UNIVERSITÄTS BIBLIOTHEK

universität wien

Factors for Enabling Sharing and Reuse of Research Data

A study performed by
NOAD Austria

OpenAIRE
Open Access Infrastructure for Research in Europe

Barbara Sánchez Solís,
Version 1.0, March, 2014

Contents

Introduction

The management of digital objects and the need to ensure their future use has long been a central issue within academic institutions. In recent years, the focus has additionally shifted to the scientific data underlying publications as well as raw data. Not only is the scientific community gradually discovering the potential of data for sharing and reuse, but funding bodies increasingly demand from their funded researchers that they issue Research Data Management Plans and make research data available to the public. European initiatives like OpenAIRE[i] and Horizon 2020[ii] explicitly emphasize in their policies and objectives that open access benefits society as a whole by increasing the competitive advantage of knowledge and the significance of national research, by providing new research partnerships, and by reducing professional isolation (Open Access and Open Data Policies and Mandates 2011).

Advanced technological infrastructures in combination with a new type of awareness towards the use of data are laying the groundwork for new research behaviour and research that is more open, more collaborative and more creative. Traditionally, the researcher had to frame his hypothesis in advance, but now the hypothesis is often generated by means of exploratory data analysis. Huge quantities of data can provide unexpected results and yet reveal relations between different sciences, leading to new research methods and even to the existence of new scientific disciplines. There is evidence that the research community is gaining a new approach to negative results. Data from failed experiments can result in an important research source and can also encourage revised research with different and new methods for the future. For the classical hermeneutic disciplines, digital data are fast gaining significance. Different forms of text representation support analysis strategies and interactive forms of texts. The linking of data with other data does not only have great potential for science, but it enables Linked Open Data and the Semantic Web. In fact, we already currently find ourselves experiencing a significant deluge. The more data are used, the more precise is the outcome of measurements. For this reason, it is all the more important to acquire awareness and skills in the handling of data. Perpetual preservation actually starts with the creation of data, and due to huge volumes, it might be necessary to consider which data deserve preservation and which not. Efficient data management is needed in education and training, and should be taught at an early phase of an academic career.

This paper focuses on the processes that enable sharing and reuse of research data in the academic environment, and sheds light on the perspectives of all involved parties at all stages of a data lifecycle, including: the views of the data providers such as researchers, the views of the institutional repository and the downstream users. Data are the result of many processes like capturing, processing, transforming, integrating and analysing. Owing to this complexity, the first chapter addresses various definitions of research data and learning objects, and a vision of the ontology of the objects as well as their dimension in terms of content type and format. The second chapter will describe the different processes of research data management, starting from the conceptualisation of a project and initial measures for curation, which span from the planning of preservation processes and perpetual preservation, to its dissemination and reuse. All stages are interconnected and involve different agents: the data provider or scientist at preingest and ingest levels, the repository management at the level of ingestion, curation and preservation of data, and the downstream users at the phases of access, interpretation and visualisation. In order to enable a smooth process, it is necessary to identify the different roles and responsibilities of stakeholders and to work out their intersections and their distinctions. The paper also includes findings of personal interviews conducted with professionals from Austrian universities. They outline the actual situation of data management in the higher education sector, the significance of research data across various disciplines, and expectations and visions for future developments.

This research study is a task that the NOAD (OpenAIRE National Open Access Desk) Austria put into execution. The reason being is that throughout the projects periods (OpenAIRE and OpenAIREplus) it became evident that most academics, researchers and even repository managers are not aware of the new challenge they have to face in regard to the perpetual use of data and open access, both for already published articles and more so for research data. This study should help to deal with these new tasks and find solutions, not only for researchers but also for repository managers and institutions involved in this topic.

1. Definitions

1.1. Definitions of research data

Research data take so many forms that it is neither possible nor the intention of this paper, to simplify the term down to one common denominator. Many different definitions can be found on the websites of universities, research institutions and funding bodies. Often, a definition of research data is sought by classifying the term into types, formats and objects, and this division will also be applied in the following chapters. More seldom, the term “data” is traced back to its fundamental meaning as a distinct piece of information, deriving from the Latin term *datum*, “a single piece of information”, as well as “a thing given”. From this perspective, data may be viewed as the lowest level of abstraction from which information and knowledge are derived (EDINA and Data Library, n.d.). Some people refer to data as anything digital which is the product of research, whereas others refer to data as any material that is collected, observed or created either to produce original research results or to validate original research results. In this respect, the term "research data" means data on which an argument or hypothesis is based. Data may be raw or processed and may be stored and shared in any format or media. Digital research data can be regarded as data created in a digital form (born digital) or converted to a digital form (digitised).

The definition of research data may have a direct impact on its preservation. It is not always clear where to draw the line: Should completed questionnaires and recordings be treated as "primary materials" and the transcripts derived from them viewed as research data that should be kept? Maybe future researchers regard the primary material as necessary for the validation of research findings. There are many questions data producers have to ask themselves at the beginning of a project, and some of these will be outlined in the second part of this paper. Another issue to be considered is that research data can be situational, meaning digital material can represent research data for some individuals and institutions, but not for others. Moreover, time always affects the significance of a piece of information, which is then always subject to change. For example, a postcard of a glacier in the 1930s may not yet be regarded as a relevant piece of scientific information, but if you collect postcards with the same prospect from following decades, it might become a useful resource for geologists in order to track glacial melting. The same or similar approach can be expressed with other topics, for example food recipes that would not count as research data at first glance might provide useful insights into cultural developments.

Data can be created by researchers for one purpose and used by another group of researchers at a later date for a completely different objective as well as for research-based teaching. Therefore, it is crucial to understand the difference between data versus information. Some data may actually represent simple facts. When said data are processed, however, that is, when they are organized or presented in a given context, they then become information. Therefore, it is of crucial importance to have a closer look at the processes organized behind the creation of data. This means that it might not be sufficient to have (raw) data. Only when these data are interpreted and processed to determine their true meaning do they become useful and can then be called "information". According to the Boston University Library website for Research Data Management, data is the computer's language, and information is our translation of this language into more codes (*What Is "Research Data"?*, n.d.). Often, a research dataset can be classified as either "static" (finalised data, which is no longer in the process of change) or "dynamic" (still in development or still undergoing some process of change). The difference between the two becomes particularly important if the data is first shared, then published and cited. Citation in the established sense can be achieved when data is no longer undergoing development (*What counts as research data?*, n.d.).

1.1.1. Classification of research data

Research data can be generated for different purposes and through different processes, and can be divided into different categories. One classification frequently seen is the following (*What counts as research data?*, n.d.):

- Observational: data captured in real time, usually unique and irreplaceable. For example, sensor data, survey data, sample data, and neurological images.
- Experimental: data from lab equipment, often reproducible, but can be expensive. For example, gene sequences, chromatograms, and toroid magnetic field data.
- Simulation: data generated from test models where model and metadata may be more important than output data from the model. For example, economic and climate models.
- Derived or compiled: resulting from processing or combining "raw" data, often reproducible but expensive. For example, text and data mining, compiled database, and 3D models.
- Reference or canonical: a (static or organic) conglomeration or collection of smaller (peer-reviewed) datasets, most probably published and curated. For example, gene sequence databanks, chemical structures, and spatial data portals.

1.1.2. Types and formats of the representation of research data

In a research data process, two phases become important: the creation phase and the transformation of data into formats, comprehensible by humans or machines. Many strategies and techniques have to work together to maintain digital information that is accessible and usable over the long term. Preserving and reusing a digital object means that the object needs to be found, identified and obtained.

The main goal of preservation is to maintain the **authenticity** of a digital object, which is more specifically broken into:

- Integrity (completeness)
- Identity (consistency: content information, contextual information and provenance)

The **properties** of a digital object that must be maintained throughout the data lifecycle are:

- Metadata
- Authorship
- Relations
- Licence
- *Persistent identifier (PID), if available*

In order for data to be shared, they must be properly curated and archived. The DANS-developed Data Seal of Approval (DSA) has set up the following quality criteria for researchers and institutions that create and use digital files (Data Archiving and Networked Services 2010, p. 2):

- The research data can be found on the internet
- The research data are accessible, while taking into account ruling legislation with regard to personal information and intellectual property of the data
- The research data are available in a usable data format
- The research data are reliable
- The research data can be cited

The subsequent chapters provide a rough overview of different types and formats of research data. Though issues for preservation are implied, this is a topic of further inquiry. The formats below are not the formats that are the most suitable for archiving and perpetual preservation, but rather the formats in which data are commonly stored by data creators. Web-based data is also excluded since this topic would then extend beyond the scope of this study. On the whole, a format specification describes the stream of bits, meaning it is the description of how data need to be stored and interpreted in the future in order to preserve the entire document. In the rough sense, there exist proprietary and open formats. With proprietary formats, the specifications are not sufficiently known, whereas the specifications of open formats are usually accessible and well documented. The specifications include instructions for the semantic interpretation, so even if a reading programme no longer exists, the information might still be extracted (Funk 2009). The characterisation and validation of formats is of course a decisive issue for the repository management.

1.1.2.1. Formats with textual content

Examples for formats, many of these proprietary are: plain text files, MS Word, Portable Document Format (PDF), Rich Text Format (RTF), SVG, TEI, LaTeX, Hyper-Text Markup Language (HTML), and Extensible Markup Language (XML).

Document and text files are probably the most common file types. Usually, research projects end in some kind of final report in the form of a text document. Text documents range from simple reports and papers to full books. These files consist of structured text but often include other elements like images, figures and tabular data. Most text documents are created in word processing packages such as Microsoft Word and OpenOffice. More recent packages have shown a move towards human-readable XML-based formats and standards. This is an attempt to standardise these formats and to allow different software packages to read non-native formats (Archaeology Data Service 2009). In their final versions, many documents are stored in a common interchange format, mostly Adobe's Portable Document Format (PDF), which allows the format and structure of a document to remain consistent across a variety of platforms. There has also been a move towards a standard (PDF/A). This means that the file is not secured, includes no javascript, video or audio content or compression and that all fonts and images are correctly embedded and that the file is correctly "tagged" (Archaeology Data Service 2009). Beyond common word processing formats and PDF files, text documents may also exist in a range of plain text or marked up formats such as SGML, HTML and XML.

