EMBL Archive Digital Preservation Framework (2020)

Purpose and Scope

The EMBL Archive strives to provide a resource that supports and documents European research, instrumentation and training in the field of molecular biology. To do so, it captures, processes, and makes accessible EMBL's scientific records and archives, and institutional documentary heritage, in both analogue and digital formats.

With regard to digital holdings, their inherent fragility and machine-dependency make it especially important that adequate digital preservation activities are undertaken in a timely manner (sometimes as soon as a new digital item is created) to ensure that digital holdings are continuously accessible and reliable in the future. It is therefore imperative for staff within the EMBL Archive to be proactive in how they preserve these materials.

This framework exists to describe the needs and strategies for preserving and ensuring long-term access to the digital holdings of the EMBL Archive. It puts forward basic guidelines for digital preservation which will form the basis for the development of solid workflows that will help ensure that digital preservation is a structured activity rather than an afterthought.

The scope of this framework is all of the digital material accessioned or created, and held by the EMBL Archive, as governed by the EMBL Archive Terms of Reference. These exist in several types, including the following:

- textual materials (e.g. word-processed documents, emails)
- images (e.g. scanned material and photographs)
- audio-visual materials (e.g. videos, recorded interviews)
- Datasets (e.g. research and administrative data)

Digital assets held by other EMBL departments (e.g. the EMBL Szilárd Library and EMBL Photolab) and scientific data sets (e.g. those managed by EMBL-EBI) are not covered by this framework.

What is Digital Preservation?

Digital preservation is defined as a series of activities that seek to ensure continued access to digital assets. It includes the policies, workflows, people and systems involved in endeavouring to make sure that digital assets are accessible and reliable in the long term.

Why is Digital Preservation Needed at EMBL?

It is an essential component of the job of staff in the EMBL Archive to ensure that all holdings are adequately preserved. A large proportion of the material received and held in the EMBL Archive is born digital, and a further tranche consists of digitized material (i.e. digital surrogates of analogue material). The inherent fragile nature of these digital assets which are machine-dependent and require active, ongoing management in order to remain readable, together with the fact that digital preservation is a relatively new and developing field, mean that developing and

applying coherent and robust workflows and strategies are challenging yet essential to ensure long-term access to digital materials and mitigate the risks of these records and archives being lost forever.

Principles

This framework rests on several principles:

- Access: The primary goal of digital preservation is to ensure long-term access to digital assets. (Access in perpetuity is not guaranteed.)
- Authenticity and provenance: Maintaining the authenticity and provenance of all holdings is a core principle of archival science and will be applied to digital assets.
- Intellectual Property: The EMBL Archive is committed to respecting the rights of any intellectual property rights holders whose material is in the EMBL Archive.
- Professional Good Practice: Digital assets will be preserved according to current standards and good professional practice.
- Resources and Sustainability: The EMBL Archive will be sure to maintain the necessary hardware, software and use these expertly to ensure long-term access to digital assets. At the same time, these resources will be managed in such a way that they are sustainable and economically viable.
- Training: The EMBL Archive will work to ensure that both its staff and relevant EMBL staff are adequately trained in areas relating to digital preservation.

Responsibilities

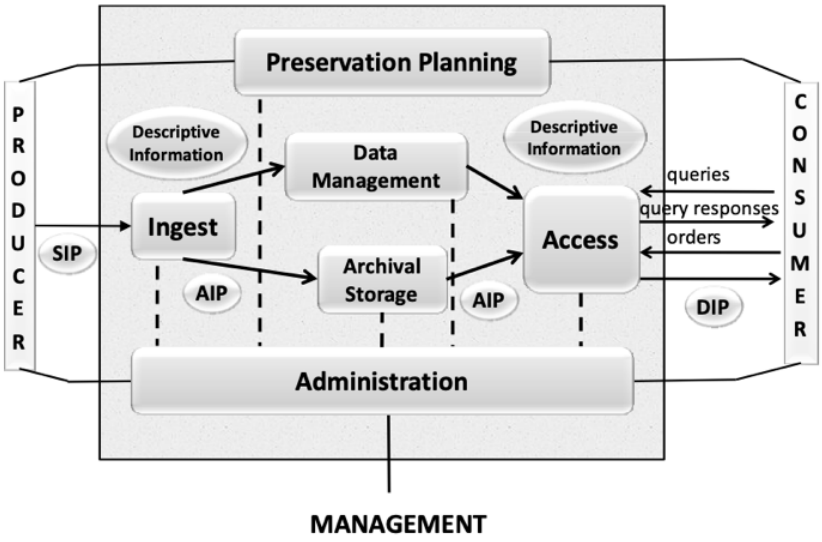
The responsibility for developing and implementing good practice in the area of digital preservation relating to the EMBL Archive falls solely on the staff of the EMBL Archive.

However, they are supported by several other entities:

- The EMBL Director General participates in and supports digital preservation through decision-making and other contributions that inform policies and practice.
- EMBL IT Services support digital preservation by ensuring and enforcing data security on all systems and software used by the EMBL Archive. They also provide maintenance of these systems.
- Creators of digital material across EMBL are responsible for ensuring that those material which will be eventually transferred to the EMBL Archive are well managed while they are in their care. They are, in particular, responsible for preventing loss of digital assets due to poor planning, software or hardware obsolescence, or file corruption. They can approach the staff of the EMBL Archive at any point for advice if needed.

Reference Models and Standards

The main conceptual model for digital repositories is the Open Archival Information Systems (OAIS) Reference Model, which describes different components of a system responsible for long-term digital preservation.[1] It therefore centers around digital assets within the context of digital repositories (or OAIS instances).



