML) is a standard which provides an XML-based framework for creating and exchanging security information between online partners (see SAML community at http://saml.xml.org/).
[104] Even though the CLARIN centres do not perform any specific operation to check the quality of the metadata, they do require these to be compliant with the Component MetaData Infrastructure (CMDI) format (see section 4.4).
[105] The protocol is available here [in Dutch]: http://www.dans.knaw.nl/sites/default/files/Provenance_document_DEF.pdf

Depositor agreement(s) / Responsibility

When a resource developer wishes to deposit a resource in a CLARIN Center, (s)he has to sign a so-called "Deposition License Agreement" (DELA) with the CLARIN Centre where the resource will be available. This Agreement is signed by the copyright curator (the CLARIN Centre) and the Copyright holder (the resource owner). CLARIN provides a set of templates, one for each CLARIN resource category:

- CLARIN PUB Deposition License Agreement: For resources which are going to be deposited as publicly available.
- CLARIN ACA Deposition License Agreement: For resources licensed to Academic use and researchers.
- CLARIN RES Deposition License Agreement: For resources with restricted access and which require individual authorization.

As indicated in the legal website of FIN-CLARIN, the Finnish branch of CLARIN, "CLARIN deposition licenses are available for curating a minimal set of usage conditions to include a resource in the CLARIN PUB, ACA or RES categories. The minimal deposition licenses can be used as checklists if you wish to use your own set of deposition licenses to curate additional usage conditions from a resource provider."

The templates provided by CLARIN aim at enabling the licensing and depositing of resources. The agreements are drafted in a clear and concise way, and it is up to the resource provider to modify the template and add further restrictions and conditions if (s)he deems it necessary. Moreover, CLARIN has foreseen a set of additional restrictions and conditions. These have also been described and listed and refer to common restrictions and conditions in licenses such as:

- Attribution (+BY): The obligation to cite a resource in publications describing research in which the resource was used.
- No commercialization (+NC): The prohibition to use the resource in commercial applications and obtain economic benefits with it.
- Share alike (+SA): The obligation to distribute the derivate resources only on the same conditions as the original work.
- Inform (+INF): The obligation to inform the copyright holder of any use of the material.
- Local (+LOC): The restriction to download the resource or use it out of the boundaries of the repository where it is stored.
- ReDeposit (+ReD): The obligation to redeposit any derivates of the resource within CLARIN.

Requirements of confidentiality

According to the CLARIN Terms of Service (TOS), the user agrees to follow the data protection policy of the CLARIN Services. Although the CLARIN ERIC does not provide anonymisation procedures, individual centres may provide several levels of assistance with such procedures.

Moreover, in those cases in which the data has not been or cannot be anonymised, CLARIN offers the possibility of using a Restricted License, which would ensure that only authenticated and authorized people have access to the resource.

IPR (rights and ownerships)

According to the CLARIN Deposition Licenses (DELAs)[106], "the ownership of the Resource remains with the original Copyright holder or holders. A copy of a Resource and the ownership of its physical carrier deposited by the Copyright holder are transferred to the Copyright curator at the time of delivery".

As far as Intellectual Property Rights and Access Rights are concerned, these are also regulated in the DELAs. Depending on the category of the licenses (public, academic or restricted), different conditions may apply.

(a) Public licenses (CLARIN PUB Deposition License Agreement):

> *"7.1 The intellectual property right and/or other rights governing the Resource subject to this Agreement belong to the Copyright holder or his licensors. Any third-party content of the Resource is identified in Appendix 2.*
>
> *7.2 The Copyright holder makes the Resource available according to one or several of the licenses enclosed in Appendix 3:*
>
> *[ ] The latest version of the Creative Commons ZERO.*
>
> *[ ] The latest version of the Creative Commons BY.*
>
> *[ ] The latest version of the Creative Commons BY-SA.*
>
> *[ ] The latest version of the Creative Commons BY-ND*
>
> *[ ] The latest version of the Creative Commons BY-NC-SA.*
>
> *[ ] The latest version of the Creative Commons BY-NC.*
>
> *[ ] The latest version of the Creative Commons BY-NC-ND*
>
> *[ ] The GPL v.2 or later.*
>
> *[ ] The LGPL v.2 or later.*
>
> *[ ] The EUPL license.*
>
> *[ ] The BSD license.*
>
> *[ ] The Eclipse Public license.*
>
> *[ ] The latest version of the Apache license.*
>
> *Additional rights to the Resource may be agreed separately in writing.*
>
> *7.3 Information about the license is to be published in conjunction with the Resource in accordance with the terms of the license. A sample End-User license agreement is enclosed in Appendix 4.*

[106] As stated earlier, the DELAs are available in the legal website of FIN-CLARIN, the Finnish branch of CLARIN: https://kitwiki.csc.fi/twiki/bin/view/FinCLARIN/FinClarinLegal

If the Resource is made available by the Copyright holder with the Creative Commons ND condition, the following rights are added: "The Resource can be modified for the personal use of the End-User or his research group, even if such a modified Resource may not be distributed. If the End-User or his research group wishes to distribute a modified Resource, this may be agreed separately with CLARIN."

If the Resource is made available with the Creative Commons NC condition, the following interpretation is added: "Government-funded or non-profit research projects, e.g. projects funded by <NATIONAL RESEARCH FUNDING AGENCIES>, are not regarded as gaining economic benefit even if a portion of the financing is contributed by companies."

(b) Academic licenses (CLARIN ACA Deposition License Agreement):

"7.1 The intellectual property right and/or other rights governing the Resource subject to this Agreement belong to the Copyright holder or his licensors. Any third-party content of the Resource is identified in Appendix 2.

7.2 The Copyright holder grants the Copyright curator a non-exclusive and perpetual (for the duration of the copyright) right to use and make copies of the Resource as modified, as not modified or as part of a compilation or derived work; to distribute Copies in the CLARIN Service to End-Users for educational, teaching or research purposes; and to publicly perform the content as modified, as not modified or as part of a compilation or derived work. The permission applies to all known or future modes and means of communication and includes a right to make such modifications that make it possible to use the Resource in other devices and formats.

[ ] LOC: The Resource may not be copied outside the servers of a trusted centre.

[ ] INF: If the Resource is used as material for a scientific work and the work is published, the Copyright holder is to be informed about the publication within reasonable time.

[ ] NC: It is prohibited to use the Resource for gaining economic benefit. However, government-funded or non-profit research projects, e.g. projects funded by <NATIONAL RESEARCH FUNDING AGENCIES>, are not regarded as gaining economic benefit even if a portion of the financing is contributed by companies.

[ ] ReD: The Parties agree that derived works of the Resource may be redeposited with CLARIN. Derived works include but are not limited to annotated or extended versions of the Resource.

Additional rights to the Resource may be agreed separately in writing.

7.3 Information about the license is to be published in conjunction with the Resource in accordance with the terms of the license. A sample End-User license agreement is enclosed in Appendix 3."

(c) Restricted licenses (CLARIN RES Deposition License Agreement):

"7.1 The intellectual property right and/or other rights governing the Resource subject to this Agreement belong to the Copyright holder or his licensors. Any third-party content of the Resource is identified in Appendix 2.

7.2 The Copyright holder grants the Copyright curator a non-exclusive and perpetual (for the duration of the copyright) right to use and make copies of the Resource as modified, as not modified or as part of a compilation or derived work; to distribute Copies in the CLARIN Service to End-Users for personal use; and to publicly perform the content as modified, as not modified or as part of a compilation or derived work. The permission applies to all known or future modes and means of communication and includes a right to make such modifications that make it possible to use the Resource in other devices and formats.

[ ] Use of the Resource presupposes that the user's research plan has been accepted by the Copyright holder.

[ ] LOC: The Resource may not be copied outside the servers of a trusted centre.

[ ] INF: If the Resource is used as material for a scientific work and the work is published, the Copyright holder is to be informed about the publication within reasonable time.

[ ] PD: The Resource includes data covered by the personal data legislation and the Copyright curator acts as an archive having the right to grant temporary use of the Resource only to registered users against an acceptable research plan.

[ ] NC: It is prohibited to use the Resource for gaining economic benefit. However, government-funded or non-profit research projects, e.g. projects funded by <NATIONAL RESEARCH FUNDING AGENCIES>, are not regarded as gaining economic benefit even if a portion of the financing is contributed by companies.

[ ] ReD: The Parties agree that derived works of the Resource may be redeposited with CLARIN. Derived works include but are not limited to annotated or extended versions of the Resource.

Additional rights to the Resource may be agreed separately in writing.

7.3 Information about the license is to be published in conjunction with the Resource in accordance with the terms of the license. A sample End-User license agreement is enclosed in Appendix 3."

Data preservation

Retention period

The CLARIN ERIC does not have specific requirements for the retention period of deposited data. However, individual CLARIN centres may have their own policy regarding the amount of time of data storage. Unless otherwise specified, a centre may retain the data for an indefinite time. The conditions of removal are usually indicated in the deposition agreement signed between the resource owner and the CLARIN centre where the resource is being deposited. For instance, the

CLARIN centre at the Institut für Deutsche Sprache (IDS) in Mannheim specifies the condition of removal of Content (i.e. digital data files) in the following terms:

"if sufficient indispensable grounds exist, the Repository has the right to remove Content from the archive wholly or in part, or to restrict or prevent access to Content on a temporary or permanent basis. The Repository shall inform the Depositor in such cases".[107]

Data file preservation

CLARIN ERIC does not have specific requirements on how datasets should be managed over time. Data management procedures may apply depending on individual centres. For example, the CLARIN deposition and license agreement of the IDS repository states that:

- "The Repository shall ensure, to the best of its ability and resources, that the deposited Content is archived in a sustainable manner and remains legible and accessible.

- The Repository shall, as far as possible, preserve Content unchanged in its original digital format, taking account of current technology and the costs of implementation. The Repository has the right to modify the format and/or functionality of Content if this is necessary in order to facilitate the digital sustainability, distribution or re-use of Content.

- If the access categories 'Restricted Access' or 'Academic Access', as specified at the end of this Agreement, are selected, the Repository shall, to the best of its ability and resources, ensure that effective technical and other measures are in place to prevent unauthorised third parties from gaining access to and/or consulting the Content or substantial parts thereof."[108]

Authenticity (fixity)

By promoting the Data Seal of Approval (DSA) as certification procedure for data management, CLARIN makes a clear statement about checking procedures regarding the authenticity (fixity) of datasets. In fact, one of the 16 criteria of this self-assessment procedure explicitly requires that "the data repository ensures the authenticity of the digital objects and the metadata".[109]

Metadata types and schemas

For the description of linguistic resources, several metadata schemes can be used (e.g. Dublin Core, OLAC, the TEI header for text, IMDI for multimedia collection). However, CLARIN requires its certified repositories to use the Component MetaData Infrastructure (CMDI). As described in the CLARIN-D User Guide, CMDI "provides a framework to create and use self-defined metadata formats. It relies on a modular model of so-called metadata components, which can be assembled together, to improve reuse, interoperability and cooperation among metadata modellers."[110] With the component-based approach, several metadata components can be combined "into a self-defined scheme" that suits the user's particular needs. Accordingly, CLARIN encourages users to share and reuse components that are already available.

[107] http://repos.ids-mannheim.de/resources/DepositorsAgreement.pdf
[108] http://repos.ids-mannheim.de/resources/DepositorsAgreement.pdf
[109] http://www.datasealofapproval.org/en/
[110] http://media.dwds.de/clarin/userguide/text/metadata_CMDI.xhtml

Access and reuse

Access to data objects

Access to the resources distributed by CLARIN may be restricted depending on the User identity and the license type (PUB, ACA, RES) that applies for individual resources. In addition, the user may be required to accept "additional licensing or usage terms for Academic and Restricted Content".[111]

Unless a resource is completely open (PUB), users must be authenticated and authorized to access any data. Proper procedures ensure that only users with the appropriate credentials get access to copyrighted data. Users are generally required to read and accept license agreements where all provisions regulating the usage of such data are specified. Thus, to obtain access to individual resources, end users must log in to identify themselves and agree to a Terms of Service agreement[112], and eventually, one or more End-User License Agreements[113] (EULA). As some of the resources are subject to additional ethical restrictions, the user may also be required to sign a Data User Agreement[114]. To a large extent, such procedures are handled through web interfaces.

The CLARIN Terms of Service (TOS) do not establish any specific embargo policy. However, individual CLARIN Centres may give users the possibility to set an embargo period for specific items (e.g. such items will then not be available for download until the specified date).

Access methods

Depending on the resource content and license type, an authenticated user may be able to download resources and language tools distributed by CLARIN, use the web services provided by the Centres, etc. Thus, the nature of each specific resource will determine whether a link to the data is provided or not, and how the access to data will be done. While a direct download may be available for public data, a request may have to be sent to access data in the case of restricted resources. As CLARIN Centres also provide access to several web services and tools, online analyses are also possible.

Use and reuse of data objects

The reuse of data is regulated in the provisions of the different license types available in CLARIN. As indicated already in 3.3 above, in the legal website of FIN-CLARIN, the Finnish branch of CLARIN, "CLARIN deposition licenses are available for curating a minimal set of usage conditions to include a resource in the CLARIN PUB, ACA or RES categories. The minimal deposition licenses can be used as checklists if you wish to use your own set of deposition licenses to curate additional usage conditions from a resource provider."

CLARIN has foreseen a set of additional restrictions and conditions. These refer to common restrictions and conditions in licenses such as:

- Attribution (+BY): The obligation to cite a resource in publications describing research in which the resource was used.
- No commercialization (+NC): The prohibition to use the resource in commercial applications and obtain economic benefits with it.

[111] https://kitwiki.csc.fi/twiki/pub/FinCLARIN/FinClarinLegal/CLARIN-TOS-v0.95.rtf
[112] https://kitwiki.csc.fi/twiki/pub/FinCLARIN/FinClarinLegal/CLARIN-TOS-v0.95.rtf.
[113] https://kitwiki.csc.fi/twiki/bin/view/FinCLARIN/ClarinEULA.
[114] See, for example, the Data User agreement of the CLARIN Center at the Saarland University, https://fedora.clarin-d.uni-saarland.de/ressources/DataUserAgreement.en.pdf.

- Share alike (+SA): The obligation to distribute the derivate resources only on the same conditions as the original work.
- Inform (+INF): The obligation to inform the copyright holder of any use of the material.
- Local (+LOC): The restriction to download the resource or use it out of the boundaries of the repository where it is stored.
- ReDeposit (+ReD): The obligation to redeposit any derivates of the resource within CLARIN.

Embargo is possible in CLARIN (see 5.1).

User statistics

The FIM model (see section 5.1) used in CLARIN allows for identification of individual users and maintenance of logs/statistics of user actions. According to CLARIN Terms of Service (TOS), "CLARIN maintains usage statistics as a measure of readership and other use of the CLARIN Services by authors and researchers. It is a violation of CLARIN policy for a party to directly or indirectly use CLARIN with a view to affecting download and other usage statistics, or to encourage others to do so. As part of its general right to refuse or terminate service and remove or edit the content of the CLARIN Services, CLARIN reserves the right in its sole discretion to limit access, remove content, and adjust usage statistics to respond to any activity that appears likely to have such an effect."

Usage of Persistent Identifiers

The CLARIN ERIC requires that all resources (metadata records and non-metadata files) have a Persistent Identifier (PID) compatible with the Handle system:

"Centres need to associate (handle) PIDs with their metadata records. These PIDs should be suitable for both human and machine interpretation, taking into account the HTTP- accept header. […]

- Non-metadata files should receive a PID or a PID in combination with a part identifier, if these files:
  - are accessible via internet
  - are considered to be stable by the data provider
  - are considered to be worth to be accessed directly (not via metadata records) by the data provider."[115]

Assigning PIDs to the resources is particularly relevant in that it allows datasets included in CLARIN centres (i) to be easily cited in a paper and (ii) automatically processed by another application or a web service.

There is an arrangement between CLARIN and the European Persistent Identifier Consortium (EPIC)[116] stipulating "that CLARIN members will be able to register PIDs and of course resolve them." The service provided by this consortium is based on the Handle system[117]. However, there is no requirement for CLARIN centres to use a specific PID service. As specified by the CLARIN PID policy summary[118], each CLARIN centre is recommended:

- to have its own prefix;
- to optionally use EPIC (alternatively, Centres can have their own handle server);

[115] hdl:1839/00-DOCS.CLARIN.EU-78
[116] http://www.pidconsortium.eu/
[117] http://handle.net/
[118] http://www.clarin.eu/sites/default/files/CE-2013-0340-PID-policy-summary.pdf

- when using EPIC, to ensure that API version 2 is used accordingly.

CLARIN has also suggested solutions that improve the compatibility of URNs to the Handle system.

Other/Technical

Closure and succession

As a distributed infrastructure, CLARIN is potentially more robust than if it were a single site. Also, as a general policy of CLARIN, data must be safely preserved. However, the application of this policy is implemented at the level of individual centres. Some centres have a safe replication policy. For instance, the Max PLANCK Data Archive in the Netherlands is carrying out data replication at a physical level to preserve the stored data. The data are replicated to a site located in another country (i.e. Germany).

CLARIN does not have a clear policy on succession arrangements.

