OPEN SCIENCE AND RESEARCH

RESEARCH DATA FILE FORMATS AND DIGITAL PRESERVATION

FINAL REPORT

This is an English translation of the report in Finnish, titled
Tutkimusaineistojen tiedostomuodot ja pitkäaikaissäilytyskelpoisuus.

Publication Research Data File Formats and Digital Preservation — Final Report	
Publisher The Open Science and Research Initiative (ATT)	**Publication Date** 10.2.2017
Author Digital preservation services development group	

Distribution
A PDF file can be downloaded at http://openscience.fi/

Contact
http://openscience.fi/
avointiede@postit.csc.fi

CONTENTS

1. SUMMARY

The Open Science and Research Initiative (ATT) was started in 2014 by the Ministry of Education and Culture of Finland to promote open science and the availability of research information. An important aspect of the initiative is the digital preservation and availability of research results and data. To ensure their usability and applicability over a time period of several dozen years, stable operational models are being developed. The research information digital preservation ensemble includes services and the technical infrastructure, which support the operational models and provide the required preservation functionalities, application programming interfaces and user interfaces.

This report is part of designing the digital preservation ensemble. It focuses on research data file formats, whose understandability, prevalence and software support are important for data reuse. The report is based on international sources and interviews with Finnish researchers. Additionally, the report presents preliminary requirements for accepting research datasets for digital preservation.

1.1. Example Datasets

Research datasets almost always consist of several files that are related to each other. For example, the dataset may contain raw data from a measurement device, metadata describing the settings of the device, a description of the conducted experiment and a publication presenting the results of the research. It is essential that the included files together as a whole are understandable to the researchers reusing the dataset.

The interviewed researchers provided eleven example datasets for digital preservation analysis. They are presented in the table below. The sample does not cover all fields of science, but gives a good overview of data types and file formats.

Abbrev.	Creator or owner	Field of science	Description of the dataset
1000Gen	International 1000 genomes project	Bio and health sciences	Human gene sequences collected through international cooperation
BrainImg	Aalto University, Brain and Mind Laboratory	Medical technology	MRI scans of the brains of persons watching a movie
ERNE	University of Turku, Space Research Laboratory	Natural sciences, space research	Measurements of cosmic radiation by the ERNE experiment
FIRE	University of Helsinki, Institute of Seismology	Environmental science, seismology	Seismological measurements of the Earth's crust in Finland
FSD	Finnish Social Sciences Data Archive	Social sciences	Surveys of Finns' media use and relationship to cultural heritage
Crystals	Aalto University, Bioeconomy Infrastructure	Natural sciences, biochemistry	Measurements of the formation of crystals in soft matter
MAXIV	MAX IV Laboratory, Sweden and University of Oulu	Natural sciences, material physics	An example file related to X-ray microscopy

Abbrev.	Creator or owner	Field of science	Description of the dataset
Planck	European Space Agency (ESA)	Natural sciences, space research	Measurements of cosmic radiation by the Planck satellite.
RITU	University of Jyväskylä, Accelerator Laboratory	Natural sciences, particle physics	Measurements of particles by the RITU separator
SMEAR	University of Helsinki, SMEAR research stations	Earth sciences	Database of measurements by several instruments and stations
Suomi24	CSC and Institute for the Languages of Finland	Social sciences, linguistics	Linguistic analysis of messages on the Suomi24 discussion forum

Almost all example datasets consisted of files in several different file formats. In total 26 different formats were present, half of which have already been approved as recommended formats or acceptable for transfer into digital preservation in the National Digital Library (NDL) of Finland. The majority of the remaining, not yet approved formats were also open and documented.

Most of the file formats in the example datasets can be accepted for digital preservation, when technical metadata requirements have been written. The datasets also need to be carefully documented. All components of each dataset will be packaged. Packaging in this context means primarily a standard method of representing the roles, relationships and metadata of the different parts of the dataset.

Information about commonly used research data file formats and databases was also gathered during the interviews. A greater variety of formats are used in research than in cultural heritage content, and many of them are specific to certain fields of science. Preserving databases is particularly challenging. Understanding the formats and data often requires knowledge of the respective field. However, international cooperation is guiding researchers to use compatible and well-documented file formats, which facilitates digital preservation efforts.

1.2. Accepting Datasets for Digital Preservation

The objective is that transferring datasets into digital preservation will be easy and convenient, in order to get them broadly and quickly available for reuse. The acceptance requirements designed in the National Digital Library project were taken as a starting point, and modifications were made, taking into account the special characteristics of research data.

Digital preservation ensures the understandability of the information during a very long time scale, over technological, methodological and other major changes. It sets fairly tight requirements for the file formats and metadata. The requirements for research datasets are similar to the requirements for preserved cultural content in the National Digital Library.

When more time is needed to decide about long-term digital preservation, ensuring the integrity of datasets is a simpler means of keeping them available for many years. This kind of safe storage is referred to as *data repository* in this document. In a data repository, the file format requirements in particular are less strict in comparison with digital preservation. However, the dataset and its parts must be appropriately documented to be usable by other researchers. Also the reuse rights and conditions need to be stated.

Requirements for accepting a dataset for digital preservation:

1. The dataset is usable by other researchers. (mandatory)
2. The files belonging to the dataset and their relationships are described according to

the digital preservation requirements. (mandatory)

3. The files are in formats that have been approved as recommended formats or acceptable for transfer. (mandatory in long-term preservation, recommended in the data repository)
4. The usage rights and conditions are stated. (mandatory)
5. The licence of the dataset conforms to the open science recommendations. (recommended)
6. The dataset is documented according to the metadata requirements. (mandatory)

Data repository requirements for file formats that have not been preapproved (see requirement 3 above):

1. The file format is supported in at least one software program that is generally available. (mandatory)
2. The structure of the file format is documented. (recommended)
3. The file format is widely used in the field. (recommended)
4. The file format has been standardised by an independent organisation or by the scientific community. (recommended)

File formats approved as recommended formats or acceptable for transfer into preservation naturally fulfil all mandatory requirements and in almost all cases also all the recommendations.

A dataset can be transferred directly into long-term preservation provided that it fulfils the requirements. Alternatively, it can be first published in the data repository service, so that the file format approval process will not delay the publication. In that case it can be decided later whether the dataset will be transferred into long-term preservation.

1.3. Extending the NDL Preservation Services to Preserve Research Datasets

The digital preservation specifications of the National Digital Library form a solid basis for the preservation of research datasets. Existing specifications can be extended to cover new content types and file formats, while making the necessary changes to processes and areas of responsibility.

The packaging model designed in the NDL is suitable for research datasets as well. It is particularly important to focus on the user friendliness of the packaging service and the metadata creation tool.

In the NDL the responsible entity is usually a museum, a library or an archive, which has a statutory mission to preserve content. Research datasets are typically produced in projects that have an ending date and no long-term responsibility for preserving the data. The organisation responsible for transferring content into preservation may be a research infrastructure that manages data within a specific field of science across university borders and has good abilities to uniformly document the datasets.

Digital preservation of research datasets is also internationally in a relatively early phase of development. Most organisations maintain a data repository that does not include all long-term preservation features. Comprehensive lists of recommended file formats as well as specifications of metadata are often missing. The national digital preservation solution gives Finland an opportunity to be a pioneer and a desirable partner for the preservation of international datasets.

2. INTRODUCTION

The Open Science and Research Initiative (ATT) was started in 2014 by the Ministry of Education and Culture of Finland to promote open science and the availability of research information. The objective is for Finland to become one of the leading countries in openness of science and research by the year 2017 and to ensure that the possibilities of open science will be widely utilised in our society. In addition to this, the ambition is to promote the trustworthiness of science and research, support the culture of open science as a way of working within the research community, and to increase the societal and social influence of research and science.

An important part of the initiative is to ensure the long-term preservation (LTP) and availability of research results and datasets. Long-term preservation of digital information means developing working methods, processes and systems that are designed to maintain the usability of the information during the next several dozen years, over changes in technology and research practices. International compatibility, in particular within scientific fields, needs to be ensured with cooperation, agreements and suitable data models on the semantic level.

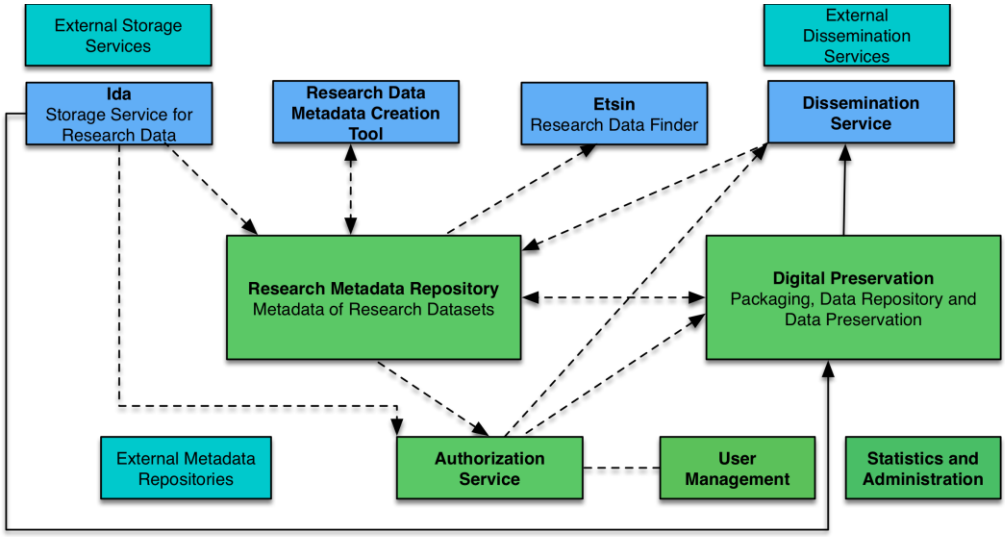

Figure 1: Long-term preservation ensemble

Accepting research datasets for long-term preservation and offering them to other researchers for reuse is a complex process. Responsibilities need to be defined, usage rights clarified and technological choices made. On the technological level, it will be ensured that the datasets are and will remain intact, understandable and suitable for reuse. The long-term preservation ensemble (Figure 1) includes services and systems that implement the preservation functionalities and provides the necessary application programming and user interfaces. According to preliminary plans, cultural information in the National Digital Library (NDL) and research datasets can be preserved using a shared technological platform.

Research datasets almost always consist of several files that are related to each other. For example, a dataset may contain raw data from a measurement device, metadata describing the settings of the device, a description of the conducted experiment and a publication presenting the results of the research. It is essential that the included files together as a whole are understandable to the researchers reusing the dataset.

This report focuses on the file formats of research data, which is one of the key factors influencing reuse. Information about commonly used file formats was gathered both using international sources and by interviewing researchers who are working with datasets in Finland. As part of the work, a closer look was taken at eleven contemporary datasets.

The report begins by presenting the used working methods and the example datasets received for analysis. The chapter entitled “Analysis of the File Formats” discusses formats used in the

example datasets as well as popular file formats and databases in various scientific fields. The properties of the formats are analysed from the reuse point of view. The “Accepting Datasets for Digital Preservation” chapter lists the preliminary requirements for accepting datasets, describes the acceptance process and evaluates the conformance of the example datasets with the requirements. Finally, conclusions and needs for future work are presented. The list of interviewed persons, interview questions and detailed tables about file format analysis can be found in the appendices.

3. WORKING METHODS

The subject was approached in two ways. Up-to-date international information was acquired from documents and Internet sites about file formats, as well as from a few collaborators. The second important source comprised interviews of Finnish researchers and research groups, who provided information about datasets of different scientific fields and national needs. The research groups were asked to provide example datasets, which were analysed in detail down to the level of individual files.

The interviewed persons and example datasets were chosen so that they provided a comprehensive overview of different types of data and file formats from the long-term preservation point of view. The sample does not cover all fields of science, but provides a solid basis for designing dataset-related long-term preservation requirements and processes. Refinements, additions and changes required by individual scientific fields can be made later.

When evaluating the suitability of file formats for long-term preservation, one source of information consisted of lists and descriptions of file formats accepted and recommended by other preservation organisations. The evaluation of the software support of various formats relied on publicly available information; the software was in most cases not tested.

Project manager Esa-Pekka Keskitalo from the National Library of Finland and Secretary General Pirjo-Leena Forsström of the Open Science and Research Initiative defined the objectives and made the key decisions during the project. Arto Teräs, a consultant specialised in long-term preservation, was responsible for conducting the interviews, gathering the information and writing the report. Juha Törnroos from CSC - IT Center for Science also participated in several of the interviews. The requirements for accepting datasets for preservation were formulated together in the digital preservation services development group. The project wants to express its gratitude to the interviewed persons (Appendix A) who had a crucial role as information sources. They had an opportunity to read the report and comment on it before publication.

4. INTERNATIONAL OVERVIEW

Research projects are increasingly based on international cooperation and research datasets are used internationally. Large international projects often collect their data in central storages and databases, which researchers both supply with data and use as information sources. To facilitate international cooperation, it is important to take international solutions and practices into consideration when designing national long-term preservation. By taking care of file format and metadata compatibility on an international level, Finnish researchers can more easily exchange information with foreign colleagues and organisations. Popular file formats can also be processed with readily available validators and other software tools, reducing the cost of preservation.

The Open Science and Research Initiative has previously published an international survey focusing on long-term preservation processes, services and their management in four organisations in different countries [ATT_KVKatsaus]. In this work a closer look was taken at international choices and recommendations on file formats. The surveyed organisations were the National Archives of Australia, the National Computing Center for Higher Education CINES in France, the Data Archiving and Networked Services (DANS) in the Netherlands, the UK Data Archive (UKDA) in Great Britain, the Library and Archives Canada and the Library of Congress in the United States.

The choices of recommended and acceptable file formats made by the surveyed organisations are largely similar to the choices made in the National Digital Library [NDL_Formats]. The NDL specification is more specific about file format versions and related metadata than most of the international recommendations and guidelines. On the other hand, it does not include many research data file formats that are approved in international recommendations.

The documents published by the National Archives of Australia [NAA_Formats] and UKDA [UKDA_Formats] are simple lists of preferred and acceptable formats, without more details about versions or metadata. CINES mentions the versions and also lists the formats which the preservation service is able to validate [CINES_Formats].

DANS Preferred Formats document [DANS_Formats] provides additional information and guidelines related to each file format and category. It includes several data types that have not yet been considered in the NDL, for example geographical information systems (GIS), computer aided design (CAD), 3D models and databases. The instructions however do not specify recommended versions of the formats or which metadata should be included.

Library and Archives Canada has already published two documents. The older of these, from the year 2010, includes a fairly comprehensive evaluation of the formats [LAC_Formats_2010]. The evaluation criteria for recommended and acceptable formats used in the NDL are based on this document. The newer list published in 2015 [LAC_Formats] is less elaborate, but still includes fairly detailed version information on the recommended formats and a general explanation of how the formats were chosen. The list includes several data types not yet considered in the NDL.

The Library of Congress document [LoC_Statement] takes another approach: it does not even try to be a comprehensive list of acceptable formats, but gives instead more general recommendations and rates formats in order of preference in different categories. Some formats are mentioned explicitly, but the document also lists more general classes such as "markup formats", "publicly documented formats" and "widely used proprietary formats". In addition to listing the formats, the document specifies requirements and provides instructions about metadata and the content of the dataset.

The Sustainability of Digital Formats website by the Library of Congress [LoC_Formats] deserves a special mention. The site provides general instructions on choosing file formats and detailed descriptions of many popular formats. The number of formats is lower than in the PRONOM format registry [PRONOM], but its detailed and professional descriptions make the Library of Congress site a more useful information source.

International recommendations were considered when evaluating the example datasets and file formats used in different scientific fields. The remarks are integrated in the text of the report as part of the format descriptions.

5. EXAMPLE DATASETS

Eleven example datasets were received for analysis. They are presented in the table below.

Abbreviation	Creator or owner	Field of science	Description of the dataset and the subset selected for analysis
1000Gen	International 1000 genomes project	Bio and health sciences	The dataset contains the gene sequences of 1000 humans, collected through international cooperation. The data is freely available online. [1000GENOMES] The gene sequence of one test subject (HG00180) in the most commonly used file formats was selected for analysis.
BrainImg	Aalto University, Department of Neuroscience and Biomedical Engineering, Brain and Mind Laboratory	Bio and health sciences, medical technology	The dataset consists of image series obtained by magnetic resonance imaging (MRI), depicting the brain functions of test subjects watching a movie. [AALTO] The image series of three test subjects and the accompanying files of the experiment, including the movie, were selected for analysis. The selected files were part of a dataset used in a preservation pilot study in 2015. [PAS_Pilots_2015]
ERNE	University of Turku, Department of Physics and Astronomy, Space Research Laboratory	Natural sciences, space research	The dataset contains measurements of energies of cosmic radiation particles hitting a measurement device used in the ERNE experiment. [ERNE] A subset of measurement results was selected for analysis. It was the same subset that the researchers had already used in a preservation pilot study in 2015. [PAS_Pilots_2015]
FIRE	University of Helsinki, Institute of Seismology	Environmental science, seismology	The dataset contains reflection seismic measurements of the Earth's crust in Finland. It was collected in the large national FIRE project during 2001-2004. [FIRE_Project] Two subsets of measurement data from different survey lines, together with the accompanying files, were selected for analysis.

Abbrevia-tion	Creator or owner	Field of science	Description of the dataset and the subset selected for analysis
FSD	Finnish Social Sciences Data Archive	Social sciences	A repository of datasets collected by the Finnish Social Sciences Data Archive. The datasets are published for reuse to researchers of the field, either freely or requiring permission. Two freely available survey datasets were selected for analysis: qualitative data of Finns' relationship with cultural heritage [FSD2981] and quantitative data of Finns' Internet and media use [FSD2985].
Crystals	Aalto University, School of Chemical Technology, Bioeconomy Infrastructure, Biohybrid materials research group	Natural sciences, bio-chemistry	The dataset contains measurement results about crystal formation in soft matter. It is part of materials research belonging to the Bioeconomy Infrastructure at Aalto University. A sample of measurement results chosen by the research group was selected for analysis.
MAXIV	MAX IV Laboratory, Lund University, Sweden / Diamond Light Source Ltd, Great Britain University of Oulu is the national coordinator in Finland.	Natural sciences, material physics, also bio and medical sciences.	The Swedish MAX IV Laboratory offers X-ray microscopy services, which can be used to examine protein structures, for example. [MAXIV_Lab] An example file in Nexus HDF5 format, recommended by the laboratory, was selected for analysis. Most files of the laboratory will be stored in that format. The example has been created by the British synchrotron science facility Diamond Light Source.
Planck	European Space Agency (ESA)	Natural sciences, space research	The dataset consists of cosmic background radiation measurements by the Planck space telescope during a period of four years. It is freely available for download in the Planck Legacy Archive [PLA]. The measurement results of one frequency during one day were selected for analysis.