The concerns of repositories include the fast changing and enhancing formats used by popular word processing programmes as well as the incompatibility between older versions of a file and the current version of the software. In terms of file formats for long-term preservation, there is a clear preference to store and preserve documents in one of the now popular standardised XML formats such as Microsoft's OOXML (.docx) or OpenOffice ODF (.odt). These are recognised, international, open standards. Both formats are also similar in that they utilise a zipped archive format to contain the separate components that make up each file. In addition to these XML-based formats, PDF/A should also be considered as a potential preservation format, though primarily in cases where the original document only exists in a PDF format. Even though it is a binary format, PDF/A is an open standard with a freely available reader and growing third-party support (Archaeology Data Service 2009).

1.1.2.2. Tabular/structured

Examples for formats, many of these proprietary: SQL database, MS Excel, CSV, SPSS, Stata, SAS, Office Open XML (OOXML) and OpenDocument Format (ODF).

Tabular formats have different functions but generally they are used to collect and store data in a similar way: data are defined in terms of records/rows and fields/columns. Many database and spreadsheet applications allow users to embed other media (especially images) within files. Spreadsheet applications such as Microsoft Excel and OpenOffice Calc allow users to embed graphs and charts generated from data, along with other images. In terms of future reuse, data creators should consider that tables or sheets may not always stay packaged together within a single file. From a preservation perspective, the key significant properties are the data values themselves and the structure (tables or sheets) in which data are held. For this reason, the use of styles and formatting (for example, font colour, cell colour, border styles and so on) that are normally used to convey meaning and highlight certain data should be avoided. These can be lost when the data is migrated. Furthermore, it is important to use controlled vocabularies where possible and adhere to consistent naming conventions. Like text programmes, spreadsheets have also seen a move towards support for common XML-based file formats and many applications such as

Microsoft Office, OpenOffice and WordPerfect Office, and support both Office Open XML (OOXML) and OpenDocument Format (ODF) file formats (Archaeology Data Service 2009).

1.1.2.3. Multimedia

Examples for formats, many of these proprietary: JPEG, TIFF, PNG, GIF, SVG, Dicom, MPEG, DivX, Quicktime, Flash Video, Bitmap, WAV, AIFF, FLAC, OGG, Geospatial vector and raster data.

Images can be created from original data capture such as digital photographs, scans and drawings for geophysical survey data and images from GIS layouts. Raster images “essentially consist of a matrix of pixels with a fixed size and resolution” (Archaeology Data Service 2009). In terms of preservation, it is important that an appropriate file format is used. Each file format possesses a range of individual features and capabilities, such as compression, colour depth, support for transparency and embedded metadata. Compression falls into either lossless type file formats (GIF and PNG) or lossy formats (such as JPG). A number of formats, such as TIFF and PNG, allow data to be stored without any compression. Vector images, in contrast to raster images, represent objects as geometric entities and vector objects, and not a grid of pixels. This can include lines, circles, rectangles and curves, all connected by points and paths (Archaeology Data Service 2009). These objects are defined by coordinates which make them scalable without loss of quality. The most common examples are 2D images/illustrations, often derived from CAD and GIS datasets.

Digital video is becoming popular as a tool to accompany, document and supplement other data or as a tool to record events related to investigation such as surveys, procedures, conferences and interviews. This also shows the ease and availability of video editing and dissemination applications, widely used for platforms on the web, such as YouTube and Facebook. Digital video can also be the result of projects using data collection and analysis techniques such as 3D modelling and virtual reality in which a video “fly-through” is produced. This allows users to view a model as if they were inside it and moving through it and thus to quickly evaluate the contents of large scale 3D datasets.

Data in “raw” high quality format can be very large. When the material is edited, it must be decided what will be preserved and in which quality. As with images, it is recommended that lossless compression is used for the original video file. Lossy compression results in the loss of data. Many video formats are container formats which include separate video and audio streams. Often files allow that metadata is embedded within the container and data providers should use this functionality to record important aspects of the data creation process. Digital audio formats can also feature lossless and lossy compression or come in uncompressed formats, and data creators should be aware of how these affect data quality and size, as well as at what stages in the workflow the files will be used. Here, metadata play a key role in documenting the creation process and contents (for example, names and dates of interviews, locations, and so on), as these elements are not as obvious as in video files.

1.1.2.4. Other content types

Other content types include models, for example 3D, statistical, similitude, macroeconomic, causal, and software source code. These can comprise contents of an application (input, output, logfiles for analysis software, simulation software, schemas) as well as algorithms. Furthermore, discipline-specific content types include:

- Flexible Image Transport System (FITS), in astronomy
- Crystallographic Information File (CIF), in chemistry
- GRIdded Binary (GRIB), in meteorology

Or instrument-specific types (EDINA and Data Library, n.d.), for example:

- Olympus Confocal Microscope Data Format
- Carl Zeiss Digital Microscopic Image Format (ZVI)

This list is of course far from complete and could be extended endlessly as the progression of content types and formats and their related software is an on-going process, as well in the private sector as in open source developments. Various forms of content types represent a science that is driven by large quantities of data from a range of sources such as sensors, scanners, MRI and telescopes, but also human-generated data from social media and digital libraries. Data are processed through statistical procedures, algorithms, and other computational measures, allowing researchers to discover unrecognised patterns, leading to new insights (Puschmann 2014).

1.1.3. Research data objects

As is obvious from the previous definitions, research data are manifested in a variety of forms as are the underlying research objects. Some materials exist in born digital form, whereas others have to be digitised. These may range from digital documents and spreadsheets to laboratory notebooks, field notebooks, diaries, questionnaires, transcripts, codebooks, audiotapes, videotapes, photographs and films. Test responses, samples, methodologies and workflows may equally serve as research data objects and contents of databases (video, audio, text, images), as well as CAD and GIS files. This does not yet include web-based data such as the whole blogosphere. In the humanities, many artefacts obtained by retro-digitization serve as research objects, including books, newspapers, folders, flyers, posters, letters, handwritten correspondence, musical scores, comics and many other items.

1.1.4. Learning objects

Learning objects are effective teaching tools in an e-learning environment. Lecturers can obtain and use a variety of freely open educational resources covering many different subjects. They can also share content that they created themselves for education and higher education communities. For students, learning objects facilitates active learning and communication. Content, usually structured into a modular form, provides material for independent work, self-study and supports complementary studies in traditional as well as blended learning processes. The integration of multimedia allows the utilization of different learning styles and may promote motivation. Depositing learning objects in repositories enables their public exposure. Facilitating learning objects to be downloaded from repositories, allows students to evaluate materials, where they can then grade and make comments with the tools available in repositories (Chiappe, Segovia & Rincon 2007). Learning objects may be shared through institutional repositories but there also exist repositories solely for learning objects, like Jorum[iii] in the UK and the Swiss Virtual Campus[iv] in Switzerland, to name a few. In their essay *"Toward an instructional design model based on learning objects"*, Chiappe et al. define a learning object as:

> A digital, self-contained and reusable entity, with a clear educational purpose, with at least three internal and editable components: content, learning activities and elements of context. The learning objects must have an external structure of information to facilitate their identification, storage and retrieval: the metadata (2007, p. 675).

Types of information that may be included in a learning object, as well as its metadata are (Churchill 2007; IEEE Standard for Learning Object Metadata 2011):

- General course descriptive data, including: course identifiers, language of content, subject area, descriptive text, descriptive keywords
- Life cycle, including: version, status
- Instructional content, including: text, web pages, images, sound, video
- Glossary of terms, including: terms, definition, acronyms
- Quizzes and assessments, including: questions, answers
- Rights, including: costs, copyrights, restrictions on use
- Relationships to other courses, including prerequisite courses
- Educational level, including: grade level, age range, typical learning time, and difficulty

A working group of Phaidra Vienna has supplemented this definition by stating that a learning object is a combination of different digital objects, with the following properties: it contains relations, the objects are structured, there is a table of contents, it has an individual set of metadata, it should have a persistent identifier, it has a defined container format for a viewer, it may have a data article and it may contain information for a layout.

E-learning contents may exist as heterogeneous multimedia types, content is often dynamic and manifested in a variety of formats. The development of content can be elaborate and expensive, so the idea is to exchange individual components with other learning objects. However, the content is always closely combined with its respective system (for example, learning management system) and often this system is proprietary which may be an obstacle to exchange. Copyright issues may also represent a barrier to sharing materials. Besides, there are components that are specific to the individual, such as lists of participants, entries in forums, chats and surveys. To allow perpetual use of the digital materials, it is advisable to use open standard formats as opposed to proprietary formats, but it is difficult to implement this requirement in an environment where the use of formats like DOC and PPT are very popular. The selection of the material can be one criterion for preservation. It can be the case that not everything is relevant for long-term use along with a possible situation where the course contents change so much over the years that there is no interest in the concept of perpetual preservation. In any case, it is essential to provide descriptive and technical metadata, and persistent identifiers (Möller-Walsdorf 2009).

The experts who were interviewed and who had experiences with learning objects, show themselves to be well aware of the challenges for sharing and reuse. One reason is that e-learning technologies are widely based on proprietary standards and less on open source technology, which makes interoperability difficult. Long-term preservation and reuse may be challenging for the same reasons. Objects that were developed years ago, and would still today be suitable for teaching and learning, might have to be used in combination with old software which is usually unsatisfactory, but importing such into new software is not always possible either. The distinct didactic focus can be another difficulty. The definition of learning objectives, the development of topics in teaching sequences and their structures, may work for one purpose but not necessarily in a different didactic context. In natural and life sciences, it might be easier to combine any teaching and learning contents because they do not depend so much on context in certain cases. In the humanities, however, interviewees have observed the existence of different cultural spaces and different systems of concepts, and these can stand in the way of a standardised exchange of learning contents. Some experts suggest that it is better to invest in different strategies like working with raw data combined with improved research methods than in the development, packaging and preservation of highly complex learning objects.

2. Process management for research data

2.1. Processes from the perspective of data providers

In the course of a research project, it is crucial to handle information effectively. Data management is a general and widely used term for how to organise, structure, archive, and care for information. It covers topics including how to deal with information on a day-to-day basis over the lifetime of a project and how to deal with the data over the long term. With respect to perpetual use and open access, good data management extends over the entire lifecycle of the data, from its creation through storing to (open) dissemination. The data will continue and can be reused after the initial research project has concluded. The way data is managed depends heavily on the type of the data involved and how it is collected. Data may undergo different conversions in different stages of a research project. For many experts, data sharing and perpetual preservation already start when data is created. The following chapters address the transformation of data within a scientific process and the measurements that must be taken to ensure reuse and interpretability. Data creators know their data, but this does not necessarily mean that others understand them. Can a researcher understand data if they were created in a different discipline? How do users know if the data are accurate, and what if they include personal details? Our "information age" requires the development of principles for interoperability on a technical, semantic, legal and ethical level (Riding the wave 2010).