References

CLARIN web page: http://www.clarin.eu/ (accessed: 03.06.2014)

CLARIN terms of services: https://kitwiki.csc.fi/twiki/pub/FinCLARIN/FinClarinLegal/CLARIN-TOS-v0.95.rtf (accessed: 30.06.2014)

CLARIN End-User License Agreement (EULA): https://kitwiki.csc.fi/twiki/bin/view/FinCLARIN/ClarinEULA (accessed: 30.06.2014)

CLARIN Deposition License Agreements (DELA):

CLARIN PUB Deposition License Agreement: https://kitwiki.csc.fi/twiki/pub/FinCLARIN/FinClarinLegal/CLARIN-DELA-PUB-v0.95.rtf (accessed: 30.06.2014)

CLARIN ACA Deposition License Agreement: https://kitwiki.csc.fi/twiki/pub/FinCLARIN/FinClarinLegal/CLARIN-DELA-ACA-v0.95.rtf (accessed: 30.06.2014)

CLARIN RES Deposition License Agreement: https://kitwiki.csc.fi/twiki/pub/FinCLARIN/FinClarinLegal/CLARIN-DELA-RES-v0.95.rtf (accessed: 30.06.2014)

CLARIN PID policy summary: http://www.clarin.eu/sites/default/files/CE-2013-0340-PID-policy-summary.pdf. (accessed: 04.07.2014)

CLARIN deposition & license agreement – IDS repository: http://repos.ids-mannheim.de/resources/DepositorsAgreement.pdf (accessed: 11.06.2014)

CLARIN Center of the Saarland University (UdS): http://fedora.clarin-d.uni-saarland.de/index.en.html (accessed: 10.07.2014)

CLARIN depositing services: http://www.clarin.eu/content/depositing-services (accessed: 10.07.2014)

CLARIN checklist: hdl:1839/00-DOCS.CLARIN.EU-78 (accessed: 10.07.2014)

CLARIN recommended standards: http://www.clarin.eu/faq/what-standards-are-recommended-clarin (accessed: 10.07.2014)

CLARIN Standardization Action Plan: http://www.clarin.eu/system/files/private/Standardisation%20action%20plan-v8.pdf (accessed: 10.07.2014)

CLARIN - Overview of CLARIN Centres: http://www.clarin.eu/content/overview-clarin-centres (accessed: 30.06.2014)

CLARIN D - The Component Metadata Initiative (CMDI): http://media.dwds.de/clarin/userguide/text/metadata_CMDI.xhtml (accessed: 30.06.2014)

CLARIN – DK: http://clarin.dk/ (accessed: 10.08.2014)

CLARIN DK - Data Seal of Approval: https://assessment.datasealofapproval.org/assessment_105/seal/html/ (accessed: 10.08.2014)

CLARIN DK – guidance for data deposit: http://info.clarin.dk/kom-godt-i-gang/deponer-resurser/vejledning/ (accessed: 10.08.2014)

CLARIN DK – authentication: http://info.clarin.dk/overblik/hvem/ (accessed: 10.08.2014)

CLARIN DK – types of resources: http://info.clarin.dk/overblik/typer/ (accessed: 10.08.2014)

CLARIN DK – license type overview: http://info.clarin.dk/overblik/licenser/ (accessed: 10.08.2014)

CLARINO: http://clarin.b.uib.no/ (accessed: 12.06.2014)

DASISH training module on Authorization and Authentication: http://training.dasish.eu/training/2/ (accessed: 01.06.2014)

Data Archiving and Networked Services: http://dans.knaw.nl/ (accessed: 30.06.2014)

DANS archivist standard protocol: http://www.dans.knaw.nl/sites/default/files/Provenance_document_DEF.pdf (accessed: 10.07.2014)

Data Seal of Approval: http://www.datasealofapproval.org/en/ (accessed: 10.06.2014)

Institut für Deutsche Sprache (IDS) Mannheim: http://www1.ids-mannheim.de/start/ (accessed: 02.06.2014)

European Persistent Identifier Consortium: http://www.pidconsortium.eu/ (accessed: 10.06.2014)

European Research Infrastructure Consortium: http://ec.europa.eu/research/infrastructures/index_en.cfm?pg=eric (accessed: 30.06.2014)

The Handle System: http://handle.net/ (accessed: 10.06.2014)

LINDAT depositing service: https://lindat.mff.cuni.cz/repository/xmlui/page/deposit (accessed: 08.08.2014)

Max Planck Data archive: http://www.mpi.nl/ (accessed: 10.06.2014)

SAML OASIS Standard: http://saml.xml.org/ (accessed: 15.06.2014)

Appendix 2: Case study 2: EUDAT

Description

EUDAT, the European Data project[119], is a pan-European data infrastructure initiated in 2011 to support multi-disciplinary research. One particular goal of the project is to address "the challenge of data proliferation in Europe's scientific and research communities"[120]. To achieve this goal, EUDAT promotes the development of "…a sustainable cross-disciplinary and cross-national data infrastructure that provides a set of shared services for accessing and preserving research data."[121] This infrastructure is conceived as a Collaborative Data Infrastructure (CDI) which functions as "a connected network of European research institutions ('community sites') and data centres, each offering one or more common EUDAT data services to both participating research communities and independent researchers."[122]

Policy Model

EUDAT has a data life-cycle approach and addresses a wide range of data preservation services[123] including sharing and storing research data, finding data objects and collections, data staging, and safe replication. Also, the infrastructure has defined various policies with respect to data management, open access to data, the use of persistent identifiers.[124]

EUDAT's model for data management[125] generally conforms to the Guidelines on Data Management in Horizon 2020[126] (H2020). According to this model, participating projects are required to have Data Management Plans (DMP) "detailing what data the project will generate, whether and how it will be exploited or made accessible for verification and re-use, and how it will be curated and preserved."[127] This requirement aims to support the data management life cycle for digital data objects collected, processed or generated by a given project. The H2020 model also specifies additional policies with respect to e.g. discoverability, accessibility, intelligibility of scientific research data as well as their interoperability to specific quality standards.

Content coverage

Scope

EUDAT supports multi-disciplinary research by providing a set of shared technical services that address the needs of various research communities from different disciplines, including[128]:

- CLARIN (Linguistics)
- diXa (Chemical Safety)
- DRIHM (Hydrometerology)
- ENES (Climate Modeling)
- EPOS (Seismology, Volcanology)

[119] http://www.eudat.eu
[120] http://www.eudat.eu/news-media/published-articles/open-access-and-data-management-planning
[121] http://www.ijdc.net/index.php/ijdc/article/view/8.1.279
[122] http://www.eudat.eu/news-media/published-articles/open-access-and-data-management-planning
[123] http://www.eudat.eu/services
[124] http://www.eudat.eu/system/files/Open-Access-and-Data.pdf
[125] http://www.eudat.eu/news-media/published-articles/open-access-and-data-management-planning
[126] http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf
[127] Ibid.
[128] http://www.eudat.eu/eudat-communities

- INCF (Neuroinformatics)
- LifeWatch (Biodiversity)
- VPH (Human Physiology)

Mandate

The mandate of EUDAT is to support a CDI,"…which will allow researchers to share data within and between communities and enable them to carry out their research effectively.

Kinds of data (selection and appraisal)

EUDAT includes "any kind of stable research data" which a user wants to "preserve and share in a safe environment"[129]. EUDAT's depositing service, B2SHARE, is designed to store the so called "long tail of data", which are "often stored on disconnected machines such as notebooks, desktops or departmental servers thereby risking scientific data loss, either because other researchers do not have easy access to the data or because such storage is often relatively insecure"[130].

Form/status of data

EUDAT does not restrict the data to be stored by its status. Data included in the infrastructure may be primary, processed, empirical and theoretical data as well as data as basis for a publication. The only requisite is for data to be included into the infrastructure is that "the data source or purpose of the data has a scientific background"[131].

Data versions / version control

In the context of EUDAT, the data repository is required to ensure "the integrity of the digital objects and the metadata"[132], including versioning of the deposited data. "New versions of archived resources can be deposited, in which case the old versions will be moved to a version archive. In the future, these old versions will also be made available to the end users but this is currently not yet the case."[133]

Data formats

EUDAT does not have restrictions on which data formats will be accepted for deposit. The research data "can have various formats - papers, spreadsheets, audio-visual media, practically any kind file and format"[134].

Size / volume of data

In general, EUDAT does not have restrictions on the number of files that a user can deposit. However, the infrastructure sets limits on the size of individual files not exceeding 2GB[135].

[129] http://www.eudat.eu/b2share-faq-generic
[130] http://www.eudat.eu/b2share
[131] http://www.eudat.eu/b2share-faq-generic
[132] http://www.eudat.eu/system/files/EUDAT-DEL-WP7-D7%202%201-Managing%20data%20curation%20and%20long-term%20preservation%20in%20a%20federated%20environment.pdf
[133] Ibid.
[134] http://www.eudat.eu/b2share-faq-generic
[135] Ibid.

Pre-ingest

Guidance for researcher

EUDAT offers information and guidance to data depositors as part of its service. Instructions for depositing data (including registration and login) are provided in form of an FAQ webpage and a user documentation.[136]

Besides, EUDAT provides training material (e.g. screencasts) to help demonstrate and teach the use of the depositing service. Also, a service helpdesk[137] is available, and can be contacted through a webform. Furthermore, EUDAT has education and training activities designated to provide assistance to the potential users of the infrastructure to be "in how to optimally use the platform of technologies, tools and services provided by the project"[138].

Data ingest

Eligible depositors

EUDAT has no restriction on the position or status of data a depositor, except that (s)he must register. Only registered users can deposit data.[139] But otherwise, a data depositor can be either an individual user or a member of a specific community. In practice, EUDAT depositing service (B2SHARE) is "open to all researchers and scientists who are affiliated to research institutions, universities as well as to individual researchers (citizen scientists)"[140].

Review/moderation of deposited data

EUDAT has a set of checking procedures to make assessment with respect to the quality/validity/accuracy of the ingested data. In accordance with these procedures, control sums are computed and checked during ingest in order to ensure the integrity of the data . Also, contents will be randomly checked "to prevent the upload of inappropriate content according to the Terms of Use, such as non-scientific or illegal data".

Depositor agreements /responsibility

When submitting data for archiving, data producers are recommended to follow EUDATs terms of use[141]. Before storing his/her data, the data provider agrees to accept the conditions of use.

Requirements of confidentiality

According to the EUDAT Terms of Use, the user agrees to "respect the legal protection provided by copyright and licensing of software and data as well as intellectual property and confidentiality agreements."

Information about anonymization procedures is not disclosed on EUDAT website.

IPR (rights and ownerships)

EUDAT has no claim over the ownership of any of data deposited into the infrastructure: "Ownership of data will remain with the contributor, although EUDAT will encourage openness from all participants and contributors."[142]

---

[136] https://b2share.eudat.eu/docs/b2share-guide
[137] http://www.eudat.eu/b2share
[138] http://www.eudat.eu/training
[139] http://www.eudat.eu/User%20Documentation%20-%20B2SHARE.html
[140] http://www.eudat.eu/b2share
[141] EUDAT's terms of use, http://www.eudat.eu/terms-use-eudat-b2share-service

Furthermore, EUDAT encourages all contributors, including stakeholders of the CDI "to adopt open licenses for access to their data collections"[143]. For instance, to those communities which wish to be part of the CDI, EUDAT recommends to adopt the two main licensing schemes:

- "Creative Commons v4.0, particularly:
  - the Creative Commons Attribution License 4.0 International (`CC BY 4.0');
- Open Data Commons, particularly:
  - the Open Data Commons Open Database License (ODbL) v1.0;
  - the Open Data Commons Attribution License v1.0."[144]

Data preservation

Retention period

Unless otherwise specified, EUDAT may retain the data for an indefinite time. The measures to be taken in the event of a closure of the data infrastructure are indicated on the website. "In the unlikely event that the B2SHARE service would draw to close in the future", there will be a guarantee to keep the data "archived and accessible for at least 2 years"[145]. Also, in such case, EUDAT is committed to helping data providers to migrate their data "to other suitable repositories"[146].

Data file preservation and security level

EUDAT has specified data management procedures which follow the Guidelines on Data Management in Horizon 2020 (H2020)[147]. According to the H2020 guidelines, participating projects are required to develop a Data Management Plan (DMP), which specifies "what data the project will generate, whether and how it will be exploited or made accessible for verification and re-use, and how it will be curated and preserved" [148].

For archiving and preservation (including storage and backup), the DMP should include a description of "the procedures that will be put in place for long-term preservation of the data. Indication of how long the data should be preserved, what is its approximated end volume, what the associated costs are and how these are planned to be covered."[149]

Additional policies specified in these guidelines require that scientific research data should be easily discoverable, accessible, assessable and intelligible, usable beyond the original purpose for which it was collected, and interoperable to specific quality standards.

Currently, EUDAT works in collaboration with the Digital Curation Centre on a version of DMPonline with the aim to develop a data management planning tool capable of both to capture the H2020

[142] http://www.eudat.eu/news-media/published-articles/open-access-and-data-management-planning
[143] Ibid.
[144] Ibid.
[145] http://www.eudat.eu/b2share-faq-generic
[146] Ibid.
[147] http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf
[148] Ibid.
[149] Ibid.

guidelines and to address the needs of European research. Also, EUDAT data management policies include using iRODS to manage large collection of digital objects, and to maintain metadata[150].

Authenticity (fixity)

EUDAT's policies for safe replication and data staging explicitly require that "checksums are computed on ingest and subsequently when the object is updated" and used "to periodically validate all replicas of a digital object"[151].

Metadata types and schemas

During deposit, data providers can fill a generic metadata form[152], which contains both mandatory and optional fields. For instance, while title and description are mandatory, creator, open access, licence, publisher, publication date and tags are optional.

In addition, EUDAT harvests metadata from metadata providers (i.e. the communities and sub-communities) using the standard OAI-PMH[153] interface. Harvested metadata is made searchable via the EUDAT metadata catalogue, B2FIND[154]. "The community itself decides which metadata is made available for EUDAT."[155]

Access and reuse

Access to data objects

The resources distributed in the EUDAT CDI can be accessed and used for free.[156] Unless a resource is subject to specific license or ownership conditions, registered users will have free access and use. This follows EUDAT fundamental principles for promoting open access defined as "the free availability of data on the public Internet, permitting any user to reproduce and redistribute them for any purpose, and in particular for the purpose of non-commercial research, without financial, legal or technical barriers. The only allowable constraint on reproduction and redistribution should be to give authors control over the integrity of their work and the right to be properly acknowledged and cited."[157] Based on these principles, EUDAT strongly recommends the communities wishing to join the CDI to adopt the Creative Commons and Open Data Commons licensing schemes, as mentioned in 3.5.

According to EUDAT guiding principles, "all data in the CDI should, in time, become full open access. Open access is the norm for CDI data"[158].

Furthermore, in the context of EUDAT, users have the possibility to set embargo periods for original producers, but on the "condition that such data become openly accessible when the embargo period expires."[159]

---

[150] http://www.eudat.eu/User%20Documentation%20-20iRODS%20Deployment.html
[151] http://www.eudat.eu/deliverables/d721-managing-data-curation-and-long-term-preservation-federated-environment
[152] http://eudat.eu/User%20Documentation%20-%20B2SHARE.html
[153] http://www.openarchives.org/pmh/
[154] http://www.eudat.eu/User%20Documentation%20-%20B2FIND.html
[155] Ibid.
[156] http://www.eudat.eu/news-media/published-articles/open-access-and-data-management-planning
[157] Ibid.
[158] http://eudat.eu/User%20Documentation%20-%20B2SHARE.html
[159] http://www.eudat.eu/news-media/published-articles/open-access-and-data-management-planning

Access methods

Both registered and unregistered may be able to search, browse and download files distributed by EUDAT via the B2SHARE graphical, web-based tool. For unregistered users, however, EUDAT has foreseen a specific restriction that they can only search for / download data sets that are under public licenses. "EUDAT do not offer data processing (beyond metadata annotation) as a core service."[160]

Use and reuse of data objects

Access and reuse are determined by the licensing schemes given in the IPR section above.

User statistics

Currently, EUDAT works on designing and implementing a collaborative Authentication and Authorization Infrastructure (AAI) in a federated scenario[161]. This AAI model should allow for:

- "Leveraging existing identification systems within communities and/or data centres.
- Establishing a network of trust amongst those involved in the AAI processes (including Identity Providers (IdPs), Service Providers (SPs), Attribute Authorities and Federations), and
- Attribute harmonization"

The strategy for the design and implementation of the AAI model consists in considering first internal EUDAT communities (e.g. ENES and CLARIN), before extending it to all interested communities. In connection with this, different options are under consideration, including the possibility to join existing federations, or to use identity provider (IdPs) from existing federations, or to use international interfederation services similar to the eduGain model.

The design of EUDAT AAI system will take various implementations (e.g. eduRoam, OpenID, and Shibboleth) and technologies under consideration[162].

Usage of Persistent Identifiers

EUDAT requires that resources in the CDI have an associated Persistent Identifier (PID) compatible with the Handle system. There is an arrangement between EUDAT and the European Persistent Identifier Consortium (EPIC) "to ensure all data objects registered in the CDI receive a unique, persistent Handle."[163] EUDAT requires from individual communities and data centres in the CDI to integrate Handle in their infrastructure. As specified in the EUDAT PID policy, individual communities and data centres "need to have a prefix" with the following two options[164]:

1. A community / data center can run its own Handle server;
2. Alternatively, details of the prefix can be passed to EUDAT partner SURFsara which will manage it on the behalf of this community/data center.

[160] Ibid.
[161] http://www.eudat.eu/authentication-and-authorization-infrastructure-aai
[162] This list of technologies and implementations under consideration has been taken from EUDAT website, http://www.eudat.eu/authentication-and-authorization-infrastructure-aai.
[163] http://www.eudat.eu/news-media/published-articles/open-access-and-data-management-planning
[164] http://www.eudat.eu/User%20Documentation%20-%20PIDs%20in%20EUDAT.html

Other / Technical

Closure and succession

To face natural disasters, EUDAT has defined a general policy-rule, which stipulates that research data must be safely preserved through via a backup or "a robust, safe and highly available replication service"[165]. To meet this vision, EUDAT has developed a safe replication service (B2SAFE), which "is based on the execution of auditable policy-rules and the use of PIDs, as offered by the EPIC Handle service"[166]. In addition to the backup or replication service, EUDAT also provides a disaster recovery plan.[167]

[165] http://www.eudat.eu/b2safe
[166] Ibid.
[167] http://www.eudat.eu/b2share

Appendix 3: Policy models from data centres

ADS

Description

The Archaeology Data Service (ADS), located at the University of York, UK, supports research, learning and teaching with freely available digital resources. It preserves digital data in the long term by promoting and disseminating a broad range of data in archaeology. The ADS was established in 1996, as one of five discipline-based service providers within the Arts and Humanities Data Service (AHDS). In 2008 the Arts and Humanities Research Council (AHRC) and JISC ceased their funding for AHDS, and AHDS Archaeology ceased to exist. However, AHRC still supports the ADS which continues to provide ongoing support for digital preservation and re-use, for research, learning and teaching for Archaeology and the Historic Environment sector. In 2010 the ADS implemented the Data Seal of Approval[168]. This was renewed in 2013 and ADS currently holds the Data Seal for 2014-2015.

Policy Model

The ADS states that it follows the OAIS reference model. They also have several internal policies and procedures that guide and inform the archiving work. The key document for their preservation and archiving routines is the ADS Preservation Policy[169]. Several of the supplementary documents are available from the ADS preservation and digital archiving webpage[170] (e.g. ADS Repository Operations, ADS Ingest Manual, Copyright Infringement Policy). In the Preservation Policy links are provided to several internal policy and strategy documents.