Abbreviation	Creator or owner	Field of science	Description of the dataset and the subset selected for analysis
RITU	University of Jyväskylä, Department of Physics, Accelerator Laboratory	Natural sciences, particle physics	The dataset consists of measurement data produced by the RITU separator, which has been developed in the laboratory, and its accompanying files. [RITU] A subset of measurement results was selected for analysis. It was the same subset that the researchers had already used in a preservation pilot study in 2015. [PAS_Pilots_2015]
SMEAR	University of Helsinki, Department of Physics, SMEAR research stations	Earth science, atmospheric sciences	The dataset consists of measurement data produced by several different devices at multiple observation locations. It contains measurements of the atmosphere, soil, forest cover and water quality. More data is continuously being collected and stored in a MySQL database. It can be freely viewed and downloaded through a web interface on the AVAA platform. [SMEAR_AVAA] The whole database was selected for analysis, including data for several years until 25 January 2016.
Suomi24	CSC and the Institute for the Languages of Finland, the Language Bank of Finland	Social sciences, linguistics	The dataset consists of messages written on the Suomi24 discussion forum during 2001-2015, linguistically analysed and annotated. [Suomi24] Three subsets of messages from different years were selected for analysis: a total of 1.5 million messages stored in 141 files.

5.1. File Formats and Size of the Example Datasets

The file formats and the size of the example datasets are summarised in the table below.

Dataset	File formats	Size of the analysed subset (GB)	Total size of the dataset
1000Gen	BAM, CRAM	34.8	several hundred terabytes
BrainImg	BIDS, JSON, NIFTI, PDF, TSV, WMV	1.3	about 8 GB
ERNE	PDF, PNG, TXT (structural)	0.8	about 22 GB
FIRE	Corel Draw, DOC, JPG, SEG-Y, TXT (structural), WMV	2.1	about 2 TB

Dataset	File formats	Size of the analysed subset (GB)	Total size of the dataset
FSD	PDF, RTF, SPSS Portable, TXT, XML	< 0.1	Variable, the dataset of a single project typically less than 0.1 GB
Crystals	PDF, XLSX	< 0.1	Result files less than 0.1 GB
MAXIV	Nexus HDF5, HTML, PDF	< 0.1	Variable, both small and large datasets (depending on the research project using the service)
Planck	FITS, HTML, PDF	1.1	about 20 TB
RITU	Java, GREAT, PDF, TXT (structural), XML	4.6	about 200 GB
SMEAR	MySQL database, SIARD, CSV, HDF5, HTML, JSON, TSV,	32.7	Size of the database as a MySQL dump file: 32.7 GB. As a database: about 10 GB. The raw data produced by the measurement devices is several hundred terabytes.
Suomi24	TXT, VRT	4.0	About 170 GB

Almost all datasets consisted of files in several different file formats. Examples of popular formats were TXT (text) and PDF (Portable Document Format) documents, PNG (Portable Network Graphics) images and WMV (Windows Media Video) videos. They are widely used in different fields and already approved as recommended formats or acceptable for transfer in the National Digital Library specifications [NDL_Formats]. Publications and complementary documents such as descriptions of the measurement devices are without exception either already in one of the recommended formats or can be easily converted to them.

Eight of the example datasets consist primarily of measurement data, either directly produced by a device or processed using a standardised method. Two of the datasets consist of statistical or other data produced by humans (FSD, Suomi24). In one dataset (Crystals), the most important results and measurements have been compiled in an Excel worksheet (XLSX file format), leaving the raw data out of the set. The data files are either structural text files (ERNE, VRT), binary files (BAM, CRAM, GREAT, NIFTI, SEG-Y, SPSS Portable) or a combination of structural text and binary data (FITS, Nexus HDF5). The parameters used in the measurements are in most cases stored in a separate file either as structured text, as key-value pairs in JSON format or in an XML file.

In addition to file formats, another notable format is the Brain Imaging Data Structure (BIDS) directory structure used by the BrainImg dataset. The BIDS structure, created by the international scientific community of brain researchers, defines not only which file formats should be used to store data, but also how to name the files and organise them in directories. There is even a validation tool to verify BIDS compatibility.

The SMEAR dataset differs from others by storing the measurement data in a database. The database structure and content can be exported into a so-called dump file, but searching for parts of the data is much faster and more flexible using the actual database. The AVAA service provides a web-based user interface, through which desired parts of the data can be selected and downloaded in CSV, HDF5 or TSV format [SMEAR_AVAA]. A JSON API is also available. For

testing purposes, the SMEAR dataset was additionally converted into the SIARD format, which is specifically developed for digital preservation of databases.

5.2. Metadata of the Example Datasets

To enable the reuse of a dataset, various kinds of metadata are needed. They include, for example, the settings of measurement devices, a description of the conducted experiment and information about the structure of data files. Upon transfer into preservation, it needs to be ensured that all essential metadata for understanding the dataset is included in the package. It should be noted that even a well-described dataset often cannot be understood by a layman; interpreting it may require deep knowledge of the scientific field in question.

The table below presents a summary of the metadata of the example datasets, and how it is presented in each dataset.

Dataset	Metadata
1000Gen	▪ Essential for understanding the dataset: how the gene sequences have been processed, the phenotype of the test subject and certain technical details about gene sequencing. ▪ The BAM and CRAM files that contain the measurement data (the gene sequence) have a header section for storing metadata. Some of the fields are obligatory, others voluntary. ▪ The internationally popular Sequence Read Archive (SRA) model is a standardised method for describing the metadata of gene samples. ▪ The phenotype is stored separately; there is no standard convention. Often the information cannot be published due to data protection requirements. In the example dataset, the only available phenotype information consisted of nationality and gender.
BrainImg	▪ Essential for understanding the dataset: the settings of the imaging device, the phenotype information of the test subjects and the description of the conducted experiment. ▪ The settings of the devices and other essential technical information are stored in JSON files. The research group has moved it there from the DICOM files produced by the imaging device in order to facilitate processing. ▪ Phenotype information is stored in a TSV file; there is no standard convention. As in the 1000Gen dataset, phenotype information often cannot be published due to data protection requirements. ▪ The standardised BIDS directory structure and file naming conventions help researchers to find the essential information and facilitate the automatic processing of data. ▪ A short description of the dataset and other standardised metadata fields is in JSON format in the dataset_description.json file (part of the BIDS specification). ▪ A more detailed description of the research project can be found in articles stored in PDF format.

Dataset	Metadata
ERNE	▪ Essential for understanding the dataset: the description of the measurement device, the settings of the device and the conducted experiment. ▪ The data files include a header section, which describes the structure and lists the physical quantities contained within the file. ▪ The settings of the measurement device, a description of the experiment and other information relevant for interpreting the data are in separate documents in PDF format. ▪ Auxiliary datasets and information that has been used during interpretation (e.g. the magnetic field of the solar wind, exact location and position of the satellite) are not included in the dataset.
FIRE	▪ Essential for understanding the dataset: the settings of the measurement devices, the coordinates of the observation points, the logbook containing information about individual measurements and the curated field report, which contains the description of the measurement, parameters and other important information. ▪ The SEG-Y data files contain a header section, where certain standardised metadata of the measurements is stored. ▪ The coordinates of the observation points and the logbook are in structured text files. There is no standardised convention in the field describing in detail how the information should be stored. ▪ The field report is a document in DOC format. ▪ The whole experiment and results are presented both in articles in PDF format and in a video in WMF format.
FSD	▪ Essential for understanding the dataset: the description of the content of the dataset, information about test subjects and/or the target group who took part in the survey, the variables used in analysis. ▪ The description of the content, list of variables and other essential metadata are stored in a machine-readable XML file conforming to the DDI 2.0 standard, and additionally in a human-readable PDF document. ▪ The example dataset has already been processed by FSD for preservation and reuse. Researchers do not typically produce the metadata in such organised structure themselves; it is common in the field of social sciences that datasets are processed and metadata homogenised by data archives.
Crystals	▪ Essential for understanding the dataset: the research method, which has been described in a scientific article in PDF format. ▪ The dataset does not include the raw data produced by measurement devices. The most important results and parameters have been compiled into a single Excel file (XLSX format). Some results are presented in graphical form in addition to numeric tables. ▪ The dataset in its current form is mainly intended to be read by humans. Machine-readable files can be produced by selecting desired parts of the results and storing them in separate tables.

Dataset	Metadata
MAXIV	▪ The example dataset does not belong to any actual research project, as the MAX IV laboratory is not operational yet. ▪ The metadata is stored in binary format in a HDF5 file according to the Nexus HDF5 specification.
Planck	▪ Essential for understanding the dataset: the description of the measurement device (the Planck satellite), the settings and the method of producing the data ▪ The FITS data files contain a header section, where settings and parameters used during the measurement are stored. ▪ Descriptions of the measurement device and the method of producing the data are in HTML format on the web site of the Planck data archive.
RITU	▪ Essential for understanding the dataset: the description of the measurement device, the configuration parameters, the observation notebook and the structure of the data file. ▪ The description of the measurement device is stored as text in a text file and as a diagram in PDF format. ▪ The configuration parameters are stored in a structured text file. ▪ The description of the structure of the data file is in a PDF document. ▪ The GREAT data format is proprietary to the manufacturer; there is no standardised convention about measurement data file formats. ▪ The results of the analysis are stored in Aida XML format, which is commonly used in the field.
SMEAR	▪ Essential for understanding the dataset: the descriptions of the measurement devices, the locations of the observation points, the measurement parameters and the structure of the database. ▪ The descriptions of the measurement devices are in HTML format on the SMEAR project web site. ▪ The locations of the observation points, short descriptions of measured physical quantities and notes on data post processing and quality control operations are in separate tables in the database.
Suomi24	▪ Essential for understanding the dataset: background information of the source data, abbreviations used in parsing the data (words and sentences). ▪ A brief description of the source data is in a text file. ▪ The actual data files in VRT format contain metadata for each analysed message in a structure resembling XML. ▪ The abbreviations used in parsing the data are not documented. According to the interviewed person, a linguistic researcher can understand their meaning.

In the National Digital Library Standard Portfolio [NDL_Standards], metadata is divided into descriptive, administrative and structural metadata. Administrative metadata is further divided into technical metadata, metadata for digital preservation and usage rights.

The Standard Portfolio includes a list of recommended formats for transferring descriptive metadata. In the case of research datasets, similar widely used metadata formats exist only in certain fields of science. For example, the DDI format is commonly used in social sciences and standardised by an international alliance of organisations, so it can be recommended for use in long-term preservation. In many other fields, it needs to be studied whether suitable descriptive metadata formats can be found and what kinds of criteria need to be set for their use. This topic will not be addressed in more detail in this report.

Technical metadata is closely connected to file formats. The mandatory technical metadata schemes in the NDL are listed in the recommended and accepted file formats specification [NDL_Formats]. In the area of research data, comparable well-defined metadata schemes are available only for a few file formats. Certain common guidelines covering all types of datasets can be provided, such as which character sets should be used in texts, but much of the metadata only concerns specific file formats, fields of science or research methods.

To store the metadata for digital preservation, usage rights and structural metadata for research datasets, it should be possible to use the PREMIS and METS formats already specified in the NDL. However, that has not been studied in detail in this report. More detailed information about the suitability of the abovementioned formats and the potential need for adjustments will be gathered during the development of the metadata creation tool and the packaging service as well as in pilot projects of packaging data sets for preservation. The pilots conducted in 2015 showed among other things that particular attention must be paid to registering information about ownership and usage rights. As an example, the actual data files within the dataset may be free to redistribute and reuse, but the publications crucial to understanding the data are covered by the copyright of the publishing company.

Requirements related to metadata and the packaging of datasets are presented in the chapter entitled “Accepting Datasets for Digital Preservation”. More information about usage rights and metadata related to their administration is available in a separate report that will be published in the near future.

6. ANALYSIS OF THE FILE FORMATS

6.1. File Formats of the Example Datasets

The example datasets contained in total 26 different file formats, half of which have already been approved as recommended formats or acceptable for transfer into digital preservation in the National Digital Library (NDL). The recommended formats, 10 in total, were HTML, Java (preservable as text), JPEG, JSON (preservable as text), PDF, PNG, TSV (preservable as text), TXT (normal and structured) and XML. Formats acceptable for transfer were DOC/DOCX, WMV and XLSX.

Of the remaining, not yet approved formats, the majority (11 in total) were open and documented: BAM/SAM, CRAM, FITS, GREAT, HDF5, MySQL dump, NIfTI, RTF, SEG-Y, SIARD and VRT. CorelDraw and SPSS Portable were the only two formats without open documentation. The BrainImg dataset was additionally organised according to the BIDS specification, which is not a file format but a directory structure.

The majority of file formats in the example datasets have already been approved for digital preservation or could be added to the list by defining the necessary technical metadata and other details. On the other hand, some of the formats (Java, JSON, TSV) are supported in the NDL only as normal text – it would be possible to improve the support. Custom-structured text formats, developed by the research group or other scientists in the field, can also be preserved as text, but the structures need to be documented first.

Of the non-documented formats it would be easy to convert the CorelDraw files of the FIRE dataset into PDF format. They are part of the documentation of the dataset, which the user needs to read, but it is not essential to be able to modify them. In the case of SPSS Portable in the FSD data, the user needs to be able to modify the files when reusing the dataset. Converting them without loss of data into an open format is not straightforward.

A comparison with recommended file formats of six other preservation organisations produces a similar result. The file formats approved in the NDL are also widely internationally accepted. Of the file formats currently not approved in the NDL, four (HDF5, RTF, SIARD and SPSS Portable) have been approved by some of the organisations in the comparison; the remaining nine are missing also from all international lists of recommended formats.

The file formats of the example datasets are analysed in more detail in Appendix C.

6.2. A Quantitative Survey of Research Data File Formats

A rough estimate of the popularity of file formats in Finnish research can be acquired by looking at the IDA storage service at CSC, which is used by many researchers. The thirty most popular file formats stored in IDA are presented in Figure 2.

The graph shows that many of the file formats accepted for preservation in the NDL are also popular in research datasets stored in IDA. For example, JPEG, PNG and TIFF images, Excel, Word and PDF documents, ASCII texts and XML files have been approved in the NDL either as recommended formats or acceptable for transfer.

On the other hand, the graph shows a large number of file formats that have not yet been considered in the NDL. The most popular type is "generic", which simply means that the automatic file type detection in IDA does not recognise the format. Many of them are probably measurement data or other data files. TAR and GZIP files also contain several different file formats, because the analysis does not look inside TAR and GZIP packages. One popular category among the recognised formats comprises source code files (C, Fortran, Java and Perl).

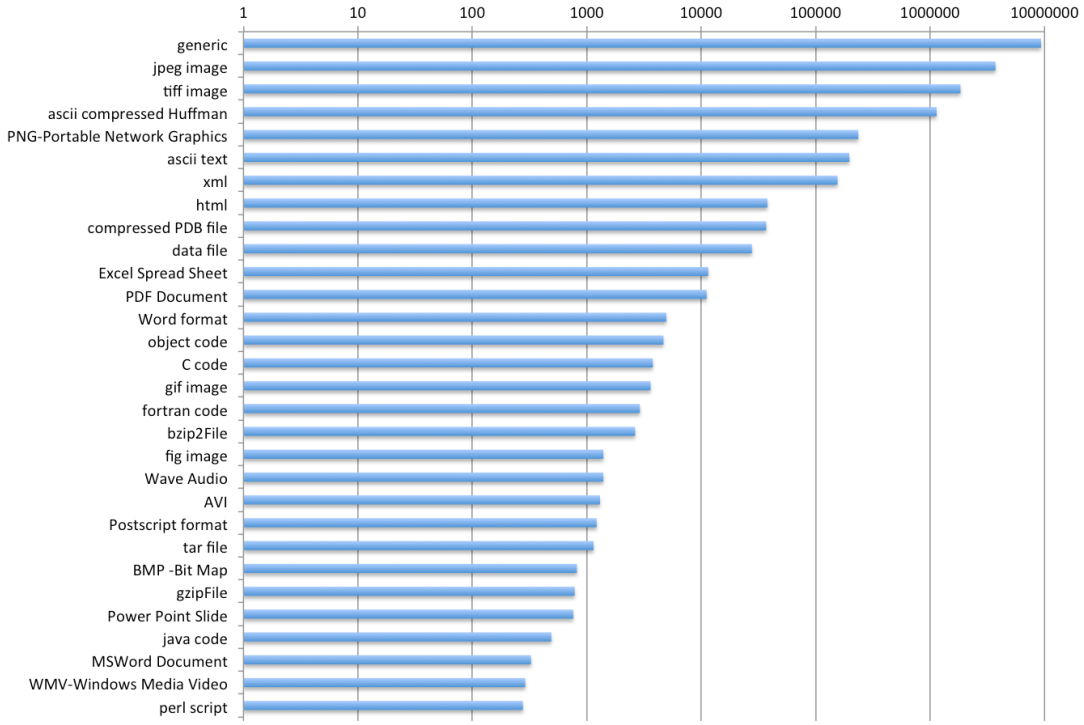

Figure 2: 30 most popular file formats in the IDA storage service

A more comprehensive analysis of the popularity of research data file formats would require a broad inquiry targeted at universities and research groups, which was not possible within this project. In Austria, an extensive survey covering the whole national scientific community has been made [Austrian_Survey] and its results are similar to those of the rough IDA-based estimate. Almost all researchers produce text, tables and images, but self-developed software (source code and binary files) and measurement data are also important file categories. Databases are also a popular category in the Austrian survey, but it remains unclear what kind of databases they are. Probably at least some of them are normal files produced by statistical software packages used in social sciences, as the multiple choice questionnaire did not contain any other suitable category for them.

Analysis based on number of files emphasises file formats where data is divided into many small files instead of one big file. Looking at the size of the files would on the other hand emphasise scientific fields that manipulate large masses of data. However, fields producing smaller datasets are equally important from the preservation and reuse point of view. The Austrian survey gives a bit more information on which file formats should be supported to serve as many researchers as possible. Still, it can only be used on the general level; to know more about specific needs, it is necessary to study formats within each field of science in detail.

6.3. File Formats Widely Used in Research Datasets

This section presents file formats that are widely used in research datasets, organised by the purpose of the formats and by scientific fields. The information is based on the interviews conducted during the project and websites. The sample does not cover all research datasets or scientific fields, but it gives a good overview of different data types and file formats from the digital preservation perspective.