2.1.1. From primary data to processed data

Initial considerations for research data management (RDM) planning involve:

- Organisation of data, clear file structure, consistent file naming and version control
- Choice of format, if not predefined by computer software, the best formats are unencrypted, non-proprietary, open, and with a documented standard
- Documentation of data, involving study background, project history, aims, hypotheses, methodologies, coding or classification schemes, instruments, formats, structure, and temporal and geographical coverage
- Future representation and visualisation of data
- Metadata: structured data with a minimum standard to describe origin, purpose, time reference, geographic location, creator, access conditions and terms of use
- Quality control throughout the scientific process: completeness of data, relationships with other data, and version control
- Data Management Plans (DMPs) aim at ensuring that data is accurate, complete, identifiable, retrievable and securely stored

The beginning of a scientific process is usually marked by the conceptualisation or the design of a project. Inseparable from this process is the handling of research data: which data is created and how. In fact, data have often been collected at this stage. Therefore, the planning for data management should already be considered from the very beginning of a project. It is useful to develop procedures for consistency and data quality and if more researchers are involved in a project, conventions and standards must be communicated to the whole research team. It is also important to ask if the most appropriate software or other tools are used to store and analyse data. Today, many universities and funding bodies provide detailed Data Management Plans (DMPs) with guidelines and checklists. The DMPs always reflect the policy of the respective institution, but basically they aim at ensuring that data is accurate, complete, identifiable, retrievable and securely stored. They ensure that the long-term preservation of data is considered and that legal and ethical requirements are met. Budget planning is also a big part of DMPs, but hereafter, the focus will be mainly on the processes that reflect formats and format conversions, aspects of sharing and reuse as well as future interpretation.

As for long-term preservation and open access, many items in DMPs are applicable and these should be emphasised in more detail below. The foundation of data management is the organisation of data, normally a familiar field to academic data creators. A clear file structure and a consistent file naming convention naturally have an impact on the retrieval of data in the short term as well as in the long term. The names should be concise and informative, whereas it might be relevant to consider ordering elements within a filename that would then allow chronological sorting. It is also important to include version information. Software tools exist that can organise data files and folders in a consistent and automated way through batch renaming.

The choice of format in which data are stored has a direct impact on perpetual preservation. In some cases, the format of data files is predefined by the used computer software, but in many cases there is a choice between a variety of formats. When choosing a file format, any discipline-specific norms or technical standards must be taken into consideration and the formats that are considered best for preservation should be selected. As pointed out earlier, many formats are at risk of obsolescence. This can be reduced by using formats that are unencrypted, uncompressed, non-proprietary, and open, with documented standard and standard representation (ASCII, Unicode). Other aspects that affect reuse are those involving legal and ethical considerations. Intellectual Property Rights (IPRs, for example, copyright or patents) have an influence on the way reusing parties can use the research outputs. If these rights are not clarified, there can be consequences like limitations on the research, its dissemination, any follow-up research, and profit and credit (Research Data Management Databris, n.d.).

Documentation and metadata are the contextual information that is required to make data intelligible and interpretable and to minimise the risk of misunderstanding or misuse. Documentation makes data user-friendly and shareable. This requires clear description, annotation and contextual information. In some sciences, reproducibility of research data is a driving force for sharing. In order to prove validity, the researcher not only needs to know the methods but also whether certain information is missing or whether there are any measuring inaccuracies. The change of legal principles and requirements can also influence data, for example when methods change or even borders. In medicine, there are clear guidelines for data creation processes and these standards are widely required by journals.

Research data need to be documented at various levels (EDINA and Data Library, n.d.). On the project level, there is documentation on the study background, project history, aims, hypotheses, data collection methodologies (for example, fieldwork and interviewer instructions) and instruments (for example, questionnaires and showcards). On the file or database level, there is documentation on formats and the dataset structure, that is, the relationships between files. Finally, there is documentation on the variable or item level, for example, how a variable was generated, whether there are missing values or label descriptions. Documentation should cover fields like the description of the dataset, authorship, date of creation, purpose, details of editing. It should also describe the methodology and methods, special researchers' practices, units of measurement, definitions of jargons, acronyms, coding and classification schemes, temporal and geographic coverage and provide references to related data. Researchers can embed certain annotations in data files: variable/value labels, worksheet information, table relationship and queries in relational database, GIS data layers and tables. For example:

- SPSS: variable attributes documented in Variable View (label, code, data type, missing values)
- MS Access: variable descriptions and attributes documented in Design View; relationships
- ArcGIS: shapefiles (layers) and tables in geodatabase; metadata created in ArcCatalog
- MS Excel: base worksheet data-related documentation

Metadata, meaning standardised, structured data, become crucial when data are shared online. Metadata serve to create a bibliographic reference and help to place a dataset in a broader context, allowing users outside an institution, discipline, or research environment to understand how to interpret the data. In the context of data management, metadata explain the origin, purpose, time reference, geographic location, creator, access conditions and terms of use of data collection. Metadata for discovery portals are often structured to international standards or schemes such as Dublin Core (DC), ISO 19115 for geographic information, Data Documentation Initiative (DDI), Metadata Encoding and Transmission Standard (METS) and General International Standard Archival Description (ISAD(G)) (UK Data Archive 2011). The ways that descriptive metadata are created or captured can vary: instrument metadata are automatically included in each data file. Often enough, however, the only metadata provided are the title and short textual description that is manually completed in the web submission forms, when depositing the datasets in a repository. In this case, an XML metadata file will be created in conformance with a minimal standard as part of the data package, along with the data files.

Providing detailed and meaningful data titles, descriptions, keywords and other information enables repositories to create discovery metadata for archived data collections (UK Data Archive 2011). Metadata ultimately are also crucial for the representation of data and content. They can be extracted and used for analysis. In the interdisciplinary exchange, however, interoperability is sometimes stretched to its limits. It is not so difficult to agree on Dublin Core as a common standard, but when it comes to details, the different requirements regarding understanding and formats of the different domains become apparent.

For reuse purposes, data need to remain their authenticity. Processing data, in one way or another, means maintaining good quality. When working with digitisations, transcriptions and coding, it is important to stay focused and work carefully. To avoid mistakes, it can be useful to use standardised and consistent procedures, data checking and verifying, and define whether the procedure is automated or manual. Data should be checked regarding their completeness and their relationships with each other. Furthermore, data producers should be in command of version control and keep track of different copies and versions of data files, as well as version control table and file history within or alongside data files. For transcriptions, it might be helpful to use templates with a unique identifier, to make use of speaker tags, and to have a document header with brief details of the interview, including date, place, interviewer name, interviewee name and interview details. Disseminating the transcripts through open access, there might be considerations, including in which format the transcript will be accessed and who will be reading the transcript.

2.1.2. From shared data to open access published data

Considerations and preparations for sharing of data involve:

- Format, software, documentation and metadata, ethics and confidentiality, consents, future rights management and licensing, possibly a timescale for release
- Infrastructure for sharing
- Benefits of sharing: enhanced visibility, increased citations, validation of results, reduction of duplication of effort, acceleration of research and discovery processes, facilitation of collaborative science, transparency and openness and public

engagement with research, facilitation of long-term studies and continuity over more research generations

Legal and ethical aspects:

- Copyright does not apply to raw data and facts themselves, but to the particular way they are presented in the dataset or database
- Licenses, like Creative Commons, clarify the conditions for accessing and reuse of data
- Participants of surveys and interviews and persons appearing in audio-visual material need to be informed about the use of data
- Anonymisation, aggregation, coding and disclosure control, editing of sensitive material in interview transcripts, consent conditions and access conditions of data
- Documentation and context to minimise risk of misunderstanding and misuse
- Sharing through cloud-based file sharing services, blogs, wikis and social media platforms not suitable for confidential data or for preservation of data

When data providers consent to sharing and giving open access to data, there are a number of issues that they need to consider. The future “shareability” of research data involves format, software, (if necessary) anonymisation, documentation, considerations on ethics and confidentiality, consents, future rights management and licensing, possibly a timescale for release and an infrastructure for sharing.

Yet the benefits of sharing are numerous. First of all, enhanced visibility leads to increased citations of the data in question and of associated papers. Researchers can get credit for high-quality research as well as recognition for their contribution to the research community. Research can be extended beyond one discipline and enable collaborations on related themes and new topics. Sharing enables validation of results, reduces duplication of effort and accelerates research and discovery processes so that public research funding can be used more effectively. Research can reinforce innovation that serves public policy and the services. Sharing also makes the best use of hard-to-obtain data and can provide valuable real-life resources for education and training (UK Data Archive 2011). The perpetual use of data preserves cultural heritage, enables long-term studies and ensures continuity of research over more research generations. The open data movement originated in the scientific research community but has recently expanded to the public sector (Brown 2013). It enhances transparency and openness and public engagement with research. A new form of reusing research data is developing with the trend of crowdsourcing where the public or a defined part of the public (people with a disciplinary background) is involved in data collection and annotations to research data. In this way, non-professionals can also be included in a research process (Oßwald, Scheffel & Neuroth 2012).

In addition, open access complies with laws and regulations and promotes the adoption of emerging standards. In some domains, a shift in how research outputs are viewed is in current development: data are increasingly gaining in value, also commercially, and major journals are looking to publish datasets alongside articles. With regard to open access, however, the emphasis is on free of charge use to the downstream users. Apparent from the "Fact sheet: Open Access in Horizon 2020" issued in December 2013, the European Commission pays particular attention to open access of research data, apart from possibly related publications: "Open access is not a requirement to publish, as researchers are free to publish or not, nor does it interfere with the decision to exploit research results commercially e.g. through patenting." For the "Open Research Data Pilot" in Horizon 2020, the following requirements are framed:

> Projects participating in the Pilot will be (i) required to deposit the research data (…) preferably in a research data repository and (ii), as far as possible, take measures to enable third parties to access, mine, exploit, reproduce and disseminate this research data. At the same time, projects should provide information about tools and instruments at the disposal of the beneficiaries, and are necessary for validating the results, for instance specialised software or software code (Fact sheet: Open Access in Horizon 2020).

Sharing and open access require ethical considerations where confidential or sensitive data are concerned and legal concerns when third-party data are involved. The copyright is an intellectual property right (IPR) which is assigned automatically to the creator and prevents unauthorised copying and publishing of an original piece of work. Copyright may also apply to research data and plays a role when reusing data. The owner is usually the author or creator of a "work". If a work has two or more authors, there is a joint copyright for two or more authors. When research material is derived from existing data, whether free or purchased, joint copyright applies. Even if existing data has been purchased, it is still under copyright. Copyright is also maintained with information taken from public sources like websites and research interviews. The individual interviewees retain copyright in the words of their particular interviews. Copyrights cover original literary, dramatic, musical and artistic work, sound recordings, films and broadcast, typographical arrangements of publications as well as computer programmes and databases. Considering the secondary use of data and copyright, it might thus be necessary to obtain copyright clearance before data can be reproduced. An exception to this, however, can be the use of data for non-commercial research, private study, teaching, quotations, criticism or review, as long as author and source are cited (UK Data Archive 2011).

