Archiving terms and definitions

G. Kahila Bar-Gal, D. Triebel

Terms and definitions

- As part of the first slot or during the workshop we will try to define some terms

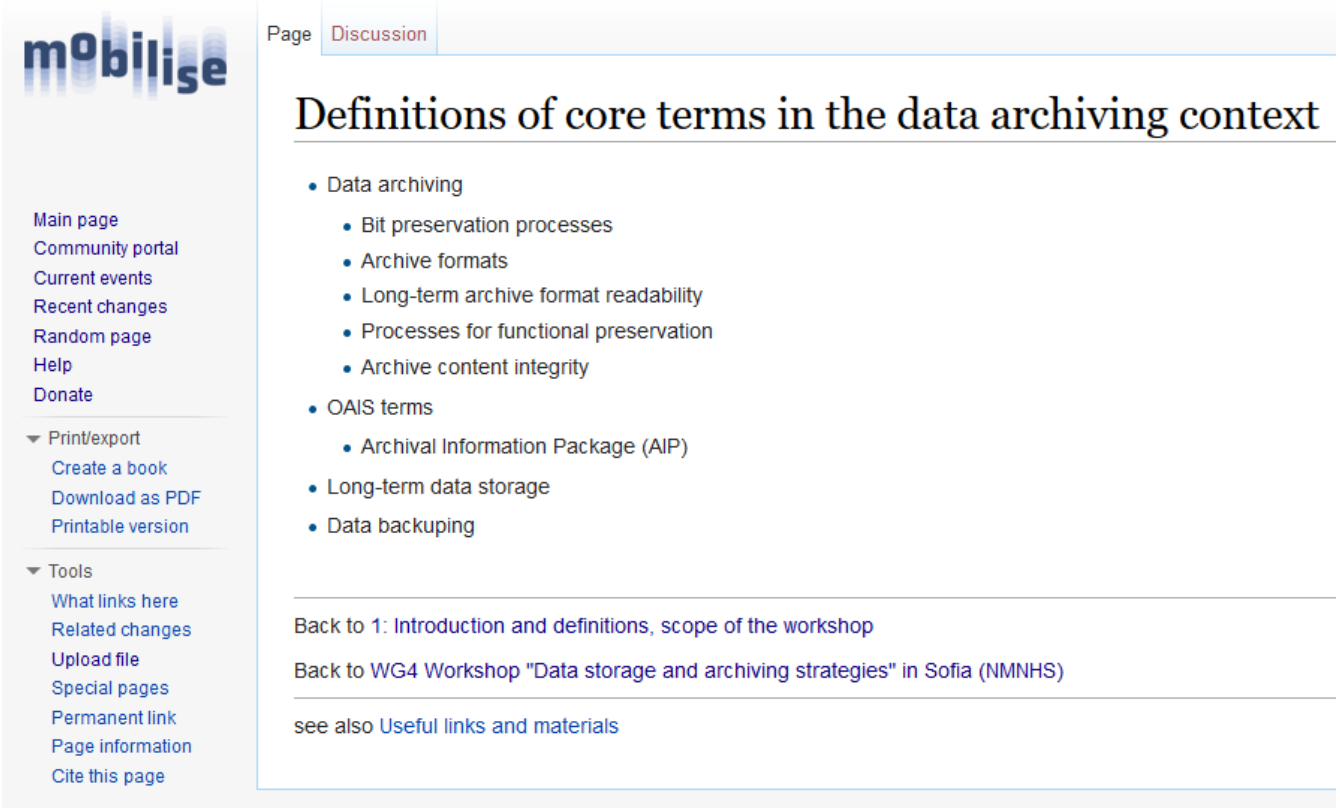
- and document this in the Wiki under https://costmobilise.biowikifarm.net/wiki/Definitions_of_core_terms_in_the_data_archiving_context

+ Add useful links

Terms and definitions

Data archiving:
Methods that stage the primary data itself to a cheaper storage media with quick data retrieval.