General Purpose File Formats for Research Data

General purpose file formats for research data offer a possibility to store tables of floating point numbers and other popular data structures in an efficient and hardware-independent manner. The files are among other things compatible between hardware architectures using different byte orders (big endian vs. little endian). The development may have begun within a certain field of science, but the structures and specifications of the formats are scientific field-independent and suitable for many different use cases.

The best-known file format belonging to this group is Hierarchical Data Format 5 (HDF5). It defines two basic elements, using which one can store almost any type of data and associated metadata, and organise the data objects in a tree structure. HDF5 is an open standard, but the specification is fairly long and complex. From the digital preservation point of view, it should be noted that simply using HDF5 does not ensure understandability; it is essential to describe the used data types and metadata.

Various more detailed specifications have been created on top of HDF5 for purposes such as describing data types used within a certain scientific field or certain types of datasets. Examples of these HDF5-based file formats are Network Common Data Form version 4 (NetCDF-4), Data Exchange [DXFile] and Nexus HDF5, which is used in the example dataset of the MAX IV laboratory and by several other synchrotrons. They are easier to manage from the digital preservation perspective than generic HDF5, because the permitted data types are more accurately specified. On the other hand, treating all HDF5 variations as separate file formats leads to a larger number of approved formats, specifications and versions.

Older, but still widely used and maintained general purpose formats for research data are Common Data Format (CDF), Network Common Data Form version 3 (NetCDF-3) and Hierarchical Data Format 4 (HDF4). They can be used to store similar kinds of datasets, but have significant differences in terms of features and internal structures. That makes the formats incompatible with each other, even NetCDF-3 and NetCDF-4 or HDF4 and HDF5 [CDF_FAQ].

From the digital preservation perspective, general purpose formats share some characteristics with custom formats developed by researchers themselves. Neither has a specified standard location for metadata, which is essential for understandability and reuse. However, standardisation on at least the generic level and openly available programming libraries to process the files are a significant advantage compared to self-developed custom formats. Therefore, general purpose formats should be preferred as well as criteria and tools developed to document the data structures and metadata well enough for preservation.

Internationally, HDF5 has been approved as a recommended or transferable format in three (CINES, DANS, LoC) of the six surveyed organisations. LoC also mentions CDF. More information about the level of support or documentation requirements was not available.

Measurement Device-specific Formats

Measurement device-specific file formats are used in many different fields of science. They differ quite widely from each other due to both differences between the devices themselves and varying practices employed by manufacturers. File formats of devices that are primarily sold to companies for production use are often closed and require special software. Files produced by devices that have been developed mainly for research are often also specific to the device or scientific field, but openly documented.

Measurement parameters and device-specific metadata are essential for the interpretation of the measurement data. They may be stored either in the same file as the data (in the header section of the file) or in separate files. In long-term preservation, it is important to recognise which parameters and metadata are essential for understandability and reuse, and to ensure that the received dataset includes those pieces of information. A description of the operating principle of the measurement device may also be necessary to understand the data.

Examples of measurement device-specific file formats are the GREAT format in the RITU example dataset and Digital Micrograph 3 (DM3) files produced by electron microscopes. The documentation of the GREAT format can be downloaded from the manufacturer website. In the case of DM3 files, users have themselves inspected the structure of the files and created a partial documentation of the format based on their observations.

Depending on the type of measurement data it may be possible to convert the files to a more easily preservable format. Commonly used formats also make it easier to take advantage of datasets across different scientific fields. For example, the DM3 files of electron microscopes

are essentially bitmap images and can be preserved for example in TIFF format, which has been approved in the NDL. That however does not preserve the DM3 file metadata, which needs to be stored separately. In converting images it is also important to make sure that the resolution and bit depth of the original image are preserved.

None of the surveyed foreign preservation organisations has a policy about preservation of measurement device-specific file formats or provides instructions about the topic.

Geospatial Data File Formats

Geospatial datasets and maps are particularly interesting for preservation, as they can be used in many different scientific fields and in cross-disciplinary research. Measurements including geographical coordinates can be compared with other data such as statistics related to countries or municipalities, and be plotted on maps for visual observations. Compatibility of essential features such as coordinate systems is particularly important when comparing datasets.

The geospatial file formats can roughly be divided into vector- and raster-based formats. Vector formats are based on coordinates connected by straight or curved lines, whereas raster formats are based on regular grids with equal distances between points. There are several different file formats for both classes. Examples of vector-based formats are Esri Shapefile (Shape), Geography Markup Language (GML) and Keyhole Markup Language (KML). GeoTIFF, JPEG2000 and PNG are popular raster-based formats. There are also a few formats that do not belong to either of these two classes, such as the LAS format used for light detection and ranging measurements, and various databases.

Esri Shapefile is a set of interrelated file formats developed by a private company selling GIS software. Due to the popularity of that software (about 40% of the GIS software market) the format has become a de facto standard in the field. It is simple, stable, fairly well documented and supported also in software not developed by Esri, which makes the format suitable for digital preservation. The Shapefile format is already being used for example in the Paituli spatial data download service that is part of the Avaa portal. For long-term preservation, it still needs to be specified which of the optional features of Shapefile are supported and which metadata is required. Additionally, it is important to ensure that all necessary components are received: the Shapefile format consists of several separate but interconnected files.

Open Geospatial Consortium (OGC) is a standards organisation focusing on geospatial data. It is based on voluntary participation and has over 500 member organisations. Members include both commercial companies and non-commercial organisations, governmental entities and research organisations. OGC has created or selected several dozen geospatial data-related standards that complement each other. They are all freely available on the website of the organisation [OGC_Standards].

The most important OGC standard for research datasets is Geography Markup Language (GML), an XML-based markup language to present various geospatial features. It is also an ISO Standard (ISO 19136:2007). In addition to the core part of the standard, GML files may include community-developed extensions. GML is open, well documented, widely supported and therefore suitable for preservation. GML extensions can be accepted into preservation as XML documents, even without specific support for them. Geography Markup Language should not be mixed with the older Graph Modeling Language format, which is used to store graphs and carries the same acronym GML.

Keyhole Markup Language is an XML-based markup language developed by Google, designed especially for annotating and visualising two- and three-dimensional maps. It is nowadays also an OGC-approved standard. KML overlaps partly with GML and there are plans to harmonise some elements or at least improve the compatibility between the two languages in the future. KML is well documented and suitable for preservation, at least as an XML document (without specific KML support), which is already a recommended format in the NDL.

GeoPackage is a third general geospatial data standard by OGC. It is technically an SQLite database and may include both vector and raster data. GeoPackage is a fairly new standard and it has been envisioned to replace both Shapefiles and GML/KML files, but it is not very widely used yet.

The most important of the raster file formats is the Tagged Image File Format (TIFF), which has already been approved as a recommended format in the NDL. In geospatial data, TIFF images can however include additional channels or small extra files including information about, for example, the position of the image and the coordinate system used. The GeoTIFF standard defines how to store geospatial metadata within TIFF image files. Both multichannel images and GeoTIFF metadata should be considered in the preservation support of TIFF images. Other raster formats commonly used in geospatial data are JPEG2000 and PNG, which are already recommended formats in the NDL.

Light detection and ranging measurements use their own file LASer (LAS) format, which has been established as a de facto standard in the field. It is a fairly simple binary format, consisting of a header and a data section. The header section includes the most important measurement-related metadata. The LAS format is open, well documented and widely supported in software used in the field, and is therefore suitable for preservation.

Geospatial datasets are being increasingly stored in various databases, which offer quick and handy methods for choosing desired parts of the datasets as well as efficient search functionalities. There is no standard format for databases, which makes them more challenging than other formats from the preservation point of view. A closer look can be found in the Databases section.

In addition to file formats, the choice of the coordinate systems is essential for the compatibility of geospatial datasets. Globally there are as many as tens of thousands of coordinate systems. In Finland, even different municipalities may use different coordinate systems. To facilitate the reuse of the datasets the number of supported coordinate systems should be as small as possible, and it should be required that preserved datasets use one of the supported systems. Finnish datasets should use coordinate systems specified in the JHS-197 recommendation, primarily ETRS-TM35FIN coordinates.

Internationally, geospatial file formats have been considered at least to some extent in all of the six surveyed organisations. The choices and recommendations differ quite a bit between them. In CINES, the only accepted format is GeoTIFF, which is also on all other lists except at NAA. The Open Spatial Consortium GML format has been approved by DANS, LAC, LoC and UKDA. ESRI Shapefile and KML can be found on the DANS, LAC and UKDA lists; NAA on the other hand recommends the Spatial Data File (SDF) format by Autodesk. The LAC list includes quite a few more formats. LoC recommends storing the original dataset in the most complete form, even if the file format would be a closed one. It additionally recommends native formats of widely used GIS software as well as formats developed or chosen by OGC.

Software Source Code and Binary File Formats

In practically all fields of science, at least some of the researchers program themselves, and the datasets include the source code and binary files of the developed software. Their preservation is useful both for ensuring the reproducibility of the research and for reuse: using software that has been developed to analyse the data often makes it possible to start further research quickly.

Source code files are in principle easy to preserve. Independent of the used programming language, they are text files, which is already a recommended format in the NDL. Metadata needs some attention; for example, the name and version of the programming language are essential information. The quality of the internal documentation of the code varies largely, but assessing the quality is practically impossible: it would require an in-depth manual look into the files and the inner workings of the program. If desired, the documentation can be automatically extracted from the code and indexed for search functions.

Executable binary files that have been compiled from the source code are convenient for users but difficult for digital preservation. They are typically dependent on both the operating system and a large number of library files, often even requiring specific versions of those libraries. It may be useful to accept executable files into preservation and offer users the possibility to download them in addition to the dataset, but their functionality in future operating system versions cannot be guaranteed.

Also, compiling source code into an executable program can be more difficult in a new system with newer libraries than in the original development environment. However, source code files can be modified, which gives a competent user a possibility to do the necessary changes to enable the compilation. Reading the source code may also help in understanding the dataset or the research method. Therefore, it is worthwhile to include source code files as part of the preserved dataset.

In international preservation organisations, source code files can be stored as text files just like in the NDL. Only the LoC document gives more detailed instructions about describing and preserving the metadata and operating system environment related to the code.

Markup Languages

Markup languages can be used for many different purposes independent of the scientific field: they can be used to store data and metadata or to write documentation. Popular markup languages include HTML, JSON, and the particularly versatile XML, which are presented in more detail in Appendix C. Other noteworthy languages are Standard Generalised Markup Language SGML, LaTeX, which is designed for writing articles and books, and YAML, which is particularly suitable for metadata.

All markup languages are structured text, so they can be accepted for preservation at least as text files. However, processing markup languages is much more convenient than processing plain text, so users should be encouraged to use them by offering advanced support for markup languages in the preservation service. Standard compliance can be validated automatically, and it is not necessary to require users to send detailed structural documentation of files that have been successfully validated. It is possible to create scientific field-specific XML or JSON schemes for storing metadata, and to offer a user interface in the metadata creation tool for filling in the information.

Internationally, HTML and XML are widely accepted as preservable formats. DANS and LoC additionally support SGML. JSON is also supported by LoC and the subset JSON-LD by DANS. In any case, all markup languages can be preserved at least as text files, which are supported by all organisations.

File Formats of Statistical Analysis Packages and Spreadsheet Applications

Statistical analysis software is popular especially in social sciences research. Each software package typically has its own file format, most of which are proprietary. One of the most popular statistical analysis packages is SPSS, a commercial solution whose file formats SAV and SPSS Portable have become de facto standards in the field. Most other software packages, including the open source PSPP, support these formats at least partially. Neither of them, however, is openly documented. The word "portable" in the latter means only that the files can be transferred between different computer architectures. SAS is another statistical analysis software that is in wide use, particularly in health sciences; it uses its own proprietary file format.

Data analysed by the statistical software packages can be converted into spreadsheet file formats or into the CSV format, which have been approved as recommended or acceptable formats in the NDL. However, some information is often lost in the conversion, and opening the files again in the statistical software for further processing may not succeed without problems. In Finland, the Social Sciences Data Archive FSD has chosen SPSS Portable as their preservation format. Data in other statistical software formats is converted to it using commercial software specially designed for such conversions. In hands-on tests, SPSS Portable

has proved itself to be well downwards and upwards compatible. According to FSD, the format can therefore be recommended for preservation. When doing analysis, it has some restrictions compared to the native formats of statistical analysis packages.

FSD is actively following the usability of SPSS Portable format and is ready to convert the files into other formats in the future if necessary. That is a good reason to consider an exception to the general approval criteria, which require open specifications of the preserved file formats. SPSS Portable is already on the preliminary list of formats soon to be approved in the NDL, with certain preconditions. Storing the data additionally in CSV format in parallel to the SPSS Portable format is also an option. The files are typically small, so from the size point of view storing them in two formats in parallel would not be a problem.

Researchers who program themselves are increasingly doing statistical analysis using the open source R statistical analysis software. The analysis commands are given using the R programming language instead of a graphical user interface like in SPSS and many other analysis packages. R supports several different open and proprietary file formats; for example, the CSV format is popular. The programming commands are stored in a structured text file.

Spreadsheet software is widely used in many scientific fields. The two most popular spreadsheets are Microsoft Excel and LibreOffice/OpenOffice Calc, which both have their own file formats. LibreOffice Open Document Spreadsheet (ODS) has been approved as a recommended format in the NDL; the Excel Office Open XML (XLSX) is also acceptable for transfer. One should, however, note that research datasets are more likely than cultural datasets to use advanced features of the software. That may lead to problems when opening the files in other software than the one that was originally used to create them, or when converting the files to some other format.

Internationally, CSV, ODS and XLSX are approved either as recommended or acceptable formats in all of the surveyed organisations. SPSS Portable is approved by DANS and UKDA. The same two organisations also support a few other proprietary statistical software file formats, at least as acceptable formats.

Computer Aided Design (CAD) and Modelling File Formats

Two- and three-dimensional computer aided design and modelling can be used in different scientific fields. In particular, 3D modelling is becoming more common. The models may be related to measurement devices or materials being researched, but also to social sciences research, which studies the influence of objects and environment on research subjects. It may be useful to preserve the models for ensuring understandability or for reuse.

Examples of popular modelling software include the commercial AutoCAD, SolidWorks and SketchUP as well as the open source Blender. Simple 2D models are often created using general vector graphics software such as Microsoft PowerPoint, LibreOffice Draw, Corel Draw and Adobe Illustrator. Three-dimensional structures can also be based on images produced by measurement devices such as magnet resonance imaging (MRI) devices or 3D scanners.

If the only objective is to ensure understandability, for example to describe a measurement device used in the research, the models can be printed as images as PDF files, already a recommended format in the NDL. However, PDF is not suitable for editing or otherwise reusing the models.

Of the general purpose vector graphics software, the formats of LibreOffice Draw and Microsoft PowerPoint have been approved in the NDL as recommended or acceptable. They are however rather unusable for reuse in modelling, particularly with respect to 3D models.

The most popular modelling software file format is the AutoCAD DWG format. Its development is controlled by Autodesk, Inc., and official documentation is not publicly available. However, the Open Design Alliance has produced a fairly accurate description of the format [ODA_DWG_Specification] and it is reasonably well supported in many software packages. If DWG is considered to be approved as a recommended or acceptable format, validation and acceptance requirements need to be based on the unofficial documentation.

Other noteworthy 2D and 3D modelling file formats are 3D Studio (3DS), AutoCAD Drawing Interchange Format (DXF), Blender format BLEND, Initial Graphics Exchange Specification (IGES), Product Representation Compact (PRC), STEP File, Wavefront OBJ and X3D. Of these, STEP (ISO 10303-21) and IGES (v. 5.3, ANSI 1996) are both official standards and well documented, but their feature sets are outdated. X3D is a newer standard developed particularly for presenting 3D content online, but it is not very well suited for reusing the models.

Autocad DXF is Autodesk's suggested exchange format between different CAD software packages. Unlike DWG, it is openly documented, but does not support all the new features. 3D Studio is the 3D modelling format developed by the same company, and like the 2D format DWG, it has become a de facto standard despite the lack of documentation. The BLEND format used by Blender is versatile and due to its open source background well documented, but has an unconventional structure and is not well supported in other software. Wavefront OBJ is a documented, fairly simple format for presenting 3D structures, and the standardised PRC (ISO 14739-1:2014) is designed for embedding 3D models in PDF files. It is however not part of the PDF 1.7 or PDF/A standards that have been approved in the NDL.

None of the above mentioned 2D and 3D modelling file formats is very well suited for digital preservation. Either the documentation or compatibility is lacking, the formats are outdated or they are only suitable for presentation; in other words, they ensure understandability but are not suited for the reuse of models. It is also unclear how widely 2D and 3D models are used in research datasets and which formats are the most popular.

Four of the six surveyed foreign organisations (DANS, LAC, NAA, UKDA) accept AutoCAD DWG and DXF formats for preservation. DANS names DXF as the recommended choice, while UKDA prefers DWG.

Gene Sequencing File Formats

Gene sequences are usually stored in BAM/SAM and CRAM file formats, which are presented in Appendix C. Other noteworthy formats are BCF/VCF and FastQ.

The raw data produced by the sequencer is typically stored in FastQ format, and the processed data in BAM format. However, BAM can also be used as a replacement for FastQ, and its structure allows for more versatile storage of metadata. Both formats are openly documented. The advantage of FastQ is simplicity, but BAM is probably a better choice for digital preservation thanks to its better-designed metadata features.

The CRAM format was introduced to save storage space – it is basically a BAM file with parts of the gene sequence information omitted in a documented and controlled way. Its additional features make the CRAM format more complicated than BAM. As the gene sequencing datasets are large, up to dozens or hundreds of terabytes, it is still justified to support CRAM as an additional preservation format.

Variant Call Format (VCF) and its binary sibling BCF are used for processed information. VCF/BCF files are not pure sequencing data but genotypes, and they may include genomes from one or more persons. It is a relatively new format but has already been established as a de facto standard in the field. It complements the BAM and CRAM formats, is openly documented and therefore also suitable for preservation.

The surveyed foreign preservation organisations do not have gene sequencing file formats on their lists of recommended and acceptable formats. Gene sequencing datasets are typically stored in specialised gene research data banks, which are actively used by the international research community. This has led to fairly good stabilisation of the file formats.

Brain Research File Formats

Brain functions are typically researched using series of images produced by magnetic resonance imaging (MRI). Other commonly used technologies include electroencephalography (EEG) and magnetoencephalography (MEG).

MRI technology provides information about both the anatomy and the functionality (functional MRI) of brains, which can be compared between different test subjects and experiment settings. Large amounts of data are often accumulated. Brain research using MRI provides a good example of how file formats and common conventions have developed as a result of technological advancements and increased cooperation between research groups.