A basic concept of copyright is that it is not ideas or information themselves that are protected, but the form in which they are expressed. For research data this means: "raw" data and mere facts are not protected by copyright, but copyright might apply to the particular form of expression being used:

> Mere information or a random collection and listing of unrelated facts or data will not be considered to be a compilation for copyright purposes. However, a factual compilation will be a literary work if it supplies 'intelligible information'. It will be protected by copyright as an original literary work if it has been produced by the application of independent intellectual effort by the author/s, which may involve the exercise of skill, judgment, knowledge, creativity or labour in selecting, presenting or arranging the information. Copyright applies not to the facts/information itself, but to the particular way the facts/information are presented in the dataset or database. "Raw" data, for example data generated from mere measurement processes, are not bestowed with copyright, but when these data are processed, selected and arranged, in a way that intellectual work is implied, copyright is attached (Copyright and Data 2009, p. 1).

The EU established a "database right" (*sui generis* database rights - SGDRs) to protect substantial investment made by database producers in obtaining, verifying and presenting database contents. The structure of the database, including the selection and arrangement of the database's contents may include intellectual creativity so that it is under copyright (UK Data Archive 2011). When the database structure or its contents are subject to copyright, reproducing, sharing, or modifying the database will often be restricted. The new version 4.0 of the Creative Commons[v] (CC) licences covers the SGDRs in addition to copyright. That means a database creator based in the European Union can use a CC 4.0 license, allowing use of the database "relieving an EU-based user of any worries that she might be violating SGDR" (What's new in 4.0 2013). The new CC version also states clearly that text mining and data mining on licensed content can be conducted.

For a data creator and downstream user, a licence clarifies the conditions for accessing and making use of a dataset. The most well-known and widely adopted of licensing systems is the already mentioned range of CC licences. The data provider has to decide for one licence model, whereas this is then mandatory and cannot be changed any longer, but in exchange it should ensure that the creator must always be named and that rules of citation have to be followed. If researchers wish to publish their results first, they have the possibility of putting a time-limited embargo on their research data in a repository. Restrictions may also be driven by commercial or political sensitivity.

Planning ahead can reduce many difficulties for data providers as well as for reusing parties when it comes to data confidentiality and access. This might comprise anonymisation, aggregation, coding and disclosure control of data, the editing of sensitive material in interview transcripts, consent conditions and access conditions of data. When doing surveys and interviews, copyright permission should be sought before sharing and archiving. The conditions need to be discussed with the participants. They need to be informed how information and data will be used, processed, shared and disposed of before they are asked to give consent. Extra care is needed with relational datasets and geo-referenced data. A dataset in combination with publicly available information can disclose further information (UK Data Archive 2011). The method of “blanking out” should be handled with care, it might be better to use pseudonyms or replacements. On the other hand, too much anonymising should also be avoided because removing information from the text can distort data and consequently make them unusable and unreliable. As for audio-visual data, subsequent digital manipulation is expensive and time consuming, so it is always advisable to obtain consent to use and share data.

Data providers must be aware that sharing of data does not yet mean open access of data. Cloud-based file-sharing services, blogs, wikis and social network sites may be suitable for sharing certain types of data and for certain research groups. Researchers who post their data on the internet increase their visibility and chances of being quoted, and this can contribute to a “self-marketing” strategy, but the social media platforms are not suitable for data that are confidential or which need to be preserved. Apart from this, users do not have control where these data are ultimately stored and what they are ultimately being used for (UK Data Archive 2011). It is also important to think about the preservation over the long term, for example, what kind of preservation is intended in five-years’ time or in ten-years’ time? Essential again for the collaborative use and reuse is the provision of metadata and context. They are the key to collaborative science as they make data accessible and contain information necessary for the comprehension and interpretation of data, and thus provide transparency.

2.1.3. Archiving data

- Many different storing practices are in place: personal computer, external hard drive, institutional and cloud server, personal and project website, and so on
- Data transfer into repositories should be encouraged: protects data from loss, maintains intelligibility and usability, facilitates discovery and reuse
- Before depositing, data need to be prepared, that is, they must be selected, converted into suitable formats, and provided with applicable licence
- In the heterogeneous landscape of data repositories, registries provide a review

Practice shows that during but also after the termination of a project, data are stored very individually and in many different places, often on the personal computer or on an external hard drive, on institutional or cloud servers, and on personal or project websites. Concluding from expert interviews, data can often no longer be accessed after a researcher has moved on to a different institution. Another observation is that data frequently disappear from the web after some years. After the termination of a project, including larger European projects, the question of data curation often remains an unresolved issue. It happens that for different reasons, the former project consortium is no longer able or does not feel responsible for the long-term preservation of data. In this way, many valuable resources then get lost.

The upload of data onto an archive or trusted repository should be firmly encouraged. Repositories are a familiar way to facilitate open access, and repository content is increasingly harvested by Google and other search engines. Besides the transfer of data into a properly managed environment, it protects them from unwanted loss and curates them in a way that maintains understandability and usability for the scientific community. It also facilitates data discovery and reuse through the development and standardisation of metadata. In addition, a repository mediates between scientific communities and digital libraries to implement the latest developments in technology and information science (ISCPR 2013). Making use of trusted platforms and repositories is exactly what recent European initiatives encourage and it should be noted that there is currently a reasonably large amount of investment and movement into developing appropriate infrastructures.

Before data are ingested in a repository, they have to be prepared in certain ways. First, selection is recommended as it is neither always possible and nor always advisable to keep everything. For repositories, managing data in the long term is very cost-intensive. Researchers might have to consider if certain data have to be kept because they are the data underlying publications, or because they have scientific or historical value, or they cannot be recreated for a certain reason. Some data might be relevant to the funding body policy and other data might have to be disposed of for legal reasons. When working with audio-visual material, the basic raw material can be very large in size. When the material is edited, the data creator must decide what should be preserved, and in which quality format. This includes considerations whether the file is the only documentation or is just supportive of other datasets (Archaeology Data Service 2009). For images and video material, it is recommended to use lossless compression as lossy compression results in the loss of data. When data are shared, more processing may be necessary, depending on the condition of the dataset and the anticipated level of reuse. Besides, different levels of access may be provided, ranging from closed or embargoed access to open access. These conditions need to be clear for the data providers at the time of deposit (Jones 2014).

Another way of preparing data is format conversions. All digital information is interpreted by computer programs and is therefore software dependent. Data are endangered by the obsolescence of the hardware and software environment. Ideally, data are processed in open standard formats that most software is capable of interpreting and that are suitable for data interchange and transformation (UK DATA Archive 2011). When converting, care must be taken not to commit conversion errors, for example the loss of internal metadata, formatting, formulas, as well as the loss of data. Using tools to create checksums helps to assure that data are complete. When compression of data is conducted, it is recommended that lossless compression is used since lossy compression results in the loss of data.

Due to disciplinary requirements, the landscape of data repositories is very heterogeneous. They vary hugely in size and scope, and it is not always easy for researchers to select an appropriate repository either for storage or for conducting searches for research. For researchers in a number of disciplines, subject-based repositories are an important part of their research environment. These are places where they look for (new) information, where they share early findings with their peers, where they look for collaborators and where they can and do deposit their own research output (Finch et al 2012). Many universities today have an institutional repository, but the "policies of neither research funders nor universities themselves have yet had a major effect in ensuring that researchers make their publications accessible in institutional repositories as a matter of routine. Levels of depositing as of yet remain low, and for journal articles in particular, most of the records in institutional repositories tend to consist of metadata rather than full text" (2012, p. 82).

In regular workflows, researchers might use a mix of institutional and external repositories. Many repositories are subject-based or community-based, some are linked to publishers, and large international initiatives are also taking place. Some research funders directly support repositories to curate the data generated from the research they have funded. Within the growing number of repositories, one recent important addition is Zenodo[vi], a free repository with the aim of sharing all research outputs from across all fields of science in Europe. Zenodo allows researchers to create their own collections and accept or reject all uploads to it. The system is further integrated into reporting lines for research that is funded by the European Commission via OpenAIRE, meaning that with the uploading of research, the repository takes care of the reporting. For research that is not under Creative Commons, the repository allows flexible licensing. The OpenAIRE initiative supports the development of a network of repositories:

> It provides a portal for access to resources stored in these repositories, and guidance to ensure that repositories are compliant with a set of Europe-wide standards, especially relating to metadata (in order to facilitate cross-searching and harvesting). It works within the context of the EU's open access pilot in the FP7 Framework programme, and the European Research Council's Guidelines for Open Access. (Finch et al. 2012, p. 83)

A useful list of research data repositories from around the world is available at Databib[vii], which shows the subject areas supported by each repository and outlines restrictions on data access and licence agreements. Another registry is Re3data[viii] which also aims at covering research data repositories from different academic disciplines. Institutional repositories have typically been developed to store publications rather than data, but their technical infrastructure can be extended to enable the curation of research data (Jones 2014).

2.2. Technical processes from the view of repository management

The following processes focus on what happens to data upon their deposit into a repository. These processes, which widely build on technology, include preingest and ingest workflows, the management of data and preservation processes from the side of the repository, representation and visualisation of data, as well as access. For digital objects, storage is not the same as preservation, but the term preservation stands for much more. All agents involved in a data lifecycle interact with each other. The previous chapters illustrate that the quality of the content is controlled by the data providers. The quality of the data is ensured by the repository system – though it still greatly depends on the quality of the uploaded data. Subsequently, there will be a broad outline of the steps that are necessary for ingesting and preserving data, and which makes them perpetually accessible. "Ingest" here means the process of transferring data to a repository.