What needs to be archived and for how long?



Data archiving

Data volume is growing exponentially day by day, and needs to retain historical data for a certain period.

Most important archival of data should perform single instance storage - stored once across the enterprise (even with multiple users).

Data archiving  **Data backup**

Copy of data made for long-term storage and reference. May include data no longer in use.

Copy of the current data, used to restore when data is lost or damaged

Archive formats

File formats:

The way that information is encoded for storage in a computer file.

https://en.wikipedia.org/wiki/List_of_archive_formats;

https://www.powergrep.com/manual/prefsarchiveformats.html

Compressed File Formats Supported by PowerGREP

These are the compressed file formats or archive formats that PowerGREP can decompress. You can search trough files in all these formats. PowerGREP can also create o
Select the format you want to configure.

Formats that PowerGREP can decompress, create, and update:

- **ZIP archives**: ZIP format used by PKZip, WinZIP, and a host of other compression utilities. This is the most popular archive format. PowerGREP uses the original Deflate when creating ZIP archives.
- **ZIPX archives**: Same as the ZIP format. The .zipx extension is used to indicate that newer compression methods are used. PowerGREP uses the PPMd algorithm when
- **7-zip archives**: 7z format used by 7-zip. This format yields the smallest files of all the archive formats that PowerGREP can create.
- **TAR uncompressed**: Uncompressed tarball (.tar file).
- **TAR GZip archives**: Tarball compressed into a GZip file (.tar.gz or .tgz file).
- **TAR BZip2 archives**: Tarball compressed into a BZip2 file (.tar.bz2 file).
- **TAR XZ archives**: Tarball compressed into an XZ file (.tar.xz file).
- **GZip file**: Single file compressed with GZip
- **BZip2 file**: Single file compressed with BZip2
- **XZ file**: Single file compressed with XZ

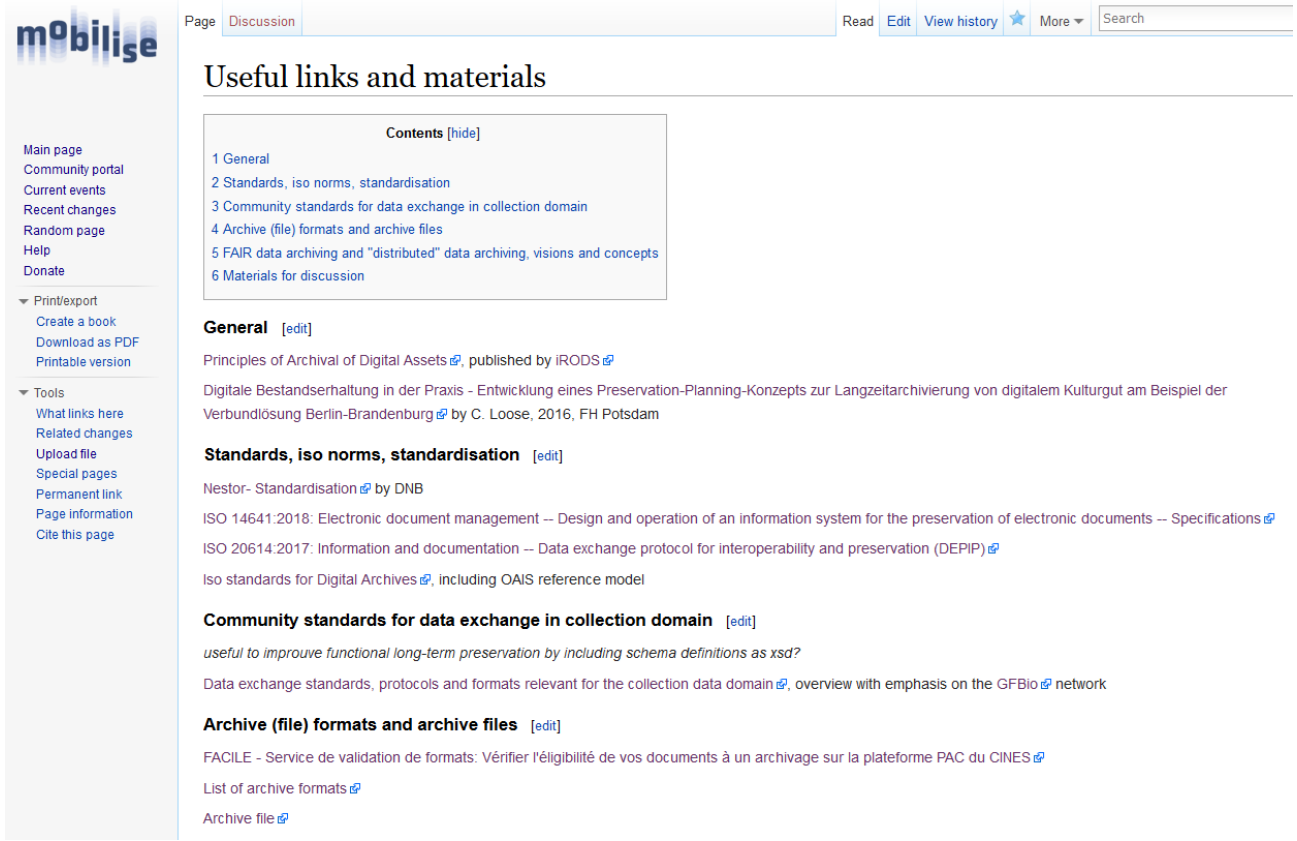
Formats that PowerGREP can read and write, but which should be treated as document files, even if they are technically archives:

- **Zipped documents**: Office Open XML (MS Office 2007) and OpenDocument Format (OpenOffice) are technically ZIP archives containing multiple XML and other files.
- **CHM files**: HTML Help files consist of compressed HTML files and other files.

Formats that PowerGREP can decompress only:

- **ARJ archives**: Files compressed with ARJ
- **CAB archives**: Microsoft Cabinet
- **DEB packages**: Debian Linux installation packages
- **FAT images**: Disk images of FAT volumes such as floppy disks
- **ISO and UDF images**: CD and DVD images
- **LHA and LZH archives**: Files compressed with LHARC
- **RAR archives**: Files compressed with WinRAR
- **RPM packages**: Red Hat Linux installation packages
- **WIM images**: Windows Imaging disk images
- **XAR archives**: Files compressed with XAR

WG 4 links useful in archiving context



mobilise

Page | Discussion | Read | Edit | View history | More | Search

Main page
Community portal
Current events
Recent changes
Random page
Help
Donate

Print/export
Create a book
Download as PDF
Printable version

Tools
What links here
Related changes
Upload file
Special pages
Permanent link
Page information
Cite this page

Useful links and materials

Contents [hide]

General [edit]

Principles of Archival of Digital Assets, published by iRODS

Digitale Bestandserhaltung in der Praxis - Entwicklung eines Preservation-Planning-Konzepts zur Langzeitarchivierung von digitalem Kulturgut am Beispiel der Verbundlösung Berlin-Brandenburg by C. Loose, 2016, FH Potsdam

Standards, iso norms, standardisation [edit]

Nestor- Standardisation by DNB

ISO 14641:2018: Electronic document management -- Design and operation of an information system for the preservation of electronic documents -- Specifications

ISO 20614:2017: Information and documentation -- Data exchange protocol for interoperability and preservation (DEPIP)

Iso standards for Digital Archives, including OAIS reference model

Community standards for data exchange in collection domain [edit]

useful to improuve functional long-term preservation by including schema definitions as xsd?

Data exchange standards, protocols and formats relevant for the collection data domain, overview with emphasis on the GFBio network

Archive (file) formats and archive files [edit]

FACILE - Service de validation de formats: Vérifier l'éligibilité de vos documents à un archivage sur la plateforme PAC du CINES

List of archive formats

Archive file

The **O**pen **A**rchival **I**nformation **S**ystem reference model

The OAIS reference model (ISO standard) is a widely accepted model. The terms, e.g. AIPs, are widely used to characterize archiving processes.