MRI devices typically produce image files in the DICOM format, which also includes the parameters used during the imaging session and other metadata. However, the parameters are manufacturer-specific and the DICOM standard allows proprietary elements within the file, which cannot be redistributed due to copyright restrictions. Therefore, DICOM files are often converted to the manufacturer-independent NIfTI format, which is presented in Appendix C. The BIDS directory structure, also presented in the same Appendix, has become a standard in brain research and requires the use of NIfTI. It also specifies file naming conventions and the storage of metadata in TSV and JSON formats.

However, the NIfTI format alone does not meet the needs of all brain researchers. New methods, which for example compare signals moving inside the brain, surface structures and interrelations between different parts of the brain, produce data that cannot be stored in NIfTI format. That has led to the birth of the GIFTI and CIFTI file formats. GIFTI files are used to store data about brain surfaces. CIFTI is an extension to store additional metadata in XML format and additional measurement data inside NIfTI. Neither of the two has been adopted as widely as NIfTI and they are not yet part of the BIDS specification. However, adoption in the well-known and respected Human Connectome project means that the two new formats are represented in the datasets of more and more brain researchers.

The file formats appear to be properly documented but are still under development. In addition to the rather new CIFTI and GIFTI formats, both NIfTI and CIFTI have received a new version within the last two years (NIfTI-2 and CIFTI-2), neither of which is fully compatible with the old version. Internal efforts to ensure the reusability of datasets within the field of brain research will probably lead to gradual stabilisation of the formats. That also applies to file formats for storing EEG and MEG scans, although there is currently less agreement on common formats than with MRI images.

From the digital preservation point of view, the NIfTI, CIFTI and GIFTI formats including their new versions are acceptable with respect to openness, documentation and software support. Before accepting the formats for preservation, required metadata fields and details of their content need to be specified. It is also rather likely that the files need to be converted into newer formats in the future in order to keep up with the rapid development of the field.

None of the six surveyed foreign data preservation organisations mention brain research file formats on their lists of recommended formats. Similarly to gene sequence data, brain research datasets are primarily stored in dedicated services within the research community, which simultaneously control the development of the file formats.

Medical Technology File Formats

In addition to brain research, which was presented in a separate section, there is plenty of other research taking advantage of medical technology. Characteristic to the field is the use of expensive measurement devices, the details of which are often trade secrets of the manufacturers. Many devices support the DICOM standard, which defines both the connection protocol and the image file format. The DICOM files are therefore documented, but certain parts of them and other files produced by the devices are often manufacturer-specific; documentation is not openly available and proprietary applications are needed to process and analyse the files.

From the digital preservation perspective, medical technology is challenging. Many fields of research have not yet started harmonising file formats in the same way as brain researchers do. Data protection requirements set additional limitations to data reuse. Supporting DICOM image files should however be considered.

Internationally DICOM has been approved by DANS and LAC. The other four surveyed organisations do not have it on their lists of recommended formats.

Linguistics File Formats

Linguistic research deals with different types of data, each of which has their own file formats. The three main groups are text, sound and video, all of which can additionally be enriched with analytic information.

The analysis of textual information is most often stored in structured text files. The structure may contain information about the syntax, morphology and semantics of the analysed text, displaying for example the elements of the sentence and the conjugation of the words. The file formats are usually open, but not always well documented. The CoNLL-U format [CoNLL-U] is better documented than most and has established itself as a de facto standard.

Most of the file formats do not have any possibility or standard location to store metadata that is critical for the understandability of the data. Furthermore, the abovementioned CoNLL-U files do not have any header section or other means of storing metadata. On a technical level, different character sets may lead to incompatibility problems, especially with older datasets. New datasets nearly always use the UTF-8 character set. On the descriptive level, the source of the text, the context and the language used are examples of essential metadata. The metadata can be stored in a separate file, for example in XML or JSON format. In the CLARIN project, different metadata schemes and formats are managed using the Component MetaData Infrastructure [CLARIN_CMDI].

The VRT format used in the Suomi24 example dataset is a mixed format, where metadata in an XML-like structure is combined with CoNLL-U type analysis in the same file. It is however not an XML format and the structure and used abbreviations are less well documented than CoNLL-U.

Text Encoding Initiative (TEI) is both a consortium and an XML-based standard developed by the consortium, designed for storing textual datasets. It enables storing both the original text, markings about the structure and metadata in one file. The TEI standard is extensive, but it has been designed in a flexible manner so that only the necessary parts of the definition can be used. A valid TEI document can contain text stored almost as is, for example only using a few XML tags to separate paragraphs similar to HTML, or the text can be enriched with very detailed markup connected to each individual word.

As an XML-based format, TEI is suitable for automatic processing and digital preservation. Validators capable of checking the syntax and conformance to the TEI schema are available. In digital preservation, it needs to be defined which metadata fields are required and the validation needs to be extended to verify those fields.

The TEI standard is flexible enough that almost all structured text files used in linguistics could theoretically be converted to TEI files. However, many readily available analysis tools do not support it and linguists who program themselves often prefer simpler forms such as CoNLL-U. Therefore, it is justified to also support those simpler forms, with the same documentation and metadata requirements as for structured text files in general.

Sound and video recordings use the same file formats as in the NDL, and the specifications already defined in the NDL can be applied. Additionally, it is important to be able to make annotations referring to specific moments of the recording. The annotations are typically stored in their own separate file in the ELAN Annotation Format (EAF) developed specifically for that purpose. It is a rather simple XML-based format, which is well suited for preservation. As is usual with new file formats, the necessary metadata fields and their details need to be defined. It also needs to be ensured that the EAF file and the relevant sound or video recording are stored together.

Internationally, TEI is on the recommended formats lists of CINES, DANS and LoC. The other abovementioned formats cannot be found on any of the lists of the surveyed organisations.

However, all of them generally accept XML-based formats, and some also give more detailed guidance about storing XML files.

Seismology File Formats

The most common file format for storing measurement data in seismology is the SEG-Y format, which is also used in the FIRE example dataset. Another widely used format is Seismic Unix, which is the format of an open source analysis software carrying the same name. Seismic Unix files can be easily converted to SEG-Y and vice versa, so it will probably be sufficient to support SEG-Y in the digital preservation. All the most important software packages in the field are able to both read and write the SEG-Y format.

In addition to the data files, essential information when interpreting seismic datasets are the coordinates of observation points, the measurement parameters, the observation logbook and the field report, which includes both the used parameters and a written description of the measurement. There is no widely agreed convention for storing this information. Some parameters can be stored in the header section of SEG-Y files, but the coordinates, the observation logbook and the field report are typically structured text files or documents written using word processing software. Their preservation needs to rely on the general requirements for structured text files. Particular attention should be paid to the compatibility of geographical coordinates with other datasets. If necessary, the coordinates of observation points should be converted to one of the coordinate systems that are supported in preservation.

Seismology file formats are not listed on the recommended formats lists of the surveyed international preservation organisations.

Earth Science File Formats

Atmospheric science and ecosystems research, or more generally Earth System research, typically uses a set of geographically distributed measurement devices. Projects are often international, which has an influence on collecting and processing data.

The most common file formats are structured text, CSV and HDF5. Remote sensing data is used as a reference and it is most commonly either in GeoTIFF or NetCDF format. NetCDF is based on HDF5. The file formats are usually open and well documented.

Databases are also commonly used in the field, in particular in international projects. In most cases databases do not directly replace the data files produced by measurement devices but instead complement them and provide interfaces that help researchers to use the datasets. International infrastructures administering the databases often focus on certain variables, and collect measurement results from several research groups all over the world. The Finnish SMEAR project sends data to several different international infrastructures, and additionally maintains a national dataset in Finland, which includes more variables but is geographically more restricted [SMEAR_AVAA].

Earth science researchers typically use datasets from several different sources. It should be noted that storage conventions often differ between fields. Ecosystems data is mostly based on the values of variables, for example the value of temperature independently of which device it has been measured with. The measurement device may be changed during data collection. In atmospheric sciences, a new dataset is started whenever the measurement device changes.

Earth system research file formats are not separately listed by the surveyed foreign preservation organisations. However, all of them designate CSV as a recommended or acceptable format (in CINES only as text without specific CSV support), GeoTIFF is approved in all but NAA and HDF in three (CINES, DANS, LoC) of the six surveyed organisations.

Space Research File Formats

Space research utilises many kinds of observation data, among other things telescope images of various objects (the Earth, the Sun, other planets and stars) using different wavelengths and environmental data of the satellites. The latter can be measurements of the plasma environment (density, temperature, flow rate), the electromagnetic field or the radiation around the satellite. Each of the observation types may use several different file formats.

Data about particle radiation is mostly stored in CDF or text format, sometimes also as HDF5. The most popular text file format is CSV; structured text files with fixed width columns are used as well. Data about plasma environment and electromagnetic fields is stored as CDF or text.

In astronomy and satellite images, FITS is the most popular file format and also the format used by the Planck example dataset. It is a relatively complex format that allows storing not only images but also many other kinds of data. FITS is an open, documented format that is suitable for preservation when metadata requirements have been defined. A more detailed description of the format can be found in Appendix C.

Satellite images destined for manual observations are distributed in general image file formats such as TIFF, PNG and JPEG, which are approved as recommended formats in the NDL.

Internationally HDF5 has been approved as a recommended or acceptable format in three (CINES, DANS, LoC) of the six surveyed organisations. In addition, LoC lists the CDF format. The FITS format is not on any of the lists of the surveyed organisations, but it is an established format in storage services within the scientific field. Text files and general image file formats are widely accepted both nationally and internationally.

Particle and Nuclear Physics File Formats

Particle and nuclear physics research typically relies on expensive measurement devices specifically developed for research purposes, as well as highly specialised software. The file formats are often software-specific but rather stable, as the research projects are long and the data may be analysed over dozens of years. The source code of the software is in most cases available and the file formats in principle open, but the documentation may be lacking. Therefore, the datasets are not very easily transferable from one software to another.

The best-known file format within the field is the ROOT format, developed at CERN and named after the software that uses it. The format is optimised particularly for high-performance computing, as the datasets are large and their analysis needs great computing capacity. The file format itself is quite well documented, but the ROOT analysis software and programming library are extensive and complex.

Other widely used file formats in the field are RadWare, MED and ENDSF. The first two are primarily software packages and the documentation of the file formats is inadequate. This answer to a question about file format structure on the RadWare FAQ illustrates the situation: "Many and various. The best (and most accurate) way to find the format is to look at the source code for routines that read / write the files that you are interested in." ENDSF is more clearly a file format and also appropriately documented. Databases are also used to a certain extent, for example in the National Nuclear Data Center in the U.S. [NNDC_Databases].

Most of the particle and nuclear physics file formats have been developed in the research organisations themselves. The storage and preservation of raw data is also mostly centralised in the same organisations. When planning digital preservation it is therefore essential to clarify whether a national preservation service would provide value for researchers, and which datasets should be stored there. Based on that, the required support for file formats and metadata can be planned. The selected file formats need to be appropriately documented.

Particle and nuclear physics file formats cannot be found on the lists of the six surveyed foreign preservation organisations. The datasets are typically stored by organisations specialising in research in the field.

6.4. Databases

Research datasets are increasingly being stored in databases, from which parts of the dataset can be searched, selected and downloaded more flexibly than using traditional files. Particularly large international datasets take advantage of databases. Either the whole dataset can be stored inside the base or the database can act as an index, helping to search and select files containing the actual data. A combination of these two approaches is also possible. Data stored in a database can be retrieved through an API or user interface in various file formats, which can be changed or adapted easily if necessary.

The complexity of the database structure has a large influence on how demanding the preservation is. Size does not necessarily tell much: a large database may have a simple structure or a small base may include many different tables, objects and relations between them. It should also be noted that databases may host many kinds of content, including binary objects. When evaluating the requirements for preservation it is necessary to consider not only the database but also all the included data types.

Automatic validation tools are particularly important in preservation of databases. Databases cannot be opened in a program and observed manually like text or image files. They also cannot be preserved directly in the format they are stored in while being used. The content needs to be exported from the database server into a separate preservation format. Visualisation tools are available for many databases, making it possible to browse information and see the structure, but ensuring the completeness of the information and the correctness of the preservation format must be based on validation tools.

Reusing databases that have been downloaded from the digital preservation service presents its own challenge. The dataset needs to be transferred from the preservation format again to a database server in order to take advantage of its versatile search and selection features. Installing the server software is difficult for the end user. A user interface may have been developed on top of the server and it may be complicated to get running. User interfaces are from the preservation point of view comparable to software source code and binaries, which were described in the previous chapter.

Relational Databases and the SIARD Format

The most widely used database type is the relational database, implementations of which are readily available from several different manufacturers. Popular database software solutions include IBM DB2, Microsoft SQL Server, MySQL, Oracle and PostgreSQL. They are in principle based on the SQL standard, but each manufacturer has its own extensions to and deviations from the standard. In particular the programming functionalities of the databases are mostly manufacturer-specific and not compatible with each other.

The content of all relational databases can be backed up into a so-called dump file using tools provided by the software manufacturer. From the dump file the stored information can be restored into a new, empty database. Restoring the information to a new version of the database software from the same manufacturer is in most cases possible, but there are no guarantees of compatibility, especially in the long term. This makes databases and their dump files challenging to preserve.

In order to preserve relational databases, the Swiss Federal Archives started to develop the SIARD format at the beginning of the 21st century [SIARD_2004]. The objective was to preserve essential information content based on the SQL standard, independently of the manufacturer-specific solutions and extensions. The SIARD format also includes fields for descriptive and technical metadata to ensure the preservation of understandability.

SIARD version 1.0 was approved in Switzerland as a national standard in 2013. Meanwhile, the Danish National Archives had already adopted the SIARDDK format, which differs slightly from the original SIARD, and the Portuguese national archive had developed a similar format called DBML. Based on experiences of these three formats, SIARD 2.0 was developed and seems to be establishing itself as the preservation format for relational databases. SIARD 2.0 was

approved as a national standard in Switzerland in June 2016 [SIARD_Standard] and it is a recommended format also in CINES in France and DANS in Denmark.

SIARD 2.0 supports all data types and constraints defined in the SQL:2008 standard. It preserves the relations between database tables, which would disappear if the tables would be stored separately in files, e.g. in CSV format. Manufacturer-specific features of different relational databases such as programming functionalities are not supported. Many datasets do not use such features or they are not critical for the preservation of the content. This needs however to be checked on a case by case basis before transferring databases into preservation.

The SIARD format is implemented as an XML file and it uses the Unicode character set, usually UTF-8. It can however contain binary elements if binary objects (BLOBs) such as images have been stored in the database. SIARD version 2.0 supports storing the binary elements in separate files, enabling them to be handled separately in the preservation processes. The SIARD file itself without binary elements could already be preserved as an XML file using the current NDL specifications, but it is better to define dedicated support for SIARD files.

SIARD files can be produced using the open source Database Preservation Toolkit (DBT) software. It supports the most popular relational databases, reading the structure and content from the base and storing them in the SIARD format. The resulting files can be transferred back to the same or another relational database using the same program.

Information cannot be searched and loaded from SIARD files using SQL commands like from the databases themselves; SIARD is meant purely as a preservation format. Furthermore, none of the widely used databases currently supports importing data directly from SIARD. The abovementioned DBT conversion tool is required. The software is not yet stable and user-friendly enough to be well suited for the typical end user. These factors complicate the use of SIARD, although the format itself seems to be well defined.

Converting the SMEAR Dataset into the SIARD Format

The SMEAR example dataset is a relatively large but structurally simple MySQL relational database. It consists of a few dozen tables that have a large number of columns, but the tables are either independent or their interrelations are easy to understand. There are no binary elements and MySQL programming functionalities are not used. The dataset should be relatively easy to preserve, at least without the web interface that has been developed on top of it.

A test environment was set up to shortly test converting the dataset to SIARD format. The operating system was Ubuntu Linux 14.04 LTS, database server MySQL version 5.5.50 and Java environment version 1.7.0_111 (OpenJDK IcedTea 2.6.7). The newest version 2.0.0-beta5 of the Database Preservation Toolkit was downloaded and installed in the environment. The dataset was first exported from the MySQL server into a file in SIARD 2.0 format. Then it was imported to the same MySQL server with another name, as well as to a PostgreSQL server, version 9.3.14.

Converting the SMEAR database into SIARD format succeeded without problems, and at least based on a short manual observation it seemed to include all the essential information. When importing the data back to the MySQL and PostgreSQL servers a few problems were encountered, which will be solved with the developers. The SIARD format itself will probably be suitable for preserving the SMEAR dataset and other similar datasets, when the problems in the error processing functionality of the conversion tools are fixed.

Other Databases

In addition to relational databases, there are also other types, often called NoSQL databases. They are based on some other data model than two-dimensional tables with relations between them, for example on key-value pairs or document or object storage. Like relational databases, NoSQL databases are often accessed using a query language, enabling the user to store, search and export data from the database. Unlike SQL, the languages are not yet

standardised. Well known NoSQL databases include Google BigTable, Amazon Dynamo and open source MongoDB.

Not much information is available on using other than relational databases for storing research datasets and none of the interviewed persons mentioned any NoSQL databases. Their preservation is not considered in more detail in this report. The topic should be looked into if valuable research datasets stored in NoSQL databases are encountered.

In addition to storage, databases can be used to implement web-based search engines, helping to find desired parts of large datasets. In that model, the dataset itself is stored conventionally in files and the database only facilitates the search. From the digital preservation point of view, it needs to be evaluated whether the database itself contains valuable information that should be preserved, or whether it is sufficient to preserve only the files of the dataset.

7. ACCEPTING DATASETS FOR DIGITAL PRESERVATION

The criteria and acceptance requirements presented in this chapter have been defined together in the project group developing the research information digital preservation services. They are still preliminary.

The objective is that transferring datasets into digital preservation will be easy and convenient, in order to get them broadly and quickly available for reuse. Simultaneously it needs to be ensured that the datasets are usable by other researchers and appropriately documented for preservation.

The acceptance requirements designed in the National Digital Library project were taken as a starting point and modified taking into account the special characteristics of research data.

7.1. Levels of Preservation

The digital preservation service for research datasets offers two levels of preservation:

1. Data repository: the dataset is published for reuse and its integrity is ensured
2. Long-term preservation: understandability and long-term availability are ensured.

The term long-term availability refers to the next several dozen years ahead, during which technology and research practices will change.

A dataset that fulfils the requirements can be transferred directly into long-term preservation, which includes all the functionalities of the data repository. Alternatively, the dataset can first be published in the data repository service and it can be later decided whether or not it will be transferred into long-term preservation.

When the dataset is accepted for preservation, it will receive a permanent identifier in the digital preservation service.

7.2. Requirements for Accepting a Dataset for Preservation

Most of the acceptance requirements are identical in the data repository and in long-term preservation.

