The EMBL Archive and IT: needs and use

EMBL Archive and Simon Wilson, 2020

Table of Contents

1. Document Statement of Purpose

The purpose of this document is to provide a clear explanation of the EMBL Archive's IT needs, especially with regard to its digital preservation activities, with the view of ensuring mutual understanding. It was prepared in collaboration with Simon Wilson as part of the setting up of digital preservation activities at the EMBL Archive.

2. Introduction

2.1. What are archives?

Archives are items that have been identified for permanent preservation because they contain information of legal, evidential or historical value and may be either analogue (paper) or digital in format. Digitised material is where an analogue item has been converted to a digital format. Born-digital material is used to identify items that originated in digital format – for example word-processing files, spreadsheets, databases, videos, emails and websites. The term digital archives is used here to refer to both digitised and born-digital material held by the archives.

A key difference between paper and digital archives is the scale of material: instead of talking about the number of boxes or linear metres, digital material is usually quantified in terms of number of files (10,000+ is not atypical) with extent measured in GB or TB. Using these metrics together provides the clearest context (e.g. 12,345 files and 6.7TB). It is also important to note that digital archives might arrive on a range of media (e.g. floppy drive or an old laptop) and being able to safely extract content from a range of carriers is another challenge faced by the Archives team.

2.2. The EMBL Archive

In alignment with EMBL's institutional mission, the EMBL Archive's vision is to provide a resource that supports and documents European research, instrumentation and training in the field of molecular biology. It does this by capturing, preserving and making accessible EMBL's institutional, scientific and other records of historical value. In so doing, the Archive supports EMBL as a transparent and accountable institute.

Internal material transferred to the EMBL Archive falls into one of three categories:

- Scientific records - including the Group's scientific notebooks, are transferred to the EMBL Archive at the point when a group or team leader departs EMBL
- Institutional records – including records created by Internal Audit, DG Office, International Relations departments and the Staff Association

- Administrative records – including records created by Human Resources, Finance, Legal Services, Purchase and Facility Management

The primary users of the EMBL Archive are scientists and social scientists interested in EMBL and its work. The EMBL Archivist provides support to interested parties and facilitates access to material and information.

With an increasing quantity of material for permanent preservation being digital in format, a digital preservation framework has been created to guide the EMBL Archive's activities as it safeguards its archival digital assets for future generations.

2.3. The importance of IT to archivists

Archivists rely on IT tools to undertake several aspects of their work responsibilities. The general workflow that shows has these steps and software work together is recapped in 7.2. The EMBL Archive Digital Preservation Workflow.

2.3.1. Cataloguing

A crucial element of the archive's work is to share details of its holdings to potential users. The EMBL Archive uses AtoM collections management software (see 3.1. AtoM for cataloguing) to document and describe, according to following international standards, the material it receives and manages. The EMBL Archive Catalogue is accessible online (https://archive.embl.org) and provides descriptive information about the collections, including any access restrictions. In line with archival practice, the online catalogue is intended for specialists and generalists alike.

2.3.2. Digital preservation

Digital preservation is a series of activities including policies, workflows, people and systems to ensure that digital archives can be reliably accessed in the long term. As digital assets have both hardware and software dependencies, they require active, ongoing management to remain readable for future generations. Digital preservation is an IT-driven activity and at EMBL, the principal software used is Archivematica (see 3.2. Archivematica for digital preservation).

Receiving and processing digital holdings requires both specialist hardware (see 4. Hardware) and software, for example to identify file formats, check for duplicates and arrange the files for cataloguing (see 3. Software used to meet the EMBL Archive's IT needs, in particular § 3.4 to 3.8).