2.2.1. Preingest and ingest workflows

Preingest and ingest workflows comprise:

- Depending on institutions' policies, deposit agreements and metadata guidelines clarify processes and responsibilities for data providers and repositories
- The more "preservation friendly" data are submitted by the data providers, the more likely is their perpetual use
- Technical controls: formats, malware check
- Monitoring and validation: Was the data handover complete? With all documentation? Are they related to each other?
- Metadata control, format control
- Metadata extraction
- Confirmation to data provider about complete and accurate transfer
- Automatic generation of the persistent identifier ensures that data can be identified and found, even when URLs change
- Licences: different levels of access, from closed or embargoed access, through various levels of restricted access to open access

Before (selected) data are uploaded into the repository, data providers have to ensure that they fulfil certain requirements. Researchers need to know what is expected of them during the handover process. According to the institutions' policies, deposit agreements and metadata guidelines help to clarify processes and responsibilities. The repository management must clearly outline technical requests whereas format requirements and minimum standards and a declaration that long-term preservation can only be ensured for defined formats. The agreement may also give the repository rights to manipulate the data because it may be necessary to migrate them to new formats for preservation (Jones 2014).

The more "preservation friendly" data providers submit their data, the more likely is their management and perpetual use. Providers need to validate their data and revise them for completeness. Metadata need to be assigned and relationships between the individual components described, if complex datasets are to be uploaded. It is fundamental that appropriate formats are used. It is advisable to use open standard formats as opposed to proprietary formats. Licensing issues should already have been clarified at this step. Different levels of access can be provided, from closed or embargoed access, through various levels of restricted access (allocation of rights for single users, user groups, departments, faculties) to open access. To summarise, preservation-friendly objects are open, well supported, standard formats for which access tools are likely to remain available in the future, complemented with documentation about objects, formats, software and agreements about their use. For this kind of data package, the reference model Open Archival Information System (OAIS) employs the concept of SIPs (Submission Information Packages).

Throughout the whole process, a trusted repository must ascertain quality control. Quality control comprises monitoring, log files and validation. Validation is achieved through checksums, metadata control and format control. Monitoring and log files are the essential tools to control data in preingest and ingest phases. They guarantee the detection of data corruption in time. It might turn out that a dataset is not yet "ready" for ingestion for various reasons, and this is then signalled in the system. With complex models, the relation between the single files must be monitored strictly and the relationship between the individual components must be described. Having completed a transfer, the repository should acknowledge receipt to the content provider. This requires confirmation that the transfer is complete and accurate and meets all of the necessary standards. Standards and methods for transfer should cover file formats, minimum documentation and transfer methods. There is either manual transfer or automated transfer which might only be feasible where data is being regularly acquired from the same provider or system. The process can be initiated at either the provider (push) or repository (pull) end (Brown 2013).

From the view of a repository, accession is the process by which new content is brought within the control of the system, and it involves a series of workflows. In the first step, the standard package of content and metadata, the SIP, is made available, and the repository assures that no malicious software or viruses or other threats are ingested into the system.

Identifying a format of a file is the key to future representation of the content. The file extension provides an indication, but this is not always reliable or not sufficiently detailed. For example, the .DOC extension usually indicates Microsoft Word format, but does not tell which of the many versions of the format it is. Validation includes checking that the structure and content of a digital object complies with the external specification. Another related process is metadata extraction to acquire additional descriptive and technical information about an object (Brown 2013).

A repository may have the technology to automatically generate a persistent identifier for every new submission. This ensures that if the persistent identifier has been included in references to the dataset, the data can be found on the web and increases the acceptance of research data as independent, quotable research objects. References to scientific information on the web are often achieved by using URLs. However, after some years, URLs tend to suffer from link rot and users receive the message “page not found”:

> Persistent identifiers provide a technique and an organizational structure for avoiding this problem. The cause of link rot lays in the fact that URLs are meant to identify a location, whereas researchers actually want to specify the resource at that location. This works fine as long as the resource can be maintained at that specified location forever, but in practice that is not feasible. Scientists need a trustworthy way of referring to scientific output on the Internet. A solution to ensure the integrity of scientific referencing is called “persistent identifiers.” Persistent identifiers allow unique naming of a resource on the Internet, independent from its location. This enables researchers to refer to the resource itself instead of its location. (Data Archiving and Networked Services 2010, p. 43)

Metadata are crucial to all aspects of digital curation. Digital objects have to be assigned with administrative, descriptive, technical, structural and preservation metadata, using appropriate standards to ensure adequate description and control over the long term. Metadata describe data. They indicate the relationships between one digital object and other objects. They provide technical information about data, what is needed to use them (for example, format, compression, encoding algorithms, encryption and decryption keys and software) and describe what is needed to represent them. They also describe what happens to data by identifying responsibility for their preservation, and they record their history. They provide information when data were created, when they were updated, migrated and converted. A fundamental requirement of metadata management is that a link be maintained between the metadata and the data they describe, and that interoperability is ensured. Often, metadata are moved between systems. This might be required when migrating a repository from one technology platform to another, or when sharing or exchanging digital objects with another organisation (Brown 2013).

2.2.2. Preservation processes

Preservation processes involve:

- Preserving the authenticity of a digital object by maintaining integrity and identity
- Bitstream preservation and content preservation
- Facilitating usability: digital objects must be capable of being accessed and represented by the current technological environment, whereas metadata must allow the record to be found, obtained and interpreted
- A repository being able to identify threats, e.g. via technology watch and community watch
- Preservation strategies such as migration or emulation

Many strategies and techniques must be in synergy so that digital information remains accessible and perpetually usable. A digital object should retain its integrity, identity and usability. For the future interpretation of data, the fundamental objective is content preservation, e.g. the preservation of the “significant properties”. That the object must retain its integrity means that it must be complete and protected from unauthorised alteration. All authorised actions must be described by metadata. An important issue for repository management is version control. If the metadata are changed or updated by the object owner, the new metadata should again be versioned and archived. In order to preserve the identity of an object, it must be able to be placed within its original context. It must be consistent and equally important, be endowed with content information, contextual information and provenance. For its usability, the object must be capable of being accessed by authorized users, it must be able to be represented by the current technological environment, and there must be sufficient metadata to allow the record to be found, obtained and interpreted.

What is also crucial is information on formats, that is, what the original submission format was, which format should be available for future users, and what the format for preservation is. Usually, the original bitstream version is kept as well as any preservation version. The repository management has to identify threats to the continued availability and accessibility of authentic digital objects. Threats can be the loss of the data object (bitstream) or loss of context. This can be triggered by outdated technologies. Threats to integrity can be the deletion of data, the alteration of data, due to hardware, software, human, and network and replication failures. Threats to usability can be technology obsolescence, the loss of representation information and even cultural obsolescence, as also languages and writing systems are subject to change (Brown 2013). Thus, a repository also has to perform constant technology monitoring as well as community monitoring.

For digital curation, there exist various preservation strategies. One strategy is to transform the source object to a new form which is capable of being rendered. This so-called migration requires the use of software tools which are able to transform data objects from one file format to another. Any process of transformation, however, includes the potential for information loss. The new format may not support the full range of significant properties to preserve the information of the original. It can also happen that the migration process cannot transform all of the properties from the original. Another strategy is to maintain the object in its original form, and instead develop ways to access it within the current technology environment. This so-called emulation involves the use of software that recreates the functionality of an obsolete technology environment on a modern platform. However, this can be complex and expensive to create and can require the user to have detailed knowledge of how to operate older technologies (Brown 2013).

Summarising, the challenges of digital curation are the "nature" of the digital objects, representing authentic digital objects and retaining them over the long term. The Archival Information Packages (AIPs) have to be preserved and constantly monitored. What makes it more complex is that each content type needs to be represented differently and that each content type has different metadata sets. Different projects require different metadata and it must be considered how these should be represented. Further challenges include how the objects are organised in a repository.

2.2.3. Access, use and reuse, including the perspective of down-stream users

- Different levels of access allow: viewing a digital object, downloading a digital object and extracting content
- The repository system transforms digital objects into current, user-friendly formats, in response to a request for access
- For some formats, viewers are available, for others a repository can provide the necessary access environment (providing technical metadata or examples of software capable of rendering a digital object)
- The system can provide tools that do not only represent the information, but allow manipulation and reuse
- Rights management: the repository manages the content in a way that respects the rights of the content owners and enables appropriate access for users

Data must be able to be discovered and accessed before being reused. Digital objects can be discovered by applying standards that ensure that appropriate metadata are present and allow interoperability. Access means different things to different people. Some users may be content with viewing objects on the screen, while others might want to download a file, and yet others may want to extract the content in a format they can use elsewhere. Some users may expect detailed documentation and guidance, while others need only minimal assistance. The repository needs to consider which forms of access it wishes to provide for its users and must always consider both capabilities and expectations (Brown 2013). In order to find digital objects, users search through some sort of catalogue, either with full-text searches, browsing or advanced search techniques. The system enables retrieval and provides information on availability. The system further transforms the AIPs into DIPs (Dissemination Information Packages), meaning into current, common, user-friendly formats, in response to a request for access.

Access to a digital object involves that the user be given access to the bitstream and that the bitstream be rendered into meaningful information for the user, depending on the content type and a combination of technologies. The most basic form of download is to provide the user with a copy of the required bitstream. The user can access copies in formats for which viewers, such as PDF and JPG, are available. A more sophisticated download would be an “informed download” (Brown 2013, p. 243), providing the user with information about the required environment. For example, a document might be accompanied by technical metadata, which explains in which format it is, and gives examples of software capable of rendering it, perhaps also with respective links. A repository can thus provide the necessary access environment to users. This might be accomplished online, for example, by providing an embedded web-based video player, or on-site, by providing workstations which have the appropriate software installed.

An increasingly popular level of access is to provide tools that go beyond simply representing the information, but also allows the user to manipulate and reuse it (Brown 2013). Access in this sense is about what users can do with the content: to analyse and manipulate it, to shift it from one format to another, to reuse it and thus facilitate the creation of new knowledge. Use and reuse, however, depend much on the formats in which content is made available: the uses of a PDF file, for example, are considerably more limited than for content that is made available in XML. For many research objectives, it would be a key requirement, to provide data in formats that are as easy to manipulate as possible, and with as little restriction as possible, with respect to what can be done with the content (Finch et al, 2012).

One of the fundamentals of reusing data is clarity of reuse permissions, terms, and conditions. The content licensing governs what users are allowed to do with the content without asking for further permission. The repository is able to manage the content in a way that respects the rights of the object owners, and enables appropriate access for users. For the repository, the most fundamental approach is to provide a clear written statement of the terms and conditions of use. This can be displayed on, or linked from a web page which the user must visit before accessing the digital object itself. The access system can require the user to accept the terms and conditions explicitly, for example, by clicking a button on a form. Rights information may also be described in metadata, either within the object, for example, by using the technique of watermarking for applicable formats, or in a separate metadata file. Under the umbrella of digital rights management (DRM), a range of technological methods have been developed. These can limit or prevent certain actions for users, such as printing, copying, altering or even accessing the content (Brown 2013). It must be taken into account, however, that digital objects that are provided with copy protection or that use a DRM system prevent effective long-term preservation, as these objects should not be changed (Schumann 2013).