To make it easier to transfer datasets into the preservation service, the file format requirements are more permissive in the data repository. However, the dataset and all its parts must be appropriately documented in both the data repository and long-term preservation.

The requirements for accepting a dataset for preservation are listed below.

1. The dataset is usable by other researchers. (mandatory)
   - The dataset must contain all the essential information for understanding the data, including the documentation of files and research practices.
   - The dataset must be self-describing so that other researchers can independently use it. It does not need to be understandable to a layman.
2. The files belonging to the dataset and their relationships are described according to the digital preservation requirements. (mandatory)
   - The description is written as a METS document, see [NDL_Standards].
   - If preferred, the METS document can be created using the packaging service.
3. The files are in formats that have been approved as recommended formats or acceptable for transfer. (mandatory in long-term preservation, recommended in the data repository)

- If some of the files are not in preapproved formats, they will need to fulfil separate file format requirements (see next section).

4. The usage rights and conditions are stated. (mandatory)
   - This information will be given using the metadata creation tool, which includes predefined selections covering the most common cases.
5. The licence of the dataset conforms to the open science recommendations. (recommended)
   - The metadata creation tool contains a list of recommended licences. The current recommendation is Creative Commons Attribution 4.0 (CC-BY 4.0).
6. The dataset is documented according to the metadata requirements. (mandatory)
   - The metadata creation tool can be used to produce the documentation.

More detailed metadata and usage rights specifications will be written later.

7.3. File Format Requirements

These requirements apply to files being transferred into the data repository, which are not in one of the preapproved file formats (see requirement 3 above). In the case of preapproved formats, conformance with the requirements below has already been checked.

1. The file format is supported in at least one software program that is generally available. (mandatory)
   - The software may be commercial and does not need to be available free of charge. If special software is required to open the files, the name of the software and a link to its homepage must be provided.
2. The structure of the file format is documented. (recommended)
   - Files in proprietary, closed formats can be transferred into the data repository, but their understandability cannot be ensured in the long term.
   - If possible, the file should be stored in an open and documented format in parallel to the closed format.
   - Self-developed custom file formats must be documented according to the documentation requirements.
3. The file format is widely used in the field. (recommended)
4. The file format has been standardised by an independent organisation or by the scientific community. (recommended)

Files fulfilling these requirements can be transferred into the data repository without preapproval. Within the preservation service, the file format will be evaluated as part of the recommended and acceptable formats selection process, which will decide whether it will be added to the list of approved formats.

7.4. Selection Criteria of Recommended Formats

To select recommended and acceptable formats, the following evaluation criteria are used.

1. The file format fulfils all requirements of the data repository. (important)
2. The file format is supported in at least one open source program. (important)

3. The file format is widely supported in different programs. (fairly important)
4. The documentation of the file format is clear and of good quality. (important)
5. The documentation of the file format is available free of charge. (not very important)
6. The file format is upwards and downwards compatible. (not very important)
7. The file format has been selected as a recommended format in at least one well-known international data archive. (fairly important)
8. The file format is stable; new versions are published rarely. (not very important)

The criteria are based on the selection criteria adopted in the NDL project. The applicability of the NDL criteria to research data file formats is evaluated in more detail in Appendix D.

7.5. Notes about Accepting Datasets for Preservation

The METS document mentioned in the requirements can be created using the digital preservation packaging service. Alternatively, the owner of the data may create the METS file in their own computing environment and send it to the preservation service together with the dataset.

The metadata requirements are partly file format-specific. However, the requirements should be harmonised as much as possible to facilitate the combination and interdisciplinary reuse of datasets. For example, it is probably reasonable to require or at least recommend using the UTF-8 character set which has become a de facto standard in nearly every field.

The fulfilment of the requirements can be partly ensured automatically using validators. To ensure the quality of the datasets, it may be necessary to also include a manual inspection of the description of the dataset as part of the acceptance and publication process. The digital preservation service should support a process where the dataset is checked and approved by another person.

Sometimes there are several alternative file formats, and the preservation service may guide users to choose the best possible ones for preservation and reuse. For example, structured text files could be accepted into preservation, but XML-based formats are recommended and preferred. Users could be encouraged to adopt the recommended formats by offering extended support for them. Recommended formats could, for example, be automatically recognised upon reception or be selected from a list in the metadata creation tool or in the packaging service, automatically including the documentation for the formats.

For some research datasets, the preservation level provided by the data repository may be sufficient. The length of the life cycle is however difficult to estimate when the dataset is ready for publication. The decision about transferring it to long-term preservation can be done later if the dataset has proven to be popular and file formats have evolved. That can be done even several years after the dataset has been published and transferred into the data repository.

A clear decision about the duration of storage in the data preservation should be made. For example, according to the Deutsche Forschungsgemeinschaft (DFG), an organisation similar to the Academy of Finland, good scientific practice requires data to be stored securely for ten years. Appropriately documented datasets can be expected to be usable for ten years without file format conversions or other major operations, as long as their integrity is ensured. The end of the set time period may not necessarily mean that the dataset will be deleted, but unlike in long-term preservation the usability of the dataset will not be monitored nor procedures to ensure its understandability undertaken.

The organisation responsible for transferring content into preservation may be a research infrastructure. Such organisations typically manage datasets within a specific field of science across university borders and have better abilities to uniformly document them than individual researchers or universities. It is also important to collaborate with research infrastructures when choosing recommended and acceptable file formats. On the other hand,

the infrastructures often have their own storage service, whose role with respect to the digital preservation service needs to be clarified.

7.6. Acceptance Process

A draft of the process of accepting datasets into digital preservation is presented in Figure 3. Among other things, it shows the role of the data repository in comparison with long-term preservation.

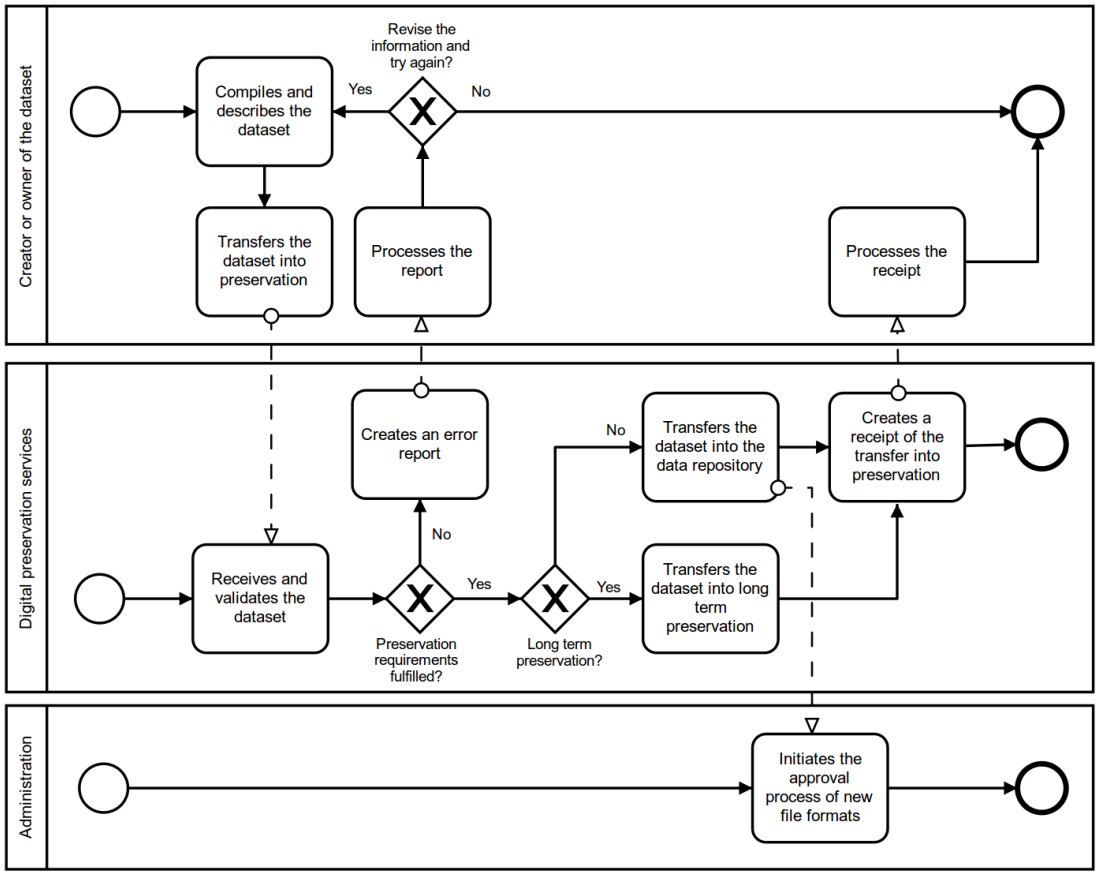

Figure 3: Draft of the acceptance process

The creators or the owners of the data compile, describe and package the dataset either in their own computing environment or using the metadata creation tool and the packaging service (not shown in the illustration) of the digital preservation ensemble. After that they transfer the dataset into the preservation service, which receives and validates it.

If the dataset fulfils the requirements and passes the validation, it will be transferred either into the data repository or into long-term preservation. The choice between the two may depend on preservation agreements or technical requirements. In the proposed model, the largest difference between the two is the preapproval of file formats, which is required only for datasets destined for long-term preservation. In the case of the data repository, the administration is notified of any new file formats present in the dataset and will initiate their approval process. The result of the approval process plays a role in the later decision on whether or not the dataset will be transferred from the data repository into long-term preservation (not shown in the illustration).

The party who has transferred the dataset into preservation is informed about the successful outcome with a receipt notification. In case the dataset does not fulfil requirements or validation fails, an error report is produced. If the reception, validation or transfer of the dataset fails due to technical problems, the administration takes action. The tasks of the administration also include user support. Those tasks are however not shown in the illustration, as the focus here is to show the normal course of the process.

It may be necessary to complement the automatic validation of datasets with human approval, such as by checking the quality of the description. Details of the process will be defined after the different actors and their responsibilities within the preservation services are clarified. In any case, both the dataset acceptance process and the preservation services in general must support delegation of tasks and responsibilities among participants. The tasks and responsibilities include approving datasets, following the development of file formats and converting outdated formats to current ones.

7.7. Readiness of the Example Datasets for Digital Preservation

This section presents an estimate of the readiness of the example datasets for digital preservation. The estimate is based on the NDL specifications and the preliminary requirements for research datasets presented in this document. The table also lists the necessary changes before the datasets could be accepted for preservation according to the preliminary requirements.

The datasets are not rated or organised in any order of preference; the table is simply an overview of what kind of work is to be expected when datasets are prepared for preservation. On average, this will probably require more work than in these example cases, as several of the example datasets had already participated in the preservation pilots [PAS_Pilots_2015]. During the pilots the datasets were already compiled with digital preservation in mind.

Dataset	Readiness of the documentation and the dataset as a whole	Readiness of the file formats
1000Gen	▪ Essential documentation for understanding the data is not included, but at least most of it could be collected from the 1000 Genomes project website. ▪ The dataset should be packaged according to the preservation requirements.	▪ The file formats are neither nationally nor internationally approved as recommended or transferable, but they are documented and widely used in the scientific field. ▪ The file formats fulfil the data repository requirements and recommendations.
BrainImg	▪ The dataset participated in the preservation pilot. It includes the essential documentation and the METS file as specified in the packaging requirements.	▪ The file formats are neither nationally nor internationally approved as recommended or transferable, but they are documented and widely used in the scientific field. ▪ The file formats fulfil the data repository requirements and recommendations.
ERNE	▪ The dataset participated in the preservation pilot. It includes the essential documentation and the METS file as specified in the packaging requirements.	▪ The data is not in a standardised or widely used file format. However, the format is documented. ▪ Documents and images are in recommended file formats. ▪ Other file formats fulfil the data repository requirements.

Dataset	Readiness of the documentation and the dataset as a whole	Readiness of the file formats
FIRE	▪ Essential documentation for understanding the dataset is included. ▪ Documentation of the parameters used during the measurement should still be improved. ▪ The dataset should be packaged according to the preservation requirements.	▪ The data file format SEG-Y is neither nationally nor internationally approved as recommended or transferable, but is documented and widely used in the scientific field. ▪ In addition to SEG-Y, most of the file formats of the dataset fulfil at least the data repository requirements. ▪ Supplementary files are mostly structured text files, which can be preserved as normal text. The structure of the files should be better documented. ▪ The final report of the experiment should be converted to PDF format for preservation.
FSD	▪ Essential documentation for understanding the dataset is included. ▪ FSD has its own homogenised method of describing and packaging the datasets. Based on that, it is easy to produce packages conforming to the preservation requirements.	▪ The file formats RTF and SPSS Portable are not yet approved as recommended or transferable in the NDL. ▪ SPSS Portable will preliminarily be approved in the near future with certain conditions. ▪ Internationally RTF is widely accepted, SPSS Portable in some organisations. ▪ Documentation and other supplementary files are all in recommended formats.
Crystals	▪ Essential documentation for understanding the dataset is included. ▪ The dataset should be packaged according to the preservation requirements.	▪ The data (results file) is in a format acceptable for transfer in the NDL. ▪ The documentation is in a recommended format.

Dataset	Readiness of the documentation and the dataset as a whole	Readiness of the file formats
MAXIV	▪ It is difficult to estimate the understandability, as the file is a Nexus HDF5 example and not a dataset of an actual research project. ▪ The dataset should be packaged according to the preservation requirements.	▪ HDF5 file format is not approved as recommended or transferable in KDK; internationally it is approved in some organisations. ▪ HDF5 format does not guarantee understandability by itself. Also the data types and metadata need to be specified. ▪ The Nexus specification limits the genericness of HDF5 and is better suited for preservation.
Planck	▪ Essential documentation for understanding the dataset could be added by a specialist by downloading the necessary files from the Planck archive (the example dataset was compiled by the author of this document). ▪ The dataset should be packaged according to the preservation requirements.	▪ The FITS data file format is neither nationally nor internationally approved as recommended or transferable, but is documented, widely used in the scientific field and maintained by an independent working group. ▪ The format fulfils the data repository requirements and recommendations. ▪ The images are in a recommended format (PNG).
RITU	▪ The dataset participated in the preservation pilot. It includes the essential documentation and the METS file as specified in the packaging requirements.	▪ The data file format has been created by the manufacturer; it is documented but neither nationally nor internationally approved as recommended or transferable. ▪ Other file formats in the dataset are either recommended or transferable.

Dataset	Readiness of the documentation and the dataset as a whole	Readiness of the file formats
SMEAR	▪ Gathering essential documentation for understanding the dataset would need some work (not all documents available from the one source). ▪ The dataset should be packaged according to the preservation requirements.	▪ Data is stored in a MySQL database, from which it can be read to a MySQL dump file or converted to SIARD format, which is better suited for preservation. ▪ MySQL dump file is documented but neither nationally nor internationally approved as recommended or transferable. ▪ SIARD files can be preserved as XML files and they have been approved as preservable in some international archives.
Suomi24	▪ Essential documentation for understanding the dataset is only partly included. However, the dataset can be mostly understood by manually observing the files.	▪ The VRT file format is a structured text file, which can be preserved as normal text. The file structure should be documented. ▪ Documentation is plain text, which is approved in the NDL as a recommended format.

Most of the example datasets could be accepted at least to the data repository with rather minor changes. Most of the file formats fulfil the data repository requirements. Improvements are needed mainly in the documentation, in particular with respect to structured text files. Approving the file formats as recommended or transferable into long-term preservation would need a more in-depth review and detailed specifications of the required technical metadata.

The descriptions of the datasets are not homogeneous or comparable with each other, as guidance for writing the descriptions and a common metadata model are missing. Evaluating the quality of the documentation is difficult without in-depth knowledge of the respective scientific fields.

The packaging according to the NDL requirements and the METS file are naturally missing from all other datasets except those that participated in the preservation pilot. However, all datasets can be packaged according to the NDL requirements.

8. CONCLUSIONS

The digital preservation specifications of the National Digital Library form a solid basis for the preservation of research datasets. Existing specifications can be extended to cover new content types and file formats. However, the special characteristics of research datasets require some significant changes in areas of responsibility, processes and technical specifications.

8.1. Conclusions about File Formats

Generally, research datasets are developing favourably from the preservation point of view. Due to increasing international collaboration and openness in science, the datasets often have more users than the original creators. This has already led to better documentation and harmonisation of file formats in several scientific fields.

Nevertheless, the variety of formats in research is larger than in cultural heritage content, and many of them are specific to certain fields of science. Evaluating file formats and selecting new recommended formats needs to be a continuous process, because formats evolve along with the development of research methods.

Unlike the recommended formats in the NDL, there are no existing metadata schemes for most of the research data file formats. Creating metadata schemes will probably require considerable resources, but the effort is paid back through the better usability of datasets as a result of the harmonisation of metadata.

It is relatively common for research datasets to include custom file formats created during the research project. Those formats must be documented before transferring them into preservation. For example, the files might be text files and therefore suitable for preservation, but their internal structure is essential for understanding the data. It is necessary to write clear documentation instructions and requirements for accepting custom file formats into preservation.

The preservation and reuse of datasets stored in databases is complicated by its workflow. To preserve the data, it needs to be exported from the database into a file, and imported back again in order to take advantage of the versatile possibilities of searching and selecting parts of the dataset. Furthermore, the file formats produced by popular database servers are manufacturer-specific. The SIARD format originally developed the Swiss Federal Archives is the best available option and seems to be establishing itself as the database preservation format of choice. However, there is as yet no easy and convenient method to offer preserved databases for reuse by end users.

8.2. Conclusions about Accepting Datasets for Preservation

The NDL model requiring preapproval of the file formats would easily lead to many datasets being left without preservation because of the slowness of the file format approval process.

Transferring datasets into preservation can be facilitated by introducing a new data repository preservation level with more permissive file format requirements. Datasets can then be received in the preservation system, and a parallel process be launched to evaluate whether the new file formats can be approved as recommended formats and to specify their metadata requirements. The datasets are more quickly secured in safe storage and can be reused, while the decision about transferring them to long-term preservation can be made later.

Unlike cultural content, understanding and using research datasets often needs in-depth expertise of the field. The datasets accepted for preservation therefore do not need to be understandable for a layman. The goal and requirements of the description and documentation should be that another researcher can understand and use the dataset.

For some datasets, the data repository preservation level may be sufficient. This allows targeting the resources of the long-term preservation to datasets that have become popular or are estimated to be particularly valuable for other reasons. However, when transferring the

dataset into the data repository, it must already be appropriately documented to be usable by other researchers. Improving the description and documentation afterwards is more difficult than technical adjustments or conversions of the file formats.

