NFDI4Earth

Deliverable D2.4.3

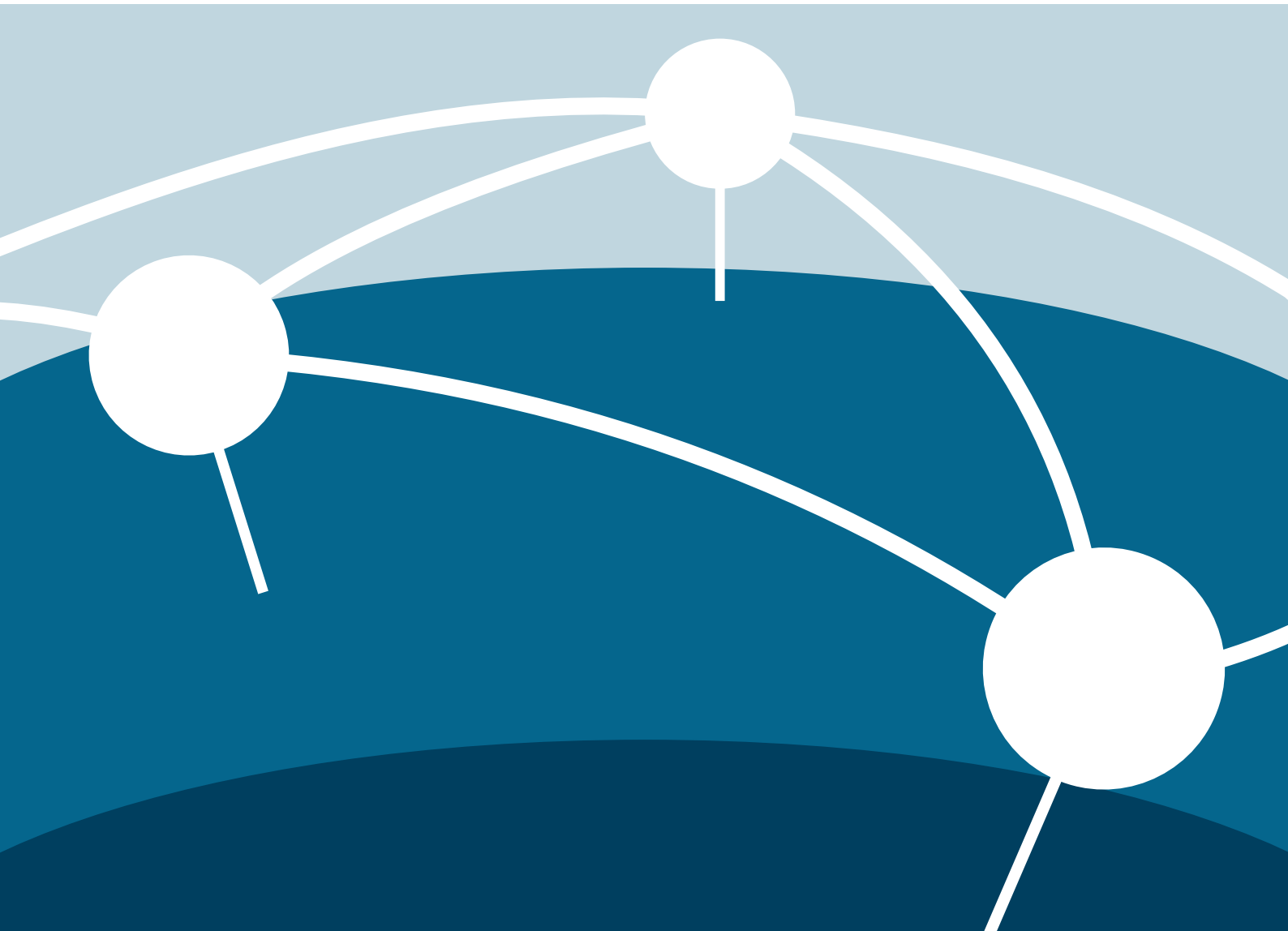
# Guidelines for the preparation of data for archiving

Peter Valena (Peter.Valena@gda.bayern.de), Tim Schürmann

2024-09

**DOI:** 10.5281/zenodo.13837990

nfdi4earth.de



**Citation**

Peter Valena, Tim Schürmann. 2024. *Guidelines for the preparation of data for archiving (NFDI4Earth Deliverable D2.4.3)*. NFDI4Earth Community on Zenodo. https://doi.org/10.5281/zenodo.13837990



**Acknowledgement**

This work has been funded by the German Research Foundation (DFG) through the project NFDI4Earth (DFG project no. 460036893, https://www.nfdi4earth.de/) within the German National Research Data Infrastructure (NFDI, https://www.nfdi.de/).

## Executive summary

With these “Guidelines for the Preparation of Data for Archiving” we in the NFDI4Earth Measure 2.4 “Data in Long-Term Storage” offer an overview of relevant aspects to be considered for the preparation of digital ESS data for long-term preservation, thus keeping data FAIR throughout time. The aim of these guidelines is twofold. On the one hand, we want to raise awareness for the challenges and requirements for a successful preservation of digital data in the long term. On the other hand, the guidelines are intended to provide recommendations and guidance for the preparation of ESS data for archiving.

The target group of these guidelines are data producers in ESS, which can be individual scientists, research projects, or research institutes. In the guidelines we cover relevant aspects for the preparation of data for archiving, from the beginning of the data life cycle up to the point at which data is being ingested into a digital archive.

Included in these guidelines is a “Checklist for the Preparation of Data for Archiving” (Appendix A as part of the document), offering a concise overview of the aspects covered in the guidelines. In addition, the “List of Sustainable File Formats for Long-Term Archiving” (Appendix B as a separate Excel document), expands on the information provided in chapter 6.

## Abbreviations

**BDSG** - Federal Data Protection Act (Bundesdatenschutzgesetz)

**CARE** - CARE Principles: Collective Benefit, Authority to Control, Responsibility, Ethics

**CC** - Creative Commons

**CF** - Climate and Forecast Metadata Convention

**DIN** - German Institute for Standardization (Deutsches Institut für Normung)

**DMP** - Data Management Plan

**ESS** - Earth System Sciences

**FAIR** - FAIR Principles: Findable, Accessible, Interoperable, Reusable

**GDPR** - General Data Protection Regulation

**GeoIDG** - Geological Data Act (Geologiedatengesetz)

**GIS** - Geographic Information System

**ISO** - International Organization for Standardization

**IPR** - Intellectual Property Rights

**M2.4** - NFDI4Earth Measure 2.4 Data in Long-Term Storage

**NFDI** - National Research Data Infrastructure (Nationale Forschungsdaten Infrastruktur)

**NFDI4Earth** - NFDI Consortium Earth System Sciences (NFDI Konsortium Erdsystemforschung)

**OAIS** - Open Archival Information System

**ODC** - Open Data Commons

**PID** - Persistent Identifier

**PREMIS** - PREservation Metadata: Implementation Strategies

**RDM** - Research Data Management

**TRUST** - TRUST Principles: Transparency, Responsibility, User Focus, Sustainability, Technology

# Contents

# 1. Introduction

Securing the long-term preservation of relevant data is a key component in moving towards a sustainable data culture in Earth System Sciences (ESS) and science in general. While in many scientific disciplines a retention period for research data of ten years[1] has been established, long-term archiving aims at securing the availability, access, interpretability, and reusability of data indefinitely. In the NFDI4Earth measure M2.4 "Data in Long-Term Storage" we develop materials and guidelines for the improvement of long-term preservation of ESS data, aiming to facilitate an easy to understand entry point on the topic.