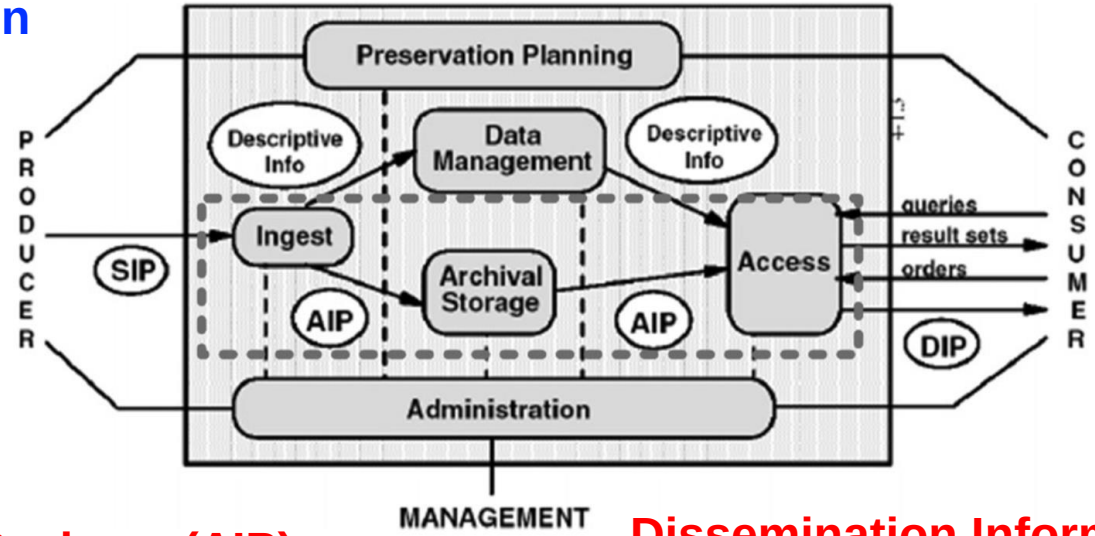
See, e.g., overview under

https://gfbio.biowikifarm.net/wiki/ISO_Standards_for_Digital_Archives

Open Archival Information System (**OAIS**)

Three types of information packages

Submission Information Package (SIP) - Delivered by the Producer to the OAIS for use in the construction or update of one or more AIPs and/or the associated Descriptive Information.



Archival Information Package (AIP) - Consists of the Content Information and the associated Preservation Description Information (PDI), which is preserved within an OAIS

Dissemination Information Package (DIP) - Derived from one or more AIPs, and sent by Archives to the Consumer in response to a request to the OAIS.

Three types of information packages

- **Submission Information Package (SIP)** = Submission information packages: archiving of original incoming files and data objects (binary data and text-based data) with documentation in detail (e. g. xls, jpg) - archival storage with timestamp
- **Archival Information Package (AIP)**= Archival information packages, e. g. renamed and newly organised information packages (binary data and text-based data) following the internal data structure (RDMS, multimedia data storage; see Relational database management systems under Technical Documentations) and information organisation of the single archives - periodical archival storage with timestamp
- **Dissemination Information Package (DIP)**= Dissemination information packages (binary data and text-based data) deviate substantially from SIP and AIP concerning, data structure, format and data content. DIPs (e. g. ABCD xml archives; SDD xml archives) have to be archived separately - (periodical) archival storage with timestamp

Data archiving as part of a cloud-based infrastructure solution

Cloud-Based solution - Pros and cons

Most basic Cloud service model is the Infrastructure as a Service (IaaS) model, that provides virtual computing equipments.

Implementing an OAIS model in a Cloud IaaS means all OAIS components are deployed into one or more Virtual Machine running in Cloud datacenters, in order to guarantee scalability and elasticity of adopted resources.

What about Security aspects?!