Users who simply wish to view an image are probably content with a widely supported format, but users who wish to manipulate an image, to publish or broadcast it, require the maximum possible resolution, and lossless file format. With a text-based document, for the ease of viewing and for being closest to the original layout, a PDF/A may serve best, but for the ease of editing, office formats, such as DOCX and ODT are more common. So, there might be discrepancies between format requirements and usability. For structured data, it may be necessary that the repository offers tools that allow interoperability with differently structured data, or transformation between those structures. The requirements are rather domain-specific (Brown 2013). What is offered or prescribed by the repository depends a lot on the respective policy. This can intervene considerably in the lifecycle of data and the underlying roles and responsibilities.

2.2.4. Representation and visualisation

- Visualisation of data enables data exploration and unexpected discoveries
- Data creators should ask themselves how they want to visualise their research data, for whom and which metadata they need to provide
- Applications beyond the repository allow the dynamic representation of complex file formats
- The content is provided by the repository, the visualisation is effected through a separate system (applications, e-learning platforms, and so on)
- Questions of future visualisation of research data need to be discussed
- It is necessary to work on reference models for data from different domains or disciplines and to define respective interface designs

Representation and visualisation of data can be anything in the digital world. Visualisation may simply mean viewing a PDF or web presentation. However, it may also take a more advanced form and represent elaborate “scientific dimensions”. Data exploration is a process that requires the researcher to locate data, visualise data and discover relationships. Visual data analysis enables the validation of expected results but also enables unexpected discoveries in science (Hansen et al. 2009).

Data providers should ask themselves if they want to visualise their data in a certain way: what the purpose is, for whom they want to visualise data, and how they want to achieve it. According to their intentions, they must provide appropriate metadata. They should have a clear picture of which representation of their digital objects is the target of preservation. What should the visualisation of survey data look like? What should the visualisation of qualitative data, including interviews and texts, look like? Maybe the goal goes beyond mere representation and the intention is to provide data for complex visual data analyses with new procedures, for example, by providing parallel coordinates, visual methods of data mining or geospatial discovery. It may be that data need to be prepared for a complex yet comprehensible visualisation. The visualisation of research data is dynamic, meaning they exist with the display of the content. They are also extensible and modifiable. Thus, downstream users might have completely different research purposes and different forms of visualisation in mind. They might need to generate new metadata from the available data and use different visualisation tools.

Facilitated by technological systems, semantic nets can be visualised. These can display relations and contexts that go beyond traditional research. It is possible to show connected topics, persons, places and motifs through timelines and maps. The ability to capture provenance is a key requirement for visualisation tools (Hansen et al. 2009). Geo data, annotations, and any forms of semantic documentation promote data enrichment and processing data for semantic technologies. In text-based research, digital objects can be processed with semantic retrieval systems in order to extract Linked Data, referring to names, dates and places. With these processes, which often include huge data quantities, standardized description is hardly possibly. With text-mining tools it is possible to analyse and process the information contained in text corpora and to extract relevant information, to manipulate it, and to generate new information. Text mining offers much potential to increase the effectiveness and quality of research and to unlock hidden information, but it also requires technical skills and methods to prepare the texts accordingly (Finch et al. 2012).

The representation of a complex structure, like data models, 3D models, and a collection, is enabled by applications beyond the repository. This can work with content management systems and other applications. A representation can be achieved through the repository itself but usually an access system is not developed in isolation. It will need to integrate with other systems such as catalogues and existing access systems. When the visualisation is effected through a separate system, the repository provides its content to the delivery system, but it is not directly linked to that system (Brown 2013). Separate delivery systems can be e-learning platforms, other repositories, portals for research output and research data like OpenAIRE plus as well as digital library portals like Europeana[ix].

It will be necessary to discuss questions of future visualisation of research data and work on reference models for data from different domains or disciplines, for example arts or natural sciences. Tools are needed and existing tools should be extended to support different forms of visualisation. Advances in visualisation techniques and systems must also be made to extract meaning from large and complex datasets derived from experiments and from simulation systems. It will be important to define the requirements of interface design with special consideration of disciplinary demands. It will further be necessary to work on a reference model for usability testing on each level of development and implementation, on a model for Big Data and Linked Data.

3. Roles, rights and responsibilities in the processes for sharing and reuse of research data

The perpetual use of research data involves technical and non-technical processes. Actually, the sharing of data has long been firmly established in research practice, however, the storing of data was usually conducted in local systems, with insufficient or no documentation and usually without the help of information experts. A recent survey with researchers at the University of Mainz states that sharing and long-term preservation of research data is favoured by the majority of the respondents. At the same time, they find it important to keep control over their data and would insofar willingly accept professional support for issues on dissemination and long-term preservation (Baumann 2014). The survey further states that for the readiness to use an appropriate technical infrastructure, the responding system must be easy to use. Therefore, the usability of a repository is an important factor for the use and acceptance within a scientific community.

In the whole process of sharing and reusing data, researchers have a double role: they are data creators and providers as well as data consumers. They can easily share data around the world, but can also protect their integrity and ownership. Data providers and downstream users gather and process data, they annotate and interpret data. Recent surveys generally show that researchers have an increasing interest in keeping their data authentic and interoperable as well as safely managed and stored. So it is even more important to emphasise that many factors are determined already at the stage of creating data. Data creators themselves are in the best position to decide what is necessary for perpetual use and thus initiate the planning of preservation, that is, data need to be interoperable on a technical, semantic and legal level, and under certain conditions this involves format, software, anonymisation, documentation, consents, future rights management and licensing, and an infrastructure for sharing.

To summarise, the data creator/provider controls:

- How understandable data are (documentation about data creation and methods)
- How data are prepared for sharing and reuse (metadata, suitably anonymised data, consents, licensing)
- How preservation-friendly data are deposited (choice of format)
- How easily and how permanently data can be found and accessed (use of adequate infrastructure, such as a trusted repository with persistent identifier system)
- How data can be visualised (providing appropriate metadata)
- How data can be manipulated (providing "actionable", marked up datasets)

Repositories provide services in various directions: to those who deposit content with them for preservation, and to those who consume that content. They also provide services to funders and institutions. The repository and support services beyond (e.g. visualisation technologies) use systems that store and identify data, execute tasks, and mine data for unexpected insights. A digital repository combines people (repository managers and IT specialists who define and maintain the technical framework of the system), processes and technologies which together provide the means to make data perpetually usable (Brown 2013).

To summarise, a repository controls:

- Ingest processes (complete handover of data, monitoring and validation)
- Preservation of authenticity (identity and integrity) of a digital object
- Preservation of bitstream and content
- Usability: keeping digital objects capable of being accessed and represented by the current technological environment; metadata to allow the record to be found, obtained and interpreted
- Tools for access, use and reuse of data
- Integration of applications (for visualisation and so on)
- Rights management

It needs to be considered, however, that the underlying policy of an institution and a repository fundamentally defines where the lines of responsibilities have to be drawn. This policy can determine which content types and formats a system accepts, if a content can be changed or deleted after the handover by a data provider or not, and of course, how far a repository is allowed to "change" a digital object (for example, by converting it into a different format).

The repository management should be familiar with current trends and standards in the technology but also of the respective research communities. This will make repositories better integrated and interoperable, and higher standards of accessibility will bring greater use by both content providers and users. Institutional repositories use a number of different software platforms, which means that users may encounter different platforms and interfaces, and cross-searching can be difficult. There are a number of international initiatives to improve interoperability between repositories, through organisations such as the Confederation of Open Access Repositories (COAR[x]). It would be advisable to use a flexible architecture, with the possibility to adapt the systems to projects and disciplinary requirements.

Institutional repositories perform a special role for their universities, as they provide a central platform for the institution’s research and support research information management systems. They preserve and provide access to research data, to theses, and to grey literature. Subject-based repositories will probably continue to extend their roles alongside publishers, especially in areas where such repositories have an established position in researchers’ regular workflows (Finch et al. 2012).

According to Adrian Brown in “Practical digital preservation”, downstream users can “be considered the most important stakeholders of all, at least in the long-term – they represent the ultimate motivation for undertaking digital preservation”. Their interests will lie principally in the use and reuse of archived content, and should shape requirements for how that content can be discovered and most usefully made available to them” (2013, p. 49).

The whole spectrum of roles, responsibilities and competences shows the necessity to develop process management plans and respective policies. The responsibility for managing research data in higher education spans over a heterogeneous group of actors who are situated both within and outside the institution (Pryor 2014a). Decision-makers, funders and publishers might as well have an influence as external suppliers, such as software vendors or organisations providing tools. At every phase in the processes, there are provisions to curate data, but only a continuous service and an organisation that provides support, can ensure continued quality and compliance to standards.

Research support is also a very heterogeneous group: support may include staff from the library, IT services, research administration services and staff located within individual research teams. Frequently, these groups have not previously worked together as coherent teams and they may know little about each other’s activities. Thrown together, they have to face new organisational dynamics that need to be addressed (Pryor 2014b).

Deriving from the previous chapters, the table below sums up in more detail which are the roles, rights and responsibilities of data creator/provider, repository and downstream for enabling sharing and reuse of research data.