Generic research data management (RDM) and data management plans (DMP) focus on making data FAIR[2] according to today's standards. However, keeping data FAIR throughout time as well as technological and cultural changes requires ongoing curation and preservation actions in a designated curation institution (digital archive).[3] For a digital archive to be able to conduct ongoing curation and undertake preservation actions, the data in turn needs to be prepared accordingly and fulfill certain requirements.

Regarding what these requirements are and the challenges of digital long-term preservation in general, there seems to be a widespread uncertainty in the scientific community.[4] The aim of these "Guidelines for the Preparation of Data for Archiving" is therefore, on the one hand, to raise awareness for the challenges and requirements for a successful preservation of digital data in the long term. On the other hand, they aim to serve as a white information source, providing recommendations and guidance for the preparation of ESS data for archiving.

An optimal preparation of data for archiving depends on the characteristics and structure of the particular data, requiring specific solutions and decisions, often in coordination with the designated digital archive. In view of the diversity of ESS, the many different sub-disciplines as well as the range of different data and data types in use in ESS, the guidelines will focus on providing abstract information and insight into relevant aspects to take into consideration for the preparation of data for archiving, facilitating the decision making and solution finding for the particular data. Included in these guidelines are two appendixes: Firstly, the "Checklist for the Preparation of Data for Archiving" (see sec. A) offers a condensed overview of relevant aspects. Secondly, the "List of Sustainable File Formats for Long-Term Archiving" (Appendix B as a separate Excel document) provides additional information and sources to the information contained in chapter 6.

---

[1] In accordance with the "Guidelines for Safeguarding Good Research Practice" from the DFG, see: (Deutsche Forschungsgemeinschaft, 2022)

[2] (Wilkinson, et al., 2016); https://www.go-fair.org/fair-principles/ (accessed: 19 September 2024)

[3] (Andreu, et al., 2023); (L'Hours, et al., 2022)

[4] (Paul-Stüve, Schürmann, & Valena, 2024)

The target group of these guidelines are data producers in ESS, which can be individual scientists, research projects, or research institutes.[5] The guidelines will cover relevant aspects for the preparation of data for archiving, from the beginning of the data life cycle up to the point at which the data is being ingested into a digital archive. Aspects concerning the preservation of data from this point onwards will be covered in separate guidelines. Furthermore, these guidelines will focus on the preparation of digital research data for archiving. Digitization of analog materials for archiving purposes or the archiving of analog materials combined with digital indexing are not explicitly covered. Also, the archiving of software, which entails specific and quite often very complex legal and technical issues, is not included in these guidelines.

A revised version of these "Guidelines for the Preparation of Data for Archiving" will be published in early 2026, giving the ESS community the opportunity to test their practicability and applicability, give suggestions on additional aspects to be covered, and provide feedback to the NFDI4Earth measure M2.4 via e-mail (nfdi4earth-measure-2-4@tu-dresden.de).

[5] For governmental agencies statutory regulations and laws might apply, including provisions for the archiving of the data. These are not explicitly covered in these guidelines. For more detail on this topic, see: (AdV-KLA-Arbeitsgruppe „Archivierung von Geobasisdaten", 2021)

# 2. Basics about Digital Long-Term Preservation[6]

Digital long-term preservation can be defined as "a series of managed activities necessary to ensure continued access to digital materials, or at least to the information contained in them, indefinitely".[7] The aim is to ensure the (re)usability and comprehensibility of digital data over an undefined period of time through securing the authenticity, integrity, access, and interpretability of digital data over time.[8]

## 2.1. Digital Obsolescence

For digital data to be human- and machine-readable it requires a software environment to interpret the information encoded in a specific file format correctly as well as to render and present its content. Software environments, in turn, require an operating system that they can run on. The operating system supports and manages a particular system infrastructure that is implemented in a hardware environment. All these components, hardware and software environments, operating systems, as well as file formats in use, are subject to continuous technological and cultural change, being at risk of becoming obsolete at some point in the future. This poses the risk of losing access to the content of digital data, leaving it unusable and unintelligible. Addressing and coping with the risk of digital obsolescence is the main challenge in preserving digital data in the long term.[9]

## 2.2. Preservation Strategies

### 2.2.1. Bitstream Preservation

A distinction can be made between two levels of preservation in a digital archive. Bitstream preservation is a basic strategy to preserve digital data (the exact bit string of a data file) as it was at the point of ingest into a digital archive. With the help of checksum checks[10] the integrity of the data file over time as well as the migration of the exact data file to newer storage mediums can be ensured. While bitstream preservation is a basic requirement for long-term preservation

[6] Adapted from (Valena & Schmalzl, 2024)

[7] (Digital Preservation Coalition, 2015, p. Glossary)

[8] In more detail see: (Valena & Schmalzl, 2024)

[9] For a concise description regarding the perception and rendering of digital data, its dependencies on soft- and hardware environments, as well as its ramifications for the preservation of digital data, see: (National Archives of Australia, 2002)

[10] https://www.merriam-webster.com/dictionary/checksum (accessed: 19 September 2024)

of digital data, it does not ensure access to the content of digital data through technological changes in software environments and file formats in use in the long term.

### 2.2.2. Content Preservation

Content preservation on the other hand addresses all aspects of digital obsolescence and ensures access and interpretability of the information contained in digital data in the long term. There are two main strategies offering content preservation of digital data:

- The migration strategy ensures continuous access to the content and interpretability of digital data as well as its usability through migrating data to newer file formats, once the currently used is in danger of becoming obsolete. This requires ongoing monitoring of technological changes in order to determine when preservation actions need to be taken. In addition, the definition of significant properties is necessary (chapter 7.2.). Despite the possibility of information loss during the migration process, this strategy is considered the most feasible so far.
- The emulation strategy on the other hand, keeps the original data as it is and aims at ensuring continuous access to it through (re)creating the original software environment on to newer hardware environments and operating systems. This approach has proven to be more difficult to implement in most cases due to technical and legal challenges.