[168] ADS, Implementation of the Data Seal of Approval: https://assessment.datasealofapproval.org/assessment_36/seal/html/
[169] ADS Preservation Policy: http://archaeologydataservice.ac.uk/attach/preservation/PreservationPolicyV1.3.1.pdf
[170] http://archaeologydataservice.ac.uk/advice/preservation

Table 3: ADS Preservation Policy elements

Id	Description	Description
1	**Principal Statement**	*General statement on activities and services of the ADS. Content based on Beagrie (2008).*
2	**Contextual Links**	*Provides a list of policy and strategy documents related to the preservation policy, both internally and externally. Among the documents mentioned are: ADS Five Year Plan; ADS Risk Register; ADS Collections Policy; ADS Preservation Strategy; ADS Disaster Recovery Plan; and ADS Access Policy (in prep).*
3	**Preservation Objectives**	*States the core objective of the long term preservation activities of the ADS. Describes the long-term preservation framework and its conformance to the OAIS model.*
4	**Identification of Content**	*Establishes the archives' relation to its user community. Includes statement on the costs of long-term preservation and how it affects the archive content.*
5	**Procedural Accountability**	*Descriptions of roles and responsibilities of staff.*
6	**Guidance and Implementation**	*Describes categories of the lifecycle of digital assets and their equivalent OAIS functional entities.*
6.1	Data creation	*Statement on the archives' role in the data creation phase of the data lifecycle. Provides links to the archives' guidance material for this period.*
6.2	Acquisition, retention or disposal	*Links to a number of documents that guide the process of ingesting a SIP into the archive. Describes how a SIP is processed*
6.3	Preservation and management	*Links to a number of documents on the ADS Wiki (internal site) that informs on the ongoing preservation and management of data.*
6.3.1	Storage and resilience	*Describes the storage procedures.*
6.3.2	Data management	*Describes the preservation strategy (technology watch, etc.) and the general management of data.*
6.4	Access and use	*This section is concerned with the access and use of the DIP; finding a resource, i.e. rights management and receiving a data collection. It is also concerned with the availability, reliability and security of delivery systems. Links to an Access Policy (which is 'in preparation').*
6.4.1	Prerequisites	*Describes the legal and regulatory framework which applies to the accessibility of resources held by the ADS.*
6.4.2	Resource discovery	*Describes the ADS' metadata scheme and strategy.*
6.4.3	Rights management	*Statement on terms and conditions for data access.*
6.4.4	Receiving data	*Statement on routines for delivery of data.*
6.4.5	Security of delivery systems	*Links to relevant security documents (e.g. Systems Overview, Disaster Recovery Plan, etc.)*
6.4.6	Consumer access analysis	*Statement on the use of statistics on consumer activity.*
6.4.7	Outage	*Statement on maintenance schedules and possible downtime of services.*
7	**Glossary**	*Glossary of abbreviations and technical terms.*

CentERdata(LISS)

Description

CentERdata is a data collection and research institute located at the Tilburg University. The institute is specialized in online survey research and in collecting and analysing (panel) data. It also specialises on policy analysis and model development with a main focus on labour market, pensions, efficiency studies, social security, and consumer behaviour.

The LISS panel (Longitudinal Internet Studies for the Social sciences) is the principal component of the Dutch MESS (Measurement and Experimentation in the Social Sciences) project. LISS consists of 5000 households, comprising 8000 individuals. The panel is based on a true probability sample of households drawn from the population register by Statistics Netherlands. Panel members complete online questionnaires every month. Part of the interview time available in the both the LISS and Immigrant panel is reserved for the LISS Core Study. This longitudinal study is repeated yearly and is designed to follow changes in the life course and living conditions of the panel members. In addition to the LISS Core Study there is room to collect data for different research purposes. Many disciplines, from linguistics to medical sciences, have taken up the opportunity to use the research infrastructure[171]. The data collected in these panels are preserved and disseminated via the LISS Data Archive which is managed by CentERdata. The data which are archived in and disseminated via the LISS Data Archive are also deposited in the online archiving system EASY of DANS. Data Users have access to the metadata via the EASY system, but are referred to the LISS Data Archive for accessing the actual data files and more detailed metadata.

The LISS data archive has implemented and complies with the 2010 version of the Data Seal of Approval.

Policy Model

CentERdata provides a Preservation and Dissemination Policy for the LISS Data Archive[172]. The policy elements and the general content and operational descriptions are similar to the preservation policies of UKDA and DANS. It presents the purpose, operations and functions of the archive, in addition to descriptions of security and long-term preservation measures of the LISS data. The data processing elements (6.2 to 6.6) are described using the OAIS model and terminology[173].

[171] About the LISS Panel: http://www.lissdata.nl/lissdata/About_the_Panel
[172] Preservation and Dissemination Policy of the LISS Data Archive: http://www.lissdata.nl/assets/uploaded/Preservation%20and%20Dissemination%20Policy%20of%20the%20LISS%20Data%20Archive_1.1.pdf
[173] Ibid.

Table 4: CentERdata/LISS preservation policy elements

ID	Policy Element	Policy description / content
1	**Introduction**	*Describes the general content of the policy.*
2	**Purpose**	*Describes the purpose of the archive through a mission statement and a description of scope and objectives.*
2.1	Mission	*General mission statement.*
2.2	Scope and Objectives	
3	**Legal and Regulatory Framework**	*Statement on the laws and regulations that the archive complies with.*
4	**Organisation**	*Describes the overall organisation of the archive through descriptions of the roles and responsibilities for the different organisational units within each of the data processing stages (production, archiving/management and consumption).*
4.1	Data Production	
4.2	Data Archiving and Management	
4.3	Data Consumption	
5	**Co-operation**	*Describes some of the third parties and co-operations which are related to the archive (DANS, VANCIS, DDI Alliance).*
5.1	DANS	
5.2	VANCIS	
5.3	DDI Alliance	
6	**Data Process**	*Describes the different tasks of the archive by applying the OAIS model, i.e. ingest, data management, archival storage, access, preservation planning and administration. In addition, a brief description of the pre-ingest processes are provided.*
6.1	Pre-ingest	
6.2	Ingest	
6.3	Archival Storage and System Architecture	
6.4	Data Management and Administration	
6.5	Access and Dissemination	
6.6	Preservation Planning and Long-Term Preservation Strategy	
7	**Data Safeguarding**	*Describes the safety measures that are involved in security, risk management, media monitoring and refreshing strategies.*
7.1	Security and Risk Management	
7.2	Media Monitoring and Refreshing Strategy	

CSDA

Description

The CSDA (Czech Social Science Data Archive), which is a department of the Institute of Sociology of the Academy of Sciences of the Czech Republic, documents, stores and disseminates research data from social science research projects within the Czech Republic. CSDA is a Service Provider for the Consortium of European Social Science Data Archives (CESSDA).

In addition to acquiring and archiving datasets from Czech social science research and making them publicly available for secondary analysis, it also provides technical and organisational support for large-scale survey research programmes, e.g. Czech participation in the International Social Survey Programme (ISSP) and the European Social Survey (ESS) or the newly established Czech Household Panel Survey (CHPS)[174]. It also provides training in data management and survey methodology.

Policy Model

The CSDA preservation policy[175] describes the activities of the information preservation system of the archive and is modelled on the data-lifecycle elements (i.e. from the pre-ingest stage to the stage of providing users access to the data). The policy states that the different stages of the preservation process are in line with the OAIS model and that each activity can be assigned to a given OAIS function.

[174] About the Czech Social Science Data Archive: http://archiv.soc.cas.cz/en/about-czech-social-science-data-archive
[175] CSDA Preservation policy: http://archiv.soc.cas.cz/sites/default/files/csda_preservation_policy_0.pdf

Table 5: CSDA Preservation Policy elements

ID	Policy element	Description
1	**Outline / Introduction (Preservation Policy)**	*General introduction to the policy structure. States that the different stages of the preservation process are in line with the OAIS model and that each activity "…can be assigned to a given OAIS function".*
1.1	Overview of job positions in the management of the CSDA preservation system	*Roles and responsibilities of the archive staff.*
1.1.1	Archive Director	
1.1.2	Acquisition & Ingest Administrator	
1.1.3	System Administrator	
1.1.4	Access Coordinator	
1.2	Shared & ad hoc activities	*Some functions of the archive are fulfilled on an "ad hoc basis". These are listed here.*
2	**Acquisitions & Ingest Administrator**	*Describes the functions of the Acquisitions & Ingest Administrator. Responsibilities include pre-ingest stages, receiving SIP and submission into an AIP, and coordinating updates of data and metadata.*
2.1	Scope of responsibility	
2.2	Planning the search for data for preservation	
2.3	Contacting producers and starting cooperation	
2.4	Concluding submission agreements	
2.5	Responding to submission requests	
2.6	Receiving submissions	
2.7	Acceptable data formats	
2.8	Primary quality assurance	
2.9	AIP generation	
2.10	Collaboration on SIP and AIP audit, reconciliation of audit reports	
2.11	Generating Descriptive Information, transferring AIP to Archival Storage	
2.12	Coordinating updates	
3	**System Administrator**	*Describes the functions of the System Administrator. Includes "…gathering, storing and making available of data and the checking of information on all activities taking place within the Archive".*
3.1	The CSDA database system	
3.1.1	System Configuration Archive	
3.1.2	Digital Dataset-Related Materials Archive	
3.1.3	Analogous Dataset-Related Materials Archive	
3.1.4	Document Archive	
3.1.5	Submission Information Package (SIP) Database	
3.1.6	Archival Information Package (AIP) Database	
3.1.7	Dissemination Information Package (DIP) Database	
3.1.8	File Versions Database	
3.1.8	User Request Database	
3.1.10	Staff List	
3.1.11	Client directory	
3.2	Request Administration	
3.3	Archive Performance Monitoring	
3.4	Query Processing	
3.5	Versions Administration	
3.6	Technology Base Administration	
3.7	Storage Media Administration	
3.8	Backup Storage Media	
3.9	Error Checking	
4	**Access Coordinator**	*Describes the functions of the Access Coordinator. The AC is responsible for "…timely and correct transfer of the Dissemination Information Package (DIP) in the required format. At the same time, the Coordinator shall continuously check DIPs for completeness and accuracy, communicate with users and monitor their requests".*
4.1	Specific activities falling under the Coordinator's responsibility	
4.2	Coordination of Access Activities	
4.3	Query Activation & Response Delivery	
4.4	Data & Documents Retrieval	
4.5	Provision of Access to Data Preserved by the CSDA	
4.6	Response to Non-Nesstar Query Requests	
4.7	DIP Generation	

DANS (EASY)

Description

DANS (Data Archiving and Networked Services) is an institute of the Royal Netherlands Academy of Arts and Sciences (KNAW) and the Netherlands Organisation for Scientific Research (NOW). DANS runs EASY (Electronic Archiving SYstem), which is an online archiving system that offers access to research datasets and online depositing of data. The collection includes datasets from the Humanities, e.g. history and archaeology, social and behavioural sciences, and geospatial sciences. In 2007 agreements were formalised for archaeologists to deposit their data in the e-depot for Dutch archaeology (EDNA[176]). Since 2010 the archive has gradually been extending its domain from social sciences and humanities to life sciences.

When DANS was established by KNAW and NWO, they assigned it the task of developing a Seal of Approval for data, to ensure that archived data can still be found, understood and used in the future. In 2008 DANS published the first edition of the Data Seal of Approval. In 2009 it was handed over to an international Board[177].

DANS EASY currently holds the 2014-2015 Data Seal of Approval.

Policy Model

DANS provides a preservation policy document that outlines the principles which underpin the main activities of DANS "…regarding sustainable identification and preservation of, as well as access to digital research data for use and re-use within its user communities"[178]. The policy generally conforms to the OAIS reference model, with alterations that are specific to the materials held within the DANS archive. Further, the policy explicitly states that it has been modelled on UKDA's preservation policy (the 2011-version).

[176] http://www.edna.nl/
[177] DSA, about: http://datasealofapproval.org/en/information/about/
[178] DANS Preservation Policy: http://dans.knaw.nl/sites/default/files/file/EASY/20140220%20Preservation%20Policy%20v1_0.pdf

Table 6: DANS Preservation Policy elements

ID	Policy Element	Description
1	**Outline**	*Outline and purpose of the policy. Outlines "…the principles which underpin the main activities of DANS (…) regarding sustainable identification and preservation of, as well as access to digital research data for use and re-use within its user communities".*
2	**Mission of the Archive**	*Mission statement and background of the archive.*
3	**Scope and objectives of the policy**	*Delimits the scope of the policy while describing the primary objectives. It states that the archive's primary objective is to "…identify, preserve and make available for use digital research data that have permanent or continuing value".*
3.1	Scope	
3.2	Objectives	
4	**Requirements**	*Lists a series of archive requirements and the legal documents that apply to the archive activities. Includes general statement on the legal and regulatory framework.*
4.1	The Archive's requirements	
4.2	Legal and regulatory framework	
5	**Roles and responsibilities**	*Statement on roles and responsibilities of all DANS staff.*
6	**Content coverage**	*Statement on data types and software/hardware. States that the preferred formats are the file formats which the Archive trusts to "…offer the best long-term guarantees for usability, accessibility and robustness".*
7	**Implementing the preservation policy**	*Segment that is structured around the main functional concepts of the OAIS reference model, i.e. ingest, archival storage, data management, access, administration and preservation planning. Also includes a segment on the pre-ingest function (not part of the OAIS model).*
7.1	Pre-ingest function	
7.2	Ingest function	
7.3	Archival storage function	
7.4	Data management function	
7.5	Access function	
7.6	Administrative function	
7.7	Preservation planning function	
8	**Integrity and security**	*General statement on integrity and security. States that the archive "…is committed to taking all necessary precautions to ensure the physical safety and security of the data it preserves. This includes a periodical technology vulnerability scan, the SLA with the data storage provider, a procedure for file fixity checking, an annual DRAMBORA Risk Assessment as well as the Declaration of Confidentiality for employees and a periodical safety inventory by the KNAW".*
9	**Sustainability plans and funding**	*Description of sustainability and funding plans. States that "…to fulfil its mission the Archive receives structural lump sum financing from both the KNAW and Netherlands Organisation for Scientific Research (NWO)".*

Dataverse

Description

The Dataverse Network Project is housed at the Institute for Quantitative Social Science (IQSS) at Harvard University. Coding of the Dataverse Network software began in 2006, building on an earlier Virtual Data Centre (VDC) project which spanned 1999-2006 as a collaboration between the Harvard-MIT Data Centre (now part of IQSS) and the Harvard University Library.

It builds on the 'replication standard' introduced in 1995 by Gary King[179]. The replication standard holds that "…sufficient information exists with which to understand, evaluate, and build upon a prior work if a third party can replicate the results without any additional information from the author."

The Dataverse Network is an open source application to publish, share, reference, extract and analyse research data. The main goal of the Network is to solve the problems of data sharing through building technologies that enable institutions to reduce the burden for researchers and data publishers, and incentivise them to share their data. By installing Dataverse Network software, an institution is able to host multiple individual virtual archives, called "dataverses" for scholars, research groups, or journals, providing a data publication framework that supports author recognition, persistent citation, data discovery and preservation[180].

It has a *centralised* software installation and data repository, with individual *distributed* data archives with its own branding; a Dataverse Network hosts multiple 'dataverses' where each dataverse contains studies or collections of studies, and each study contains cataloguing information that describes the data plus the actual data and complementary files[181].

Policy Model

Due to the self-curation nature of the Harvard Dataverse Network, owners or distributors of individual datasets have control over selection of materials, documentation, access policies and data user agreements of their datasets. However, the Harvard Dataverse encourages "…good curation practices through metadata, proper documentation and versioning to enable data discovery and reuse"[182]. Hence, instead of a specific preservation policy, the Dataverse Network provides a data management plan (DMP) that describes how the project conforms to NSF (National Science Foundation) Policy on dissemination and sharing of research results. Dataverse provides an outline of recommended elements for consideration in most data management plans (see below). The Dataverse outline is based on a comparison of data management checklists produced by funders, "prominent data archives" (does not state which ones), and library associations; a review of sample data management plans from funders and data archiving organizations; in addition to Library of Congress preservation format recommendations[183].

Additionally, they provide a template for a data management plan which is tailored to the National Science Foundation's (NSF) requirements, and is appropriate for data that is relatively small in size and complexity.

---

[179] Gary King's homepage at Harvard: http://gking.harvard.edu/category/research-interests/applications/informatics-and-data-sharing
[180] Wikipedia: http://en.wikipedia.org/wiki/Dataverse
[181] The Dataverese Network Project; About the project: http://thedata.org/book/about-project
[182] Data Backup and Preservation terms: http://thedata.org/book/data-backup-preservation-terms
[183] Library of Congress, Sustainability of Digital Formats: http://www.digitalpreservation.gov/formats/index.shtml

Table 7: Dataverse preservation policy/DMP elements

ID	Policy Elements	Description / Definition
1	**Project Information**	
2	**General Data Management Plan Information**	
3	**Data description**	
3.1	Nature of data {generated, observed, experimental information; samples; publications; physical collections; software; models}	
3.2	Scale of data	
4	**Existing Data** [If applicable]	
4.1	Description of existing data relevant to the project	
4.2	Plans for integration with data collection	
4.3	Added value of collection, need to collect/create new data	
5	**Audience**	
5.1	Potential secondary users	
5.2	Potential scope or scale of use	
5.3	Reasons not to share or reuse	
6	**Access and Sharing**	
6.1	Plans for depositing in an existing public database	
6.2	Access procedures	
6.3	Embargo periods	
6.4	Access charges	
6.5	Timeframe for access	
6.6	Technical access methods	
6.7	Restrictions on access	
7	**Formats**	
7.1	Generation and dissemination formats and procedural justification	
7.2	Storage format and archival justification	
8	Metadata and documentation	
8.1	Metadata to be provided	
8.2	Metadata standards used	
8.3	Treatment of field notes, and collection records	
8.4	Planned documentation and supporting materials	
8.5	Quality assurance procedures for metadata and documentation	
9	Data Organization [if complex]	
9.1	File organization	
9.2	Naming conventions	
10	Quality Assurance [if not described in main proposal]	
10.1	Procedures for ensuring data quality in collections, and expected measurement error	
10.2	Cleaning and editing procedures	
10.3	Validation methods	
11	**Storage, backup, replication, and versioning**	
11.1	Facilities	
11.2	Methods	
11.3	Procedures	
11.4	Frequency	
11.5	Replication	
11.6	Version management	
11.7	Recovery guarantees	
11	**Security**	
12.1	Procedural controls	
12.2	Technical Controls	
12.3	Confidentiality concerns	
12.4	Access control rules	
12.5	Restrictions on use	
13	**Responsibility**	
13.1	Individual or project team role responsible for data management	
14	**Budget**	
14.1	Cost of preparing data and documentation	
14.2	Cost of permanent archiving	
15	**Intellectual Property Rights**	
15.1	Entities who hold property rights	
15.2	Types of IP rights in data	
15.3	Protections provided	
15.4	Dispute resolution process	
16	**Legal Requirements**	
16.1	Provider requirements and plans to meet them	
16.2	Institutional requirements and plans to meet them	
17	**Archiving and Preservation**	
17.1	Requirements for data destruction, if applicable	
17.2	Procedures for long term preservation	
17.3	Institution responsible for long-term costs of data preservation	
17.4	Succession plans for data should archiving entity go out of existence	
18	**Ethics and privacy**	
18.1	Informed consent	
18.2	Protection of privacy	
18.3	Other ethical issues	
19	**Adherence**	
19.1	When will adherence to data management plan be checked or demonstrated?	
19.2	Who is responsible for managing data in the project?	
19.3	Who is responsible for checking adherence to data management plan?	

Dryad

Description

Dryad is a curated repository of data, served by North Carolina State University, that makes the data underlying scientific publications and peer reviewed articles accessible and reusable for other researchers. Dryad is governed by a non-profit membership organisation which originates from an initiative among a group of leading journals and scientific societies in evolutionary biology and ecology to adopt a joint data archiving policy (JDAP)[184] for their publications and the recognition that community-governed data infrastructure was needed to support such a policy. Their services have since expanded beyond biology and ecology and Dryad now provides a general-purpose platform for several types of data, mainly with the international scientific and medical literature. Dryad coordinates with journals to integrate article and data submission. The repository is community driven, governed and sustained by a consortium of scientific societies, publishers, and other stakeholder organisations and funded partly by awards from the US National Science Foundation (NSF) and data publishing charges (DPCs). Although the content is available free of cost to researchers, educators or students, irrespective of nationality or institutional affiliation, Dryad is dependent on financial support from members and data submitters to provide free access. Dryad's DPCs are "…designed to sustain its core functions by covering the basic costs of curating and preserving data". The prices range from $25 per published research article to $80 per data package, depending on payment plan and membership status[185]. Dryad encourages organisations (e.g. publishers, scientific societies, libraries, funders) to cover the costs of DPCs on behalf of their community of researchers.