	Roles	Rights & Responsibilities
Data creator/provider	▪ Create (raw) data ▪ Gather data ▪ Process data ▪ Analyse data ▪ Provide data for reuse	▪ Manage data ▪ Comply with policy (institutional, funder's) ▪ Meet standards (disciplinary, institutional) ▪ Consider legal and ethical aspects ▪ Prepare data for sharing and reuse (documentation, metadata, licensing, use of suitable formats) ▪ Consider future visualisation of data ▪ Consider manipulation of data ▪ Select, dispose of, convert data ▪ Deposit data for perpetual preservation
Repository	▪ Curate data ▪ Give access to data ▪ Provide appropriate tools ▪ Integrate applications	▪ Specify data management policy ▪ Ensure ongoing maintenance of system ▪ Manage data for perpetual use ▪ Preserve authenticity of data ▪ Facilitate usability ▪ Meet standards ▪ Protect rights of data provider, enable appropriate access for reuser ▪ Provide tools for access and reuse of data ▪ Provide visualisation tools ▪ Provide support and training ▪ Promote the services
Downstream user (reuser)	▪ Gather data ▪ Access data (view, download, extract) ▪ Analyse data ▪ Extract, manipulate and reuse data ▪ Access metadata for usability and citation ▪ Visualise data	▪ Abide by licence conditions ▪ Acknowledge data creators ▪ Manage derived data effectively

4. Findings of experts' interviews

- Findings are based on interviews with experts in the fields of: engineering, scientific computing and statistics, translation studies, geophysics and meteorology, communication studies, history and digital humanities
- Research data are seen as highly significant across disciplines
- Visualisation technologies are increasingly seen as tool for new research methods, for representing research in an intelligible way and communication tool for the public, for funders
- Processes of generating data and metadata standards depend largely on the respective discipline
- There is little experience with DMPs, whereas after the completion of a project it is not always clear who is responsible for the curation of data or what to do with data not used for publications
- Sharing strengthens cooperative science and creates new quality but in practice, many researchers are still reluctant to share data
- In some disciplines, data from industrial and subject-based data centres as well as permanent monitoring services are gathered, usage might have to be requested case-by-case
- Advantages of institutional repositories are acknowledged, provided that organisational and institutional structures exist
- Due to huge data volumes, it must be considered which data should be preserved

Part of this study included personal interviews with experts from Austrian universities with the purpose of gaining a closer look into the current practices of research processes as well as RDM in the higher education. The interviews were conducted according to a structured questionnaire, but aimed for a more conversational approach to understand the researchers' current field of research and context. The interviewees are from selected institutes: engineering, scientific computing, translation studies, geophysics and meteorology, communication studies, history and the Austrian Center for Digital Humanities in Graz. Hereafter is a summary of the talks, focusing on the significance of research data for the respective disciplines, procedures of generating data, documentation and metadata, the current status of RDM, the handling of legal and ethical issues, aspects of sharing and open access, and important factors for an appropriate e-infrastructure. The interviews reflected individual, discipline-specific views, but also common notions in terms of RDM, sharing and long-term preservation.

Though the definition of research data has always been considered with relation to the respective discipline, the significance is generally considered high and extremely high. Some sciences traditionally work in a data-oriented method but the increased significance of data, their representation, as well as their opening for new research fields and methods affects all disciplines. For some sciences, for example medicine, repeatability of experiments and verification of results are the impetus for sharing data. It was mentioned that only desired results are published, and yet it is also possible to learn from failed experiments. In the classical hermeneutic disciplines, research data are also rapidly gaining in significance. Forms of textual representations support analysis strategies and have interactive components, that is, knowledge can be explicated and contextualised. In this way, new forms of research data are created. Visualisation technologies of data are widely understood as a tool for new research methods. They can reveal relations which cannot be demonstrated with traditional research methods and function as a means to represent research in a plausible and intelligible way. For many sciences, this has become an important tool for communication, both to the public and to the funder for whom it may facilitate decision making. Some experts think that the concept of Linked Open Data offers great opportunities for research. The prevailing opinion is that the more data there is, the more precise is the outcome or are the technologies, for example, translation tools. Using data quantities as a basis, there are new discoveries and discoveries of relations, but at the same time, the task of handling the huge data volumes is considered a great challenge.

The processes of generating data depend largely on the discipline. In data-intensive sciences like statistics or geophysics and meteorology, there exist a variety of data sources, many of them commercial or semi-commercial data centres. Meteorology basically deals with two different types of data: measurement data and simulation models and these again can be classified into data which are created within a definite project or data that are permanently monitored, for example, by domestic and international weather services. Statistical offices frequently offer aggregated data for free, but more detailed data are often liable for costs. In translation studies, some parts of research are also closely related to industrial science. In the humanities, research data are generated from all forms of (indexed) texts, from the medieval manuscript which contains metadata down to the glyph level to a text corpus that makes it possible to search for grammatical structures or for certain terms.

The semantics of data is regarded as crucial, for reuse in general and especially where different sciences come together, like archaeology, history of art, or architecture, just to name a few. According to the disciplines, many different metadata standards exist. However, the respective standards are not automatically understood by non-specialists and the interdisciplinary exchange might hence face some challenges. In natural sciences, standards are also supported by e-journals. Due to huge data volumes, special standards have become prevalent in the commercial translation business and the same applies to data of geophysics and meteorology. International exchange and interoperability of systems are an immanent part of these sciences and so there is also a whole range of standards. Much development is taking place in the implementation of these standards as tools that are to be used automatically. There is also direct collaboration between data centres and journals and in joint conferences and events, where new concepts of publishing research data are presented.

In the humanities, texts can have literary enhancements, meaning that interpretations can be annotated in the texts and provided with metadata. Knowledge that only used to exist in the minds of the scientists can be explicated and contextualised, and thus create forms of data that speak for themselves. Images often require extensive description in order for users to understand the iconography which is often the basis for contextualisation. However, the subjective element in a scientific process in the humanities cannot be neglected. Research processes are often determined by personal motivation or aesthetics which are not so easily captured in standardised metadata. In the digital humanities, metadata is a concept that goes beyond description. As a rule, the content is captured with metadata. There may be meta-information (tags) on single glyphs in a manuscript or on a grammatical structure in a text. The number of attempts at automating these processes is increasing. The representation and visualisation of text is becoming very important. Consequently, in the digital humanities, it is not enough to describe data, but the data must be machine-readable. The potential of text-mining tools is to analyse and process the information contained in text corpora in order to extract relevant information, to manipulate it, and to generate new information.

The current status of data management processes provides an unbalanced picture. Throughout all disciplines, huge amounts of digital objects are being produced, but frequently these data disappear from the web within a foreseeable amount of time. There is generally little experience with DMPs, but the interviewees overall agree that RDM should be started early on in higher education. In translation sciences, where one part of research focuses on industrial training, students are confronted with formats and metadata standards from the start. In social sciences and humanities, this awareness seems to depend much on the lecturers and thesis or dissertation advisers. Consequently, the archiving practices vary significantly. In meteorological research projects, which are often international projects and involve several research teams, there is usually a structured way of generating and describing data, and a consortial agreement as for the use of data. DMP is of course an integral part of projects to such an extent, but it does not necessarily mean that researchers who handle a small part of the project issue their own DMPs. In larger, measure-oriented joint projects with various research teams involved, it is often the case that only a small percentage of the collected data is used for publications. It is not always clear, however, what to do with the other data. Data that is meant to be for open access often stay on an institutional server where access is limited or they are displayed on institutional webpages where long-term access is not guaranteed either.

Regarding sharing, open access and reuse of data, the interpretations are quite diverse. The reluctance of researchers to share their data was mentioned repeatedly. The reasons are various, from a strong sense of ownership over effort and lacking resources to legal restrictions. Legal aspects, embargo periods and DRM are generally considered of utmost importance. Researchers can decide for a preferred licence model when creating their own data. In some disciplines, however, data are also gathered from data centres or permanent-monitoring services. Often, data can be used for scientific purposes, but not necessarily for the public. If used for publication, licences must be requested on a case-by-case basis. Frequently, data from the commercial sector are confidential and cannot be used for regression analyses. The cooperation between data centres and institutional repositories is seen as diffuse. Some data centres have sections for free public access and commercial sections at the same time as they are expected to work profitably. There is a lot of debate which data should be public and which not. Nevertheless, there is also a large public sector, especially that in the European Union. Engaged professionals address the problems of provenance and citation already in introductory lectures. Disciplines that work with images, audio and video material already deal with agreement forms and emphasise the significance of copyright. All experts name documentation and metadata as a basic requirement for sharing.

Advantages of an institutional repository are defined as follows: data are visible, do not get lost, are curated, are migrated to new formats if necessary, and they are stabilised with a persistent identifier. Data are moreover visible in portals like OpenAIRE plus or Europeana and projects can get publicity. Sharing of data creates a new quality and complexity and strengthens cooperative science. Requirements are quality assurance, consistency, relations to related datasets, maybe in another repository or data centre. It should be possible in a connected world, to describe these relations in the systems. Long-term preservation is not only storing but also long-term provision of content, and it is widely agreed that these two things should be considered together. Archiving starts with the creation of data. Data providers must be aware of the formats they use in the processes and for storing and how they want their data to be visualized. Long-term preservation needs organisational and institutional structures that keep pace with the change of data. Huge quantities of data give way to ever more precise experiments and methods and predictions, but still it is necessary to think about what it is worth to be preserved and what not. In the opinion of many professionals, it is not an advisable nor affordable strategy to preserve everything, but who defines the selection criteria? Also, which time horizons do we mean when we talk about long-term preservation?

More challenges are seen with the interpretation of data: researchers who do not have expertise knowledge about a certain discipline, do not necessarily know how to approach the data, nor which kind of data they actually need for their research project. In terms of formats, there is a lot of standardisation. However, what is still missing are standardisations for the representation of data. The dynamic representation of data often depends on singular, proprietary technical solutions. DMPs would certainly help, but the individual research institutions must prepare themselves for the new requirements and it would be helpful if they develop an infrastructure together. In Austria, there is not yet a chair of digital humanities but the Center of Digital Humanities is very active in lecturing and in the development of reference curricula. Libraries, as part of research institutions, could play an important and pro-active role in collaborative research support.

Conclusion

The preservation of research data and their reuse is a process, not an end state, where data is simply handed over to a repository at the completion of a project. Preservation is not only storing but also perpetual provision of content, and these two things have to be considered together. Research data need to be interoperable on a technical, semantic and legal level, and apparently many of these factors lie in the hands of data creators who consequently initiate the planning of preservation. They are responsible for access rights, release, validation, licensing, metadata and quality assurance. Data creators fundamentally decide what is necessary for reuse and open access. IPRs and licences have an influence on the way reusing parties can use the research outputs. If these rights are not clarified there can be consequences like limitations on the research, its dissemination, and any follow-up research.

For the sharing of data, it is necessary to provide quality-assured metadata, suitably anonymised data or consent for sharing, thorough documentation about data creation and methods, persistent identifiers, formats that are validated and suitable for dissemination, guidance for data providers and reuse, and, last but not least, commitment for perpetual preservation. As for the enhanced reuse of data, it requires technical skills and methods: open standards and "actionable" marked-up datasets might be necessary to allow data visualisation and manipulation.

Repositories combine people, processes and technologies which together facilitate making data usable and reusable. They provide services to those who deposit content with them for preservation, and to those who consume that content. For researchers, repositories are a familiar way to facilitate open access. In practice, many researchers might use a mix of institutional and external, subject-based repositories. The concerns for repositories are ensuring the ongoing maintenance of the system, managing data and dealing with fast changing and enhancing formats, the incompatibility between older versions and current versions of the software, preserving the authenticity of data and facilitating usability. They protect rights of data providers and enable appropriate access for reusers. The repository management needs to be familiar with standards in the technology to make other repositories and applications integrated and interoperable, and enable high standards of accessibility. They also need to make systems compliant with funders' requirements. It is a continuous necessity to promote the services to the scientific community and offer support and training.