Some necessary jargon

The Open Archival Information Systems (OAIS) Reference Model is the principal conceptual model for digital repositories. It describes different components (see 7.2. The OAIS functional entities) of a system responsible for long-term digital preservation. The OAIS model is responsible for some of the key terms used in digital preservation, namely the following three which define the types of information packages which bring together information about the content, metadata embedded in the digital object, descriptive metadata about the material and file checksum:

- Submission Information Package (SIP): this is created as part of the data-preparation process prior to ingesting the content into the digital preservation system
- Archival Information Package (AIP): the version of the package created when it is stored by the archive within a digital preservation system and is usually enhanced with extra metadata
- Dissemination Information Package (DIP): this is the version that is made available to the end user so may include files in a different format and a subset of the metadata

2.3.3. Storage

Storage is crucial in the digital preservation workflow but different types of material have different needs. Predicting precise storage requirements can be difficult as the acquisition of material into the archives will vary from year to year.

Once material has been ingested into Archivematica there is a need to store the original files, but the priority will be on restricting access to these files rather than on immediate retrieval. As part of its normal procedure Archivematica will create derivatives of the files ingested – for example an image in .tiff format may then be converted to a lower quality .jpg as the access version. The EMBL Archive uses the Archivematica Storage Service (see 3.3. Archivematica Storage Service for storage), which is linked to Archivematica, to manage archival storage, transfer locations and packages.

There may also need to be temporary storage, for example a copy of the born-digital files, after appropriate virus checking, placed on a network drive to allow archives staff to review the material to identify copyright or sensitivity issues before ingesting it into Archivematica.

2.3.4. Access

There are two models of providing access to born-digital archives: using a locked-down laptop in the archives or managed access online (often after a login process to confirm or establish a user's identity). Some digital content might also sometimes already be accessed via the online catalogue.

The intention will be to review options and approaches to providing access to born-digital material once a significant body of work has been processed.

3. Software used to meet the EMBL Archive's IT needs

3.1. AtoM for cataloguing

Artefactual's Access to Memory (AtoM) is a web-based, open source application for archival description. All items in the EMBL Archive are described according to the International Council of Archives ISAD(G) standard which has made significant improvements to support the exchange of data including third-party aggregation services (e.g. Archives Portal Europe). The standard has also made it far easier for users to interpret archive catalogues from different archive services more easily. It can be configured to work with Archivematica for digital preservation.

3.2. Archivematica for digital preservation

Artefactual's Archivematica is a web-based, open-source digital preservation system. It brings together a number of open-source software tools to support the digital preservation workflow from ingest to access in accordance with the OAIS reference model. It supports several metadata standards including METS, PREMIS, Dublin Core and the Library of Congress BagIt and can be integrated with AtoM (see 3.1 AtoM for cataloguing) and other digital services.

3.3. Archivematica Storage Service for storage

The Archivematica Storge Service allows the configuration of all local or remote storage spaces used. Archivematica manages both the physical location of the file(s) but also the purpose of that space within the workflow – allowing distinction between locations used for ingesting files or for storing AIPs or DIPs (see Some necessary jargon). This aspect is highly configurable so you could use the same location for multiple purposes or have multiple locations with different purposes.

3.4. DROID for identifying digital material

DROID is a software tool developed by the UK National Archives and released on an open-source basis. It is a file format identification tool that identifies the file type and version from information embedded within the file itself rather than relying on the file suffix. The software can be used to create a file manifest of the collection prior to any processing or ingest. It will also create a checksum for each file which can be used to check that the file contents remains unchanged over time – changing a single character in a 20 page report would generate a different checksum.

There are regular updates to the service with individuals and organisations actively contributing research to the 1700 file formats in the underlying data registry. Each file format is assigned a unique reference (PUID) which can then be used by third-party tools and services, for example to convert all MS Word 2007 files to .pdf for preservation and access purposes.

3.5. Archifiltre for processing digital material

Archifiltre is a free, open-source tool that supports the appraisal of digital archives by creating file trees, flagging duplicates and allowing the reorganisation of directories. Its visual representation of folder sizes and file types is especially helpful when faced with a large quantity of material to review. It is often used at the point of transfer, together with the transferring department, before it is accessioned by an archive. It is developed and maintained within the French government's Fabrique numérique des ministères sociaux.

3.6. FTK Imager for processing digital material

AccessData's FTK Imager, a free component of the FTK Toolkit, can be used in conjunction with forensic write blockers (see 4.2. Write blockers) to create forensic images of files, folders or entire media. In the digital preservation workflow it plays a crucial role in the transfer of born-digital archives from a range of portable media to more stable storage.