Among the options available for the data submitters is the opportunity to update datafiles (without overwriting the older version) and set limited-term embargoes post-publication. Submitted data are linked both to and from the corresponding publication and, where appropriate, to and from external data repositories. It also assigns Digital Object Identifiers (DOIs) to data so that researchers can gain credit through proper data citation. Regarding long-term preservation, it is stated that Dryad are applying migration techniques when file formats become obsolete and they are partnering with CLOCKSS[186] to guarantee "indefinite access" and to its content.

Data users (and re-users) can download the data packages directly with full data and metadata descriptions, and DOI links to the original publication are provided on the data access page. It is stated that contents are curated to ensure the validity of the files and metadata. Contents are free to download and have no legal barriers to reuse, as the authors (data submitters) have "…waived all

[184] The Joint Data Archiving Policy describes a requirement that supporting data be publicly available. This policy was adopted in a joint and coordinated fashion by many leading journals in the field of evolution in 2011, and JDAP has since been adopted by other journals across various disciplines. JDAP consists of the following text:

" *<< Journal >> requires, as a condition for publication, that data supporting the results in the paper should be archived in an appropriate public archive, such as << list of approved archives here <<. Data are important products of the scientific enterprise, and they should be preserved and usable for decades in the future. Authors may elect to have the data publicly available at time of publication, or, if the technology of the archive allows, may opt to embargo access to the data for a period up to a year after publication. Exceptions may be granted at the discretion of the editor, especially for sensitive information such as human subject data or the location of endangered species*".

[185] http://datadryad.org/pages/pricing

[186] The CLOCKSS Archive (Controlled LOCKSS) is a private LOCKSS network. http://www.clockss.org/
LOCKSS is treated in separate segment.

copyright and related or neighbouring rights to their data" (by conforming to the "CC0 1.0 Universal" (public domain) and the "CC BY 3.0" (attribution unported) of the Creative Commons). By downloading files, users agree to the Dryad Terms of Service.

Policy Model

Dryad's policy is embedded in their Terms of Services[187], which contain multiple types of policies, like submission, content, payment, usage and privacy policies.

Table 8: Policy elements in the Dryad model

Id	Policy element	Description
1	**Introduction and Binding Agreement**	*The introduction is a mission and vision statement. The second segment is a statement that users must agree to before using Dryad content.*
1.1	Introduction	
1.2	Binding Agreement	
2	**Definitions**	*List of definitions that apply to the document.*
3	**Publication Policies**	*Sets the limits to the Dryad submission activities (acquisition and designated community) and the embargo options available for the submitters.*
3.1	Content Criteria	
3.2	Embargos	
3.3	Expression of Concern, Retraction and Removal of Data Files	
3.4	Large Files	
4	**Dryad Obligations, Representation & Warranties to Purchasers and Submitters**	*Explains and delimits the curation activities of the organisation, ranging from ingest, preservation, digest and distribution of data (though not applying ingest/digest terminology).*
4.1	Curation	
4.2	Content Distribution	
4.3	Preservation	
4.4	Representation and Warranties	
5	**Purchaser Obligations, Representations and Warranties to Dryad**	*Sets up the responsibilities and warranties for the buyers of DPCs.*
5.1	Purchaser Obligations	
5.2	Purchaser Representations and Warranties	
6	**Submitter Obligations, Representations and Warranties to Dryad**	*Sets up the responsibilities/permissions and warranties for the data submitters. A submitter may not be the same as the purchaser of a DPC, as the buyer is often an organisation while the submitter is a person/group (within the organisation).*
6.1	Permission	
6.2	Representations and Warranties	
7	**Payment**	*Describes the payment framework of the Dryad services.*
7.1	Charges of Acceptance	
7.2	Hierarchy of Patterns	
7.3	Additional charges	
7.4	Refunds	
8	**Usage**	*Cover the usage of Dryad data and ownership / property rights connected to usage.*
8.1	Content Usage	
8.2	Prohibited Uses Generally	
8.3	Ownership and Use of Other Intellectual Property	
9	**Privacy**	*Cover issues of privacy protection and data security.*
9.1	Information Automatically Collected and Stored via the Website	
9.2	Personally Provided Information	
9.3	Security and Intrusion Detection	
9.4	Disclosure of Personally Identifiable Information	
9.5	Privacy Protection Limits	
9.6	Privacy Concerns	
10	**General Provisions**	*Cover areas of more general terms of use.*
10.1	Termination	
10.2	Disclaimer	
10.3	Limitation and Release of Liability	
10.4	Indemnity	
10.5	Notices	
10.6	Changes to Terms of Service	
10.7	Arbitration/Governing Law	
10.8	General Miscellaneous Provisions	

[187] Dryad, 2013. Terms of Services: http://datadryad.org/themes/Mirage/docs/TermsOfService-Letter-2013.08.22.pdf

GESIS

Description

The GESIS Data Archive is a department of GESIS – Leibniz-Institute for the Social Sciences, Germany's biggest research-based social sciences infrastructure institution (http://www.gesis.org/en). Founded in 1960, it is one of the oldest archives in Germany to actively curate and preserve digital research data for the long term. As a CESSDA archive it cooperates closely with other European Social Science data archives.[188]

The archive preserves quantitative social research data to make it available to the scientific research community. All data are preserved for the long-term and documented in accordance with international standards. Data is free to archive, and free to access. Currently, the archive collection comprises over 5,100 studies.

In an effort to create more transparency, the archive is currently undertaking a series of tiered certification and audit procedures in accordance with the European Framework for Audit and Certification of Digital Repositories. In 2014, the GESIS Data Archive received the Data Seal of Approval.[189]

Policy Model

The preservation policy of the GESIS Data Archive, published in 2013, states the general principles and strategies governing digital curation and preservation activities carried out by the archive.[190] The policy, addressed at all stakeholders of the Data Archive, is an expression of "GESIS's awareness of responsibilities and measures required to ensure adequate preservation and access" (Preservation Policy, p. 3).

[188] http://www.cessda.net

[189] https://assessment.datasealofapproval.org/assessment_116/seal/html/

[190] http://www.gesis.org/fileadmin/upload/institut/wiss_arbeitsbereiche/datenarchiv_analyse/DAS_Preservation_Policy_eng.pdf

Table 10: Preservation policy elements in GESIS

ID	Policy Elements	Description / Definition
1	**Introduction (Purpose of the policy)**	*Defines the scope and objectives of the preservation policy and the review frequency (annually).*
2	**Organisational framework**	*Introduces GESIS, the parenting organisation, and the GESIS Data Archive itself*
2.1	Mission	*Mission of the parenting organisation as it pertains to the activities of the Data Archive*
2.2	Selection and acquisition	*Collection focus as derived from GESIS's mission and statutes*
2.3	Access and use	*The Data Archive promotes data sharing and offers data for re-use to its designated community; it takes measures to make data accessible, re-usable, and citable. Related document: Usage regulations*
3	**Challenges**	*Overview of the most important challenges met in curation and preservation of digital research data*
4	**Principles**	*Overview of the principles guiding the preservation of data at the GESIS Data Archive*
4.1	Strategy	*Commitment to active preservation management and frequent review of preservation activities; relevance of further strategic goals to these activities*
4.2	Standards and co-operation	*Use of standards by the Data Archive; national and international co-operation activities in the fields of standards and digital preservation*
4.3	Well-documented and traceable processes	*Documentation of all archival processes and any changes to the archived data*
4.4	Transparency and trustworthiness	*Commitment to transparency and audits/certification processes.*
4.5	Technical infrastructure	*Security and risk management; back up; protection of data integrity/authenticity;*
5	**Responsibilities**	*Responsible teams, departments, and roles.*
6	**Related documents**	*Related documents relevant to the policy content*

ICPSR

Description

ICPSR, the Inter-university Consortium for Political and Social Research, is an international consortium of more than 700 academic institutions and research organisations and provides leadership and training in data access, curation, and methods of analysis for the social science research community. It also maintains a data archive. ICPSR is a unit within the Institute for Social Research at the University of Michigan.

ICPSR currently holds the 2014-2015 Data Seal of Approval.

Policy Model

The ICPSR model "…provides an outline for constructing the digital preservation policy framework for ICPSR and offers a step towards identifying core components of a digital preservation policy framework to encourage a community standard for digital preservation policy documents"[191]. The proposed model for a "digital preservation policy framework" is intended to be used by any organization that wants to develop its own policy framework. ICPSR conforms to this framework in their current preservation policy[192]. In addition to the preservation policy ICPSR also provide an Access Policy Framework, a Collection Development Policy, a Redistribution Policy, a Policy on Co-distribution of ICPSR Member-Funded Data, an Accessibility Policy, a Privacy Policy and a policy on Roles and Responsibilities[193].

The policy model of ICPSR is based on findings of the Cornell Digital Preservation Management workshop curriculum development project[194] (which was co-developed by Anne R. Kenney and Nancy Y. McGovern, authors of the "Five Organizational Stages of Digital Preservation"[195] with funding from the National Endowment for the Humanities); lessons learned in the development of the Cornell University Library Digital Preservation Policy Framework; and samples of policy frameworks developed by organizations that participated in the Cornell DPM workshop, e.g., the Library and Archives of Canada, N.C. State Library. It also builds heavily on the Open Archival Information System (OAIS) Reference Model (the 2002 version), the Attributes of a Trusted Digital Repository: Roles and Responsibilities (2002), and the Audit Checklist for Certifying Digital Repositories (2006 version).

[191] ICPSR, 2007. *Version 2.0 Digital Preservation Policy Framework: Outline.* Prepared by Nancy Y. McGovern, Digital Preservation Officer, ICPSR: http://www.icpsr.umich.edu/files/ICPSR/curation/preservation/policies/dp-policy-outline.pdf

[192] ICPSR Digital Preservation Policy Framework, Created April 2007; last revised June 2012: http://www.icpsr.umich.edu/icpsrweb/content/datamanagement/preservation/policies/dpp-framework.html

[193] Digital Preservation Policies and Planning at ICPSR: http://www.icpsr.umich.edu/icpsrweb/content/datamanagement/preservation/policies/index.html

[194] Digital Preservation Management: Implementing short-term strategies for long-term problems: http://www.dpworkshop.org/

[195] Anne R. Kenney and Nancy Y. McGovern, 2003. "The Five Organizational Stages of Digital Preservation," in *Digital Libraries: A Vision for the Twenty-first Century, a festschrift to honour Wendy Lougee*. Available from the University of Michigan Scholarly Monograph Series website. Also, see DASISH *D4.1 – Roadmap for Preservation and Curation in the SSH* for more info on this model and how it inspired the "five-step maturity model".

Table 11: Preservation policy elements in ICPSR

Id	Policy Elements	Description
1	**OAIS Compliance**	*Consists of an explicit statement of the intent of the digital preservation program to comply with the Open Archival Information System (OAIS) Reference Model approved as ISO 14721 in 2003. The digital preservation plan delineates the specifics of OAIS compliance and the self-assessment results for the digital preservation program documents the status of the program's OAIS compliance*
2	**Administrative Responsibility**	*Makes an explicit commitment to digital preservation and to compliance with prevailing standards and practice.*
2.1	Purpose	*Makes explicit the intentions of an institution and defines the essential role a digital preservation program plays in fulfilling the mission to protect the organization's digital assets. This section defines the rationale for the framework, identifies responsible parties and stakeholders, indicates the intended audience for the document, and places the document in the context of organization-wide efforts. The purpose statement might range from broad to narrow, reflecting the variations in intention for different types of digital archives.*
2.2	Mandate	*Stipulates the authority, jurisdiction, or governance upon which responsible parties have developed the digital preservation program, e.g., laws, legislation, policies, and mission. This section may also address requirements that are not specifically identified as preservation, e.g., legal admissibility, authenticity, FOIA, ADA, Data Protection Acts, copyright legislation, public records acts, E-Government, National Grid for Learning (UK).*
2.3	Objectives	*States the high-level aims and targets of the organization for collecting, managing, preserving, and sustaining access to digital content. This section identifies the benefit of the program to an institution and its relationship to other objectives, goals, and policies.*
3	**Organizational Viability**	*Addresses the legal status as well as human and other resources needed to establish and maintain a digital preservation program.*
3.1	Scope	*Establishes the overall timeframe, levels of responsibility, boundaries, extent, limitations, and priorities of the digital preservation program. This section delineates what the organization's digital preservation program will do and, as importantly, will not do. The scope statement may be brief or extensive, depending on the nature of the program. The scope provides useful metric for measuring the effectiveness of the digital preservation program.*
3.2	Operating Principles	*Defines the key principles, models, processes, and assumptions upon which the digital preservation program is developed and implemented. This section is particularly important in establishing system-wide benchmarks for distributed programs when multiple operational and technical processes are implemented. Common principles include adherence to standards (in particular OAIS) and other accepted indicators of good practice, support for life cycle management, interoperability, evidence-based requirements, and preferred methods of preservation.*
3.3	Roles and Responsibilities	*Describes key stakeholders and their respective roles in digital preservation, including creators, producers, digital repository staff, administrators, financial managers, user groups, advisors, other repositories, and collaborators. This section makes an explicit statement that digital preservation is shared responsibility requiring participants within and beyond the organization. It describes broad categories of roles and responsibilities and cites documents containing more specific descriptions.*
3.4	Selection and Acquisition	*Provides the rationale and processes for developing and retaining collections based on specific parameters (e.g., formats, types of records, geographic scope). A clear articulation is critical to the success of a digital repository and ensures that collections support the institutional mission and priorities, and that requisite resources are made available for digital preservation. One aspect of auditing a digital archive is to verify that the stated mission and intended scope of a digital archive matches its actual content. Specific policies logically follow from the conceptual statement in the framework to further collection development aspects, e.g., submission guidelines.*
3.5	Access and Use	*Identifies the designated communities for the digital preservation program and the barriers and/or restrictions to use of the digital content for which the*

		digital preservation program is responsible. Specific policies should be developed to further articulate access and use requirements and restrictions.
3.6	Challenges and Risk	*Identifies the organization's risks, difficulties, sense of urgency, and incentives for developing a digital preservation program. This section provides evidence that even though the full process may not be clearly understood, the need to act now is strong.*
4	**Financial Sustainability**	*Documents the tangible basis for sustaining the digital preservation program.*
4.1	Institutional Commitment	*Confirms and synthesizes the support for the program and the resources available to sustain the digital preservation program.*
4.2	Cooperation and Collaboration	*Acknowledges that the organization's effort exceeds or will exceed available resources and may not guarantee the safety of all vital assets. This section places the digital preservation programs into a broader context that recognizes the program's dependencies on other partners and on the community at large.* *Collaborations and partnerships may require formal, legally binding agreements that delineate explicit roles and responsibilities of each party.*
5	**Technological and Procedural Suitability**	*This component summarizes the preservation approach, strategies and techniques that are employed by the digital preservation program to achieve stated objectives. This section states the general philosophy of the digital preservation program and points to relevant requirements, policies, standards, guidelines, and practice. It makes a tangible link to the preservation planning component of the digital preservation program and to the organization's preservation plan.*
6	**System Security**	*Specifies the organization's commitment and approach to ensuring the accuracy, completeness, authenticity, integrity, and long-term protection of the organization's digital assets.*
7	**Procedural Accountability**	*Acknowledges the need for and stipulates the means for ensuring the transparency and accountability of the digital preservation program's policies and operations.*
7.1	Audit and Transparency	*Explicitly commits the organization to periodic self-assessments and audits to evaluate, measure, and adjust the policies, procedures, preservation approaches, and practices of the digital preservation program. Transparency enables self-assessments and audits. Self-assessments and audits improve internal operations, facilitate external reviews, and contribute to the development of effective partnerships and collaboration.*
7.2	Framework Administration	*Describes the organization's policies and practice pertaining to the development, approval, maintenance of the policy framework over time, e.g., frequency of updates and reviews, maintenance roles, expiration dates. The framework has little value if it has not received the appropriate approvals and has not been implemented. At minimum, the date and source of approval and the review cycle should be provided.*
7.3	Definitions	*Identifies terms and concepts that may be needed to understand the framework and may be instrumental in strategies for securing institutional commitment. This is an optional section, but one that can be very important. It is particularly important to include legally required and other mandated terminology and definitions. The section may either provide or point to requisite definitions.*
7.4	References	*Provides citations for or pointers to key resources that were informed the development and application of the framework.* *This section identifies more detailed documents, both internal and external, that provide a deeper expression of the mission, underlying principles, illustrative processes, and sustaining roles. It may contain citations for these documents or point to a current list of relevant community standards and guidance.*

LOCKSS

Description

The LOCKSS Program is an open-source, library-led digital preservation system built on the principle that "lots of copies keep stuff safe." The LOCKSS system allows librarians and publishers at each involved institution to take custody of and preserve access to the e-content to which they subscribe; using their computers and network connections, librarians can obtain, preserve and provide access to purchased copies of e-content. The idea is that through a LOCKSS distributed network, libraries are cooperating with one other to ensure their preserved content remains authentic and authoritative. This collaboration measure and validates the integrity of the participants' holdings. As a result, involved institutions "…are self-reliant and self-sustainable" in their communities. This distributed approach is characterised by the fact that there is no human intervention; there are no "trigger events" that require human intervention.

Policy Model

Preservation at LOCKSS states that it follows a few principles considered vital to successful long-term preservation. The approach and principles has been developed after research into the best practices, and greatest risks, of long-term preservation. The principles are as follows:

- Decentralised and distributed preservation (lots of copies keeps stuff safe)
- Give libraries local custody and control of their assets
- Preserve the publisher's original authoritative version
- Perpetual access – guaranteed and seamless
- Affordable and Sustainable

LOCKSS officially conforms to the OAIS reference model. A formal statement[196] from LOCKSS states that it conforms to the requirements set out in Section 1.4 of the ISO, which consist of two parts, namely to support the model of information described in Section 2.2 and to fulfil the responsibilities listed in Section 3.1 (of the OAIS model). This is what is referred to as a trust maturity "level 1" in the five-level trust maturity development model that was laid out in DASISH report D4.1[197].

The standard defines what conformance involves:

- A conforming OAIS Archive implementation shall support the model of information described in 2.2 (OAIS Information Definition section). The OAIS framework recognizes a clear definition of information as central to the ability of an OAIS to preserve it.

- A conforming OAIS Archive shall fulfil the responsibilities listed in 3.1 (Mandatory Responsibilities).