The packaging model designed in the NDL is suitable for research datasets as well. It is particularly important to focus on the user friendliness of the packaging service and the metadata creation tool. Special challenges are posed by the great differences in datasets between different fields of research, and the vast size or number of files in some datasets.

8.3. Conclusions about Actors and Responsibilities

In the NDL, the responsible entity is usually a museum, a library or an archive, which has a statutory mission to preserve content. In the field of research the situation is less clear. Research datasets are typically produced in projects that have an ending date and no long-term responsibility for preserving the data. Many datasets are collected through international cooperation and not owned by any single organisation. In any case, there is an increasing motivation to preserve and publish datasets, and this is also required by more and more research funders. There is a clear need for a research information digital preservation service.

The organisation responsible for transferring content into preservation may be a research infrastructure. Such organisations typically manage datasets within a specific field of science across university borders and have better abilities to uniformly document them than individual researchers or universities. It is also important to collaborate with research infrastructures when choosing recommended and acceptable file formats. On the other hand, the infrastructures often have their own storage service, whose role with respect to the digital preservation service needs to be clarified.

Digital preservation of research datasets is also internationally in a relatively early phase of development. Some organisations have already specified criteria for preserving research datasets and approved a number of file formats, but none of them have a comprehensive list of formats with detailed specifications. Most of the organisations focus on maintaining a data repository that does not include all long-term preservation features. The national digital preservation solution gives Finland an opportunity to be a pioneer and a desirable partner for the preservation of international datasets. International collaboration is in the case of preserving research datasets even more important than in the NDL, as datasets are increasingly produced internationally and their use is global.

9. FUTURE WORK

There is still a great deal of work to be done on many levels in planning the digital preservation of research information, from defining high-level responsibilities to developing services and specifying various technical details. Below is a list of tasks that are especially related to research data file formats and the readiness of datasets for preservation. The list is not in the order of importance.

- Metadata model and instructions for describing the datasets. It is important to design or select a common metadata model for storing basic information on all datasets. To facilitate international collaboration, an existing model, for example the CERIF model [CERIF] recommended by the EU, should be used.
- Instructions for documenting research methods. It is not possible to define strict rules or provide ready-made forms for documenting methods, but instructions and examples can be provided to facilitate the creation of good quality and understandable documentation.
- Approving file formats as recommended and acceptable formats and defining related requirements. Evaluating and approving new formats is a continuous process, because formats evolve along with the development of research methods. The work should be started with popular and established file formats, which are used in existing datasets. As in the NDL, the accepted versions, technical metadata scheme and its obligatory and optional fields need to be defined for each format.
- Documentation instructions for structured text files and custom binary file formats. The structures of the files are essential for understanding the datasets. Documentation instructions can facilitate the preparation of datasets for preservation and help to harmonise conventions, which promotes reuse.
- Developing the research data metadata creation tool, the packaging service and the validation of the file formats. Development of the services and pilot use should already start in parallel with the writing of the specifications to ensure that the services correspond to user needs. When the services are taken into use, the functionality of specifications and processes will be tested in practice.
- A systematic survey of research data file formats. In this project the topic was approached through examples, which does not as yet cover all file formats in Finnish research datasets. One method is a national survey like the one conducted in Austria [Austrian_Survey]; an alternative is to study file formats on a field-by-field basis by approaching organisations and research groups representing each scientific field.

10. REFERENCES

[1000Genomes] The 1000 Genomes Project. http://www.1000genomes.org/

[AALTO] Lahnakoski, J. M., Salmi, J., Jääskeläinen, I. P., Lampinen, J., Glerean, E., Tikka, P., & Sams, M. (2012). Stimulus-Related Independent Component and Voxel-Wise Analysis of Human Brain Activity during Free Viewing of a Feature Film. PLoS ONE, 7(4), e35215. Public Library of Science. doi:10.1371/journal.pone.0035215 http://dx.doi.org/10.1371/journal.pone.0035215

[ATT_KVKatsaus] Long-term preservation of research data, international survey document (only in Finnish): Tutkimusdatan pitkäaikaissäilytys: Kansanvälinen katsaus. Open Science and Research Initiative, 2.11.2015. https://avointiede.fi/documents/10864/12232/Tutkimusdatan+pitk%C3%A4aikaiss%C3%A4ilytys+Kansainv%C3%A4linen+katsaus+2015.pdf

[Austrian_Survey] Bauer, Bruno; Ferus, Andreas; Gorraiz, Juan; Gründhammer, Veronika; Gumpenberger, Christian; Maly, Nikolaus; Mühlegger, Johannes Michael; Preza, José Luis; Sánchez Solís, Barbara; Schmidt, Nora; Steineder, Christian (2015): Researchers and their data. Results of an Austria survey – Report 2015. Version 1.2. DOI: 10.5281/zenodo.3400. https://phaidra.univie.ac.at/o:40931

[CDF_FAQ] CDF Frequently Asked Questions list. http://cdf.gsfc.nasa.gov/html/FAQ.html

[CERIF] Common European Research Information Format (CERIF) metadata model. http://www.eurocris.org/cerif/main-features-cerif

[CINES_Formats] CINES FACILE - Service de validation de formats, version 3.4.4. https://facile.cines.fr/

[CLARIN_CMDI] CLARIN Component MetaData Infrastructure (CMDI). https://www.clarin.eu/content/component-metadata

[CoNLL-U] Description of the CoNLL-U file format. http://universaldependencies.org/format.html

[DANS_Formats] Data Archiving and Networked Services (DANS) Preferred Formats, September 2015, version 3.0. https://dans.knaw.nl/en/deposit/information-about-depositing-data/DANSpreferredformatsUK.pdf

[DBPTK] Database Preservation Toolkit software. http://www.database-preservation.com/

[DFG_Praxis] Deutsche Forschungsgesellschaft: Sicherung guter wissenschaftlicher Praxis / Safeguarding Good Scientific Practice, 25.10.2013.
Print-ISBN 978-3-527-33703-3.
DOI 10.1002/9783527679188.oth1.
http://doi.org/10.1002/9783527679188.oth1

[DXFile] Data Exchange file format.
http://dxfile.readthedocs.io/en/latest/

[ENA_SRA] European Nucleotide Archive (ENA): Sequence Read Archive (SRA) metadata model.
http://www.ebi.ac.uk/ena/submit/metadata-model

[ERNE] Description of the ERNE project and instrument.
http://www.srl.utu.fi/projects/erne/

[FIRE_Project] Finnish Reflection Experiment (FIRE) project
http://www.seismo.helsinki.fi/english/research/projects/fireEng.html

[FSD2981] National Board of Antiquities & Finnish Local Heritage Federation: Cultural Heritage of Finland 2014 [dataset]. Version 1.0 (2015-01-15). Finnish Social Science Data Archive [distributor].
https://services.fsd.uta.fi/catalogue/FSD2981

[FSD2985] Finnish Broadcasting Company (YLE) & 15/30 Research: Internet Use of Finns 2013 [dataset]. Version 1.0 (2015-06-30). Finnish Social Science Data Archive [distributor].
https://services.fsd.uta.fi/catalogue/FSD2985

[Human_Connectome] Human Connectome project home page.
http://www.humanconnectome.org/

[LAC_Formats_2010] Library and Archives Canada (LAC), Local Digital Format Registry (LDFR): File Format Guidelines for Preservation and Long-term Access, Version 1.0, 2010.
http://www.councilofnsarchives.ca/sites/default/files/LAC%20File%20Format%20Guidelines%20for%20Preservation%20and%20Long-term%20v1_2010-12_0.pdf

[LAC_Formats] Library and Archives Canada: Guidelines on File Formats for Transferring Information Resources of Enduring Value, 2015-02-05.
http://www.bac-lac.gc.ca/eng/services/government-information-resources/guidelines/Documents/file-formats-irev.pdf

[LoC_Formats] Sustainability of Digital Formats, Planning for Library of Congress Collections. http://www.digitalpreservation.gov/formats/

[LoC_Statement] Library of Congress Recommended Formats Statement 2016-2017.
https://www.loc.gov/preservation/resources/rfs/RFS%202016-2017.pdf

[MAXIV_Lab] MAX IV Laboratory. https://www.maxiv.lu.se/about-us/

[NAA_Formats] National Archives of Australia, Preservation file formats. http://naa.gov.au/Images/Preservation-File-Formats_tcm16-79398.pdf

[NDL_Formats] National Digital Library: File Formats, v. 1.4.0, 19.1.2016. http://www.kdk.fi/images/tiedostot/NDL-File-Formats-v1.4-en.pdf

[NDL_Standards] National Digital Library of Finland Standard Portfolio, v. 1.1.0, 17.5.2016. http://www.kdk.fi/images/tiedostot/NDL_standardportfolio_v1.1-en.pdf

[NetCDF] Network Common Data Form (NetCDF) file format home page. http://www.unidata.ucar.edu/software/netcdf/

[Nexus_HDF5_Talk] Presentation about the Nexus HD5 file format, Eugen Wintersberger, DESY, 27.05.2013. http://www.desy.de/dvsem/SS13/wintersberger_talk.pdf

[NNDC_Databases] National Nuclear Data Center Databases. http://www.nndc.bnl.gov/databases/databases.html

[ODA_DWG_Spec] Open Design Specification for .dwg files, version 5.3. http://opendesign.com/files/guestdownloads/OpenDesign_Specification_for_.dwg_files.pdf

[OGC_Standards] List of the Open Geospatial Consortium Standards. http://www.opengeospatial.org/standards

[PAS_Pilots_2015] Research data long-term preservation service pilots document (only in Finnish): Tutkimuksen PAS-palvelun pilotit 2015: loppuraportti (8.4.2016) http://avointiede.fi/documents/10864/12232/Tutkimuksen+PAS-palvelun+pilotit+2015+Loppuraportti/3c0324c5-fb8c-4318-8582-c1600683dd78

[PLA] Planck Legacy Archive. http://pla.esac.esa.int/pla/

[PRONOM] PRONOM file format registry. http://www.nationalarchives.gov.uk/PRONOM/

[RADWARE_FAQ] RadWare analysis software Frequently Asked Questions (FAQ). http://radware.phy.ornl.gov/faq.html

[RITU] RITU gas filled recoil separator. https://www.jyu.fi/fysiikka/en/research/accelerator/nucspec/RITU

[ROOT] ROOT analysis software home page. https://root.cern.ch/

[SIARD_2004] Heuscher, Jaermann, Keller-Marxer, Moehle: Providing Authentic Long-term Archival Access to Complex Relational Data. Proceedings PV-2004: Ensuring the Long-Term Preservation and Adding Value to the Scientific and Technical Data, 5-7 October 2004, ESA/ESRIN, Frascati, Italy. Report number ESA WPP-232, pp. 241-261. http://arxiv.org/abs/cs/0408054

[SIARD_Standard] SIARD Format Specification v. 2.0, Swiss National Standard eCH-0165. http://www.ech.ch/vechweb/page?p=dossier&documentNumber=eCH-0165

[SMEAR_AVAA] SMEAR dataset in the AVAA service. http://avaa.tdata.fi/web/smart/smear

[Suomi24] Suomi 24 corpus. http://urn.fi/urn:nbn:fi:lb-2015040801.

[UKDA_Formats] UK Data Archive File Formats Table. http://data-archive.ac.uk/create-manage/format/formats-table

APPENDIX A. INTERVIEWED PERSONS

The persons interviewed during the project are listed in the table below. Most of the interviews were conducted on site or as video meetings. Two of the interviewees preferred to answer the questions by email.

Organisation	Interviewed persons
Aalto University, Department of Neuroscience and Biomedical Engineering, Brain and Mind Laboratory	Postdoctoral Researcher Enrico Glerean
Aalto University, School of Chemical Technology, Bioeconomy Infrastructure	Vice Dean Sirkka-Liisa Jämsä-Jounela University Teacher Jukka Kortela
Biocenter Finland	Director Olli Jänne Planning Officer Marianna Jokila
University of Jyväskylä, Department of Physics, Accelerator Laboratory	Senior Researcher Panu Rahkila
University of Helsinki, Department of Physics, Observational Cosmology Group	Professor Hannu Kurki-Suonio Academy Research Fellow Elina Keihänen
University of Helsinki, Department of Physics, Division of Atmospheric Sciences	Principal Investigator Ari Asmi Postdoctoral Researcher Pasi Kolari
University of Helsinki, Department of Geosciences and Geography, Institute of Seismology	Research Director Pekka Heikkinen Application Designer Kari Komminaho
University of Helsinki, Department of Modern Languages	Researcher Jussi Piitulainen
Lund University (Sweden), MAX IV Laboratory	IT Strategist Krister Larsson
University of Oulu, Department of Physics, Nano and Molecular Systems Research Unit	Professor Marko Huttula
CSC - IT Center for Science Ltd	GIS Coordinator Kylli Ek Application Specialist Pekka Järveläinen Development Manager Ilkka Lappalainen
University of Turku, Department of Physics and Astronomy, Space Research Laboratory	Professor Rami Vainio
Finnish Social Science Data Archive FSD	IT Services Specialist Tuomas Alaterä Development Manager Mari Kleemola

APPENDIX B. INTERVIEW QUESTIONS

This appendix presents the questions of the interviews conducted in the project. They were sent to the interviewees in advance. If no suitable example dataset was available for analysis, the focus was on part B of the questions.

Questions, part A: Sample dataset

This part focuses on a sample dataset that the interviewee or his/her research group is or has been working on, for example in a current or recently finished research project. The dataset may consist of files in several different formats, it can be a database or a combination of both. We ask you to propose a suitable dataset that you think would be valuable for reuse in the long term.

If permitted by data protection and copyright restrictions, we would like to have a copy of the sample dataset or a small subset thereof, for example the files related to one experiment or measurement and the related descriptive information. The copy can be made during the interview for example on a USB stick. The project group working on the report will study the files and their features with long-term preservation and reuse in mind.

1. The contents of the sample dataset

Which files, file formats and/or database(s) does the sample dataset consist of?

Is the data stored in a specific directory structure or another structure that is important in order to interpret the data?

How large is the quantity of data in each format?

2. Metadata

Where is the metadata of the dataset (e.g. the structure of the files, settings of measurement devices, description of the measurements/experiments, etc.) stored?

Is everything included in the data files or is essential information partly elsewhere, such as in publications, in separate description documents or in non-written sources (such as undocumented information which only the people working on the data are aware of)?

3. Openness, documentation and standards

Are the file formats and structures open and sufficiently documented?

Is there a standard for the file formats and/or structures?

Are there standards available on your scientific field, which have been taken into account when choosing the file formats, or when producing metadata and the documentation?

4. Software

Which software do you use to process and analyse the data?

Is there any other software available that could be used to process or analyse the data?

5. Stability

When has one of the file formats of the dataset changed the last time?

How often do you estimate that the file formats generally change?

6. Compatibility

Is the version number of the format marked in the data files?

Is the most recent version upwards and/or backwards compatible with the previous version(s)?

7. Integrity

Does the dataset include checksums of files or some other mechanism in order to detect possible corruption?

How significantly would a corruption of the file (e.g. the change of a few bits) affect the interpretation of the data?

8. Reuse

Where is the data currently stored?

Is the same data used in any other research group or organisation?

Are the same file formats used in any other research group or organisation?

Which factors do you consider particularly significant, when/if another researcher or research group would use the dataset?

Do you estimate that there would be users of the dataset after 5, 10 or 50 years?

9. Other

Anything else to note about the example dataset?

Questions, part B: File formats in your scientific field

These questions focus on either a field of science (such as physics) or a subfield (e.g. material physics, nanophysics) in which the interviewed person(s) are working. The goal is to get information on the file formats, structures, and databases of the scientific field on a broader scope than in part A, which focuses on the details of a sample dataset.

1. Commonly used file formats

Which file formats are commonly used in your field?

Do you think you can name all/most of the essential file formats used in the field, or only a small part of them?

Which other sources (websites, people, other) could we use to get more information?

2. Commonly used software

Which software are commonly used in your field?

Is the software developed by commercial software manufacturers, in cooperation by several researchers and organisations in the field, or by single researchers or research groups?

3. Openness of software and file formats

Is the source code of the software programs used in your field (usually) available?

Are the interfaces (e.g. how to connect with or extend a software program or a database in order to access/process data) well documented?

Are the file formats well documented?

4. Compatibility

Are the file formats in your field uniform and/or compatible, or is heterogeneity a problem?

Have you encountered a situation where you cannot open a file or a dataset, for example because the file format is not compatible or because the file is corrupted?

5. Metadata

How is the metadata related to datasets (e.g. the structure of the files, settings of the measurement devices, description of the measurement/experiments, etc.) in your field usually stored?

Which information do you need to understand a dataset if it has been produced by another researcher or research group?

6. Databases

Are databases (commonly) used in your scientific field?

Are there search interfaces for the databases through which also others than the original creators of the database can search and access the data (e.g. through a web page on the Internet)?

Do you personally use data that is stored in databases?

7. Standards, regulations and guidelines

Are there standards related to software, file formats or metadata in your field (either official or de facto standards)?

Does your own organisation set regulations or give guidelines related to software, file formats or metadata?

Is there some other authority in your field (e.g. an international organisation) that sets regulations or issues guidelines?

8. Organisations

Which are the most important organisations in your scientific field in Finland, Europe and the world?

Do you have cooperation with or contacts in the organisations mentioned above?

9. Reuse of datasets

Are datasets in your scientific field somewhere available for reuse? If yes, under which terms of use?

APPENDIX C. ANALYSIS OF THE FILE FORMATS IN EXAMPLE DATASETS

This section contains a description of all the file formats in the example datasets from the digital preservation point of view. Attention was paid especially to the structure of the files, the quality of the documentation, standardisation, software support, support for automatic processing and human readability.

The information is based on interviews, detailed observations of the example datasets and web-based sources, in particular the file formats library of the Library of Congress [LoC_Formats]. The conformance of the files with the documentation and standards was checked only superficially, without automatic validation. Remarks about software support are based on manufacturer statements and other publicly available information; the programs were in most cases not tested. It was checked on a per format basis whether or not they are included on international lists of recommended formats [CINES_Formats] [DANS_Formats] [LAC_Formats] [LoC_Statement] [NAA_Formats] [UKDA_Formats].