3.7. TeraCopy for processing digital material

TeraCopy is a free file manager utility that includes a file verification element, using the file checksum to check that the file is exactly the same. It retains the original file date stamp which is not always the case with file manager software. Visit preferences after installing the software to set the 'always test after copy' option. It is a good tool to use if you wanted to create a reliable copy of born-digital archives to undertake a content review or assessment ahead of any forensic processing.

3.8. Antivirus software for processing digital material

Installing anti-virus software on your forensic workstation (see 4.1. Forensic workstation) is crucial. Exactly how you protect your workstation is very much dependent upon your particular set-up and the technical support you have to call-on. Networked devices will be covered through institutional anti-virus programmes but most forensic workstations tend to be disconnected from the network to avoid introducing a virus from amongst your born-digital archives. The two most frequently cited anti-virus packages used in digital preservation workflows are ClamAV and AVG.

Most workflows suggest conducting two tests 30 days apart before ingesting any content assuming a monthly update of the antivirus software. With good practice advocating only one collection being processed at a time this quarantine period can be difficult to accommodate. An alternative approach that is being adopted by many is to test the same content with two different software packages.

If a virus is detected and can be deleted this should be done and the entire collection re-scanned for verification. In some cases it might be possible to secure another copy of the file(s). Whatever actions are undertaken make sure that a note is added to the collection. If a large number of files have been

removed from the collection due to virus or malware it might be appropriate to re-run DROID (see 3.4. DROID for identifying digital material) to create an updated manifest.

3.9. ePadd for processing e-mail

ePadd is a free open source tool developed by Stanford University Special Collections and University Archives. It supports the appraisal, processing, preservation and discovery of e-mail archives from mbox or IMAP services assuming you have the appropriate login and password details.

Once a mailbox has been imported you are able to browse the email by date or recipient, review any attachments or run a series of reports. The processing module allows you to organise and edit the archive, you can annotate, redact or flag sensitive or key content for future activities. A stand-alone discovery module supports public access to the content based on any access restrictions imposed in the processing module. The tool utilises a lexicon, which can be edited locally, to support searches for specific terms or analysis and uses a sentiment lexicon to provide an additional route into the underlying content.

Use of the tool across the digital preservation community is increasing across North America and Europe and a community forum has been established to encourage further development of the tool.

4. Hardware

Apart from the usual workstations needed to carry out cataloguing and usual duties in the EMBL Archive, several pieces of hardware are specifically needed for digital preservation purposes.

4.1. Forensic workstation

A forensic workstation is a dedicated PC or laptop that is used solely for processing born-digital archives and forms a key component in the digital preservation workflow. A range of software, mostly open-source, can be used to support processing tasks but can also be used by researchers to gain new perspectives into the material in archival care. One essential characteristic of a forensic workstation is that it off-line.

4.2. Write blockers

A write blocker is a piece of hardware that sits between a forensic workstation and the media containing born-digital archives. It allows data to be read but critically not modified to protect the metadata embedded within the file and supports the safe transfer of data to more stable storage.

Using a write blocker (like the CRU ComboDock 5.5 right) with forensic software like FTK Imager ensures that date created, for example, remains unchanged and can form key contextual information for future researchers.

5. The Archival and Digital Preservation Community Online

The archive and digital preservation communities are a tremendous source of advice and support, with individuals and services happy to share experiences. Many of the tools used by the digital preservation community receive input and suggestions for improvement from their users. There are many examples of projects and initiatives aimed at particular types of record (eg email) or types of organisation (eg APARSEN (Alliance Permanent Access to the Records of Science in Europe Network) project 2012-2104).

A few useful resources include:

- The Digital Preservation Coalition: www.dpconline.org

The Digital Preservation Coalition (DPC) develops and maintains such resources as the DPC Handbook and technology watch reports, with regular blogs by members sharing practical experiences. It organises many events, most of which are for DPC members only, but the slides and recordings are usually available for members and non-members alike. The DPC also maintains the DPC Rapid Assessment Model (DPC-RAM) which is a digital preservation maturity modelling tool that enables benchmarking of an organization's digital preservation capability. The EMBL Archive is using the DPC RAM model to structure and develop its digital preservation activities.