These conforming principles involve recognition of the information processes (and OAIS conceptualisation of these processes) involved in the data preservation. That is, accepting information from appropriate producers (ingest), controlling, managing and preserving the received

[196] LOCKSS Formal statement of conformance to ISO 14721:2003: http://www.lockss.org/locksswp/wp-content/uploads/2011/11/OAIS-LOCKSS-Conformance.pdf
[197] DASISH Deliverable D4.1: *Roadmap for Preservation and Curation in the SSH*. http://dasish.eu/publications/projectreports/D4.1_-_Roadmap_for_Preservation_and_Curation_in_the_SSH.pdf

information (archiving), and make this information available (disseminate) to relevant users, the designated community. As such, to support the OAIS Information Model it is necessary to distinguish between an *Information Package* that is preserved by an OAIS and the Information Packages that are submitted to, or disseminated from, an OAIS. Basically, the model consists of a *Submission Information Package* (SIP) that is sent to an OAIS by a producer, where the SIPs (one or more) are transformed into *Archival Information Packages* (AIPs) for preservation.

The OAIS Reference Model does not define or require any particular method of implementation of these concepts.

Although the LOCKSS formal statement on OIAS does not work as an explicit policy as such, it constitutes, along with the preservation principles stated above, the general policy framework of the LOCKSS preservation system.

Table 12: Preservation policy elements in LOCKSS

Element	Definition of element (from OAIS model)	Description in LOCKSS policy
Content information	A set of information that is the original target of preservation or that includes part or all of that information. It is an Information Object composed of its Content Data Object and its Representation Information	*In the LOCKSS system the Knowledge Base of the Designated Community is embodied in web browsers. The Content Information consists of bit streams with associated HTTP header information including MIME types sufficient for browsers to render the bit stream.*
Preservation Description Information	The information which is necessary for adequate preservation of the Content Information and which can be categorized as Provenance, Reference, Fixity, Context, and Access Rights Information.	*In the LOCKSS system Provenance is provided by the URL from which the content was collected, Context is provided by the links embedded in the content, Reference is provided by the original URL and by the availability of the text and the metadata it includes to search engines, and Fixity is provided by the mutual auditing protocol which supplies regular assurance that the content agrees with other replicas.*
Packaging Information	The information that is used to bind and identify the components of an Information Package.	*In the LOCKSS system Packaging Information is encoded in instances of Java classes implementing the LOCKSS plugin API. In most cases this is a generic implementation driven by XML files.*
Submission Information Package (SIP)	An Information Package that is delivered by the Producer to the OAIS for use in the construction or update of one or more AIPs and/or the associated Descriptive Information.	*In the LOCKSS system SIPs are created by the publisher, who places a "publisher manifest page" containing metadata on their website and publishes the URL. Individual LOCKSS system administrators direct their systems to preserve this page and the content it describes. Their LOCKSS system collects the page and the content it describes.*
Archival Information Package (AIP)	An Information Package, consisting of the Content Information and the associated Preservation Description Information (PDI), which is preserved within an OAIS.	*Internally, the LOCKSS system preserves content in a repository defined by a set of Java classes. The AIP consists of instances of these classes, representing the content itself, metadata obtained from the publisher manifest page and the HTTP headers, and an instance of a Java class implementing the LOCKSS plugin API encapsulating metadata not obtained from these sources. This instance is normally driven by externalized metadata in the form of XML files.*
Dissemination Information Package (DIP)	An Information Package, derived from one or more AIPs, and sent by Archives to the Consumer in response to a request to the OAIS.	*The LOCKSS system disseminates information by acting as an HTTP proxy, making it appear to the Designated Community that the SIP is still available from its original URLs (with any changes required by preservation operations such as form at conversion). The entire SIP, including the publisher manifest page with its metadata, is available. Thus the LOCKSS DIP is the same as the LOCKSS SIP.*
Negotiate for and accept appropriate information from information Producers		*An organization's LOCKSS system will, as directed by the authorized administrator, collect content from information Producers in the form of an appropriate SIP. The LOCKSS SIP must contain a "publisher manifest page" instantiated as an HTML page that describes and links to the relevant content and contains a statement that institutional subscribers have permission to collect and preserve that content.*
Obtain sufficient control of the information provided to the level needed to ensure Long-Term Preservation.		*An organization's LOCKSS system will as directed by the authorized administrator, collect via HTTP the entire SIP containing the content and the "publisher manifest page" and store it together with all available HTTP header information (including MIME type). This information is sufficient at the time of collection for a browser to render the content.*
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		*The LOCKSS "publisher manifest page" permission includes permission for the institution's reader's to access the material subject to the institution's subscription agreement. The Designated Community is thus the institution's reader.*
Ensure that the information to be preserved is Independently Understandable to the Designated Community		*At the time of collection, the SIP collected is sufficient for a browser to render the content; because it is collected in exactly the same way that a browser would access it. The LOCKSS system's DIP replicates the SIP exactly by acting as a proxy for the original SIP to the Designated Community, so the information preserved is Independently Understandable.*
Follow documented policies and procedures which ensure that the information is preserved against all reasonable contingencies, and which enable the information to be disseminated as authenticated copies of the original, or as traceable to the original.		*LOCKSS systems preserving the same SIP cooperate to audit and repair it, ensuring that the information is preserved against all reasonable contingencies. LOCKSS systems preserving the same SIP collect it independently from the*

	Producer, audit their independently collected SIPs and come to consensus as to the SIP's content. This audit allows the SIP to be authenticated and traceable to the original Producer.
Make the preserved information available to the Designated Community.	*An organization's LOCKSS system's DIP replicates the SIP exactly by acting as a proxy for the original SIP to the Designated Community.*

Odum

Description

The Odum Institute for Research in Social Science is a unit within the University of North Carolina Chapel Hill and provides services for researchers in managing, archiving and preserving social science data. It is part of the Dataverse Network (Odum Institute Dataverse Network[198]) which "…provides access to data collections curated by the Odum Institute as well as collections owned by other institutions and individual scholars". The Odum Institute's data curation software is really a combination of the Dataverse Network catalogue and custom tools. The Dataverse Network (DVN) serves as a catalogue for Odum's own studies as well as those received from faculty and students at the University of North Carolina at Chapel Hill. However, the DVN also serves as a catalogue for studies from partner organizations such as the National Archives and Records Administration and the Institute for Qualitative Social Science at Harvard University[199].

The data archive of the institute has attained the 2014-2015 Data Seal of Approval (DSA)[200]. The University (UNC) is connected to the DMPTool where it has its own customized data management plan template.

Policy Model

The archive provides a set of policies, ranging from appraisal, metadata, preservation, access and use policies; to policies on security, terms of use and legacy. Here we will focus on the preservation policy. The Odum preservation policy framework[201] is divided into two sections: one for higher level statements (organizational infrastructure) and a second section covering the more detailed preservation policy aspects. The preservation section addresses the seven attributes outlined in the Trusted Digital Repositories: Attributes and Responsibilities.

[198] http://arc.irss.unc.edu/dvn/

[199] The Odum Institute, 2009. Background Paper, International Data Technology Alliance Workshop: http://www.irss.unc.edu/content/pdf/Odum_background_paper_Final.pdf

[200] Odum Institute Data Archive, Implementation of the Data Seal of Approval: https://assessment.datasealofapproval.org/assessment_93/seal/html/

[201] Odum Digital Preservation Policies: http://www.irss.unc.edu/odum/contentSubpage.jsp?nodeid=629

Table 13: Preservation policy elements in ODUM

id	Element	Description / Policy statement
		Organizational Infrastructure
1	**Mission statement**	*General mission statement. "The Odum Archive's mission is to support and advance education and research in the social sciences through access to digital and legacy materials…".*
2	**Governance**	*Describes the governance structure of the The Odum Archive: "… is part of the H. W. Odum Institute for Research in Social Science founded in 1924.[…] and is under the governance of the UNC-Chapel Hill Vice Chancellor for Research".*
3	**Staffing**	*Describes staff structure and staff positions.*
4	**Organizational Chart**	*Provides an organisational chart: "…the Institute is structured as follows: Administration; Data archive services;Grant services; Research design and data collection services; Statistical and computing services; Institute programs".*
		Digital Preservation Policy Framework
5	**OAIS Compliance**	*Formalises the archives' conformance to the OAIS reference model: "The Archive is committed to developing policies and procedures which comply with the current standards and practices outlined in the Open Archival Information System (OAIS) Reference Model (ISO 14721:2003)…".*
6	**Administrative Responsibility**	*Describes organisation commitment, mandate, purpose, responsibilities and objectives.*
6a	Mandate	
6b	Objectives	
7	**Organizational Viability**	*Describes the scope of the organisation and the roles and responsibilities connected to each aspect of the data lifecycle.*
7a	Scope	
7b	Operating Principles	
7c	Roles and Responsibilities	
7d	Selection and Acquisition	
7e	Access and Use	
7f	Challenges and Risks	
8	**Financial Sustainability**	*Lays out the funding model of the archive: "…the H.W. Odum Institute derives funding from a variety of primary and secondary sources. First and foremost, the Institute is funded by the state of North Carolina, enabling the physical placement of the Odum facilities on the UNC-Chapel Hill campus…".*
9	**Technological and Procedural Suitability**	*Describes the procedures connected to data processing and curation: "…upon receipt of new digital content, the Archive processes the data and documentation, assesses that proper confidentiality concerns are addressed, fixes errors if necessary, and converts data formats. The Archive has adopted both normalization of file formats and migration on ingest and performs data migration when necessary as determined by periodic reviews".*
10	**System Security**	*Describes security measures: "…The Archive's Data Deposit Form addresses the need to authenticate digital content through requesting supporting documentation, data sets and a signature for submission […] …the distributed nature of the Data-PASS shared catalogue and the replication and policy-driven audits performed by the PLN via the SafeArchive system works as a disaster plan ensuring the long-term protection of digital assets".*
11	**Procedural Accountability**	*Lays out the organisational commitments to trust, accountability audit and transparency.*
11a	Audit and Transparency	

UK Data Archive

Description

The UK Data Archive is curator of the largest collection of digital data in the social sciences and humanities in the United Kingdom. UKDA manage the UK Data Service which is a portal for research resources, where they host national and international survey data collections, international databanks, census data and qualitative data. UKDA are engaged in a number of data management and preservation initiatives, supported by the ESRC, MRC, JISC and the EU, in addition to providing data curation for other organisations.

In 2005 the UK Data Archive was designated a Place of Deposit by The National Archives. This status meant that the Archive "…had to modify a number of its procedures to ensure that its previous emphasis on usability with reliability, and levels of integrity has been replaced with a much stronger emphasis on authenticity, integrity and reliability, while not ignoring usability".[202]

UKDA currently holds the 2010 Data Seal of Approval.

Policy Model

The UKDA preservation policy outlines the principles which underpin the main activities of what it calls its "active preservation" of digital resources for use and re-use within its core user community[203]. The policy generally conforms to the OAIS Reference Model[204], with additions and alterations which are specific to the materials held within the archive.

The first formal UKDA preservation policy was published in 2003 and revised in 2005. These earlier versions of the policy were informed more by internal practice than by outside influences[205], as opposed to later revisions which built more on external standards and recommendation. The latest version is from October 2012. Some of the formal standards they used in later policy revisions (apart from the OAIS model) include for example the BS ISO 18492 Long term preservation of electronic information, the Information security standards (BS ISO 27001 and 27002) and the Records management standards (BS ISO 15489). Earlier version also looked at trust planning tools like the TRAC Criteria and Checklist (Trustworthy Repositories Audit and Certification), the nestor catalogue of criteria, the Digital Preservation Coalition's Handbook and the DRAMBORA toolkit[206].

[202] UKDA Preservation Policy: http://www.data-archive.ac.uk/media/54776/ukda062-dps-preservationpolicy.pdf

[203] UK Data Archive Preservation Policy: http://www.data-archive.ac.uk/media/54776/ukda062-dps-preservationpolicy.pdf

[204] CCSDS: Reference Model for an Open Archival Information System (OAIS): http://public.ccsds.org/publications/archive/650x0m2.pdf

[205] *Constructing a Preservation Policy: the case of the UK Data Archive*. Presentation made by Matthew Woollard, UK Data Archive, at Digital Preservation Planning: Principles, Examples and the Future with Planets, British Library Conference Centre, London, 29th July 2008: http://www.planets-project.eu/docs/presentations/matthew_woollard.pdf

[206] See DASISH Deliverable 4.1 for a discussion of these and other similar resources.

Table 14: Preservation policy elements in UKDA

ID	Policy Elements	Description / Definition
1	**Purpose**	*Defines the core activities, primary functions and goals of the organisation.*
2	**Scope and Objectives**	*Defines and limits the scope of the policy and outlines the main objectives of the policy and of the archive activities.*
3	**Requirements**	*Lists the series of requirements which the archive "...strives to ensure are followed as closely as possible". That is, how the data sets and collections are processed and curated. In addition it lists the other core activities and documents which the requirements are dependent on (co-exists with). Among the documents listed here are the Strategic Plan, the Collections Development Policy, the Information Security Policy, the Records Management Policy and the Information Security Management Policy.* *Sub-section: Legal and Regulatory Framework. Defines and clarifies the legal and regulatory framework which the archive operates under. Lists laws and agreements the Archive follows. For example, among the listed items are the Copyright, Design and Patents Act, 1988; the Data Protection Act, 1998; the Freedom of Information Act, 2000; the EU Copyright Directive, 2001; the Environmental Information Regulations, 2004; English or UK law for commercial agreements and contract law; and current best practice.*
4	**Roles and responsibilities**	*Lies out and defines the functions and responsibilities of the different sections and head staff within the archive.*
5	**Model**	*Specifies the model/framework which defines the relationship between the entities of the archive.* *As the archive follows the broad guidance given in the OAIS reference model, the model description is divided into three subsections: Pre-ingest, Ingest and Archival storage function.*
5.1	Pre-ingest function	*Lays out the benefits of the pre-ingest function. Although the pre-ingest function is not explicitly specified in the OAIS model, UKDA consider it as having considerable benefits within a preservation model.*
	Ingest function	*Defines and explains the content of the ingest component.*
5.3	Archival storage function	*Describe the functional component that ensures that "...what is passed from the ingest process remains identical and accessible".* *"In the Archive this function receives AIPs and DIPs from the ingest function and adds them to the permanent storage facility, oversees the management of this storage, including media refreshment and monitoring".*
5.3.1	Physical data preservation and storage	
5.3.2	Media monitoring and refreshing strategy	
5.3.3	Compression	
5.4	Data management function	*Describes the elements that fall under the data management function: "…it works in conjunction with the Archival Storage function. It maintains databases of descriptive metadata; supports external finding aids; and manages administrative metadata which support internal operations, including change control".*
5.4.1	Administrative preservation database	
5.4.2	Version control/change procedures	
5.4.3	Data collection withdrawal	
5.5	Access function	*Describes the function that deals with access. That is, the way users interact with the organization to find, request and receive data collections / data sets. Key elements identified are: finding data collections; requesting data collections; and delivering data collections.*
5.6	Administrative function	*Refers to how the day-to-day operations of the organization are managed.*
6	**Preservation planning and strategy**	*"The UK Data Archive has chosen to implement a preservation strategy based upon open and available file formats, data migration and media refreshment. Preservation decisions at the Archive must always be made within the context of its Collections Development Policy, balancing the constraints of cost, scholarly and historical value, and user accessibility alongside the requirements of levels of authenticity and legal admissibility".*
	Preservation strategy overview	
6.2	Integrity measures	
6.3	Monitoring, review and feedback	
7	**IT Architecture**	*Describes the IT infrastructure principles of the organization. Deals with factors like hardware and software (upgrades), on-site/off-site storage, storage capacity, etc.*
8	**Security**	*"The UK Data Archive is committed to taking all necessary precautions to ensure the physical safety and security of all data collections that it preserves: fire prevention and protection system; physical intruder prevention and detection systems; and environmental*

		control systems".
9	**Co-operation**	*«The Archive has established productive working relationships with other institutions and organisations in order to address the Archive's preservation needs. The Archive recognises the need for communication with groups active in formulating national preservation policies and programmes. It also acknowledges the need to participate in activities and programmes in the area of digital preservation".*
10	**Funding and resource planning**	

Appendix 4: Policy models from guidelines and best practices

Beagrie

Description

The JISC-funded[207] study/report by Charles Beagrie Limited[208] is the result of a call by the JISC, issued in January 2008, for a study that would assist UK higher and further education institutions to formulate policies relating to the preservation of their digital assets[209]. JISC had at the time noted that "…'the costs and benefits of developing a coherent, managed and sustainable approach to institutional preservation of digital assets remain unexplored"[210]. Hence, the report emerges from a community for which the development of institutional preservation policies were 'sporadic' and "…digital preservation issues rarely considered in key strategic plans".

Though the study was primarily aimed at helping institutions in the UK Higher and Further Education sector to understand, develop and implement relevant digital preservation policies, it's findings and recommendations are broad in scope and the results and guidelines from the study has been referred to and implemented in several institutions and organisations in the years succeeding its release.

Policy Model

The report provides a practical guide for developing an institutional digital preservation policy. It contains strategic policy advice supported by further reading sections which select and provide brief descriptions of existing resources to assist policy implementation using specific strategies and tools. The policy model build on analyses of existing (in 2008) examples of preservation policies including guidance on policy frameworks; case studies, technical strategies and resources available from among others, JISC, the Digital Curation Centre (DCC), the National Archives (TNA), the Digital Preservation Coalition (DPC); a sample of online key institutional policies for research, teaching and learning, information, and UK library and records management; and other relevant digital preservation literature and resources (among them DRAMBORA and the RLG/NARA Trusted Repository checklist).

Among the particular and explicitly mentioned policies/documents of note for the study are the UK Data Archive (Woollard, 2008); the former Arts and Humanities Data Service (James 2004); the JISC/NPO Beagrie-Greenstein strategic framework for creating and preserving digital resources (Beagrie and Greenstein 2001); the ICPSR (McGovern 2007); the Canadian Heritage Information Network (Canadian Heritage Information Network 2004); University of Columbia (Columbia University 2006); and the Cedars Guide to collection Management (The Cedars Project 2002).

---

[207] Joint Information Systems Committee, a UK body concerned with information and communications technology in education: http://jisc.ac.uk/

[208] An independent management consultancy company specialising in the digital archive, library, science and research sectors: http://beagrie.com/

[209] JISC/Beagrie, 2008: *Digital Preservation Policies Study. Part 1: Final Report*: http://www.jisc.ac.uk/media/documents/programmes/preservation/jiscpolicy_p1finalreport.pdf

[210] Pennock, M., 2008: *JISC Programme Synthesis Study: Supporting Digital Preservation & Asset Management in Institutions:* http://www.jisc.ac.uk/media/documents/programmes/preservation/404publicreport_2008.pdf

The model itself is divided into two sections, Policy and Implementation. The *policy* clauses are set at a 'higher level' and are less technically detailed. It highlights some key points of consideration needed at the beginning of a digital preservation policy. The more technical *implementation* level is considered as either a significant part of the digital preservation policy itself, and/or to be part of a separate set of detailed procedures which are developed to accompany the main policy. It is noted that the policy model is intended to provide a general framework and work as guidance. It allows for a selective approach to meet the particular needs of specific organisations.

Table 15: Preservation policy elements in the Beagrie model - Policy clauses.