## 2.3. Roles and Responsibilities

Long-term preservation of digital data requires ongoing curation in a digital archive and is associated with continuous financial and human resource allocation. With the ingest of the digital object into a digital archive the responsibility for long-term preservation passes on to the designated digital archive (respectively the institution responsible for maintaining the digital archive). However, the preparation of data for archiving is most of the times considered to be the responsibility of the data producer, which can be an individual scientist, a research project, or research institute. The long-term cost of digital long-term preservation depends significantly on the degree and quality in which the digital data has been prepared for archiving. The following chapters will outline in more detail relevant aspects to be considered for the preparation of digital data for archiving.

# 3. Choosing a Suitable Digital Archive

The preservation of your data in the long term is going to be secured in a digital archive. Therefore, it is paramount to select an appropriate digital archive that will be capable of preserving your specific data in the long term.

## 3.1. Digital Repositories and Archives

The terms (digital) repository and (digital) archive are often used interchangeably and are not clearly distinguished from one another. Repositories, the more frequently used term within the scientific community, are in general storage locations for digital objects, offering either open access or restricted access to a specific group of users.[11] Repositories often specialize in holding a specific type of digital object (e.g., research data, programming code, software, scientific publications) and/or limit the acceptance of digital objects to subject-specific relevance. Some repositories function as generic and cross-disciplinary publication platforms (e.g., Zenodo, many university and institutional repositories), accepting different types of digital objects.

A digital archive aims at securing the preservation of a digital object throughout time and technological changes, considering the risk of digital obsolescence, for as long as the digital object is needed, potentially for an indefinite timeframe. Digital archives might vary in the degree of access offered to their holdings, ranging from open access to provision of data on demand and/or exclusively on site.

However, not all repositories function as a digital archive. Repositories vary in the degree and duration of curation, levels of preservation, as well as preservation timeframe they offer for digital objects. While bitstream preservation, thus securing the technical integrity of the digital object from the point of ingest onwards, is a common standard at most repositories, content preservation required for long-term preservation of a digital object is not offered by all repositories.

## 3.2. OAIS and TRUST Principles

Finding and choosing an appropriate repository, offering adequate long-term preservation suitable for your data, can be challenging. Adherence and implementation of the Open Archival Information System (OAIS) Reference Model[12] as a de facto standard regarding the

[11] https://forschungsdaten.info/themen/veroeffentlichen-und-archivieren/repositorien/ (accessed: 18 September 2024)

[12] (The Consultative Committee for Space Data Systems, 2012). This text is identical to the standard ISO 14721:2012.

conceptual structure of a digital archive in a repository is a clear indicator that digital long-term preservation is being addressed adequately. Furthermore, the TRUST Principles[13] outline and advocate the following five abstract criteria for trustworthy and sustainable digital repositories, offering some guidance in the selection process of suitable repositories.

- **Transparency**: "To be transparent about specific repository services and data holdings that are verifiable by publicly accessible evidence." This includes information about implemented preservation policies and strategies, guaranteed minimum preservation timeframes or the ability to handle sensitive data.
- **Responsibility**: "To be responsible for ensuring the authenticity and integrity of data holdings and for the reliability and persistence of its service."
- **User Focus**: "To ensure that the data management norms and expectations of target user communities are met." This requirement includes monitoring and adapting to the changing expectations and needs of the community over time.
- **Sustainability**: "To sustain services and preserve data holdings for the long-term." This requires an adequate governance and funding scheme, securing the ability to undertake necessary preservation actions over time to keep the data understandable, accessible and usable in the future.
- **Technology**: "To provide infrastructure and capabilities to support secure, persistent, and reliable services."

## 3.3. Certification Models for Trustworthy Digital Repositories

Certification models offer a helpful orientation by assessing the trustworthiness, sustainability, and the existence of adequate long-term preservation services in repositories according to clearly defined criteria and requirements. These certifications give guidance, transparency, and quality assurance on the level of preservation and curation, adherence to preservation standards, including organizational and technical processes, as well as capabilities offered by a repository.

Most certifications are based on a self-evaluation by the repository according to defined criteria and requirements, followed by an assessment by external experts, based on the submitted documentation or, more rarely, on site. Obtaining a certificate is in general a quite complex, time-consuming and fee-based process.

There are different certification models in place, varying in the number and the content of

[13] (Lin, et al., 2020)

criteria being verified as well as the depth of evaluation. CoreTrustSeal[14] offers a core level certification for repositories and is the most widely used certification model by repositories. Further, more extensive certification models are the nestor-seal according to the DIN-standard 31644[15] and the certification according to the ISO-standard 16363[16].

### 3.4. Subject-Specific Repositories

In general, it is recommended to submit digital data for long-term archiving to a subject-specific repository, since long-term preservation of digital data requires ongoing curation, including expertise on subject-specific standards and methods as well as an understanding on the specific data type and content. A helpful overview for selecting a suitable repository is offered by the Registry of Research Data Repositories[17]. It includes general information on repositories, its disciplinary focus, its supported metadata standards as well as information on obtained certifications and the existence on publicly available policies of repositories. An overview on ESS-relevant repositories and archives can be also found in the NFDI4Earth OneStop4All.

### 3.5. Consider Specific Requirements of the Repository and your Data

It is recommended to select a suitable repository for long-term archiving of your digital data as early on as possible within a research project. Repositories offering long-term archiving of digital data often have specific requirements and standards regarding the structure and quality of data they accept, supported file formats as well as required documentation and metadata. These need to be taken into consideration during the process of preparing data for archiving. Likewise, it is important to consider whether the requirements and standards of the repository align to the characteristics and requirements of your specific data, ensuring the repository's capability to preserve your data in the long term. If in doubt, it is highly recommended to get in touch with the selected repository as early as possible.

[14] https://www.coretrustseal.org/ (accessed: 19 September 2024); (CoreTrustSeal Standards and Certification Board, 2022)

[15] https://www.langzeitarchivierung.de/Webs/nestor/DE/Zertifizierung/nestor_Siegel/siegel.html (accessed: 19 September 2024); (nestor-Arbeitsgruppe Zertifizierung, 2024)

[16] (The Consultative Committee for Space Data Systems, 2011)

[17] www.re3data.org (accessed: 18 September 2024)

# 4. Appraisal of Data

Not all data created by scientific institutions or within research projects need to be preserved for an indefinite period of time. Appraisal means the process of determining the lasting value of data (for society and science) and selecting the data that should be archived for an indefinite period of time.

## 4.1. Short- and Long-Term Cost of Archiving

The decision to archive a digital object always entails both short-term and long-term costs. The short-term costs include preparing the selected data for long-term archiving, which involves creating sufficient documentation and metadata, migrating to a sustainable file format (if necessary), as well as clarifying legal and ethical aspects, all of which requires the allocation of human resources. The long-term costs include securing sufficient storage space as well as maintaining ongoing curation of the digital object and undertaking required preservation actions in a digital archive. Therefore, careful consideration must be given to which data should be archived for an indefinite period of time and which data can be stored only for a specific period of time or disposed of entirely.