BAM / SAM

Full name:	Binary Alignment/Map (BAM), Sequence Alignment/Map (SAM)
Most recent version:	Version 1 (18.11.2015) http://samtools.github.io/hts-specs/SAMv1.pdf
Openness:	Open, documented, developed and maintained by a non-commercial working group
Compatibility:	Unknown, only one version published to date
Software support:	Several different applications support the format. See for example the ELIXIR Tools and Data Services Registry, https://bio.tools/
Validation:	Validators available. They apparently do not validate all fields. Documentation inadequate. http://genome.sph.umich.edu/wiki/BamUtil:_validate http://broadinstitute.github.io/picard/command-line-overview.html#ValidateSamFile
Integrity:	A md5 checksum in header section (optional)
LoC link:	No description in the LoC format library
PRONOM:	No description in the PRONOM format library
Datasets:	1000Gen example dataset, other gene sequence datasets
Notes:	▪ BAM is a binary version of the SAM format, compressed with BGZF. Otherwise the formats are identical. ▪ One of the widely used file formats in the field (others include for example FastQ and CRAM). ▪ BAM is not directly human readable. When uncompressed e.g. using gzip, the header section is human readable. ▪ The header section has only a few obligatory fields according to the standard. In digital preservation it needs to be defined which optional fields should be filled before accepting the dataset into preservation, and the details related to those fields. ▪ No mentions on the recommended file formats lists of the surveyed foreign organisations.

BIDS

Full name:	Brain Imaging Data Structure (BIDS)
Most recent version:	1.0.0-rc2 http://bids.neuroimaging.io/bids_spec1.0.0-rc2.pdf
Openness:	Open and documented, maintained by an international working group
Compatibility:	At present, only one version available
Software support:	Not yet integrated in most software. BIDS is a directory structure and not a file format. Most users browse the structure using standard operating system tools in the same way as other directories and files.
Validation:	Validator available. https://github.com/INCF/bids-validator
Integrity:	No checksums. The validator verifies the integrity of the structure compared to the specification and warns about exceptions or missing values.
LoC link:	No description in the LoC format library
PRONOM:	No description in the PRONOM format library
Datasets:	BrainImg example dataset, datasets containing MRI images
Notes:	▪ BIDS is not a file format but a specification, which defines the directory structure, file naming conventions, file formats and metadata of research datasets containing MRI images. ▪ Widely used and accepted within the scientific field, designed to make reuse of datasets easier. ▪ A fairly concise set of obligatory files and metadata, a considerably larger set of optional ones (e.g. different parameters and other information about the imaging hardware). ▪ Suitable for both automatic processing and manual browsing. ▪ Complementary files not part of the specification may be stored in the same directory structure. ▪ The METS structure map file can probably be generated largely automatically for datasets conforming to the BIDS specification. ▪ No mentions on the recommended file formats lists of the surveyed foreign organisations.

CorelDraw (CDR)

Full name:	CorelDraw
Most recent version:	X8 / version 18 (March 2016)
Openness:	Proprietary manufacturer-specific format. Documentation not publicly available.
Compatibility:	Downwards compatible
Software support:	Only the CorelDraw software has full support of the format. Partial support in the open source LibreOffice software.

Validation:	No validators available.
Integrity:	No mechanisms to ensure integrity.
LoC link:	No description in the LoC format library
PRONOM:	Described in the PRONOM format registry, different versions separately. http://www.nationalarchives.gov.uk/pronom/fmt/430 (version X5)
Datasets:	FIRE example dataset
Notes:	▪ Proprietary, commercial vector image format, which is difficult for preservation and reuse ▪ The format is probably used in research datasets mostly to create illustrations for publications or other documents ▪ CDR format images can be transferred e.g. to PDF or SVG format without losing essential information for viewing the image (the possibility to edit the image is lost) ▪ DANS recommends opening CDR files with the Adobe Illustrator program and converting them to SVG format ▪ No other mentions on the recommended file formats lists of the surveyed foreign organisations.

CRAM

Full name:	CRAM
Most recent version:	3.0 (June 2015) http://samtools.github.io/hts-specs/CRAMv3.pdf
Openness:	Open, documented, developed and maintained by a non-commercial working group
Compatibility:	Downwards compatible with older CRAM files and the BAM format.
Software support:	Several applications support the format. However, it is not as widely supported as the BAM/SAM format.
Validation:	No validators available.
Integrity:	Checksums in use.
LoC link:	No description in the LoC format library
PRONOM:	No description in the PRONOM format library
Datasets:	1000Gen example dataset, other gene sequence datasets
Notes:	▪ File format developed from the BAM/SAM format, with the goal to support more efficient compression methods to save space, to support all BAM features and to offer an easy migration path from BAM to CRAM ▪ Used typically with lossy compression, which discards parts of the gene sequence information in a controlled fashion. ▪ Somewhat more complicated than BAM/SAM ▪ Becoming more popular, supported in many software libraries, but not yet as widely as BAM/SAM ▪ No mentions on the recommended file formats lists of the surveyed foreign organisations.

DOC / DOCX

Full name:	Microsoft Word Document (DOC), Office Open XML Document (DOCX)
Most recent version:	ISO/IEC DIS 29500 (2012)
Openness:	DOC proprietary, DOCX documented and standardised
Compatibility:	Downwards compatible
Software support:	Supported in several different applications. Fully functional support of all features only in Microsoft Word.
Validation:	No validators available.
Integrity:	No mechanisms to ensure integrity.
LoC link:	http://www.digitalpreservation.gov/formats/fdd/fdd000397.shtml
PRONOM:	http://www.nationalarchives.gov.uk/pronom/fmt/412
Datasets:	FIRE example datasets; probably widely used in many other research datasets (scientific field-independent format).
Notes:	▪ The file format used by the Microsoft Word word processing software; at least partly supported by many other programs ▪ Approved in the NDL as a format acceptable for transfer, starting from Word software version 97 (file format version 8.0). [NDL_Formats] ▪ Internationally widely approved as an acceptable format (DANS, LAC, LoC, NAA, UKDA).

FITS

Full name:	Flexible Image Transport System (FITS)
Most recent version:	3.0 (July 2008) http://fits.gsfc.nasa.gov/standard30/fits_standard30aa.pdf
Openness:	Open, well documented, maintained by an independent working group and used by the most significant organisations in the field (e.g. NASA and ESA).
Compatibility:	Downwards compatible
Software support:	Development libraries available for several different programming languages
Validation:	Validator available (FITSVerify) http://fits.gsfc.nasa.gov/fits_verify.html
Integrity:	Possibility to add a checksum in the header section. A registered convention, but not part of the FITS standard. http://fits.gsfc.nasa.gov/registry/checksum.html
LoC link:	http://www.digitalpreservation.gov/formats/fdd/fdd000317.shtml
PRONOM:	http://www.nationalarchives.gov.uk/pronom/x-fmt/383
Datasets:	Planck example dataset, other research datasets including astronomy data

Notes:

- Developed already 30 years ago, continues to be widely used in storing astronomy data
- Fairly complex structure, which allows storing many kinds of data (not only images)
- Header section structured text and human readable, the actual data binary
- There can be several header and data sections in one file
- The technical metadata required by digital preservation can be stored in the header section. Mandatory and optional fields as well as their details need to be defined. To be decided how to handle files with several header sections, and which extensions are supported.
- In storing the upcoming Euclid satellite data, a migration from FITS to HDF5 is being considered, mainly because HDF5 offers more efficient compression methods.
- No mentions on the recommended file formats lists of the surveyed foreign organisations.

GREAT

Full name:	The GREAT / TDR Data Format
Most recent version:	3.2.2 (October 2014) http://npg.dl.ac.uk/documents/edoc504/edoc504.html
Openness:	The file format is documented, but its development is not open. The manufacturer publishes new versions or revisions if necessary.
Compatibility:	Downwards compatible.
Software support:	GRAIN software developed in the Accelerator Laboratory (source code available)
Validation:	Validator developed in the laboratory.
Integrity:	No mechanisms to ensure integrity.
LoC link:	No description in the LoC format library
PRONOM:	No description in the PRONOM format library
Datasets:	RITU example dataset

Notes:

- A binary format developed by the manufacturer of the GREAT spectrometer, used in the research project.
- Documentation available from the manufacturer website
- The file format itself does not have a place for storing metadata. In preservation, it needs to be ensured that all necessary metadata and documentation to understand the dataset are included. This has been mostly already done as part of the LTP pilot in 2015.
- No mentions on the recommended file formats lists of the surveyed foreign organisations.

HDF5

Full name:	Hierarchical Data Format 5 (HDF5)
Most recent version:	HDF5 1.10, Specifications document version 2.0 https://www.hdfgroup.org/HDF5/doc/H5.format.html
Openness:	Open and documented, maintained by a non-commercial organisation (HDF Group)
Compatibility:	Mostly both down- and upwards compatible within the different versions of HDF5. Certain extensions are not compatible. Previous major revision HDF4 is completely different and incompatible with HDF5. Conversion tools HDF4->HDF5 and HDF5->HDF4 are available.
Software support:	Supported in many different software packages, many of which do not, however, support all the features of HDF5. Most of them rely on the open source C library developed by the HDF Group to read the files.
Validation:	Validator available (HDF Group)
Integrity:	Possibility to store checksums, not an obligatory feature. Tolerance of corruption is generally poor; a small change may make the whole HDF5 file unusable.
LoC link:	http://www.digitalpreservation.gov/formats/fdd/fdd000229.shtml
PRONOM:	http://www.nationalarchives.gov.uk/pronom/fmt/807
Datasets:	MAXIV and SMEAR example datasets, widely used in many other research datasets (the format is independent of the scientific field)
Notes:	▪ HDF5 is a general purpose file format that allows storing almost all kinds of data ▪ Two types of base elements can be stored: multidimensional tables and groups, to both of which attributes can be attached. Using the base elements it is possible to store images, vectors, networks and metadata, as well as to organise the objects in a tree-like structure as desired. ▪ The flipside of genericness is complexity; the standard is long and supporting all its features is demanding. Additionally, there are various extensions and additional specifications, such as for storing images. ▪ Different projects have created additional specifications on top of HDF5, describing the data types used in the dataset(s) relevant to the project. The Nexus HDF5 used in the MAXIV dataset is one such example. ▪ From the digital preservation point of view it should be noted that simply using HDF5 does not ensure understandability; it is essential to describe the used data types and metadata, as well as to validate them when they are received in the digital preservation system. ▪ Internationally approved as a recommended or transferable format in some of the surveyed organisations (CINES, DANS, LoC).

HTML

Full name:	HyperText Markup Language (HTML)
Most recent version:	HTML 5.0 (standard) / HTML 5.1 (draft) https://www.w3.org/TR/html5/ https://www.w3.org/TR/html51/
Openness:	Open, documented, maintained by a non-commercial organisation (W3C)
Compatibility:	Downwards compatible, mostly also upwards compatible
Software support:	Widely supported in different applications, in some cases only partially. Differences in the visual representation of HTML documents
Validation:	Validators available.
Integrity:	No mechanisms to ensure integrity.
LoC link:	No description in the LoC format library
PRONOM:	http://www.nationalarchives.gov.uk/pronom/fmt/471
Datasets:	MAXIV, Planck and SMEAR example datasets, other datasets that present at least some information on web pages (scientific field-independent format).
Notes:	▪ HTML is a markup language designed to create and present web pages ▪ HTML files typically include links to many other files (pictures, sound recordings, source code, etc.) that can be in any format. The linked files are needed to form an understandable ensemble. ▪ The HTML file itself may contain source code in the Javascript language, which has a separate definition ▪ In research datasets HTML files are mostly used to store documentation. ▪ HTML version 4.01 has been approved as a recommended format in the NDL. Another related approved format is the Web ARChive Format (WARC), which gathers the HTML files and linked files together [NDL_Formats]. ▪ When preserving research datasets, in many cases a good alternative is to convert HTML documentation into PDF/A format, which is also one of the recommended formats in the NDL. ▪ Internationally largely approved as a recommended or acceptable format (DANS, LoC, NAA, UKDA).

Java

Full name:	Source code file of the Java programming language
Most recent version:	Java SE 8
Openness:	Open and documented. The development of the Java language is controlled by Oracle Corporation. The community has a limited chance to participate.
Compatibility:	Downwards compatible

Software support:	Several different Java implementations
Validation:	No validators available.
Integrity:	No mechanisms to ensure integrity.
LoC link:	No description in the LoC format library
PRONOM:	http://www.nationalarchives.gov.uk/pronom/x-fmt/422
Datasets:	RITU example dataset, other datasets where the Java programming language is used (scientific field-independent format)
Notes:	▪ A text file containing source code in the Java programming language ▪ The file can be preserved as a plain text file (approved in the NDL as a recommended format), which can be understood by a user who knows the Java language ▪ To compile the source code to an executable program may require a certain Java version, which is challenging to preserve ▪ Internationally also approved to be preserved as text, no special mentions of Java support.

JPEG

Full name:	Joint Photographic Experts Group (JPEG)
Most recent version:	Version 1.02 (September 1992) https://www.w3.org/Graphics/JPEG/jfif3.pdf
Openness:	Open and documented, ISO standard
Compatibility:	Only one version
Software support:	Widely supported in different applications
Validation:	Validator available (jpeginfo) https://github.com/tjko/jpeginfo
Integrity:	No mechanisms to ensure integrity.
LoC link:	http://www.digitalpreservation.gov/formats/fdd/fdd000018.shtml
PRONOM:	http://www.nationalarchives.gov.uk/pronom/fmt/44
Datasets:	FIRE example dataset, probably also used in many other research datasets (scientific field-independent format)
Notes:	▪ An image file format using lossy compression, approved as a recommended format in the NDL [NDL_Formats] ▪ Internationally widely approved as a recommended or acceptable format (CINES, DANS, LAC, LoC, NAA, UKDA).

JSON

Full name:	JavaScript Object Notation (JSON)
Most recent version:	Version 1.0 https://tools.ietf.org/html/rfc7159
Openness:	Open, documented, IETF standard
Compatibility:	Only one version
Software support:	Widely supported, in particular in web applications
Validation:	Several validators available
Integrity:	No mechanisms to ensure integrity.
LoC link:	http://www.digitalpreservation.gov/formats/fdd/fdd000381.shtml
PRONOM:	http://www.nationalarchives.gov.uk/pronom/fmt/817
Datasets:	BrainImg and SMEAR example datasets, increasingly popular in research datasets (scientific field-independent format)
Notes:	▪ JSON is a standardised, structured text file format, which is particularly suited to exchanging information between applications and storing various kinds of metadata such as measurement parameters ▪ The format was originally developed as part of the JavaScript programming language, but is nowadays supported also in many other programming languages and libraries ▪ Both human and machine readable ▪ The JSON standard defines only the syntax. Additionally, the fields and values to be stored in the JSON file need to be specified on a case-by-case basis. The JSON Schema may be useful for presenting the specifications (http://json-schema.org/). ▪ Can be used as a basis for derived file formats, where certain mandatory fields and the syntax of their values has been specified (e.g. GeoJSON) ▪ Can be preserved at least as a text file. Mentioned separately in the formats recommended by LoC; the subset JSON-LD is included also in the DANS recommendations.

MySQL dump

Full name:	MySQL dump file
Most recent version:	5.7.15
Openness:	Open and documented. There is no separate document about the file format, but it is based on documented commands that are used to insert information in the MySQL database. The format develops in step with the development of the database server software. The development is controlled by Oracle Corporation.
Compatibility:	Downwards compatible

Software support:	The open source mysqldump software, which is part of the MySQL server package.
Validation:	No validator available
Integrity:	No mechanisms to ensure integrity.
LoC link:	No description in the LoC format library
PRONOM:	No description in the PRONOM format library
Datasets:	SMEAR example dataset
Notes:	▪ A file format designed for backing up MySQL databases ▪ Includes a short header section which is structured text ▪ The rest of the file is a list of SQL commands, which can be used to restore the tables and information of the original database into an empty database. The commands are well described in the MySQL documentation. ▪ Dump files created from an older database can at least usually be restored in a newer version (downwards compatible) ▪ The format is MySQL-specific and does not work with the databases of other manufacturers. Using the -compatible switch in the Mysqldump tool, it is possible to produce dump files that are partly compatible with other databases. However, they do not usually work directly without manual modifications. ▪ No mentions on the recommended file formats lists of the surveyed foreign organisations.

NIfTI

Full name:	Neuroimaging Informatics Technology Initiative (NIfTI)
Most recent version:	NIfTI 1.1 (2007) http://nifti.nimh.nih.gov/nifti-1 NIfTI 2.0: https://www.nitrc.org/docman/view.php/26/1302/Approved%20NIfTI-2%20Format%20document
Openness:	Open and documented, maintained by an international working group
Compatibility:	NIfTI 1.1 is both down- and upwards compatible with version 1.0 NIfTI 2.0 is not compatible with version 1.1
Software support:	Several software packages support the format. Both closed source and proprietary software available.
Validation:	Validator available. https://github.com/INCF/bids-validator
Integrity:	No mechanisms to ensure integrity
LoC link:	No description in the LoC format library
PRONOM:	No description in the PRONOM format library
Datasets:	BrainImg example dataset, datasets containing MRI images

Notes:	▪ A file format developed for storing MRI images, in particular series of images produced by brain MRI scans. Typically created by converting the DICOM files of the MRI scanner into NIfTI format using automatic conversion software. ▪ BIDS directory structure specification requires the use of NIfTI format (either version 1.0/1.1 or 2.0) ▪ Machine readable format, not human readable ▪ Only a few fields in the header section which are marked obligatory in the standard. It needs to be specified which of the optional fields are required in files accepted for digital preservation, and to define the details of their contents. ▪ No mentions on the international recommended file formats lists.

PDF

Full name:	Portable Document Format (PDF)
Most recent version:	PDF 1.7 (July 2008) https://wwwimages2.adobe.com/content/dam/Adobe/en/devnet/pdf/pdfs/PDF32000_2008.pdf PDF/A-3 (October 2012) http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=57229
Openness:	Open and documented, development mainly controlled by Adobe, Inc.
Compatibility:	Downwards compatible
Software support:	Widely supported in different applications
Validation:	Validators available
Integrity:	No mechanisms to ensure integrity
LoC link:	http://www.digitalpreservation.gov/formats/fdd/fdd000277.shtml (PDF 1.7) http://www.digitalpreservation.gov/formats/fdd/fdd000360.shtml (PDF/A-3)
PRONOM:	http://www.nationalarchives.gov.uk/pronom/fmt/276 (PDF 1.7) http://www.nationalarchives.gov.uk/pronom/fmt/479 (PDF/A-3a)
Datasets:	BrainImg, ERNE, FIRE, FSD, MAXIV, Planck, and RITU example datasets, widely used in many other research datasets (scientific field-independent format)
Notes:	▪ In addition to the general PDF format there is a variant called PDF/A that has been specially developed for digital preservation. PDF/A does not include all features of the general PDF. PDF/A-2 and PDF/A-3 are based on PDF version 1.7. ▪ PDF/A versions 1 and 2 are approved in the NDL as recommended formats and PDF versions 1.2-1.7 as acceptable for transfer [NDL_Formats]. Therefore, most of the PDF files in research datasets are probably transferable into preservation without changes.