Summary Table of Policy Clauses		
Id	**Policy element**	**Description**
1	**Principle statement**	*Address how the digital preservation policy can serve the needs of the organisation and the benefits it will bring. Should include an example or key section of the organisation's mission statement or mandate if needed, along with statements on high level synergies or links with other organisations. Can also include statements on the current standards the organisation adhere to (if any).*
2	**Contextual links**	*Highlight how the policy integrates into the organisation and how it relates to other high level strategies and policies.*
3	**Preservation Objectives**	*Information about the preservation objectives and how they will be supported.*
4	**Identification of content**	*Outline what the policy's overall scope is in terms of content and its relationship to collection development aims*
5	**Procedural Accountability**	*Identify high level responsibilities for the policy and provide recognition of the most important obligations faced in preserving key institutional resources.*
6	**Guidance and Implementation**	*Guidance and implementation clauses on how to implement the preservation policy and/or identification of where additional guidance and procedures are available in separate documentation or from staff. The clauses and issues in the Implementation section (see below) can be used as required either to insert here and/or provide the framework for separate documentation.*
7	**Glossary**	*List of definition, if required.*
8	**Version Control**	*History and bibliographic details of the version. Date of the policy, and its intended duration and review process.*

Table 16: Preservation policy elements in the Beagrie model. Policy Implementation clauses.

Summary Table of Implementation Clauses		
Id	**Policy element**	**Description**
1	**Financial and Staff Responsibility**	*This section should be about who is responsible for digital preservation within the organisation. It should also be about financial sustainability and how the policy sits within the organisational financial plan.*
2	**Intellectual Property**	*This clause shows awareness of copyright issues and how the institution plans to recognise and tackle these key issues.*
3	**Distributed Services**	*In some situations it may be more convenient or cost effective to outsource some or all preservation activities*
4	**Standard Compliance**	*Lists the standards the archive is committed to.*
5	**Review and Certification**	*A description of how often a review of the policy is carried out, for example, bi-annually.*
6	**Auditing and Risk Assessment**	*Procedures for carrying out standardised auditing and recognition of risks facing the policy.*
7	**Stakeholders**	*Identification of all parties involved in the policy and its implementation procedures.*
8	**Preservation Strategies**	*A guidance table on preservation strategies adopted and technical implementation of the policy.*

DCC / UC3

Description

The DCC (Digital Curation Centre) in the UK has worked with research funders and universities to produce a tool that assists researchers to produce a data management plan "…to cater for the whole lifecycle of a project, from bid-preparation stage through to completion"[211]. It provides expert advice and practical help to anyone in UK higher education and research wanting to store, manage, protect and share digital research data[212]. The centre is funded by Jisc and staff is based at the Universities of Edinburgh, Glasgow and Bath.

Policy Model

In 2010, based on the findings in a curation policy report[213], the DCC released a preservation policy template for repositories. The 2009 report found that "…templates for data management and sharing plans, institutional policy statements and preservation policies were most needed"[214].

The policy template / model is intended for repositories to assist in the definition of a digital preservation policy. It is based on four external policies / guidelines, namely the AHDS[215] Collections preservation policy (v1.0, 2004), the DataShare Policy Model, the OpenDOAR policies tool and the UKDA Data Archive Preservation Policy (v.3.10., 2009).

DCC has also released a tool – the DMPonline – which has been developed to help researchers meet the funders' requirements for data management plans. In similar fashion as the DMPTool of UC3 (see below) the DMPonline tool provides several templates that represent the requirements of different funders and institutions; the users are asked three questions at the outset so that the appropriate template can be determined and displayed (e.g. the ESRC template when applying for an ESRC grant).

The tool has its roots in the first DCC Data management checklist that was created in 2009. This checklist has currently reached v 4.0 (released 2013[216]) and is one of the default templates that can be selected in the DMPonline tool (policy elements listed below). Templates also include requirements from, among others, the Arts & Humanities Research Council, Biotechnology and Biological Sciences Research Council, Wellcome Trust, Economic and Social Research Council, the European Commission (Horizon 2020) and the National Science Foundation (US).

DCC is also a contributor to the DMPTool. DMPTool is a service of the University of California Curation Center (UC3)[217] of the California Digital Library (CDL) and is a data management tool mainly

[211] DMPonline, about: https://dmponline.dcc.ac.uk/about_us

[212] DCC, about: http://www.dcc.ac.uk/about-us

[213] Sarah Jones (DCC, Glasgow, 2009): A report on the range of policies required for and related to digital curation, version 1.2.: http://www.dcc.ac.uk/sites/default/files/documents/reports/DCC_Curation_Policies_Report.pdf

[214] DCC Preservation Policy Template: http://www.dcc.ac.uk/sites/default/files/documents/Preservation%20policy%20template.pdf

[215] The Arts and Humanities Data Service (AHDS) was a United Kingdom national service aiding the discovery, creation and preservation of digital resources in and for research, teaching and learning in the arts and humanities. It was established in 1996 and ceased operation in 2008 (although the website and related digital collections are still accessible). Source: Wikipedia: https://en.wikipedia.org/wiki/Arts_and_Humanities_Data_Service; Website: http://www.ahds.ac.uk/

[216] DCC, 2013: *Checklist for a Data Management Plan. v.4.0*. Edinburgh: Digital Curation Centre: http://www.dcc.ac.uk/resources/data-management-plan

[217] University of California Curation Center: http://www.cdlib.org/services/uc3/

aimed at researchers, but which also contains tools for 'administrators' where one can create and edit institutional DMP templates.

The DMPTool began in January 2011 with eight institutions partnering to provide in-kind contributions of personnel and development. The effort was established in response to demands and requirements from funding agencies, such as the National Science Foundation (NSF) and the National Institutes of Health (NIH), that researchers provide a plan for managing their research data. It currently supports 23 public funding agencies and private foundations, including 12 NSF directorates and divisions[218]. Although it currently has a US funder approach, the project still "…holds a broader vision of a tool that serves as a coordinating hub between the management of data across many disciplines, many funding agencies, many institutions, and many countries."[219]

Among the original contributing institution (apart from UC3, CDL and DCC) are DataOne, the Smithsonian Institution, in addition to several US Universities and University Libraries.

As the DMPTool builds on requirements from various funders it does not provide one single policy/DMP with a standard set of elements. Hence the set of DMP elements are dependent on the selected funder or institutional template one chooses when generating the DMP. It is also possible to copy an existing DMP that are either publicly shared by any user, shared within the institution of the user by other DMP creators, or plans that the user have previously created.

Table 17: Preservation policy elements in the DCC Preservation Policy Template.

Id	Policy Element	Description
1	**Aim**	A clarification of the mission to preserve.
2	**Standards**	What standards, frameworks and models for digital preservation will be used?
3	**Content coverage**	What type of material can be deposited / will be preserved?
4	**Overview of preservation strategy**	Explanation of the main preservation approach(es) adopted e.g. normalisation on ingest, migration, emulation, media refreshment. Are certain data formats preferred? Will only specific formats be preserved?
5	**Methods / levels of preservation**	What different types of preservation service will be offered and why? Are there some things which the repository does not guarantee to preserve? How long will deposited items be retained?
6	**Implementing the strategy (operational details)**	Procedures for preservation; Security, authenticity and integrity; Media refreshment; Versioning; Withdrawal of collection.
7	**Sustainability plans**	What will happen if the repository is closed or funding reduced? Will services be cut or have ongoing costs been planned for? Are plans in place to transfer repository content if necessary?

[218] Presentation by Stephen Abrams on the UC3's DMPTool, presented at the ESA 2014 Meeting in Sacramento CA on 12 August 2014: http://www.slideshare.net/UC3/esa-ignite-talk-on-the-dmptool-by-s-abrams

[219] *DMPTool: supporting the data lifecycle.* Position paper for the 2011 NSF Research Data Lifecycle Management Workshop at Princeton University: http://www.columbia.edu/~rb2568/rdlm/Sallans_UV_RDLM2011.pdf

*Table 18: elements in the DCC Checklist / DMPonline tool**

Id	Policy element	Description (text taken from the directly from checklist description)
1	**Data collection**	*Give a brief description of the data, including any existing data or third-party sources that will be used, in each case noting its content, type and coverage. Outline and justify your choice of format and consider the implications of data format and data volumes in terms of storage, backup and access.* *Outline how the data will be collected/created and which community data standards (if any) will be used. Consider how the data will be organised during the project, mentioning for example naming conventions, version control and folder structures. Explain how the consistency and quality of data collection will be controlled and documented. This may include processes such as calibration, repeat samples or measurements, standardised data capture or recording, data entry validation, peer review of data or representation with controlled vocabularies.*
2	**Documentation and Metadata**	*Describe the types of documentation that will accompany the data to help secondary users to understand and reuse it. This should at least include basic details that will help people to find the data, including who created or contributed to the data, its title, date of creation and under what conditions it can be accessed. Documentation may also include details on the methodology used, analytical and procedural information, definitions of variables, vocabularies, units of measurement, any assumptions made, and the format and file type of the data. Consider how you will capture this information and where it will be recorded. Wherever possible you should identify and use existing community standards.*
3	**Ethics and Legal Compliance**	*Ethical issues affect how you store data, who can see/use it and how long it is kept. Managing ethical concerns may include: anonymisation of data; referral to departmental or institutional ethics committees; and formal consent agreements. You should show that you are aware of any issues and have planned accordingly. If you are carrying out research involving human participants, you must also ensure that consent is requested to allow data to be shared and reused.* *State who will own the copyright and IPR of any data that you will collect or create, along with the licence(s) for its use and reuse. For multi-partner projects, IPR ownership may be worth covering in a consortium agreement. Consider any relevant funder, institutional, departmental or group policies on copyright or IPR. Also consider permissions to reuse third-party data and any restrictions needed on data sharing.*
4	**Storage and Backup**	*State how often the data will be backed up and to which locations. How many copies are being made? Storing data on laptops, computer hard drives or external storage devices alone is very risky. The use of robust, managed storage provided by university IT teams is preferable. Similarly, it is normally better to use automatic backup services provided by IT* *Services than rely on manual processes. If you choose to use a third-party service, you should ensure that this does not conflict with any funder, institutional, departmental or group policies, for example in terms of the legal jurisdiction in which data are held or the protection of sensitive data. If your data is confidential (e.g. personal data not already in the public domain, confidential information or trade secrets), you should outline any appropriate security measures and note any formal standards that you will comply with e.g. ISO 27001.*
5	**Selection and Preservation**	*Consider how the data may be reused e.g. to validate your research findings, conduct new studies, or for teaching. Decide which data to keep and for how long. This could be based on any obligations to retain certain data, the potential reuse value, what is economically viable to keep, and any additional effort required to prepare the data for data sharing and preservation. Remember to consider any additional effort required to prepare the data for sharing and preservation, such as changing file formats.* *Consider how datasets that have long-term value will be preserved and curated beyond the lifetime of the grant. Also outline the plans for preparing and documenting data for sharing and archiving. If you do not propose to use an established repository, the data management plan should demonstrate that resources and systems will be in place to enable the data to be curated effectively beyond the lifetime of the grant.*
6	**Data Sharing**	*Consider where, how, and to whom data with acknowledged long-term value should be made available. The methods used to share data will be dependent on a number of factors such as the type, size, complexity and sensitivity of data. If possible, mention earlier examples to show a track record of effective data sharing. Consider how people might acknowledge the reuse of your data. Outline any expected difficulties in sharing data with acknowledged long-term value, along with causes and possible measures to overcome these. Restrictions may be due to confidentiality, lack of consent agreements or IPR, for example. Consider whether a non-disclosure agreement would give sufficient protection for confidential data.*
7	**Responsibilities and Resources**	*Outline the roles and responsibilities for all activities e.g. data capture, metadata production, data quality, storage and backup, data archiving & data sharing. Consider who will be responsible for ensuring relevant policies will be respected. Individuals should be named where possible.* *Carefully consider any resources needed to deliver the plan, e.g. software, hardware, technical expertise, etc. Where dedicated resources are needed, these should be outlined and justified.*

**The table lists the elements from the DCC checklist (default/generic alternative in the online tool). In addition to the elements listed, the checklist also includes fields for more project 'internal' information (like project name, project description, etc.) that are excluded from the tool and from our table.*

*Table 19: elements in the DMPTool (UC3)**

Id	Policy element	Description
1	**Roles and responsibilities**	*Explain how the responsibilities regarding the management of your data will be delegated. This should include time allocations, project management of technical aspects, training requirements, and contributions of non-project staff - individuals should be named where possible. Remember that those responsible for long-term decisions about your data will likely be the custodians of the repository/archive you choose to store your data. While the costs associated with your research (and the results of your research) must be specified in the Budget Justification portion of the proposal, you may want to reiterate who will be responsible for funding the management of your dat*
2	**Expected data**	*Give a short description of what data will mean the context of your research project. Explain what types of data you plan to generate, including size, file formats, and number of files. Briefly describe your methods for collecting data.*
3	**Period of data retention**	*Describe how long you plan to retain the data produced or used in your research. If you plan to embargo the data for a period of time after the research is completed, describe why this is necessary.*
4	**Data format and dissemination**	*Describe the format of your data. Ideally, data formats will be chosen that are openly and freely available, and/or non-proprietary in nature*
5	**Data storage and preservation of access**	*Describe your long-term strategy for storing, archiving and preserving the data you will generate or use.*
6	**Additional possible data management requirements**	*Any additional program-specific data management requirements. If none exist you may leave this section blank.*

**The table shows the policy elements that are generated when selecting the NSF-SBE (Social, Behavioural, and Economic Sciences) template. The descriptions are taken from the guidance texts that are included with each element in the template generator.*

DISC-UK DataShare

Description

A key deliverable of the JISC-funded DISC-UK DataShare project (2007-2009) [220] was a report that aimed to work as a guideline for policy-making in repositories[221]. The project, led by EDINA[222] and the Edinburgh University Data Library, with partners at the Universities of Southampton and Oxford, arose from the DISC-UK (Data Information Specialists Committee), a UK consortium of data support professionals working in departments and academic libraries in universities, and built on an international network with a tradition of data sharing and data archiving dating back to the 1960s in the social sciences.

The overall aim of the DataShare project was to "…contribute to new models, workflows and tools for academic data sharing within a complex and dynamic information environment which included increased emphasis on stewardship of institutional knowledge assets of all types; new technologies to enhance e-Research; new research council policies and mandates; and the growth of the Open Access / Open Data movement"[223]. The final report of the project is a distilled result of these efforts and aims to share the project experience "…with the wider community, as more institutions expand their digital repository services into the realm of research data to meet the demands of researchers who are themselves facing increasing requirements of funders to make their data available for continuing access" while also "…articulate the benefits of sound data management practices, as well as the goals of data sharing and long term access"[224].

Policy Model

The report presents a guide formed as a set of relatively broadly described data-related topics. The report states that these topics are compiled from multiple sources that focus on research data quality, management, and preservation. The guide is largely based upon the online OpenDOAR Policy Tool, the OAIS Reference Model and the TRAC checklist (OCLC, 2007). The initial focus of the report was on social science datasets, but its general scope makes it relevant for many other preservers and providers of research outputs.

The guide does not cover the "…value-added services that should be offered within a curatorial environment, details of selection and appraisal, nor does it cover advocacy, researcher requirements and data management considerations surrounding funders' mandates".

[220] DataShare Project: http://www.disc-uk.org/datashare.html

[221] JISC/DISC-UK DataShare, 2009: Policy-making for Research Data in Repositories: A Guide. http://www.disc-uk.org/docs/guide.pdf

[222] EDINA is the Jisc-designated national data centre at the University of Edinburgh: http://edina.ac.uk

[223] DataShare Project, aims and objectives: http://www.disc-uk.org/datashare.html

[224] JISC/DISC-UK DataShare, 2009.

Table 20: Policy elements in the DISC-UK/DataShare model

Id	Policy Element	Description
1	**Content Coverage**	
1a	Scope: subjects and languages	*What subject areas will be included or excluded? Are there language considerations? Will translations be included or required? (will text within data files, metadata or other documentation in other languages be translated into English, for example?)*
1b	Kinds of research data	*What kinds of research data will be included? (e.g. Scientific experiments, Models and simulations, Derived data, Canonical or reference data, Accompanying material, etc.)*
1c	Status of the research data	*Is the inclusion of the data into the repository determined by its status in the research process / lifecycle? Such as: 'raw' or preliminary data, data that are ready for use by designated users, data that are ready for full release, as specified in access policies, summary/tabular data (could be associated with a publication), 'derived' data.*
1d	Versions	*Policy considerations for the deposit of multiple versions of a dataset, and for version control.*
1e	Data file formats	*Consider what formats will be accepted for deposit, and which are preferred.*
1f	Volume and size limitations	*Consider any restrictions on the number of files per study or overall size of the study in advance of deposit.*
2	**Metadata**	
2a	Access to metadata	*Considerations: anyone may access the metadata free of charge; access to some or all of the metadata is controlled.*
2b	Reuse of metadata	*Considerations: May the metadata be reused in another medium without prior permission provided there is a link to the original metadata and/or the repository is mentioned; Will it be permissible to reuse the metadata for commercial purposes? Is formal permission required?; Will the repository system allow metadata harvesting of dataset descriptions by other institutions following the OAI-PMH guidelines, or other harvesting protocols?; What level of metadata is re-usable? Dataset descriptions? Full descriptive metadata (e.g. DDI XML record)?; Are data providers required to allow reuse of metadata?*
2c	Metadata types and sources	*The repository must make choices about what kinds of metadata will be required within the repository and from where each type will be produced. Includes entries for descriptive, administrative and structural metadata.*
2d	Metadata schemas	*Repositories may need to put in place additional metadata schemas to support the ingest, management, and use of data in their collections. Some repositories implement additional or extended metadata schemas for domain specific datasets.*
3	**Submission of data**	
3a	Eligible depositors	*Will eligibility be restricted by status (e.g. accredited members, academic staff, registered students, employees of the institution, department, subject community or delegated agents; Data producers or their representatives ('self deposit'); Only repository staff) or content ?*
3b	Moderation by repository	*Considerations: Are submissions checked to ensure that data integrity has been fully maintained during the transfer process? If so, spot checks, or all submissions?; The repository checks metadata records for accuracy; The repository adds Digital Object Identifiers (DOIs) or another persistent identifier, such as the Handle system; Does the repository's administration review items for the following: eligibility of authors/depositors? relevance to the scope of the repository? valid formats? exclusion of spam?*
3c	Data quality requirements	*Responsibility and Quality assessment.*
3d	Confidentiality and disclosure	*Requirements for the data depositors to ensure that data meet requirements of confidentiality and non-disclosure for data collected from human subjects.*
3e	Embargo status	*Descriptions of technical measures that enables embargo or the ability to sequester access to data until the content has been approved for release to the public. Agreements about the embargo – its length and what triggers its ending – need to be made between the repository and its contributor.*
3f	Rights and ownership	*License agreement with the depositor upon transfer of the data item through a written or click-through Depositor Agreement. The Agreement should cover copyright issues.*
4	**Access and reuse of data**	
4a	Access to data objects	*Considerations regarding the accessibility of data / access methods (e.g. open, controlled, restricted, and/or registration based access).*
4b	Use and reuse of data objects	*Policy element informing users of possible limitations to data. Prior to downloading data, the user may be required to agree to the terms of an online Terms of Use statement. Considerations/issues here include citation (e.g. will users of the data be required or requested to cite the dataset/s)?) and copies (e.g. what restrictions, if any, will be placed on making copies of the data and accompanying materials?).*

4c	Tracking users and use statistics	*Considerations: Will all access mechanisms be sufficiently granular to allow the identification of individual users in order to maintain logs of actions performed by users?; Will all actions relating to access to the material be recorded?; What repository use statistics will be made available and to whom?*
5	**Preservation of data**	
5a	Retention period	*Defines a dataset retention period (e.g. items will be retained indefinitely; items will be retained for at least xxx years from the date of deposition; items will be retained for the lifetime of the repository; retention periods may be set for individual items, as required).*
5b	Functional preservation	*It may not be possible to guarantee the readability of some file formats due to software obsolescence, but the repository may choose to promise to maintain the usability and understandability of the specific file formats over time.*
5c	File preservation	*The earlier section on data file formats (1.e) covers which file formats will be accepted for deposit. This section deals with how the repository will manage datasets over time.*
5d	Fixity and authenticity	*Fixity checks such as checksums, message digests, and digital signatures are used to verify that a digital object has not been changed between two points in time or events. Information created by these fixity checks provides evidence for the integrity and authenticity of the digital objects.*
6	**Withdrawal of data and succession plans**	*Sets out the conditions for withdrawal of datasets, and closure and succession plans.*

InterPARES

Description

The International Research on Permanent Authentic Records in Electronic Systems (InterPARES) 3 Project is an international collaborative project composed of several regional, national and multinational teams. Major funding for the InterPARES Project is provided by The Social Sciences and Humanities Research Council of Canada's Community-University Research Alliances (SSHRC-CURA)[225]. The overarching goal of the project is to "…translate the theory and methods of digital preservation (…) into concrete action plans for existing bodies of records that are to be kept over the long term by archives".