## 4.2. Criteria for Appraisal

The following criteria should be considered in an appraisal process and can indicate that data has a lasting value and thus should be selected for archiving:

- Scientific significance and relevance: Data with high scientific importance, contributing to the understanding and gain of knowledge in a discipline.
- Uniqueness: Data that is unique and cannot be reproduced. This applies to most geospatial measurement data. It should be considered whether the data holds sufficient significance as well as whether the complete raw data itself, a selection of data or aggregated data is the most feasible approach for archiving.
- Legal requirements: There might be legal requirements in place, determining that certain data needs to be archived (e.g., Geologiedatengesetz - GeoIDG[18]).
- Reusability: Data that has a high potential for reuse across different projects and disciplines.

[18] https://www.gesetze-im-internet.de/geoldg/ (accessed: 24 September 2024)

- Historical and cultural importance: Data that provides significant information about the reality of life and environment in a spatial-temporal context (e.g., environmental monitoring data).

In general, it should be avoided to archive the same data multiple times. It is recommended to check whether the data is (perhaps in a different form) already archived elsewhere.

## 4.3. Subject-Specific and Cross-Disciplinary Considerations

The appraisal process should be guided by subject-specific standards and requirements, while keeping the potential and relevance for cross-disciplinary approaches and research questions in mind. Most digital archives have a collection policy in place, indicating which kind of data they deem of lasting value. If in doubt, getting in touch with the chosen digital archive is advisable.

It is recommended to address the topic of archival appraisal of data in the early stages of a research project. The decision and reasoning for archiving should be documented in a DMP and, when necessary, adapted over the course of the project.

More extensive guidelines for the appraisal of ESS data for long-term preservation and archiving are being conceptualized within the measure M2.4 and will be published separately in the near future.[19]

[19] A workshop on this topic (NFDI4Earth-Workshop zur Entwicklung von Kriterien für die Bewertung der Archivwürdigkeit von Daten der Bio-/Geowissenschaften) has been held at the Helmholtz-Zentrum Potsdam, Deutsches GeoForschungszentrum GFZ on 14th December 2023. The results of this workshop are published on Zenodo: (zu Castell, et al., 2024)

# 5. Legal and Ethical Aspects

The legal aspects of long-term preservation of ESS data are important for ensuring that data remains accessible, usable, and legally compliant over extended periods of time. Legal considerations include a range of issues, from intellectual property rights and copyright to privacy laws and usage rights. These aspects must be carefully managed and documented to prevent legal disputes and to support effective preservation and reuse of data in the long term.

The complexity of legal requirements in the context of ESS data stems from the various sources and types of data involved, as well as the cooperative nature of scientific research, sometimes making it difficult to keep a complete overview. Therefore, researchers, institutions, and data curators must navigate various legal frameworks, including national and international laws, institutional policies, and agreements between collaborators.

## 5.1. Clarification of Rights

### 5.1.1. Who Holds the Rights on the Data?

Understanding and documenting the ownership of data is crucial for long-term preservation. The rights holder can be an individual researcher, a team, or an institution. The proper documentation of rights is necessary to manage and enforce the correct use and distribution of the data.

### 5.1.2. Intellectual Property Rights and Copyright

Intellectual property rights (IPR) and copyright laws govern the use and distribution of data[20]. These laws protect the creators of the data by usually granting them exclusive rights to use, distribute, and modify their work. Researchers need to ensure that the data they're using complies with these laws as well as the assigned licenses to avoid legal disputes. Appropriate documentation in this area should include the names of the authors and creators of the data, the institutions responsible for data management, and any agreements or licenses associated with the data. This not only protects the rights of the data creators but also clarifies the terms under which the data can be used by others.

[20] https://www.gesetze-im-internet.de/englisch_urhg/ (accessed: 24 September 2024)

#### 5.1.3. Personal Rights Affected

If the data includes personal information, it must comply with relevant privacy laws and regulations (e.g., GDPR[21], BDSG[22]). This includes obtaining clear consent from individuals whose data is collected and ensuring that personal data is anonymized or pseudonymized where possible. By complying with these regulations, researchers protect the privacy of individuals and adhere to legal standards that prevent misuse of personal data.

#### 5.1.4. Impact of Insufficient Legal Clarification

Failing to clarify and document the legal aspects of data ownership and use can significantly hinder the preservation process. Repositories that lack the legal documentation may face difficulties in managing and sharing data, potentially leading to restricted access or misuse. Insufficient legal clarity can also result in legal disputes, impeding the data's availability for all future research.

### 5.2. Ethical Aspects

#### 5.2.1. Ethical Use of Data

Making sure that data is used ethically requires the setting of clear guidelines and standards for data use, promoting responsible research practices[23]. This involves clear guidelines outlining policies that describe acceptable and unacceptable use of data[24]. Those policies should include aspects such as data collection, analysis, sharing, preservation, and reporting. Preventing misuse requires researchers to actively address unethical practices such as data manipulation, fabrication, and improper sharing etiquette. The ethical use of data is promoted by clearly attributing ownership and ensuring that data is accompanied by informed consent from potential participants that specifies how their data will be used, stored, and shared. Participants should be aware of their rights with respect to their data.

#### 5.2.2. Applicability of Additional Ethical Principles

While the above-mentioned aspects are a general way to handle data in an ethical way, for certain data there is sometimes the need to take into consideration additional principles. These

[21] https://gdpr-info.eu/ (accessed: 24 September 2024)

[22] https://www.gesetze-im-internet.de/englisch_bdsg/ (accessed: 24 September 2024)

[23] (Deutsche Forschungsgemeinschaft, 2022)

[24] (All European Academies (ALLEA), 2023)

needs can be very specific and only apply when certain conditions are met. One such set of particular principles are the CARE[25] Principles of the Global Indigenous Data Alliance. Outlined are practices to secure the interests of Indigenous people in a growing global data space.

## 5.3. Usage Rights Defining Reuse

### 5.3.1. Defining Who Can Use the Data and for What Purposes

It is essential to define and document who is permitted to use the data and for what purposes. This includes several specifications. One is the delineation between academic, research, and commercial use. In any case, it is important to define the conditions under which the data can be accessed and used. Outlining the terms for profit focused utilization is equally important. Access policies are able to define the level of data availability for the general public or other involved parties.

### 5.3.2. Data Trustees

A relatively novel idea is using the services of data trustees to ensure safe and ethical handling of data. In medicine e.g., patient data is very protected. The data would be very valuable for researchers and scientists to use. A solution for accessible patient data would be the anonymization of patients with a data trustee handling the pseudonyms. There is no established process on how such a model needs to work yet.[26] Nonetheless, a data trustee model could also be implemented in the context of ESS data where privacy concerns are often a roadblock to the usability of certain data.