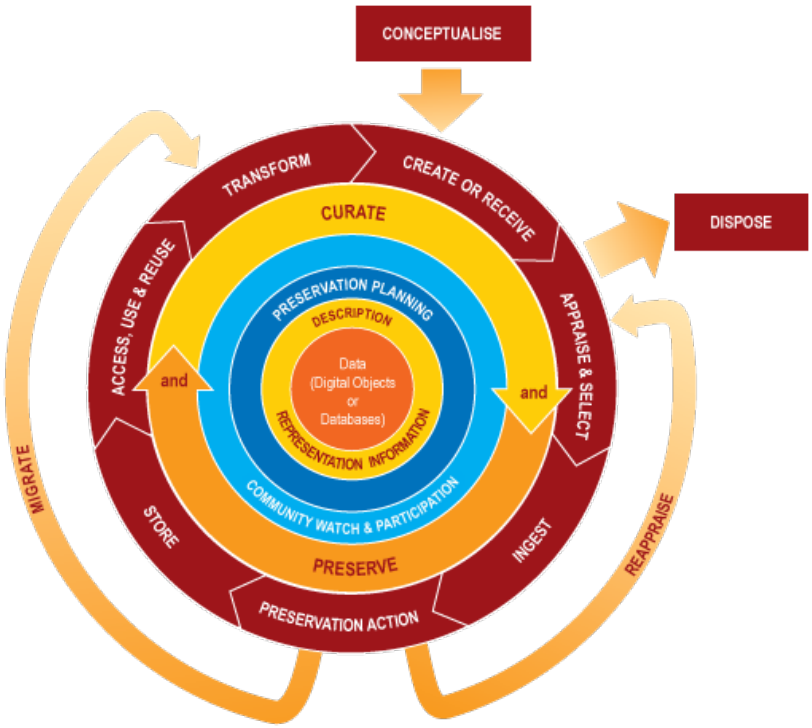
The five functional entities of the OAIS model are:

- The Ingest function takes in information in the form of a Submission Information Package (SIP) and prepares it for storage and management. This function also generates the Archival Information Packages (AIP) and transfers it to the Archival Storage function and its metadata to the Data Management function.
- The Archival Storage is responsible for taking in, storing, maintaining ad retrieving AIPs. Storage, maintenance and retrieval (in collaboration with the Access function) of the AIPs held by the archive.
- The Data Management function oversees the management of the metadata associated with AIPs.
- The Administration function is responsible for the ongoing management of the OAIS instance, for example by creating and maintaining policies, standards and workflows, and performing audits.
- The Access function ensures that users are able to identify and access the digital assets they seek. It is responsible for generating and delivering DIPs (Dissemination Information Packages) and/or metadata, and enabling users to access these through an appropriate interface.

Complementing the OAIS model, the Digital Curation Centre (DCC)'s Digital Curation Lifecycle model considers digital assets within a wider context, positing in particular that efficient digital preservation often requires actions to be undertaken at the time of or shortly after the creation of new digital assets.[2] It provides sequential activities which are required in the process of preserving digital assets.

[1] Consultative Committee for Space Data Systems (CCSDS), "Reference Model for an Open Archival Information System (OAIS)", https://public.ccsds.org/Pubs/650x0m2.pdf.

[2] Digital Curation Centre (DCC), "Curation Lifecycle Model", http://www.dcc.ac.uk/resources/curation-lifecycle-model.



Specifically, the sequential actions to be taken are:

- Create or Receive: Create data including administrative, descriptive, structural and technical metadata. Preservation metadata may also be added at the time of creation OR Receive data, in accordance with documented collecting policies, from data creators, other archives, repositories or data centres, and if required assign appropriate metadata.
- Appraise and Select: Evaluate data and select for long-term curation and preservation. Adhere to documented guidance, policies or legal requirements.
- Ingest: Transfer data to an archive, repository, data centre or other custodian. Adhere to documented guidance, policies or legal requirements.
- Preserve: Undertake actions to ensure long-term preservation and retention of the authoritative nature of data. Preservation actions should ensure that data remains authentic, reliable and usable while maintaining its integrity. Actions include data cleaning, validation, assigning preservation metadata, assigning representation information and ensuring acceptable data structures or file formats.
- Store: Store the data in a secure manner adhering to relevant standards.
- Access, Use and Reuse: Ensure that data is accessible to designated users on a day-to-day basis. This may be in the form of publicly available published information. Robust access controls and authentication procedures may be applicable.

- Transform: Create new data from the original, for example, by migration into a different format (e.g. due to format or hardware obsolescence), by creating a subset of data, by selection or query, to create newly derived results for further reuse or publication.

Stakeholders

As well as the staff of the EMBL Archive, stakeholders in digital preservation by the EMBL Archive are all those who create and use digital assets whose long-term preservation will be the responsibility of the EMBL Archive. In particular:

- EMBL alumni and other relevant parties who may hold or create historical scientific and institutional digital assets.
- Current EMBL scientific staff, who create scientific records in digital format (e.g. electronic laboratory notebooks).
- EMBL institutional departments, who create and maintain digital records of institutional importance (e.g. EMBL Council Secretariat, EMBL International Relations).
- EMBL administrative departments, who create and maintain digital records of administrative importance.
- Users of the EMBL Archive (internal or external) who need access to the digital holdings of the EMBL Archive.

Revision of this Framework

This framework was first prepared in March 2020 and will be reviewed on an annual basis.