The first InterPARES project ran from 1999-2001 and was launched as a result of the findings of the Preservation of the Integrity of Electronic Records research project, which was carried out at the University of British Columbia's School of Library, Archival and Information Studies (1994-97)[226]. The stated goal of the InterPARES 1 Project was to "…develop the theoretical and methodological knowledge essential to the permanent preservation of electronically generated records and, on the basis of this knowledge, to formulate model strategies, policies, and standards capable of ensuring their preservation**"**.[227]

The second project, InterPARES 2, was carried out between 2002 and 2006 and aimed to "…develop a theoretical understanding of the records generated by experiential, interactive and dynamic systems, of their process of creation, and of their present and potential use in the artistic, scientific and governmental sectors, and, on the basis of that understanding, to formulate methodologies".[228]

InterPARES 3 ran from 2007 to 2012, and among its many goals it aimed to deliver "…policies, strategies and procedures for small and medium sized archival organizations or programs, and guidelines for the records creators whose records fall under their responsibility".[229] This resulted in a report[230] on a template for policy and procedures, presented in the table below.

Policy Model

The above mentioned report defines two separate sets of policy elements, one for *policy* and one for *procedure.* However, the elements within each policy set are more or less overlapping (a later report[231] within the project simplifies to one set of policies – the main elements are kept intact). The policy elements (sections and headings) specified in the template were drawn from "…existing policy templates available on the Internet; and policies that have been developed to-date by InterPARES 3 case studies"[232]. It is stated that the template is not meant as a prescription for polices, rather it sets out to describes content that should be included in policies, with suggested titles based on the content of each section/element.

---

[225] http://www.sshrc.ca/web/apply/program_descriptions/cura_e.asp

[226] InterPARES 1 Project Book: *The Long-term Preservation of Authentic Electronic Records: Findings of the InterPARES Project*: http://www.interpares.org/book/index.cfm

[227] Ibid.

[228] InterPARES 2 Book: Experiential, Interactive and Dynamic Records: http://www.interpares.org/ip2/book.cfm

[229] InterPares 3, expected products: http://www.interpares.org/ip3/ip3_expected_products.cfm

[230] InterPARES 3, 2011: General Study 11: *Policy and Procedures Templates*: http://www.interpares.org/ip3/display_file.cfm?doc=ip3_policy_procedure_templates_final_report.pdf

[231] InterPARES 3, 2012: General Study 12: *Policy and Procedures Templates: Digital Records Management/Preservation Education Modules* – Module 2: *Developing Policy and Procedures for Digital Preservation*: http://www.interpares.org/ip3/display_file.cfm?doc=ip3_canada_gs12_module_2_july-2012_DRAFT.pdf

[232] InterPARES 3, 2011

As mentioned earlier, in addition to the policy template InterPARES 3 also provides a 'procedure document template', which provide "…instructions as a best practice guide on how to implement a policy". The elements listed are Purpose/Objective, Scope, Procedural Statements, Roles and Responsibilities, Definitions, Related Sources, Contact Information and Version Control. The elements are identical to the policy set (except for Mandate, Areas of Coverage and Review) and should be considered as implementation tools that are helpful for *executing* a policy. Further, it is stated that "…procedure documents should be expected to undergo revision more frequently than policy documents, as they are adapted to reflect changes in technology, feedback form implementers and other factors"[233].

Table 21: Policy elements in the InterPARES model

Id	Policy Elements	Description
1	**Purpose/Objectives**	*An introductory section/element that aligns the goals and objectives of the policy with the goals and objectives of the organization. Thus, the policy should reflect the organization's mission and mandate.*
2	**Scope**	*The scope section of a policy should indicate the objects (e.g., records, digital objects) that are covered by the policy and the individuals or department(s) covered by the policy.*
3	**Mandate**	*The mandate of the organization or department issuing the policy should be stated. Including the mandate will indicate that the department issuing the policy has the authority to do so, and, the policy supports the department and/or organization's business needs.*
4	**Policy Statements**	*Provides high-level guidance on how digital records are created, maintained and preserved. Specific directives on how to implement the policy statement should be drafted in separate procedural or guidance documents, referred to in the policy.*
5	**Areas of Coverage**	*This section(s) will include guidance statements more specific than the policy statement, relating to aspects of records management (e.g., records creation, retention and disposal, digital preservation). These may be elaborated within a single policy document or be expanded in separate, cross-referenced policy documents.*
6	**Roles and Responsibilities**	*This section ties the policy into the overall organization structure. It identifies stakeholder groups and assigns responsibility for the development (archivists and/or records managers) and implementation (IT, staff) of the policy which reflect the accountability structure of the institution.*
7	**Definitions**	*This section should provide a glossary of domain- or organization-specific terms used in the policy, especially if the use of those terms differs from usage in common English.*
8	**Related Sources**	*Policies must adhere to relevant national or local legislation and may adhere to relevant standards and best practices. These laws, policies, standards and best practices should be reference in the policy. Identifying relevant legislation, standards and best practices helps to add authority to the policy. When citing related sources, it may be useful to include a statement identifying the purpose of the source and how it relates to the policy.*
9	**Contact Information**	*The policy should include a statement identifying the department issuing the policy. It may also include contact information for the department if further guidance or clarification is needed.*
10	**Policy Review**	*Receiving policy approval from senior individual(s) or department(s) indicates that the policy has received their authority. If necessary, receiving policy approval from legal counsel will ensure that the policy adheres to any relevant legislation, such as those regarding records, access to information and privacy. Policies should be reviewed periodically, to ensure that they continue to function within and support the organization's goals.*
11	**Version Control**	*Each policy should contain version control, to ensure that staff members are following the most up-to-date policies. Information needed to support version control include: Version number of the policy; Date the policy is effective; If policy has been superseded, date policy has been superseded; and If policy has been superseded, reference to updated version.*

[233] InterPARES 3, 2011

nestor

Description

nestor is the German competence network for digital preservation. It is a cooperation of libraries, archives, museums and leading experts in the field to support the preservation and curation of our digital cultural heritage. In 2012, a nestor working group concerned with the topic of preservation policies was founded. In a recently published guideline[234] the group makes recommendations for the composition and development of institutional preservation policies (English version forthcoming).

Policy Model

The authors of the nestor guideline acknowledge that the content of a preservation policy strongly depends on the individual context, requirements and objectives of the respective organisation. Accordingly, they refrain from defining mandatory policy elements, but describe possible policy content to provide orientation for organisations working on their policies. For this purpose the guideline also includes a generic policy template.

Table 22: Policy elements in the nestor model

ID	Policy Elements	Description / Definition
1	**Purpose, scope and objectives**	
1.1	Organisation	*Defines the core activities, primary functions and goals of the organisation.*
1.2	Scope and objectives	*Function, scope and objectives of the preservation policy in relation to the organisation's purpose and mission.*
2	**Principles and objectives of digital preservation**	
2.1	Digital preservation challenges	*General remarks on important challenges in the field of digital curation and preservation.*
2.2	Designated community	*Definition of relevant user groups as well as their needs and expectations and how these will be met by the archive/organisation.*
2.3	Community watch	*How will the designated community and its (changing) needs be monitored?*
2.4	Accessibility	*Measures to keep assets accessible and findable*
2.5	Integrity and authenticity	*How are the integrity, authenticity, readability and completeness of digital assets maintained?*
2.6	Persistent identifiers	
2.7	Metadata	*Used metadata and metadata standards*
2.8	Trustworthiness	*Measures to promote diligence, trustworthiness and transparency*
2.9	Standards	*National and international standards the archive conforms to*
2.10	Accountability	*Documentation of processes*
2.11	Active development	*Development and improvement of workflows*
2.12	Preservation strategy and planning	
2.13	Roles and responsibilities	
2.14	Technical infrastructure	
2.15	Resources	*Human, technical, and other resources employed in the digital preservation process*
2.16	Cooperation	*Information on co-operations used by the institution to improve its processes*
2.17	Confidentiality	*Confidential treatment of archived information*
3	**Sustainability of the policy**	
3.1	Responsible organizational unit	
3.2	Review	*Frequency of policy review*
3.3	Related documents	
3.4	Commitment	*Commitment to the policy*
3.5	Outlook	*Future plans*

[234] nestor-Arbeitsgruppe Policy: Leitfaden zur Erstellung einer institutionellen Policy zur digitalen Langzeitarchivierung (nestor-Materialien 18). Frankfurt am Main: nestor c/o Deutsche Nationalbibliothek, 2014. http://nbn-resolving.de/urn:nbn:de:0008-2014052004

OpenDOAR

Description

OpenDOAR – the Directory of Open Access Repositories – is a directory of academic open access repositories. OpenDOAR is maintained by SHERPA Services[235], based at the Centre for Research Communications at the University of Nottingham. OpenDOAR is a structured information service, cataloguing, describing and maintaining a comprehensive list of institutional and subject-based Open Access repositories. It also encompasses archives set up by funding agencies like the National Institutes of Health in the USA or the Wellcome Trust in the UK and Europe.

In addition to the structured list of repositories it also provides the OpenDOAR Policies Tool[236]. The tool stems from a 2006 survey[237] where it was discovered that about two thirds of Open Access repositories did not have publicly stated policies for such basic archive functions as re-use, submission, long term preservation, etc. To improve the situation, OpenDOAR developed a tool to assist repository administrators to formulate and/or present their repository's policies. It provides a series of check boxes and picks lists for key policy options.

The tool provides the user with auto-fill recommended options for minimum or optimum compliance with the Open Access movement. For example, the *minimum* policy recommends allowing re-use of metadata for not-for-profit purposes, but prohibits commercial re-use. On the other hand, the *optimum* policy also allows free commercial re-use because the extra exposure given to the material is considered as outweighing any disadvantages.

It should be noted that the tool is not intended to generate policy statements for legal purposes. The emphasis is on clear plain language for repository users. Consequently, legal statements should be treated separately and be published on separate web pages.

The tool produces text and source code that can be copied and paste into the archive/repository web page. Files can be edited and amended but because the OpenDOAR project encourages the use of standard policies when possible, they recommend not editing the document (except where the text is overly generic terms – in places one wants to be more specific).

Policy Model

The core ideas of the optimum policy are that the metadata policy allows for unlimited reuse of metadata as the increased visibility is considered to outweigh the 'exploitation'; that the data policy allows for multiple copying for educational purposes and for full harvesting (LOCKSS-like preservation); and that the submission policy includes mandatory deposition of metadata and mandatory deposition of thesis full texts.

---

[235] SHERPA Services compiles and maintains the RoMEO service, which gives summaries of the archiving rights that different publishers allow authors to retain. To complement this, SHERPA Services also runs the JULIET service, which summarises the archiving responsibilities and requirements that funding agencies give as a condition of funding grants. *Open*DOAR is the third part of this repository service, listing available open access repositories.

[236] OpenDOAR Policies Tool: http://www.opendoar.org/tools/en/policies.php

[237] Peter Millington (2006) Moving Forward with the *Open*DOAR Directory, *8th International Conference on Current Research Information Systems, Bergen, 11th-13th May 2006.*

Table 23: Policy elements in the OpenDOAR model

Policy	Elements
Metadata policy	*For information describing items in the repository. Contains information on: Access to metadata; Re-use of metadata.*
Data policy	*For full-text and other full data items. Contains information on: Access to full items; Re-use of full items*
Content policy	*For types of document and dataset held. Contains information on: Repository type; Type of material held; Principal languages*
Submission policy	*Concerning depositors, quality and copyright. Contains information on: Eligible depositors; Deposition rules; Moderation; Content quality control; Publishers' and funders' embargos; Copyright policy*
Preservation policy	*Contains information on: Retention period; Functional preservation; File preservation; Withdrawal policy; Withdrawn items; Version control; Closure policy.*

For each policy it is possible to auto-fill the content with the predefined categories of "minimum recommended options" or "optimum recommended options". The proposed minimum data policy consists of the following items (examples)[238]:

Metadata policy:

- *Anyone may access the metadata free of charge.*
- *The metadata may be re-used in any medium*
  - *without prior permission for not-for-profit purposes*
  - *provided the OAI Identifier and/or a link to the original metadata record are given.*
- *The metadata must not be re-used in any medium*
  - *for commercial purposes without formal permission.*

Data policy:

- *Anyone may access full items free of charge.*
- *Single copies of full items can be:*
  - *reproduced & displayed or performed in any format or medium*
  - *for personal research or study, educational, or not-for-profit purposes*
  - *without prior permission or charge.*
- *Full items must not be harvested by robots*
  - *except transiently for full-text indexing or citation analysis*
- *Full items must not be sold commercially*
  - *in any format or medium*
  - *without formal permission of the copyright holders.*

Submission policy:

- *Items may only be deposited by accredited members of the organisation, or their delegated agents.*
- *Authors/Depositors may archive only their own work.*
- *The administrator only vets items for the exclusion of spam*
- *The validity and authenticity of the content of submissions is the sole responsibility of the depositor.*
- *Any copyright violations are entirely the responsibility of the authors/depositors.*
- *If the repository receives proof of copyright violation, the relevant item will be removed immediately.*

[238] Peter Millington (2006) Moving Forward with the *Open*DOAR Directory, *8th International Conference on Current Research Information Systems, Bergen, 11th-13th May 2006.*

RDA

Description

The Research Data Alliance (RDA) aims to enable data sharing across barriers through focused Working Groups and Interest Groups, formed of various experts from academia, industry and government. Participation in RDA is open to anyone who agrees to its guiding principles of "openness, consensus, balance, harmonisation, with a community driven and non-profit approach"[239]. It was started in 2013 by a group of interested agencies – the European Commission, the US National Science Foundation and National Institute of Standards and Technology, and the Australian Government's Department of Innovation. RDA has several Working Groups and exploratory Interest Groups that aim to "…exchange knowledge, share discoveries, discuss barriers and potential solutions, explore and define policies and test as well as harmonise standards to enhance and facilitate global data sharing"[240].

Policy Model

The Practical Policy Working Group of the RDA has collected and registered a series of practical policies by conducting a survey of production data management systems to elicit the types of policies that are being enforced[241]. The types of data management applications included archives, digital libraries, data grids for data sharing, and processing pipelines. The 30 surveyed sites used more than ten different data management systems[242].

The survey identified the highest priority policies in the surveyed sites, and based on these results the study identified eleven generic policies that were of interest to a majority of the institutions and are common to almost all data management systems[243].

The survey and succeeding report provides policy templates for the production of policies. Each policy template contains policy name; example constraints that control application of the policy; state information that is needed to evaluate the constraint; example operations that are performed by the policy; and state information that is needed to execute the operations.

[239] About RDA: https://rd-alliance.org/about.html
[240] Ibid.
[241] Two-page paper from the RDA Working Group Practical Policy, available at the RDA file depot: https://www.rd-alliance.org/filedepot?cid=104&fid=557

[242] The surveyed sites listed include the integrated Rule Oriented Data System (iRODS), dCache, Tivoli Storage Manager, Xrootd, CLASS, AFS, GPFS, Data Direct Networks Web Object Scalar, Fedora Commons, Dataverse, LOCKSS – Lots of Copies Keep Stuff Safe, and XSEDE

[243] RDA: *Outcomes Policy Templates: Practical Policy Working Group, September 2014 (version August 29, 2014)*: https://www.rd-alliance.org/filedepot?cid=104&fid=557

Table 24: Policy elements in the RDA model

ID	Policy Elements /Policy Area	Description / Definition
1	**Contextual metadata extraction**	*This policy area focuses on metadata associated with files and collections; the creation of provenance and descriptive metadata defines a context for interpreting the relevance of files in a collection. The template illustrates types of constraints, the metadata needed to evaluate the constraint, and types of operations that may be applied.*
2	**Data access control**	*This element provides access controls that limit the ability to modify or add files to a collection, while allowing the public to read public data. The policy template includes operations to establish unique names for users, files, collections, and role; operations to set access controls by file or through inheritance from a collection; operations to handle access to replicas; and operations to audit which access controls have been established.*
3	**Data backup**	*A backup corresponds to a copy of a collection that is made at a specific date. Typical state information includes defining the backup time interval, where the backups should be created, and when the backups should be checked.*
4	**Data format control**	*Many collections restrict the types of data formats that will be acceptable for ingestion. Policies that identify data formats that are not allowed can either send warning messages, or move the file to a staging area, or attempt to transform the data format. Policies can be written to manage staging areas based on the type of data format, sorting selected data formats into specified collections. The associated operations include creating metadata to list the file format type, checking file formats, and verifying file formats.*
5	**Data retention**	*Policies that control the retention of data. May include retention based on a data expiration date; Cache management based on the age of files; or retention based on migration.*
6	**Disposition**	*Once files have been identified that have exceeded a retention period, a disposition policy can be applied to either delete or archive the files. For the above example for a data expiration policy, a disposition policy can be created that migrates the expired files to an archive collection, or that deletes the expired files.*
7	**Integrity (including replication)**	*The integrity policies may include verification of integrity on ingestion through validation of a checksum; replication of the file across multiple storage locations to ensure the ability to replace a corrupted file; or Periodic verification that the files are not corrupted, that the required number of replicas exist, and that the replicas are correctly distributed.*
8	**Notification**	*Defines policies for notifications that may be triggered by events (e.g. e-mail to administrator, creation of a new collection, change of access control permission on a collection, deposition of a file into a collection, deletion of a file, etc.).*
9	**Restricted searching**	*Restricted searching policies can work as a form of restricted access control (e.g. restricting the ability of users to see any files except their own files).*
10	**Storage cost policies**	*Policies that generate usage reports and the associated storage cost. Can be integrated with the use of quotas to limit the maximum allowed storage usage, which in turn limits the maximum cost.*
11	**Use agreements**	*Only element in the RDA model that is not implemented as a computer actionable rule. Use agreement policy is typically negotiated at the time an account is established for a user, and involves the receipt of a signed document.*

RSP

Description

The Repositories Support Project (RSP) was a 7-year JISC-funded initiative running from 2006 to 2013. Its main goal was to strengthen repository capacity knowledge and skills by providing guidance and advice, primarily within UK higher education institutions. The aim of the project was to progress the vision of a deployed network of interoperable repositories for academic papers, learning materials and research data across the UK.