Downstream users expect different levels of access and as few restrictions as possible. Access today is not just about the ability to read and interpret data, but also about what users can do: to analyse and manipulate data, to shift them from one format to another and to create new knowledge. Researchers want the maximum freedom to use the latest tools and services to make the best use of the information to which they have access.

The interviews have shown that despite all progress in data sharing, there is a discrepancy between claiming for open access and the reluctance or uncertainty with regard to the sharing of own data. To create a willingness for sharing, organisational and technical structures must exist. The infrastructure for sharing and reusing research data must be flexible and also reliable, secure and also open. Data should be openly available but still be protected if necessary by certain constraints. After the termination of a project, the question of data curation often remains an unresolved issue. DMPs do not only reflect the policy of the respective institution and clarify responsibilities, but they principally aim at ensuring that data is accurate, complete, identifiable, retrievable and securely stored. RDM involves technical and non-technical processes and it is necessary to develop adequate services, concepts and tools. Intermediaries develop and invest in such services, and they also face demands for greater customer focus. Libraries have competences in the fields of metadata, long-term preservation and documentation, and might therefore become important actors in the RDM processes.

One of the great challenges today is dealing with the growing body of information. Various forms of content types stand for a science that is driven by large quantities of data from a range of sources such as sensors, scanners, MRI and telescopes, but also human-generated data. The introductory chapters have shown how impalpable the term “research data” is and that the definition always depends on (disciplinary) context. The contents in a repository are highly heterogeneous and reflect the diversity of people and communities that capture and use the data. There are different content types and formats, licenses and terms of use. For repositories, managing data in the long term, is very cost-intensive. It will be necessary to discuss selection criteria and definitions of long-term and perpetual preservation.

The ongoing curation of data reveals a heterogeneous group of actors who have their own interests and tasks in the process, and they may be situated both within and outside the institution. This study has demonstrated the responsibilities of three important actors in the whole process: data creators/providers, repositories and downstream users, and it might serve as a basis for expanding the roles. This would then touch on a conceptual framework for how different stakeholders interact within this system. It is important to have a dialogue between researchers, institutional repositories, subject-based repositories, data centres, funders, publishers, legal experts, libraries and IT services. And it might be necessary to overcome established roles. Collaborative science, after all, is based on a collaborative data infrastructure. Data management planning but also process management planning are integral parts of preservation policies. In this regard, this paper may contribute to the development of a model for an appropriate institutional infrastructure and a corresponding policy.

References

Archaeology Data Service / Digital Antiquity 2009, *Guides to Good Practice*, Archaeology Data Service. Available from: http://archaeologydataservice.ac.uk/. [3 March 2014].

Baumann, S 2014, „Langzeitarchivierung innerhalb Virtueller Forschungsumgebungen im Bereich Digital Humanities“ in *Berliner Handreichungen zur Bibliotheks- und Informationswissenschaft*, 353, Institut für Bibliotheks- und Informationswissenschaft der Humboldt-Universität zu Berlin, Berlin. Available from: http://edoc.hu-berlin.de/series/berliner-handreichungen/2014-353/PDF/353.pdf. [3 March 2014].

Brown, A 2013, *Practical digital preservation: a how-to guide for organizations of any size*, Facet, London.

Chiappe, A, Segovia, Y & Rincon, Y (2007), "Toward an instructional design model based on learning objects", in *Educational Technology Research and Development*, Springer, Boston pp. 671-681. Available from: http://dx.doi.org/10.1007/s11423-007-9059-0. [3 March 2014].

Churchill, D 2007, “Towards a useful classification of learning objects” in *Educational Technology Research & Development*, 55(5), Springer, pp. 479-497. Available from: http://link.springer.com/article/10.1007%2Fs11423-006-9000-y. [3 March 2014].

Data Archiving and Networked Services 2010, *Preparing Data for Sharing: Guide to Social Science Data Archiving*, Amsterdam University Press, Amsterdam. Available from: http://www.dans.knaw.nl/sites/default/files/file/DANS%20Data%20Guide%208%20Preparing%20Data%20for%20sharing.pdf. [3 March 2014].

EDINA and Data Library n.d., *MANTRA Research Data Management Training*. Available from: http://datalib.edina.ac.uk/mantra. [3 March 2014].

Fact sheet: Open Access in Horizon 2020, 2013, European Commission. Available from: https://ec.europa.eu/programmes/horizon2020/sites/horizon2020/files/FactSheet_Open_Access.pdf. [3 March 2014].

Finch, J., Bell, S., Bellingan, L., Campbell, R., Donnelly, P., Gardner, R., … Jubb, M. (2012), “Accessibility, sustainability, excellence: how to expand access to research publications. Executive summary”. Available from: http://www.researchinfonet.org/wp-content/uploads/2012/06/Finch-Group-report-FINAL-VERSION.pdf. [3 March 2014].

Funk, SE 2009, „Digitale Objekte und Formate“ in *Nestor Handbuch : eine kleine Enzyklopädie der digitalen Langzeitarchivierung* ; [im Rahmen des Projektes: Nestor - Kompetenznetzwerk Langzeitarchivierung und Langzeitverfügbarkeit digitaler Ressourcen für Deutschland], eds H Neuroth et al., Hülsbusch et al., Boizenburg, chapt. 7:7-7:12.

Hansen, C, Johnson, CR, Pascucci, V & Silva, CT 2009, “Visualization for Data-Intensive Science” in *The fourth paradigm: data-intensive scientific discovery*, eds T Hey, S Tansley & K Tolle, Microsoft Research, Redmond, Wash. pp.153-163.

IEEE Standard for Learning Object Metadata - Corrigendum 1, 2011. Available from: http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=01032843. [3 March 2014].

(ISCPR) Inter-university Consortium for Political and Social Research 2013, *Sustaining Domain Repositories for Digital Data: A Call for Change from an Interdisciplinary Working Group of Domain Repositories*. Available from:
http://www.icpsr.umich.edu/files/ICPSR/pdf/DomainRepositoriesCTA16Sep2013.pdf. [3 March 2014].

Jones S 2014,“The range of components of RDM infrastructure and services” in *Delivering research data management services: fundamentals of good practice*, eds. G Pryor, S Jones & A Whyte, Facet, London, pp. 89-114.

Möller-Walsdorf, T 2009, „Langzeitarchivierung und –bereitstellung im E-Learning-Kontext“ in *Nestor Handbuch : eine kleine Enzyklopä¬die der digitalen Langzeitarchivierung* ; [im Rahmen des Projektes: Nestor - Kompetenznetzwerk Langzeitarchivierung und Langzeitverfügbarkeit digitaler Ressourcen für Deutschland], eds H Neuroth et al., Hülsbusch et al., Boizenburg, chapt. 17:63-17:67.

Open Access and Open Data Policies and Mandates 2011, OpenAIRE. Available from: https://www.openaire.eu/en/open-access/mandates-a-policies. [3 March 2014].

Oßwald, A, Scheffel & Neuroth, H 2012, „Langzeitarchivierung von Forschungsdaten. Einführende Überlegungen“ in *Langzeitarchivierung von Forschungsdaten: eine Bestandsaufnahme*, eds H Neuroth et al., Hülsbusch, Boizenburg, pp. 13-21.

Pryor G 2014a,“Who’s doing data? A spectrum of roles, responsibilities and competences” in *Delivering research data management services: fundamentals of good practice*, eds. G Pryor, S Jones & A Whyte, Facet, London, pp. 41-58.

Pryor G 2014b,“A patchwork of change” in *Delivering research data management services: fundamentals of good practice*, eds. G Pryor, S Jones & A Whyte, Facet, London, pp. 1-19.

Puschmann, C 2014, “(Micro)Blogging Science? Notes on Potentials and Constraints of New Forms of Scholarly Communication”, in *Opening Science*, eds S Bartling & S Friesike, Cham: Springer International Publishing, Cham pp 89–106. Available from: http://dx.doi.org/10.1007/978-3-319-00026-8_6 . [3 March 2014].

Research Data Management n.d., *Databris*, University of Bristol. Available from: http://data.bris.ac.uk/research/. [3 March 2014].

Riding the wave. How Europe can gain from the rising tide of scientific data. Final report of the High Level Expert Group on Scientific Data. A submission to the European Commission 2010, *Digital Agenda for Europe*, European Commission. Available from: http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf. [3 March 2014].

Schumann, N 2012, „Einführung in die digitale Langzeitarchivierung“ in *Langzeitarchivierung von Forschungsdaten. Standards und disziplinspezifische Lösungen*, eds R Altenhöner & C Oellers, Sivero, Berlin pp. 39-48.

UK Data Archive 2011, *Managing and Sharing Data*, UK Data Archive. Available from: http://www.data-archive.ac.uk/media/2894/managingsharing.pdf. [3 March 2014].

"What Counts as Research Data?", n.d. *Databris*, University of Bristol. Available from: http://data.bris.ac.uk/research/bootcamp/data/. [3 March 2014].

"What Is Research Data”? Research Data Management | Boston University, n.d., *Research Data Management RSS*. Available from: http://www.bu.edu/datamanagement/background/whatisdata/. [3 March 2014].

What’s New in 4.0 2013, Creative Commons. Available from: http://creativecommons.org/version4. [3 March 2014].

Interviewees

Bobrowsi, Manfred, *Department of Communication, University of Vienna*

Bokelmann, Götz and staff members, *Department of Meteorology and Geophysics, University of Vienna*

Budin, Gerhard, *Centre of Translation Studies (CTS) at the University of Vienna, UNESCO Chair in Multilingual, Transcultural Communication in the Digital Age*

Gasteiner, Martin, *Department of History, University of Vienna*

Hudec, Marcus, *Faculty of Computer Science, Research Group Data Analytics and Computing (DAC), University of Vienna*

Stigler, Johannes Hubert, *Zentrum für Informationsmodellierung - Austrian Centre for Digital Humanities, University of Graz*

Tjoa, A Min, *Institute of Software Technology and Interactive Systems at the Vienna University of Technology, Austrian National Competence Center for Security Research*

Websites

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

[ii] http://ec.europa.eu/programmes/horizon2020/

[iii] http://www.jorum.ac.uk/

[iv] http://www.virtualcampus.ch/

[v] https://creativecommons.org/

[vi] https://zenodo.org/

[vii] http://databib.org/

[viii] http://www.re3data.org/

[ix] http://www.europeana.eu/

[x] https://www.coar-repositories.org/