- The main change in PDF/A version 3 compared to version 2 is support for embedded files. However, the standard does not define the preservability of the embedded files.
- Both PDF/A-2 and PDF/A-3 standards have subversions a, b and u, which set different requirements for the structure of the document. All three subversions of PDF/A-2 are approved in the NDL.
- Internationally widely approved as a recommended format (CINES, DANS, LAC, LoC, NAA, UKDA); the supported versions vary.

PNG

Full name:	Portable Network Graphics (PNG)
Most recent version:	ISO/IEC 15948:2003 (November 2003), corresponds mainly to version 1.2 https://www.w3.org/TR/PNG/
Openness:	Open, documented, ISO standard (ISO/IEC 15948:2003)
Compatibility:	At least downwards compatible
Software support:	Widely supported in different applications and development libraries
Validation:	Validators available, for example pngcheck http://www.libpng.org/pub/png/apps/pngcheck.html
Integrity:	CRC-32 checksums in use
LoC link:	http://www.digitalpreservation.gov/formats/fdd/fdd000153.shtml
PRONOM:	http://www.nationalarchives.gov.uk/pronom/fmt/13
Datasets:	ERNE example dataset, probably widely used in many other research datasets (scientific field-independent format)
Notes:	▪ An image file format using lossless compression ▪ Approved in the NDL as a recommended format ▪ Internationally widely approved as a recommended format (CINES, DANS, LAC, LoC, NAA)

RTF

Full name:	Rich Text Format (RTF)
Most recent version:	1.9.1 (March 2008)
Openness:	Open and documented, development controlled by Microsoft Corporation
Compatibility:	Downwards compatible
Software support:	Fairly widely supported in word processing software
Validation:	No validators available
Integrity:	No mechanisms to ensure integrity

LoC link:	No description in the LoC format library
PRONOM:	http://www.nationalarchives.gov.uk/pronom/fmt/355
Datasets:	FSD example dataset, probably also used in other research datasets, particularly older ones.
Notes:	▪ For a long time, it was a fairly widely used file format for storing text including formatting and images ▪ Machine readable, partly also human readable ▪ In principle supported in many applications and well suited for exchanging information. In practice there are often small compatibility issues, and the use of the format is in decline. Therefore, FSD is gradually migrating away from the format. ▪ Not approved in the NDL as a recommended or acceptable format ▪ Internationally widely approved as a recommended or acceptable format (DANS, LoC, NAA, UKDA).

SEG-Y

Full name:	SEG Y rev 1 Data Exchange format
Most recent version:	1.0 (May 2002)
Openness:	Open and documented, maintained by an international working group
Compatibility:	Downwards compatible with version rev 0
Software support:	Widely supported in seismology software packages
Validation:	No validators available
Integrity:	No mechanisms to ensure integrity
LoC link:	No description in the LoC format library
PRONOM:	http://www.nationalarchives.gov.uk/pronom/fmt/363
Datasets:	FIRE example dataset, other seismology datasets
Notes:	▪ A binary format used in the field of seismology since the 1970s ▪ The format includes an optional structured text header, which was not used in the FIRE example dataset ▪ Includes a binary header section; the standard defines a large number of fields whose data can be stored in it (most of them optional) ▪ In digital preservation it needs to be specified which optional fields are required in files received into preservation, and the details of the content of the fields. A validator would be useful, as it is impossible to verify the correctness of the files by manual inspection ▪ No mentions on the recommended file formats lists of the surveyed foreign organisations.

SIARD

Full name:	Software Independent Archiving of Relational Databases
Most recent version:	2.0
Openness:	Open and documented, a standard maintained by the Swiss government
Compatibility:	Downwards compatible with version 1.0
Software support:	The Database Preservation Toolkit [DBPTK] developed by the projects that have also developed the file format. Not yet supported in other relational database software.
Validation:	Validator available (http://coptr.digipres.org/KOST-Val). Apparently does not yet support validating version 2.0 of the format.
Integrity:	No mechanisms to ensure integrity
LoC link:	http://www.digitalpreservation.gov/formats/fdd/fdd000426.shtml
PRONOM:	http://www.nationalarchives.gov.uk/pronom/fmt/161 (version 1.0)
Datasets:	SMEAR example dataset (converted from the original MySQL format)
Notes:	▪ An XML-based file format specially developed for preserving relational databases ▪ Developed originally in the Swiss Federal Archives. Nowadays also some other European organisations participate in the development. ▪ Supports all the SQL:2008 standard data types and essential features. Manufacturer-specific functionalities of different relational databases (in particular programming functionalities) are not supported. ▪ A SIARD file may contain binary sections, if binary objects have been stored in the database ▪ The published open source toolkit (DBPTK) supports the most popular relational databases for exporting the information from the base into SIARD format and back. The target for restoring the information can be a database of a manufacturer other than the original one. ▪ Based on short testing, it was noticed that the conversion tools still have bugs (in particular when converting from SIARD back to the databases) ▪ The SIARD format itself is well documented and seems to be establishing itself as the de facto format for preserving relational databases ▪ Approved as a recommended format in some international organisations (CINES, DANS).

SPSS Portable

Full name:	Statistical Package for the Social Sciences (SPSS) Portable file format
Most recent version:	24.0 (March 2016, version of the software) The file format has not changed in many years. Information about the last date of change not available.
Openness:	Proprietary format, documentation not publicly available. FSD has old documentation originally received from SPSS Inc.
Compatibility:	Downwards and upwards compatible
Software support:	Supported in most commercial statistical analysis packages at least partly; compatibility problems may occur
Validation:	No validators available
Integrity:	No mechanisms to ensure integrity
LoC link:	No description in the LoC format library
PRONOM:	No description in the PRONOM format registry
Datasets:	FSD example dataset, other social sciences datasets
Notes:	• Used by the commercial IBM SPSS statistical analysis software • The "Portable" in the name means portability between different computer architectures • Supported also in many other commercial statistical analysis packages; compatibility problems occur e.g. related to character sets • Challenging to preserve due to missing documentation and the lack of open source software supporting the format • In practical tests, the format has been found well downwards and upwards compatible and will preliminarily be approved in the NDL as a recommended format with certain reservations • Approved as a recommended format in some international organisations (DANS, UKDA).

TSV

Full name:	Tab Separated Values (TSV)
Most recent version:	No version information
Openness:	Open, very simple format, no standardisation. A semi-official document about the format specification is available: https://www.iana.org/assignments/media-types/text/tab-separated-values
Compatibility:	TSV files are in principle both downwards and upwards compatible, but as only the separator of the fields is defined, they may have for example character sets that are incompatible with each other.
Software support:	Supported in many applications; support easy to implement when developing new programs

Validation:	Only a very simple validation is possible due to the simplicity of the format: it is possible to detect whether a file is in TSV format and count whether each row has the same number of fields
Integrity:	No mechanisms to ensure integrity
LoC link:	No description in the LoC format library
PRONOM:	No description in the PRONOM format registry
Datasets:	BrainImg and SMEAR example datasets; probably widely used in many other research datasets (scientific field-independent format).
Notes:	▪ Easy to create and use, machine and human readable ▪ Simplicity poses a challenge for preservation, because the format leaves open characteristics that would be useful to harmonise between datasets (e.g. the used character sets) ▪ No metadata can be stored inside the file in a standardised way, so a separate file containing metadata should be defined and transferred together with the TSV data file into preservation ▪ On the other hand, datasets in TSV format can be easily converted to CSV format, which has already been approved as a recommended format in the NDL, and the metadata can be stored in ADDML format [NDL_Formats]. ▪ Approved as a recommended or accepted format in some organisations (LoC, NAA, UKDA); in others it can be preserved as text or converted to CSV format

TXT (normal)

Full name:	Plain text (TXT)
Most recent version:	No version information
Openness:	Open, no structure so also no documentation or standardisation.
Compatibility:	The files are compatible with each other if they use the same character set.
Software support:	Widely supported in different applications
Validation:	Not possible to validate. There are validators available that check the used character set, but due to technical reasons the validation is not always reliable.
Integrity:	No mechanisms to ensure integrity
LoC link:	No description in the LoC format library
PRONOM:	http://www.nationalarchives.gov.uk/pronom/x-fmt/111
Datasets:	FSD and Suomi24 example datasets; widely used in many other research datasets (scientific field-independent format).

Notes:
- A human readable file format that can be used to store simple documentation without formatting or images
- Approved in the NDL as a recommended format, provided that ISO 8859-15 or UNICODE (UTF-8, UTF-16 or UTF-32) character set is used
- Internationally widely approved as a recommended format (CINES, DANS, LAC, LoC, NAA, UKDA), typically either in UNICODE or ASCII character set

TXT (structured)

Full name:	Text file (Plain text). Can also be named otherwise, depending on the structure. The file suffix may vary.
Most recent version:	Generally no version information
Openness:	Open. The structure may be documented or undocumented.
Compatibility:	Files having different structure are not compatible with each other.
Software support:	Supported in all text editing applications for manual observation and editing. The structures are not as widely supported.
Validation:	Depends on the structure. In most cases no validator available.
Integrity:	No mechanisms to ensure integrity
LoC link:	No description in the LoC format library
PRONOM:	http://www.nationalarchives.gov.uk/pronom/x-fmt/111
Datasets:	ERNE, FIRE and RITU example datasets; popular as different variations in many other research datasets. Typically scientific field-specific structures that are not compatible with each other.

Notes:
- Text files are used in research datasets not only for unformatted text but for storing various kinds of structures
- The structures may be for example measurement parameters, key-value pairs, tables (with values separated from each other using spaces) or a combination of those.
- Human readable and editable, usually also relatively easy to process when programming. Existing routines are however typically not available in development libraries due to the variety of structures.
- If some specific structure is widely used in some scientific field, it can be seen as a separate file format (see e.g. VRT)
- Structured text files are approved in the NDL as a recommended format as normal text, provided that ISO 8859-15 or UNICODE (UTF-8, UTF-16 or UTF-32) character set is used. It is recommended to describe the structure using the ADDML metadata scheme.
- In some cases it may be wise to convert structured text files into a better machine readable format. For example, tables could be converted into CSV format and the key-value pairs into JSON format.
- Internationally widely approved as a recommended format as normal text (CINES, DANS, LAC, LoC, NAA, UKDA), typically in UNICODE or ASCII character set. No specific instructions concerning structured text.

VRT

Full name:	Verticalised Text (VRT). (The file format used by the Corpus Workbench software, not the Geospatial Data Abstraction Library (GDAL) Virtual Format, which uses the same abbreviation VRT.)
Most recent version:	Not known / no version information
Openness:	Open. Documentation inadequate.
Compatibility:	At least mostly down- and upwards compatible.
Software support:	Supported in the IMS Open Corpus Workbench (http://cwb.sourceforge.net/) and to a varying degree in software developed by linguists themselves
Validation:	No validators available
Integrity:	No mechanisms to ensure integrity
LoC link:	No description in the LoC format library
PRONOM:	No description in the PRONOM format registry
Datasets:	Suomi24 example dataset
Notes:	▪ Text file whose structure resembles XML, but differs from it in certain aspects. The structure also includes tables with fields separated by spaces. ▪ The structure itself is fairly clear and understandable, but it uses tags and abbreviations that are not documented. According to the interviewed person, a linguistics specialist can deduce their meaning. ▪ Could in principle be preserved as a text file, which is approved as a recommended format in the NDL, but both the structure and the used abbreviations should be documented to ensure understandability ▪ No mentions on the recommended file formats lists of the surveyed foreign organisations.

WMV

Full name:	Windows Media Video (WMV)
Most recent version:	WMV 9
Openness:	Version 9 of the file format is open, documented and standardised (SMPTE 421M). There is however an option to use encryption and Digital Rights Management (DRM) extensions, which are neither open nor part of the standard.
Compatibility:	Downwards and upwards compatible

Software support:	Widely supported in different applications (the standardised, non-encrypted version)
Validation:	No validators available. The integrity of the file can however be partly checked by processing it with a program that supports the format (e.g. ffmpeg, https://www.ffmpeg.org/) and seeing if that produces errors.
Integrity:	No mechanisms to ensure integrity
LoC link:	http://www.digitalpreservation.gov/formats/fdd/fdd000091.shtml
PRONOM:	http://www.nationalarchives.gov.uk/pronom/fmt/133
Datasets:	BrainImg and FIRE example datasets; probably widely used in many other research datasets (scientific field-independent format).
Notes:	▪ A video file format using lossy compression ▪ Version 9 accepted in the NDL as a recommended format ▪ Older versions or DRM extensions which do not belong to the standard should not be used in files transferred into preservation ▪ Internationally approved in some organisations (LAC, NAA).

XLSX

Full name:	Office Open XML Spreadsheet (XLSX)
Most recent version:	ISO/IEC DIS 29500 (2012)
Openness:	Documented and standardised
Compatibility:	Downwards compatible
Software support:	Supported in several different applications. Full support for all features only in Microsoft Excel.
Validation:	No validators available
Integrity:	No mechanisms to ensure integrity
LoC link:	http://www.digitalpreservation.gov/formats/fdd/fdd000398.shtml
PRONOM:	http://www.nationalarchives.gov.uk/pronom/fmt/214
Datasets:	Crystals example dataset; probably widely used also in many other example datasets (scientific field-independent format).
Notes:	▪ Format used by the Microsoft Excel spreadsheet application; at least partly supported in many other applications ▪ Approved in the NDL as an acceptable format for transfer ▪ Internationally widely approved as a recommended format (DANS, LAC, LoC, NAA, UKDA)

XML

Full name:	Extensible Markup Language (XML)
Most recent version:	XML 1.0 Fifth Edition (November 2008) – the most popular format XML 1.1 Second Edition (August 2008) – for special purposes where the new features of version 1.1 are required
Openness:	Open, documented and standardised
Compatibility:	Downwards and upwards compatible
Software support:	Widely supported in different applications
Validation:	Several validators available
Integrity:	No mechanisms to ensure integrity
LoC link:	http://www.digitalpreservation.gov/formats/fdd/fdd000075.shtml
PRONOM:	http://www.nationalarchives.gov.uk/pronom/fmt/101
Datasets:	FSD and RITU example datasets; probably widely used in many other research datasets (scientific field-independent format)
Notes:	▪ Markup language that can be used to store both documentation, metadata and data ▪ The used structure can be formally defined using XML schemes ▪ Both human and machine readable ▪ Version 1.0 approved in the NDL as a recommended format ▪ Internationally widely approved as a recommended format (CINES, DANS, LoC, NAA, UKDA).

APPENDIX D. THE NDL SELECTION CRITERIA OF RECOMMENDED FORMATS AND THEIR APPLICABILITY TO RESEARCH DATA FILE FORMATS

Starting Point

A starting point for selecting recommended formats is provided by the NDL evaluation criteria, which are presented in the file formats document, Appendix A [NDL_Formats]. They are openness/transparency, adoption as a preservation standard, stability/compatibility, dependencies/interoperability and standardisation. This section assesses to which extent the characteristics of research datasets require additions or changes in the criteria.

Openness/Transparency

Ideally the specification of the file format has been created and is distributed by a standards organisation or another international organisation with open membership. In the case of research datasets it is rather common to have an open specification, which however has been created either by a single university or an unofficial collaboration of scientists, which is not a real organisation. The specifications are almost always available without cost, but possibly from only one location. On the other hand, copying the specifications is usually allowed.

It is good to prefer standardised file formats, but the most essential approval criteria should be that the format specification is openly available. This enables taking advantage of the dataset in ways completely different from the original purpose, such as new types of analysis using programs developed by the researchers themselves. The number of locations where the specification is available is not particularly important. As the long-term availability of the specification cannot be guaranteed, in particular if it is developed by an unofficial collaboration, a copy of it should be stored together with the dataset in the digital preservation system.

Research datasets also often use file formats created by researchers or research groups themselves, or formats specific to a measurement device, which do not have a published specification. In those cases it should be required that a document describing the file format will be created before approving it as a recommended format.

Adoption as a Preservation Standard

The NDL estimate is based on how many cultural organisations are using or planning to use the format in digital preservation. This does not apply to the research datasets, because universities do not have a statutory mission to preserve content like libraries, museums and archives do, and they have therefore not evaluated the suitability of file formats for preservation.

In principle, similar criteria could be developed based on how many data archives have internationally approved the file format as a recommended format. This is however difficult in practice, as few organisations have published lists of approved formats and the lists are not comprehensive. In addition, the levels of preservation differ, for example in terms of the period of time that the organisation is committed to preserve datasets, and how much attention is paid to preserving understandability.

Stability and Compatibility

The NDL criteria for stability and compatibility are in principle suited also for research data file formats. In practice it is difficult to get reliable information about, for example, the down- and upwards compatibility of the formats. The number of versions and the age of the newest version are usually fairly easy to find out.

The resilience to corruption depends both on the file format and the analysis method. In some cases a tiny change may render the file completely useless, whereas another method is less vulnerable to data errors. In some scientific fields, new formats less resilient to corruption

have been adopted, because they compress the data better and therefore require less storage capacity. Especially in the case of large datasets, the cost savings may be significant.

Resilience to corruption as part of the file format structure is not a particularly important feature, as the digital preservation system will in any case take care of the integrity of the files after they have been transferred into preservation.

Dependencies and Interoperability

Evaluating dependencies and interoperability is based on how strongly the file format is tied to certain hardware or software. The NDL estimates are not exact numbers but terms like "high", "medium" and "low" dependence or interoperability.

The criteria are also fairly well suited for evaluating research data file formats. The support of the format in at least two different programs is a significant advantage from the dependence and interoperability point of view. It also suggests that the specification of the file format structure is adequate for transferring files between programs. On the other hand, the level of the support is hard to evaluate without a thorough study. Even if the file format is listed as supported, the program may support only a part of its features.

It is justified to take into account whether the software supporting the file format is open source. Format support implemented as open source and under a licence permitting reuse is more valuable than support in closed source software, as the openly licenced code can be used as a basis by researchers writing their own analysis tools.

Standardisation

Evaluating standardisation in the NDL is based on what kind of process is used to develop and maintain the file format. The applicability of these criteria for research data file formats is very limited. For the majority of file formats there is no defined process, but the format is used as long as it serves the research community well. When new research methods require changes, researchers often make extensions themselves. An updated version of the file format for data exchange may be created based on the suggestion of the research group that has developed the extension, for example, or based on feedback gathered in a major conference.

The evaluation might instead look into whether the file format is controlled by a commercial company, a research organisation (e.g. a university) or the research community. In the two latter cases it is more likely that the development of the format serves the needs of the international research community. As collaboration between research groups continuously increases, there is a clear trend towards using commonly agreed file formats in all scientific fields.