- The Open Preservation Foundation: openpreservation.org

A legacy from the EU Planets programme (2006-2010), this European community supports community tools including Jhove, Fido and VeraPDF. It posts regular blogs and briefings, though much of the content is for members only.

- The US Library of Congress: www.loc.gov/preservation/digital

This is a rich source of information including recommended file formats and PREMIS, a preservation metadata standard.

- Artefactual products groups and wikis:

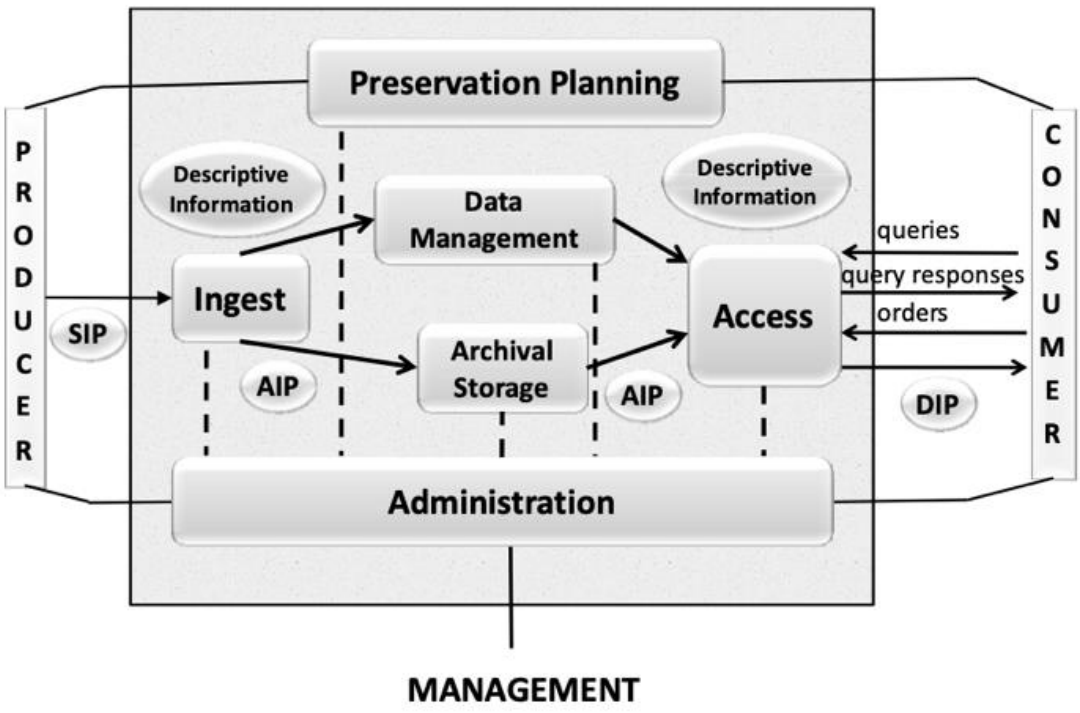
Several online, community-led resources and user groups:

  - The AtoM User Group forum: groups.google.com/forum/#!forum/ica-atom-users
  - The Archivematica User Group: groups.google.com/forum/?fromgroups#!forum/archivematica
  - The Archivematica wiki: wiki.archivematica.org

The channels include informal groups of users sharing their experiences, talking about work-in-progress and lessons learnt.