### 5.3.3. Recommendation for Open Licenses

Whenever possible, it is recommended to facilitate the use of open licenses (e.g., Creative Commons[27], Open Data Commons[28]) to maximize the potential for reuse and collaboration. Open licenses provide terms under which data can be used, shared, and modified, promoting several of the FAIR Principles.

[25] https://www.gida-global.org/care (accessed: 24 September 2024)

[26] (Lauf, et al., 2023)

[27] https://creativecommons.org/ (accessed: 23 September 2024)

[28] https://opendatacommons.org/ (accessed: 23 September 2024)

## 5.4. Usage Rights and Digital Long-Term Preservation

### 5.4.1. Enabling Necessary Preservation Actions

Long-term preservation requires the digital archive to be able to perform necessary preservation actions throughout time, which include but are not limited to file format migration, data integrity checks, and metadata updates. This oftentimes involves modifying the data, which requires explicit rights to do so. The legal aspects, including licenses and usage agreements, need to allow such modifications without infringing on the interests of the rights holder.

### 5.4.2. License Provisions for Preservation Actions

When assigning licenses, it is therefore important to include terms that allow for data transformation, deriving other or new versions from the data, and ensuring the accessibility of the data. These are key legal aspects data producers can address to significantly enhance the long-term preservation of their data.

# 6. Sustainable File Formats for Long-Term Archiving

Choosing sustainable file formats is a critical aspect of long-term preservation of ESS data. Such file formats ensure that data remains accessible, usable, and interpretable over extended periods of time, trying to reduce the impact of changes in technology. Sustainable file formats are chosen based on their ability to preserve the integrity and usability of data, support long-term access, and facilitate data sharing and reuse.[29]

The use of sustainable file formats mitigates the risks associated with digital obsolescence originating in technological advancements, where data becomes unusable in different ways due to outdated software or hardware environments. By adhering to best practices in file format selection, data producers can significantly enhance the longevity and long-term interoperability of their data, thereby supporting the goals of scientific research and collaboration[30].

## 6.1. Characteristics of a Sustainable File Format

### 6.1.1. Generic Criteria and Recommendations

Sustainable file formats share several defining characteristics that make them suitable for long-term preservation[31]:

- Open and non-proprietary: The file format specifications are publicly available and free to use, ensuring that data can be accessed and used without proprietary restrictions.
- Avoiding dependencies: Choosing file formats that are not dependent on one single software environment contributes to securing the long-term access to the data.
- Well-documented: Comprehensive documentation and specifications are available, facilitating understanding and use by future researchers and systems.
- Widespread use in the community: File formats that are widely adopted within the scientific community are more likely to be supported and maintained over the long term.

[29] (Neuroth, Oßwald, Scheffel, Strathmann, & Huth, 2010)

[30] (The Consultative Committee for Space Data Systems, 2012)

[31] (Wilkinson, et al., 2016)

## 6.2. Sustainable File Formats for Specific Data Types

### 6.2.1. Categorization of Data Types

ESS data encompasses a wide range of data types, each with unique (technical) requirements and best practices for preservation. Understanding the specific characteristics and preservation needs of each data type is essential for selecting appropriate file formats.

### 6.2.2. Widely Used Formats in ESS

Several file formats have been identified as particularly suitable for various types of ESS data due to their sustainability and community adoption. While software such as Photoshop is very popular, its proprietary output formats are not well suited for long-term archiving. Popularity is not always a sure sign that a format is suitable while proprietary does not always constitute exclusion. The following chapters will list file formats that fulfill the above-mentioned criteria for sustainable file formats suitable for long-term archiving. Further information, including sources, is provided in Appendix B, file NFDI4Earth_D2.4.3_Appendix_B_sustainable_file_formats.xlsx.

### 6.2.3. Textual Data

Probably the most common datatype would be plain text. While mostly also the easiest to archive, there are still some pitfalls to be avoided. Proprietary formats like Microsoft Word (.doc) or also advanced technical integrations like including macros in a document introduces problems when it comes to preserving the data in the long term.

| Format Name | File Extension | Comments and additional information |
|---|---|---|
| Extensible Markup Language | .xml | With added scheme- and letter-codes |
| Plain text | .txt<br>.asc | ASCII coded UTF-8 or UTF-16<br>7 bit |
| Portable Document Format | .pdf | Avoid media or script integration, preferably use PDF/A-1 or PDF/A-2 |
| Open Document Text | .odt<br>.odp | Avoid macros or auto-updating fields (e.g., Date, time) |
| JavaScript Object Notation | .json<br>.geojson | Efficient packaging and readability<br>Includes geometry features |

### 6.2.4. Images

Image data today has a wide variety of sources. From simple observation photos to extensive georeferenced digital orthophotos. Accessibility is important while archiving such data. That is why previously mentioned proprietary software formats like Adobe Photoshop, Adobe Illustrator (.psd and .ai) or raw formats directly from photography equipment should be avoided. One of the most common image data formats would be JPEG. Unfortunately, there is no lossless compression for .jpeg data, therefore it is not recommended for archiving purposes.

| Format Name | File Extension | Comments and additional information |
|---|---|---|
| Tagged Image File Format | .tiff<br>.tif | Uncompressed and preferably TIFF 6.0 |
| Portable Network Graphics | .png | Suited for migration from .gif formats |
| JPEG2000 | .jp2<br>.jpg2 | Especially suited for photos |
| Scalable Vector Graphics | .svg<br>.svgz | Without JavaScript binding and fonts;<br>Especially suited for diagrams/graphs |

### 6.2.5. Audio/Video Data

Audio and video data covers a lot of different use cases. Archiving those formats can be tricky due to a range of codecs and versions. Proprietary formats like Windows Media Audio and Video (.wma, .wmv) or web-based formats (.webm) are not recommended.

| Format Name | File Extension | Comments and additional information |
|---|---|---|
| Waveform Audio | .wav | Uncompressed and pulse-code moduled; audio only |
| Free Lossless Audio Codec | .flac | Audio only |
| MP4 | .mp4 | Multiple audio and video tracks possible |
| QuickTime Movie | .mov | Uncompressed |
| Audio Video Interleave | .avi | Uncompressed |
| Matroska Multimedia Container | .mkv<br>.mka<br>.mks | FFV1 Codec |

### 6.2.6. Tabular Data

Similar to textual data, tabular data is simple on first glance. Recommendations on what to avoid are again about extended functionalities included in the tables and proprietary

formats like Microsoft Excel (.xls). XLS data should always be converted to XLSX. Although the Office Open XML format DOCX is published by the same company as XLSX, DOCX was not recommended due to the distinctly stronger alternatives when it comes to textual data formats.

| Format Name | File Extension | Comments and additional information |
|---|---|---|
| Comma-separated values | .csv | Comma or semicolon delimited text |
| Office Open XML Spreadsheet | .xlsx | Avoid macros |
| OpenDocument Format Spreadsheet | .ods | Avoid macros |