The RSP consulted with the community at large and with JISC programme managers in order to ensure institutions could be effectively supported whatever their repository type or stage of maturity. A databank of expertise, know-how and best practice was built up. It was delivered with different views for target audiences and repository types, tailored to specific needs and available in multiple formats. Support materials concentrated on four broad themes:

- Technical: software selection and installation, technologies, metadata, interoperability
- Organisational: staffing, business requirements and incentives, copyright clearance and digital rights management
- Repository management: policies, workflows, archiving and preservation
- Advocacy: advocating to different stakeholders and advising on advocacy within institutions

Policy Model

The RSP policy model consists of a set of advisory policies that aims to provide a framework for managing the repository and legal compliance/issues. Some of the elements mentioned in the policy model are based on project management models like the JISC Digital Repositories infoKit[244] and the PRINCE2 standard[245]. The preservation specific elements are built on models suggested by the DCC and the DPE[246], especially PLATTER[247] and DRAMBORA[248]. Other elements build on policy items from the OpenDOAR policy model.

[244] JISC InfoKits, Project Management: http://www.jiscinfonet.ac.uk/infokits/project-management/
[245] PRINCE2 (PRojects IN Controlled Environments): http://www.prince-officialsite.com/
[246] Digital Preservation Europe: http://www.digitalpreservationeurope.eu/
[247] http://www.digitalpreservationeurope.eu/platter/
[248] Digital Repository Audit Method Based on Risk Assessment: http://www.repositoryaudit.eu/

Table 25: Policy elements in the RSP model

ID	Element	Description
1	**Content**	*Advice on creating policies for defining the type of content that will be stored in the repository*
2	**Submission**	*Defining policies for getting content into the repository.*
2.1	Self-archiving	
2.2	Mediated deposit	
3	**Data re-use**	*Advice on specifying how the content in a repository can be used by others.*
3.1	Metadata re-use	
3.2	Access to metadata	
3.3	Full-items re-use	
3.4	Access to full-items	
4	**Preservation**	*Help for considering how to define the preservation approach for your repository*
4.1	Retention period	*How long the repository undertakes to retain items for (i.e. indefinitely or not).*
4.2	Functional preservation	*What are your intentions to ensure to continued readability and usability of the items in your repository? What technical steps are you taking to fulfil your intentions, either by yourself or with partners?*
4.3	File preservation	*How are you backing up your repository files, in what form, and how often?*
4.4	Withdrawal policy	*Do you allow items to be withdrawn? If so, what reasons are acceptable?*
4.5	Withdrawn items	*How are items withdrawn? Are they deleted entirely, or do you just remove them from public view? Do the original URLs remain valid, and if so, do they point to 'tombstone' citations or to replacement items?*
4.6	Version control	*Do you allow items to be changed after they have been committed to the repository? Do you allow multiple versions? Can addenda and corrigenda be accommodated?*
4.7	Closure policy	*Heaven forfend that your repository be closed down, but just in case, what would happen to the material deposited in it?*
5	**Copyright**	*Handling copyright can be one of the most difficult tasks for repository managers and administrators. This section provides advice on creating copyright policies.*
6	**Take-down**	*It is important for repositories to have a robust policy to deal with disputes over items that have been submitted. This section provides advice on defining take-down policies.*
6.1	Risk management	
6.2	Policy triggers	
6.3	Policy process	
7	**Embargoes**	*Advice on creating policies for content where an embargo has been imposed.*

SCAPE

Description

SCAPE (Scalable Preservation Environments) was an EU-funded project (2011-2014, under FP7) which was "…directed towards long term digital preservation of large-scale and heterogeneous collections of digital-objects" and its main focus was to "…develop scalable services for preservation planning and preservation actions on an open source platform"[249]. It has developed a policy-based preservation planning tool together with an automated watch system that aims to ensure the implementation of institutional preservation strategies.

One of the tools coming out of the project is the Catalogue of Preservation Policy Elements[250] which is part of the SCAPE Policy Framework developed to support the preservation functions "Planning and Watch" to make use of automated policy compliant workflows. Tools to supply the automated workflows were also developed (see C3PO, Plato and Scout), but here we will focus on the preservation policy model and the specific policy elements.

Policy Model

The policy model of SCAPE builds on several other European projects that have investigated preservation policies. Explicitly mentioned are the DL.org project[251], PLANETS[252] and the SHAMAN project[253]. From these projects SCAPE specifically identifies "interoperability" (DL.org) and a "preservation guiding document" (PLANETS) as particularly important means to enable digital libraries / data centres to get the most value out of their collections and to enable sharing and "building by re-use"[254]. It also draws on the findings from the SHAMAN project which defined a number of catalogues and processes needed in digital preservation from the business governance viewpoint.

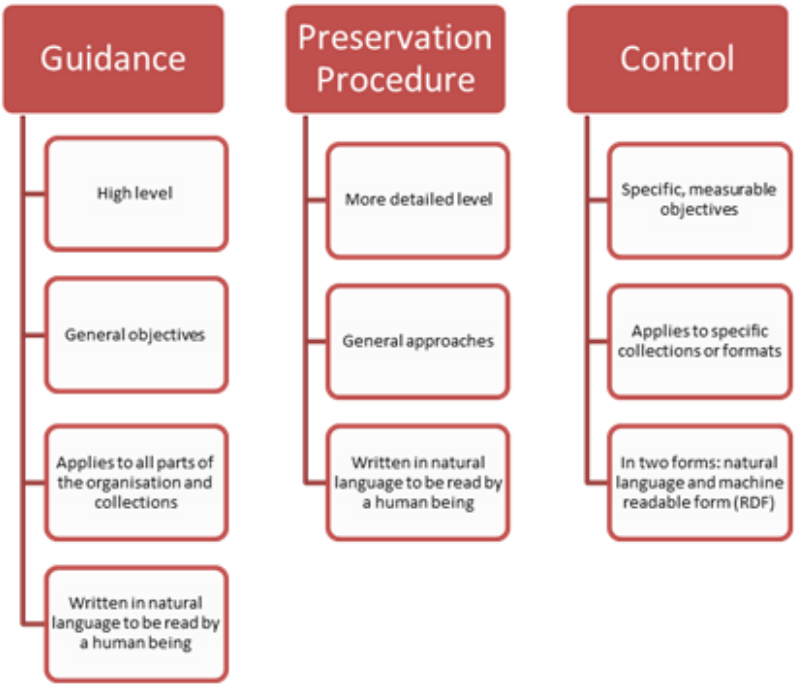
The SCAPE Preservation Policy Model consists of three preservation policy levels that aim to support an organisation to create their preservation policies. The three levels of policies identified in SCAPE are[255]:

- **High level or guidance polices**. On this level the organisation describes the general long-term preservation goals of the organisation for its digital collection(s). As an example is mentioned an organization that decides to act according the OAIS model.
- **Preservation Procedure policies**. These policies describe the approach the organisation will take in order to achieve the goals as stated on the higher level. They will be detailed enough to be input for processes and workflow design but can or will be at the same time concerned with the collection in general. It is stated that these are likely to be made publically available.

[249] SCAPE project webpage: http://www.scape-project.eu/about
[250] About the Catalogue of Preservation Policy Elements: http://wiki.opf-labs.org/display/SP/Introduction
[251] DL.Org - Digital Library Interoperability, Best Practices and Modelling Foundation: http://www.dlorg.eu/.
The DL.org Booklets are based on core set of outputs: Digital Library Manifesto, Digital Library Checklist, Digital Library Cookbook, Digital Library Reference Model - In a Nutshell.
[252] PLANETS - Preservation and Long-term Access through Networked Services: http://www.planets-project.eu/
[253] SHAMAN – Sustaining Heritage Access through Multivalent Archiving: http://www.shaman-ip.eu/
[254] Preservation Policy Levels in SCAPE. IPRES 2013 – Proceedings of the 10th International Conference on Preservation of Digital Objects: http://purl.pt/24107
[255] Ibid.

- **Control policies**. On this level the policies formulate the requirements for a specific collection, a specific preservation action, for a specific designated community. This is the level that according to the SCAPE model should be machine readable (in addition to human readable). Thus it can be used in automated planning and watch tools (like C3PO, Plato and Scout) to ensure that preservation actions and workflows chosen meet the specific requirements identified for that digital collection. It is stated that these are likely to be kept internally within the organisation.

Figure 3: Preservation Policy levels in the SCAPE



Source: http://wiki.opf-labs.org/display/SP/SCAPE+Policy+Framework

The Catalogue of Preservation Policy elements give a description of the second level of policies in the SCAPE Policy Framework, namely the preservation procedure policies. The policy elements consist of ten thematic headings, each with a sub-set of policy elements.

Table 26: Preservation Policy levels in the SCAPE model

Policy ID	Policy Elements
1	**Authenticity**
1.1	Integrity
1.2	Reliability
1.3	Provenance
2	**Bit preservation**
2.1	Define Bit Preservation
2.2	Define Bit preservation levels
2.3	Decide on Ingest activities
2.4	Develop Integrity Measures
2.5	Persistent Identifiers
2.6	Decide on number of copies, geographical distribution and organisational distribution
2.7	Define Policy for Disaster recovery
3	**Guidance Policy Functional Preservation**
3.1	Plan Functional Preservation
3.2	Define preservation strategies
3.3	Define ingest activities and preservation actions
3.4	Keep track of versions when performing migrations
4	**Guidance Policy Digital Object**
4.1	Original Object
4.2	Deletion of Objects
4.3	Keep track of developments of file formats
4.4	Take-down policy
4.5	Define significant properties
5	**Guidance Policy Metadata**
5.1	Management of metadata
5.2	Original metadata
5.3	Descriptive metadata
5.4	Preservation metadata
5.5	Structural metadata
6	**Guidance Policy Rights**
6.1	Comply with national legislation and contracts with business partners
6.2	Document object creator and copyright holder
6.3	Enter into deposit and archiving agreements
6.4	Clarify legal context for preservation actions
7	**Guidance Policy Standards**
7.1	Principle on the use of standards
7.2	Reference model
7.3	Use of specific standards
8	**Guidance Policy Access**
8.1	Usability
8.2	Digital Rights Management
8.3	Design of Dissemination Information Package
8.4	Understandable for Designated Community
8.5	Search facilities and resource discovery
8.6	Designated Community and communities identified
9	**Guidance Policy Organisation**
9.1	Staffing
9.2	Risk management
9.3	Budgets
9.4	Preservation Cost Assessment
9.5	Roles and responsibilities
10	**Guidance Policy Audit and Certification**
10.1	Standard for audit and certification
10.2	Audit preparations

Appendix 5: EU Horizon 2020

Description

Horizon 2020 is the current EU Framework Programme for Research and Innovation. It runs from 2014 – 2020 and provides a total amount of about 80 billion Euro project funding over this period.[256]

The programme aims to tie research and innovation closer together, as innovation is seen as the major factor to sustain Europe's competitiveness in the future. "The goal is to ensure Europe produces world-class science, removes barriers to innovation and makes it easier for the public and private sectors to work together in delivering innovation."[257] The three core areas of Horizon 2020 are: "Excellent Science", "Industrial Leadership" and "Societal Challenges".[258]

One of the changes in comparison to previous Framework Programmes is an explicit emphasis on Open Access to research results. "[B]eneficiaries must ensure that peer-reviewed scientific publications resulting from Horizon 2020 funding are deposited in repositories and made open access" and "must also aim to deposit at the same time the research data needed to validate the results presented in scientific publications."[259] In addition, the "Open Research Data Pilot" is introduced as another means to promote and improve access to and re-use of research data.

Policy Model

The scope of the "Open Research Data Pilot" is laid out in the "Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020".[260] The Pilot applies to projects funded under a number of programme areas[261]. Projects may opt out for several reasons (e.g. because the data produced is sensitive data), on the other hand, projects from areas not covered by the "Open Research Data Pilot" may voluntarily take part.

Projects participating in the "Open Research Data Pilot" have to provide a description of the data generated or collected and the management of that data in the form of a "Data Management Plan (DMP)". Details, including a template for the DMP, are given in the "Guidelines on Data Management in Horizon 2020".[262] They define the DMP as follows:

> "A DMP describes the data management life cycle for all data sets that will be collected, processed or generated by the research project. It is a document outlining how research data will be handled during a research project, and even after the project is completed, describing what data will be collected, processed or generated and following what methodology and standards, whether and how this data will be shared and/or made

[256] See: http://ec.europa.eu/programmes/horizon2020/en/what-horizon-2020
[257] Ibid.
[258] See: http://ec.europa.eu/programmes/horizon2020/en/h2020-sections
[259] European Commission 2013c, p. 19.
[260] European Commission 2013b, p. 8 et seqq.
[261] In the 2014-2015 work programme (European Commission 2013c), these are:
- Future and Emerging Technologies,
- Research infrastructures – part e-Infrastructures,
- Leadership in enabling and industrial technologies – Information and Communication Technologies,
- Societal Challenge: Secure, Clean and Efficient Energy – part Smart cities and communities,
- Societal Challenge: Climate Action, Environment, Resource Efficiency and Raw materials – except raw materials,
- Societal Challenge: Europe in a changing world – inclusive, innovative and reflective Societies,
- Science with and for Society.

[262] European Commission 2013a.

open, and how it will be curated and preserved. The DMP is not a fixed document; it evolves and gains more precision and substance during the lifespan of the project."[263]

A first version of the DMP is to be delivered within the first six months of a project. More advanced versions can/should be submitted later.

The DMP template contains five sections that have to be filled in with free text:

- Data set reference and name
- Data set description
- Standards and metadata
- Data sharing
- Archiving and preservation (including storage and backup)

Annex 2 of the Guidelines on Data Management gives additional guidance (in the form of questions that should be addressed in the DMP) with regard to five requirements that data generated or collected in the project should fulfil. The data should be easily

- discoverable
- accessible
- assessable and intelligible
- useable beyond the original purpose for which it was collected
- interoperable to specific quality standards.

[263] Ibid., p. 4.

Table 27: Preservation Policy levels in the Horizon 2020 Data Management Guidelines

Policy ID	Policy Elements	Description / Definition
1	**Data set reference and name**	*"Identifier for the data set to be produced." (DMP template)*
2	**Data set description**	*"Description of the data that will be generated or collected, its origin (in case it is collected), nature and scale and to whom it could be useful, and whether it underpins a scientific publication. Information on the existence (or not) of similar data and the possibilities for integration and reuse." (DMP template)*
2.1	Origin	*- If data is collected, where from?*
2.2	Nature and Scale	*- What data are produced / collected (data type, format) using which method?* *- How are data analysed, processed over the research process?* *"[A]re the data and associated software produced and/or used in the project assessable for and intelligible to third parties in contexts such as scientific scrutiny and peer review (e.g. are the minimal datasets handled together with scientific papers for the purpose of peer review, are data […] provided in a way that judgments can be made about their reliability and the competence of those who created them)?" (Additional guidance)*
2.3	Potential for (re-)use	*"[A]re the data and associated software produced and/or used in the project discoverable (and readily located), identifiable by means of a standard identification mechanism (e.g. Digital Object Identifier)?" (Additional guidance)* *"(A)re the data and associated software produced and/or used in the project useable by third parties even long time after the collection of the data (e.g. is the data safely stored in certified repositories for long term preservation and curation; is it stored together with the minimum software, metadata and documentation to make it useful; is the data useful for the wider public needs and usable for the likely purposes of non-specialists)?" (Additional guidance)*
2.4	Reference in publication	*- Does the data underpin a scientific publication?*
2.5	Similar data	*- Is there similar data?* *- Can it be integrated / re-used?*
3	**Standards and metadata**	*"Reference to existing suitable standards of the discipline. If these do not exist, an outline on how and what metadata will be created." (DMP template)*
3.1	(Discipline-specific) standards	*"[A]re the data and associated software produced and/or used in the project interoperable allowing data exchange between researchers, institutions, organisations, countries, etc. (e.g. adhering to standards for data annotation, data exchange, compliant with available software applications, and allowing re-combinations with different datasets from different origins)?" (Additional guidance)*
3.2	Metadata to be produced	*- If not covered by 3.1: What metadata will be produced and how?*
4	**Data sharing**	*"Description of how data will be shared, including access procedures, embargo periods (if any), outlines of technical mechanisms for dissemination and necessary software and other tools for enabling re-use, and definition of whether access will be widely open or restricted to specific groups. Identification of the repository where data will be stored, if already existing and identified, indicating in particular the type of repository (institutional, standard repository for the discipline, etc.)." (DMP template)*
4.1	Repository / data archive	*- repository / archive where the data will be stored*
4.2	Access conditions	*"[A]re the data and associated software produced and/or used in the project accessible and in what modalities, scope, licenses (e.g. licencing framework for research and education, embargo periods, commercial exploitation, etc.)?" (Additional guidance)*
4.2.1	Licenses	*- only for research and education?, for commercial exploitation as well?* *- embargo periods?*
4.2.2	Technical procedures	*- how is access provided (e.g. via web interface, on data carrier, on-site)?*
4.2.3	Access policy	*- open or restricted to specific groups?* *- different access levels for specific groups?*
4.3	Reasons for not sharing	*"In case the dataset cannot be shared, the reasons for this should be mentioned (e.g. ethical, rules of personal data, intellectual property, commercial, privacy-related, security-related)." (DMP template)*
5	**Archiving and preservation (including storage and backup)**	*"Description of the procedures that will be put in place for long-term preservation of the data. Indication of how long the data should be preserved, what is its approximated end volume, what the associated costs are and how these are planned to be covered." (DMP template)*
5.1	(trustworthy) Repository / data archive	*"[I]s the data safely stored in certified repositories for long term preservation and curation"? (Additional guidance)*
5.2	Preservation period	*- how long is the data to be stored?*
5.3	Volume	*- estimated volume of the data to be stored*
5.4	Costs	*- estimated costs and how they will be covered*

References

European Commission (2013a): Guidelines on Data Management in Horizon 2020. Version 1.0, 11 December 2013. 2013a.
http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

European Commission (2013b): Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020, Version 1.0, 11 December 2013.
http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf

European Commission (2013c): Horizon 2020 Work Programme 2014-15. Table of Contents and 1. General Introduction. 2013b.
http://ec.europa.eu/research/participants/portal/doc/call/h2020/common/1597683-part_01_introduction_v1.1_en.pdf