Starting with 2020-04-30 this Google Doc is closed. Further editing is done by the COST MOBILISE WG4 group in a separate document!

Archiving guideline: General topics

DRAFT; Timeframe: ca. 4,5h (i.e. slot 3-5) during Warsaw workshop

Towards an archiving guideline: Text creation on the following questions:

General topics:

- For what target group is the guideline?
  The target group for this guideline is primarily individual researchers and curators, institutional management bodies of natural scientific collections like musea and IT personnel less familiar with archiving principles and routines.
- What do we want to have as key messages?
  - Preserve National and World heritage
  - FAIR principles
  - Adopt standards for digital archiving
  - recommendation of platforms / standard "package" interoperability
  -

- What is the scope of long-term storage and archiving?
  "Archiving is a purposeful activity for preserving the assets of an organization, with the explicit intention that they be useful for some defined uses in the future. It is distinct from backup, which is primarily there to protect against mishaps and is intended to recover 'lost' files for a relatively short period. The intention of archives is that archive assets remain "usable" , for a potentially indefinite period, and for any of a variety of reasons, including legal, for re-use, or for some other considered reason."
  (https://irods.org/uploads/2014/07/Principles-of-Archival-of-Digital-Assets.pdf)

  There are two distinct processes with one needs to consider in archiving; that is bit preservation and functionality. Bit preservation means "that the [archived] assets can be returned, with provable guarantees as to their integrity", while functionally means that the content can be used … [by] future generations of users. (https://irods.org/uploads/2014/07/Principles-of-Archival-of-Digital-Assets.pdf)

Based on Lannom et al. (2019), when digitizing and archiving of natural scientific collections, one should apply the digital object architecture. This architecture includes the following units: digital object, digital specimen and digital collection.

A digital object consists of digital data, metadata associated with it including an identifier of the physical specimen from which the data were generated, and an unique and persistent identifier for the digital object itself. Digital data can be simple or complex. For example, they can be an unique and persistent identifier to a physical specimen, a plain text file, an image file, an audio file, a multimedia file, DNA sequences, or a 3D model.

A digital specimen is the collection of all digital objects associated with a physical specimen (e.g., a plant, DNA, a stone, seed), which was present at the time of the generation of the digital object.
A physical specimen can be destroyed due to voluntary or involuntary circumstances, the digital specimen still persists. For example, loss of parts or entire collections due to disaster (e.g., fire of the Brazilian museum). Another example would be that the specimen is destroyed as part of the analytical procedures (e.g., DNA extraction, TEM preparation, carbon dating).

A digital collection is a collection of digital specimens.

Explanation of the ISO 14721 standard OAIS and SIP, AIP, DIP
Archiving what the institution has under control as a physical specimen.

- FAIRness: Archiving for reuse
  Define data curation procedures. Develop and perform training materials for data curation for researchers, or employ a data curators.
- Linked Open Data and archiving
- Limitations of archiving, data archiving challenges (e.g. validity of external references, semantic drift, software archiving)
  Memory requirements of digital objects. For example, TIFF files are of higher quality and lose less information than JPEG files, but also in consequence larger in memory. Similarly, 3D models are usually generated from stacks of TIFF images. As a model it contains less information than the original TIFF files, but at the same time requires less memory. On the other hand, future applications might allow us to reconstruct better 3D models from the same stack of images. Hence, the questions might arise between
  Relying on external sources, where one has no control over for archiving, to link to them within one's own archive or to store them oneselves.
  Institutional/national decision between cost efficiency and the value of the physical object itself. Provide examples (rare and valuable specimens, type specimens). Can one reproduce the same or very similar data from the same specimen again in the future, then it might not be worth it to archive the relevant associated data (e.g. raw sequence reads, image stacks from 3D models, CT scans). For some types of

physical specimens (e.g., type species), this might be of global importance and hence should be required for being part of a consortium or certified.
Archiving is a long term commitment for several years, ideally decades. Hence, which requires stable financial support for long time periods for both technical infrastructure and human resources. Accordingly, financing should not depend on insecure funding sources like third-party funding, but should receive financial resources above the institutional level.

Glossary (see below under "Definitions of core terms in the data archiving context - COST-MOBILISE Wiki")

OAIS - Open archival information system - reference model for archiving

AIP - Archival Information Package

Members:

- Fabien Caviere (NHM, ICT, mass digitisation, images)
- Boro Jakimovski (ICT)
- Torsten Struck (NHM, molecular, DNA/tissue bank)
- Gila Kahila Bar-Gal (NHM, DNA/tissue bank)

Chapters in the guideline:

Overview

Digital objects

General topics

Functional topics

Costs

Best practices and user aspects

Back to WG 4 Warsaw: Topics and Documentation

Back to Archiving guideline: Functional topics

Back to Archiving: Best practices, certification and user aspects

---

Back to WG4 Workshop Warsaw: General Info and Participants

Back to Useful links and materials

see also

Definitions of core terms in the data archiving context - COST-MOBILISE Wiki