### 6.2.7. Geospatial Data

A lot of data could be classified as geospatial. This is a try to confine the category into the most common formats. There is an abundance of discipline specific formats to the degree where it's not feasible to include all of them here. In ESS it may be impossible to avoid certain formats due to highly specialized software and hardware environments. Proprietary formats like Esri's ArcInfo (.e00) should not be used nonetheless. The shapefile format can be viable, due to its relatively open design and being widely adopted, even though it is proprietary.

| Format Name | File Extension | Comments and additional information |
|---|---|---|
| Geospatial Tagged Image | .geotiff | Georeferenced TIFF-image via metatags |
| Geographic Markup Language | .gml | XML based with georeferencing |
| Keyhole Markup Language | .kml<br>.kmz | XML notation for earth viewing purposes |
| Shapefile | .shp<br>.dbf<br>.shx | A Shapefile includes additional data (.prj, .sbx, ...) |
| Network Common Data Form | .nc<br>.cdf | Binary or ASCII |
| GRIdded Binary | .grib<br>.grb<br>.gb | Self-describing data object |
| GeoPackage | .gpkg | No symbology embedding |
| Hierarchical Data Format | .hdf<br>.hdf5 | Self-describing data object |

# 7. Documentation and Metadata

A sufficiently detailed documentation and metadata description of the data designated for long-term archiving is essential. Without it, an ongoing authentic access to the information contained in a digital object and its reusability cannot be guaranteed in the long term. The FAIR Principles[32] provide guidance for metadata to enable the findability, accessibility, interoperability, and reusability of data according to today's standards. However, the digital archive might need additional documentation and metadata for future preservation actions required to keep digital data preserved in the long term and secure its FAIRness throughout time.[33]This chapter provides an overview of the aspects that must be taken into account with regard to documentation and creating metadata description for long-term archiving of digital ESS data.

## 7.1. Purpose of Documentation

The purpose and goal of documenting data for long-term archiving is to keep the data understandable on a semantic as well as on a technical level, securing its human- and machine-readability throughout time. Digital data designated for long-term archiving must therefore be independently understandable without the assistance of the data producer (or other individuals) or other not widely available specialized resources. This means that all relevant information about the data as well as any additional documentation material (e.g., breakdown of used abbreviations, attributes or codes, etc.) must be archived together with the data or - if that information is already preserved and available elsewhere - adequately referenced to. This is required in order for a future user community to be able to interpret, understand, and (re)use the data, as well as for the digital archive to undertake all necessary preservation actions over time.[34]

Which information and documentation is required for digital data to be kept independently understandable varies naturally on the type of data, its content and structure, as well as its technical specifications. In general, it includes information about the context, provenance, methodology and spatial-temporal details of the data's creation, a description and explanation of its content as well as its structure. In addition, information on legal aspects, such as copyright holder and usage rights, as well as information on technical aspects, such as file format specifications, required software, or settings and parameters of a measurement are required.

[32] (Wilkinson, et al., 2016); https://www.go-fair.org/fair-principles/ (accessed: 19 September 2024)

[33] (Andreu, et al., 2023); (L'Hours, et al., 2022)

[34] (The Consultative Committee for Space Data Systems, 2012); (Digital Preservation Coalition, 2015)

## 7.2. Significant Properties

Part of the data documentation process should also be the definition of significant properties. The concept of significant properties aims at defining the essential characteristics of a digital object designated for long-term archiving that needs to be preserved over time throughout preservation actions, such as e.g., the migration to a newer file format, in order to ensure its ongoing access, usability, and authentic meaning.[35]

Significant properties can be grouped into five categories[36]:

- **Content** designates how content is conveyed in the data, e.g., through text, image, audio, etc.
- **Context** refers to information describing the context and provenance of the creation of the data, e.g., who, when, why, etc.
- **Appearance** refers to information about the relevance of the appearance of the content (e.g., visual or audible) of the data to the user, e.g., font, color, layout, resolution, bit depths, etc.
- **Structure** applies to information about how the content of the data is structured and organized within a file and/or in relation to each other in the case of multiple files, e.g., embedded files, pagination, table columns, etc.
- **Behaviour** refers to functionalities and interactions (e.g., with software) required to understand and use the data, e.g., hypertext links, updating calculations, geographic information systems (GIS), etc.

While the concept of significant properties is initially a theoretical one, the implementation as well as the approach and methodology to determine significant properties and their conversion into specific actionable technical specifications for preservation actions varies from digital archive to digital archive. It is therefore recommended to get in touch with the selected digital archive and determine the significant properties in consultation with experts from the digital archive.

Furthermore, it is advised to consider and document the question of significant properties in the early stages of a research project or data creation/collection, since the determined significant properties might also have an impact on the appropriate file format best suited to preserve the specific digital object over time.

For less complex digital data, determining significant properties might be straightforward. E.g., for simple textual data, questions considering the importance of the preservation of

[35] (Wilson, 2018); (PREMIS Editorial Committee, 2015, pp. 50-55)

[36] (Wilson, 2018). There are different concepts for classifying categories of significant properties, with minor differences. (National Archives and Records Administration (NARA), 2009)

aspects such as layout or color might be relevant. Significant properties might also be generically allocated to data types preserved e.g., in the file format PDF/A. But for more complexly structured data where the preservation of specific functionalities is essential, it is recommended to determine significant properties for individual digital objects and include them in the metadata description.

## 7.3. Metadata

Metadata, which can be defined as "data that provides information about other data"[37], plays a crucial role in providing sufficient documentation for the long-term archiving of digital data. As far as possible, all relevant metadata for the documentation of digital data designated for long-term archiving should be provided in a structured and standardized form, suitable for human- and machine-readability.

In general, it is recommended to create metadata for documentation from the very beginning of a research project or data creation/collection onwards, utilizing DMPs, in order to avoid any loss of relevant information throughout the project. Furthermore, wherever possible and applicable, the use of common discipline-specific controlled vocabularies, authority files, ontologies, taxonomies, etc.[38] is advised. From a long-term preservation perspective, it is important to adequately reference the used controlled vocabulary or authority files, including the version, since they might be subject to changes over time or get out of use entirely in the future.

Metadata can be grouped into the following four categories that should be covered at the point of ingest of the data into the digital archive: descriptive metadata, structural metadata, administrative metadata, and technical metadata.

### 7.3.1. Descriptive Metadata

Descriptive metadata contain information about the context of the creation or collection of the data as well as about the content of the data. This should include, if applicable, details on the data creator, contributors, a description of the project the data was created or collected in, participating institutions, as well as a description of the content, the provenance of the data itself, and the methodology used for the creation, collection or processing of the data. For ESS particularly important are metadata on geospatial and temporal coverage of the data.