6. Appendices

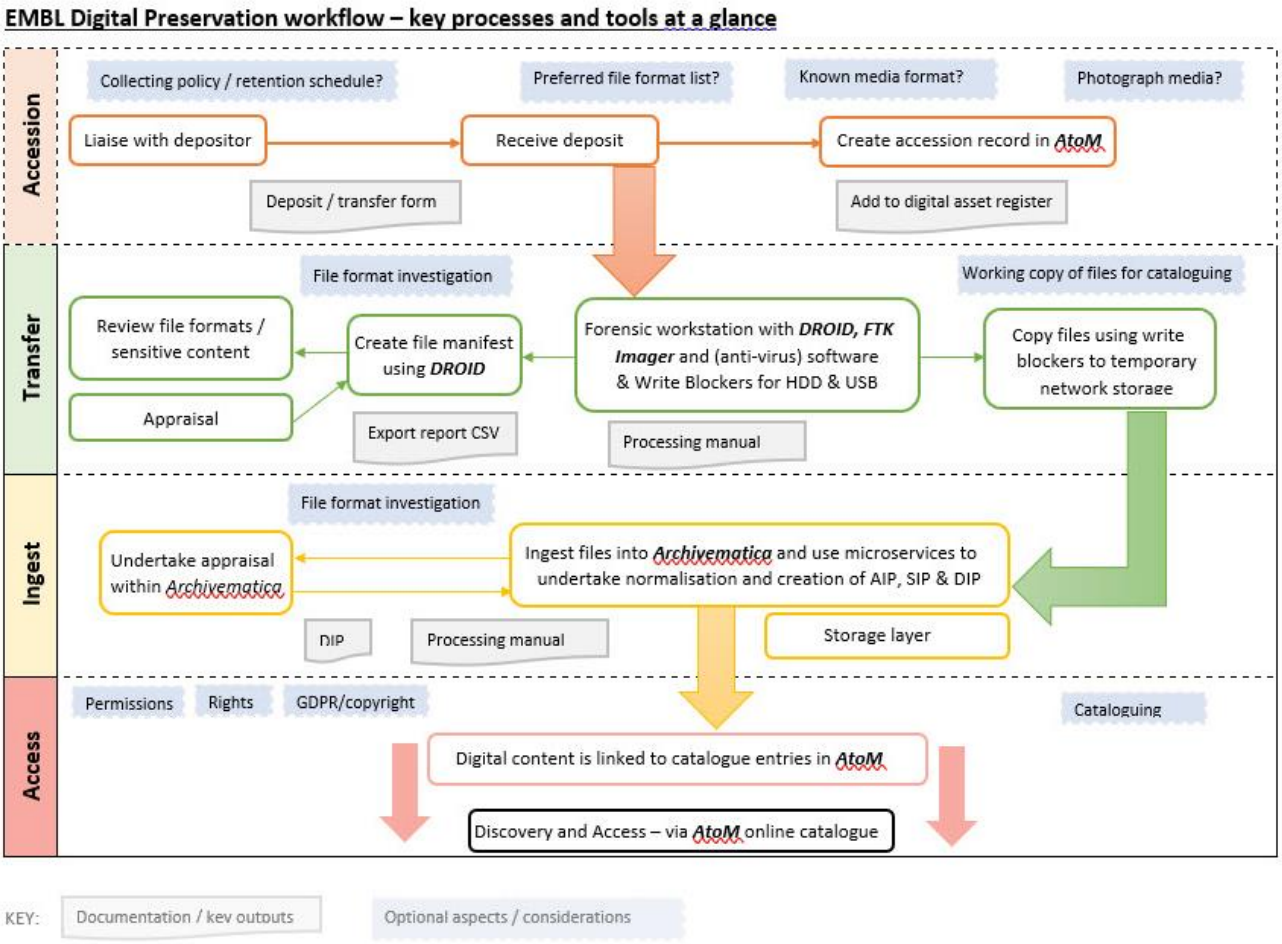
6.1. The OAIS functional entities



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 met.

6.2. The EMBL Archive Digital Preservation Workflow



In the Accession phase, material is prepared for transfer to the archives, details of ownership are documented and an accession record created in AtoM.

In the Transfer phase, forensic workstation and write blockers are used to read media and extract files. DROID creates a list of files and file types, and the archivist and check for known issues.

During Ingest, following a sensitivity review, files are ingested into Archivematica and access copies are automatically generated.

In the Access phase, material is described in AtoM and linked to the holdings in Archivematica. Discovery of the holdings is done the EMBL Archive Catalogue.