[37] https://www.merriam-webster.com/dictionary/metadata (accessed: 19 September 2024)

[38] A searchable registry is offered by e.g., TIB Terminology Service (https://terminology.nfdi4earth.de/ (accessed: 19 September 2024)) or Basic Register of Thesauri, Ontologies & Classifications (BARTOC) (https://bartoc.org/ (accessed: 19 September 2024))

Furthermore, keywords, an abstract describing the data, and the usage of persistent identifiers (PIDs) secures that the data is kept findable and citable.

#### 7.3.2. Structural Metadata

Structural metadata provide information on how the data designated for long-term archiving is organized. On the one hand, this includes information on the internal structure of a single data file, e.g., the organization of tabular data. On the other hand, if the data contains several data files and/or folders, information on the folder structure and the relation of the single data files to each other is required.

#### 7.3.3. Administrative Metadata

Administrative metadata include all relevant information required for the digital archive to manage access to the archived data over time. This primarily concerns legal aspects such as information about the rights holder, terms of use, licences, and possible restrictions due to legal or ethical reasons, in order for the digital archive to be able to determine who is allowed to be granted access to the data and under what conditions it can be reused.

#### 7.3.4. Technical Metadata

Technical metadata comprise information that is needed to keep the data understandable and reusable on a technical level. This includes specifications on encoding, the file format, required software to interpret and render the data as well as used hardware. Depending on the data type, additional details such as e.g., device settings and parameters for measurement data or resolution for modelling data are required.

### 7.4. Metadata Standards

It is highly advised to use a metadata standard when creating metadata. Metadata standards provide a framework for standardized and structured metadata, increasing interoperability and reusability as well as facilitating the administration and curation process of archived data in the long term. Most digital archives and repositories require the submission of metadata according to a specific metadata standard.

The use of discipline-specific metadata standards has the advantage of enabling discipline-specific information to be captured in a structured way. Regarding geographic information, the standard ISO 19115 is an internationally widely adopted metadata standard. Another

example of a more broadly used metadata standard in ESS is the Climate and Forecast (CF) Metadata Convention. Examples for widely used generic metadata standards are DataCite and DublinCore. The standard PREservation Metadata: Implementation Strategies (PREMIS)[39] on the other hand, is specifically designated to capture all relevant information for long-term preservation and is adopted in many digital archives.

An overview of existing metadata standards is provided by the Metadata Standards Catalogue[40]. The NFDI4Earth offers a white paper on best practices and recommendations for the harmonisation of metadata schemas in ESS.[41]

[39] (PREMIS Editorial Committee, 2015)

[40] https://rdamsc.bath.ac.uk/ (accessed: 19 September 2024)

[41] (Bernard, et al., 2024)

# 8. Outlook

With these “Guidelines for the Preparation of Data for Archiving” we aim on the one hand to raise awareness for the challenges and requirements of data preparation for a successful preservation of ESS-data in the community. On the other hand, this document is intended to serve as an easily understandable entry point regarding questions and uncertainties as well as relevant aspects on the preparation of data for archiving, contributing to the improvement of digital long-term preservation of ESS-data. We encourage the ESS community to test these guidelines for their applicability and practicability. Feedback in this regard and the guidelines in general is very welcome and can be provided via e-mail (nfdi4earth-measure-2-4@tu-dresden.de). Based on the community feedback, a revised version of these “Guidelines for the Preparation of Data for Archiving” will be published in early 2026.

Due to the diversity of ESS and the data used and created in ESS, these guidelines provide abstract guidance and recommendations as well as an overview of relevant aspects that need to be taken into consideration in the process of preparing data for archiving. For the community to move forward in regard to improving the long-term preservation of ESS data, we identified also the need for best practice examples, covering the challenges and requirements for the preparation for archiving of specific data and data types widely used in ESS in more detail, for which these guidelines can serve as a basis.

In the upcoming year the measure M2.4 will provide two further extensive guidelines on specific topics on long-term preservation, addressing different stakeholders. On the one hand, the “Guidelines for Repository Providers” will give recommendations on implementing and improving strategies, policies, data management, and processes for long-term preservation of ESS data in repositories. On the other hand, the “Guidelines for the Appraisal of ESS Data for Long-Term Preservation and Archiving” will provide more detailed guidance on how to approach the evaluation of the lasting value of ESS data and the selection process of data for long-term archiving.

## Acknowledgements

The conceptualization and creation of these "Guidelines for the Preparation of Data for Archiving" took place in close consultation with colleagues within the NFDI4Earth and beyond, especially within the NFDI4Earth Interest Group Long-Term Storage and Archiving (IG LTA).

For the very valuable discussions, suggestions, critical feedback and/or review, we want to thank (in alphabetical order): Ivonne Anders, Thomas Bentler, Peter Braesicke, Hans-Peter Grossart, Stephan Hachinger, Sibylle Haßler, Lina Hörl, Sheraz Khan, Eszter Kiss, Andrea Lammert, Andreas Lüschow, Giada Matheisen, Daniel Nüst, Thilo Paul-Stüve, Anke Penzlin, Michael Puchta, Hannes Thiemann, and Alois Wieshuber.

## References

AdV-KLA-Arbeitsgruppe „Archivierung von Geobasisdaten". (2021). Leitlinien zur bundesweit einheitlichen Archivierung von Geobasisdaten. (überarbeitet). Retrieved from https://www.bundesarchiv.de/assets/bundesarchiv/de/Downloads/Erklaerungen/leitlinien-geobasisdaten-kla.pdf

All European Academies (ALLEA). (2023). The European Code of Conduct for Research Integrity – Revised Edition 2023. Retrieved from http://www.doi.org/10.26356/ECOC

Andreu, T., Anglada, L., Antos, D., Bähr, T., Brzeźniak, M., Burgi, P.-Y., . . . Wyns, R. (2023). Recommendations Consultation. EOSC-A Long Term Data Preservation Task Force. Retrieved from https://doi.org/10.5281/zenodo.10014698

Bernard, L., Degbelo, A., Grieb, J., Henzen, C., Heß, R., Klammer, R., . . . Weiland, C. (2024). Recommendations for Earth System Sciences Metadata Provision. Retrieved from https://doi.org/10.5281/zenodo.10604587

CoreTrustSeal Standards and Certification Board. (2022). CoreTrustSeal Requirements 2023-2025 (V01.00). Retrieved from https://doi.org/10.5281/zenodo.7051012

Deutsche Forschungsgemeinschaft. (2022). Guidelines for Safeguarding Good Research Practice. Code of Conduct. Retrieved from https://doi.org/10.5281/zenodo.6472827

Digital Preservation Coalition. (2015). Digital Preservation Handbook. (2). Retrieved from https://www.dpconline.org/handbook

Lauf, F., Scheider, S., Friese, J., Kilz, S., Radic, M., & Burmann, A. (2023). Exploring Design Characteristics of Data Trustees in Healthcare — Taxonomy and Archetypes. Retrieved from https://www.researchgate.net/publication/370060215/_Exploring/_Design/_Characteristics/_of/_Data/_Trustees/_in/_Healthcare/_-/_Taxonomy/_and/_Archetypes

L'Hours, H., Kleemola, M., von Stein, I., van Horik, R., Herterich, P., Davidson, J., . . . Huber, R. (2022). FAIR + Time: Preservation for a Designated Community. (2). Retrieved from https://doi.org/10.5281/zenodo.5797776

Lin, D., Crabtree, J., Dillo, I., Downs, R., Edmunds, R., Giaretta, D., . . . Westbrook, J. (2020). The TRUST Principles for digital repositories. *Scientific Data, 7*(144). Retrieved from https://doi.org/10.1038/s41597-020-0486-7

National Archives and Records Administration (NARA). (2009). Significant Properties. Retrieved from https://www.archives.gov/files/era/acera/pdf/significant-properties.pdf

National Archives of Australia. (2002). An Approach to the Preservation of Digital Records. Retrieved from https://www.naa.gov.au/sites/default/files/2020-01/An-Approach-to-the-Preservation-of-Digital-Records.pdf

nestor-Arbeitsgruppe Zertifizierung. (2024). Erläuterungen zum nestor-Siegel für vertrauenswürdige digitale Langzeitarchive. (2.2). Retrieved from https://d-nb.info/1334021635/34

Neuroth, H., Oßwald, A., Scheffel, R., Strathmann, S., & Huth, K. (Eds.). (2010). *nestor Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung.* Retrieved from https://nestor.sub.uni-goettingen.de/handbuch/nestor-handbuch/_23.pdf

Paul-Stüve, T., Schürmann, T., & Valena, P. (2024). Survey covering status of long-term preservation and long-term archiving in ESS in Germany (NFDI4Earth Deliverable D2.4.1). Retrieved from https://doi.org/10.5281/zenodo.11200016

PREMIS Editorial Committee. (2015). PREMIS Data Dictionary for Preservation Metadata. (3). Retrieved from https://www.loc.gov/standards/premis/v3/premis-3-0-final.pdf

The Consultative Committee for Space Data Systems. (2011). *Audit and Certification of Trustworthy Digital Repositories. Recommended Practice CCSDS 652.0-M-1.* Retrieved from https://public.ccsds.org/pubs/652x0m1.pdf

The Consultative Committee for Space Data Systems. (2012). *Reference Model for an Open Archival Information System (OAIS). Recommended Practice CCSDS 650.0-M-2.* Retrieved from https://public.ccsds.org/Pubs/650x0m2.pdf

Valena, P., & Schmalzl, M. (2024). Introduction to digital long-term preservation. *NFDI4Earth Living Handbook*. Retrieved from https://git.rwth-aachen.de/nfdi4earth/livinghandbook/livinghandbook/-/blob/main/docs/LTA/_Introduction.md

Wilkinson, M., Dumontier, M., Aalbersberg, I., Appleton, G., Axton, M., Baak, A., . . . Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data, 3*. Retrieved from https://doi.org/10.1038/sdata.2016.18

Wilson, A. (2018). InSPECT Significant Properties Report (WP2.2). Retrieved from https://doi.org/10.6084/m9.figshare.7137812.v1

zu Castell, W., Lammert, A., Ott, F., Schmalzl, M., Süß, K., Unger, M., & Valena, P. (2024). Bericht zum NFDI4Earth-Workshop zur Entwicklung von Kriterien für die Bewertung der Archivwürdigkeit von Daten der Bio-/Geowissenschaften. Retrieved from https://doi.org/10.5281/zenodo.13330901

# A. Checklist for the Preparation of Data for Archiving

The following checklist is intended to provide a condensed overview of aspects that should be taken into consideration for the preparation of data for archiving as outlined and described in more detail in the guidelines. Are all aspects adequately being considered, your data should be in a good condition to be able to be archived and preserved in the long term. The order of the segments in the checklist is identical with the chapters in the guidelines. It is not intended to be in a chronological order, since many aspects are interdependent and need to be considered as a whole.

In general, it is recommended to get in touch with the selected digital archive in the early stages of your research project or data creation/collection, since an ideal preparation of data for archiving also depends on the requirements and preservation strategies being in place at a specific digital archive. These might vary from digital archive to digital archive.

## A.1. Choosing a Suitable Digital Archive

- Does the collection policy of the digital archive match your type of data and the content of your data?
- Does the selected digital archive provide information and transparency on the kind of preservation and preservation strategies offered as well as on the preservation duration guaranteed?
- Does the selected digital archive adhere to preservation standards and principles, such as the OAIS reference model or the TRUST Principles, and/or holds a certification (e.g., CoreTrustSeal)?
- Is there sufficient discipline-specific knowledge in the selected digital archive available to understand your data and undertake any necessary preservation actions in the future?
- Does your data comply with the standards and requirements of the digital archive?
- Are the requirements of the digital archive applicable and suitable for your specific data?

## A.2. Appraisal of Data

- Does your data hold significant scientific relevance and importance?
- Is your data unique and cannot be reproduced?
- Are there legal requirements for archiving the data?

- Does your data have a high potential for reuse in different projects and disciplines?das fertige PDF für dieses Deliverable findet ihr hier:
- Is your data of historical and/or cultural importance, providing significant information about the reality of life and environment in a spatial-temporal context?
- Have you ensured that your data or parts of your data are not archived elsewhere already?
- Does all of your data need to be archived indefinitely? The conclusion of this question can also be to dispose of the data or parts of the data that have not been deemed of lasting value or the preservation for a shorter retention period (e.g., 10 years).

## A.3. Legal and Ethical Aspects

- Does the data comply with relevant copyright requirements?
- Are all required rights and permissions for the archiving of the data obtained and documented?
- Does the data comply with requirements regarding personal rights (e.g., GDPR, BDSG)?
- If applicable, have all potential ethical implications of archiving your data been taken into consideration?
- Are the usage rights for reuse of your data clearly defined (e.g., open licenses like creative commons)?
- Has the digital archive been granted all required rights for undertaking preservation actions in the future?

## A.4. Sustainable File Format for Long-Term Archiving

- Is the chosen file format suitable for your specific type of data and content?
- Does the chosen file format fulfill the following criteria for a sustainable file format?
  - open and non-proprietary
  - avoiding dependencies
  - well documented
  - widespread use in the community
- If the file format does not fulfill these criteria, migrating your data to a sustainable archival file format is advised.

## A.5. Documentation and Metadata

- Is the documentation of your data complete and contains all relevant information in order for the data to be independently understandable?
- Are the significant properties of your data that need to be preserved throughout time and preservation actions defined and documented?
- Does your metadata contain all relevant information for the following categories?
  - Descriptive metadata
  - Structural metadata
  - Legal metadata
  - Technical metadata
- Does your created metadata comply with a metadata standard?