IJDC | Conference Paper

Building LABDRIVE, a Petabyte Scale, OAIS/ISO 16363 Conformant, Environmentally Sustainable Archive, Tested by Large Scientific Organisations to Preserve their Raw and Processed Data, Software and Documents

David Giaretta
Giaretta Associates Ltd

Teo Redondo
LIBNOVA SRL

Abstract

Vast amounts of scientific, cultural, social, business and government, and other, information is being created every day. There are billions of objects, in a multitude of formats, semantics and associated software. Much, perhaps the majority, of this information is transitory but there is still an immense amount which should be preserved for the medium and long term – perhaps even indefinitely.

Preservation requires that the information continues to be usable, not simply to be printed or displayed. Of course, the digital objects (the bits) must be preserved, as must the "metadata" which enables the bits to the understood which includes the software.

Before LABDRIVE no system could adequately preserve such information, especially in such gigantic volume and variety.

In this paper we describe the development of LABDRIVE and its ability to preserve tens or hundreds of petabytes in a way which is conformant to the OAIS Reference Model and capable of being ISO 16363 certified.

Submitted date 12 March 2022 ~ *Accepted* date 22 April 2022

Correspondence should be addressed to David Giaretta, 5 Batten Drive, Sherborne, DT9 4GD, UK. Email: david@giaretta.org

This paper was presented at International Digital Curation Conference IDCC22, Edinburgh, 13-16th June 2022

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

Introduction

The ARCHIVER Project (Archiving and Preservation for Research Environments) is the only EOSC related H2020 project focussing on commercial long-term archiving and preservation services for petabyte-scale datasets across multiple research domains and countries.[1]

On 29 January 2020, the CERN-led ARCHIVER project launched its Pre-Commercial Procurement Request for Tenders with the purpose to award several Framework Agreements and work orders for the provision of R&D for hybrid end-to-end archival and preservation services that meet the innovation challenges of European Research communities, in the context of the European Open Science Cloud.[2] Five consortia were selected to start.

The process has proceeded through an agile process for development, prototyping and finally piloting; at the end of the first two stages the number of consortia has been reduced. The LIBNOVA-led consortium is one of the two finalists in the Pilot phase.

This paper describes some of the challenges which have been overcome to develop what existed previously into what is now being implemented.

Background

In order to set the proper context, several decades of directed research can be summarized in the following way:

- Theoretical work on digital preservation of all types of encoded information resulted in the OAIS reference model (2020).
- The theoretical solutions were tested in the CASPAR project, which used as examples many objects in each of scientific, cultural and contemporary performing arts domains.[3]
- The requirements of users were sought in PARSE.Insight, which conducted the largest survey of users across the world, across disciplines and backgrounds, to identify the most widely held concerns in digital preservation. There were enough responses to be sure of the conclusions because there was significant agreement between groups of disciplines and even groups across countries.[4]
- Specific challenges of linked data and ontologies were investigated in the PRELIDA project.[5]
- The application of OAIS concepts to existing archives was carried out in the SCIDIP-ES project.[6]

1 ARCHIVER Project Fact Sheet, H2020, CORDIS, European Commission: https://cordis.europa.eu/project/id/824516

2 ARCHIVER Pre-Commercial Procurement request for tenders: https://cordis.europa.eu/article/id/413444-archiver-launches-its-pre-commercial-procurement-tender

3 CASPAR project: https://cordis.europa.eu/project/id/033572

4 PARSE.Insight: https://cordis.europa.eu/project/id/223758

5 PRELIDA project: https://cordis.europa.eu/project/id/600663

6 SCIDIP-ES project: https://cordis.europa.eu/project/id/283401

- Looking at the broader context of digital preservation, and in particular identifying how to justify, control and provide funding for preservation was undertaken in the APARSEN project.[7]

- Standards for the ISO Certification of archives (ISO 16363[8] and 16919[9]) were created, and OAIS was updated, based on the lessons learned.

- LIBNOVA has focused, for more than a decade, on providing the most advanced digital preservation platform, constantly researching and innovating to create the LIBSAFE technology. Year after year, the boundaries of what is possible in digital preservation have been pushed, incorporating innovations that empower the organizations to preserve their content in an easier and more efficient way, including, for example, a machine-learning project that applies the latest developments in neural networks helping users to be more efficient in several phases of the digital preservation cycle, specifically generating values for metadata fields based on multi-class classification of natural language information.

The LIBNOVA consortium, leading work within ARCHIVER, shows the application of all this research and experience to PETABYTE scale and beyond to provide solutions for challenging scientific and documentary archives, supporting the preservation, access and re-use of digitally encoded information.

ARCHIVER Challenges

The ARCHIVER challenges included:

- Support of tens of PBs of scientific data volume with linear growth over the years and sustained data ingest rates capabilities from 1-10 GB/s.

- Follow best practices for storage infrastructure as foreseen in ISO 27001[10]/27040[11]/19086[12] and ISO 14721[13]/16393[14], ISO 26324[15] and CoreTrustSeal[16] for long-term preservation of data.

- Support of vendor independent standards (such as PREMIS[17], METS[18] and Bagit[19]), interfaces and generic APIs (such as OpenAPI[20], RESTful API[21], etc.) to allow implementation of exit plans and prevent vendor lock-ins.

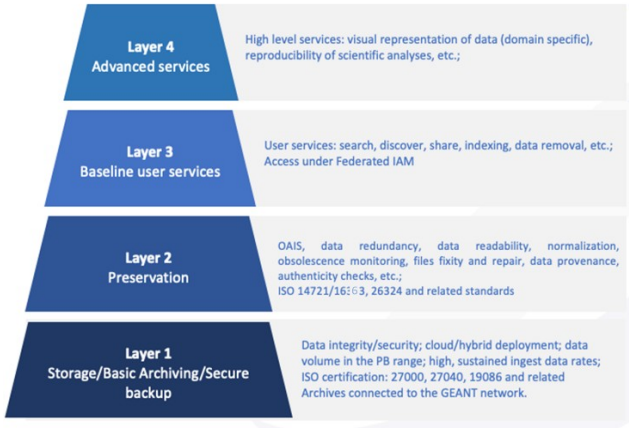
- Take into account practices as foreseen in the SWIPO Codes of Conduct[22].

7 APARSEN project: https://cordis.europa.eu/project/id/269977
8 https://public.ccsds.org/Pubs/652x0m1.pdf or later
9 https://public.ccsds.org/Pubs/652x1m2.pdf or later
10 https://www.iso.org/isoiec-27001-information-security.html
11 https://www.iso.org/standard/44404.html
12 https://www.iso.org/standard/67545.html
13 https://www.iso.org/standard/57284.html or https://public.ccsds.org/Pubs/650x0m2.pdf
14 https://public.ccsds.org/Pubs/652x0m1.pdf and https://www.iso.org/standard/56510.html
15 https://www.iso.org/standard/81599.html
16 CoreTrustSeal certification requirements:
https://www.coretrustseal.org/why-certification/requirements/
17 https://www.loc.gov/standards/premis/
18 https://www.loc.gov/standards/mets/
19 https://www.rfc-editor.org/rfc/rfc8493.html
20 https://swagger.io/specification/
21 https://restfulapi.net/
22 https://scope-europe.eu/en/projects/swipo-code-of-conduct

- Demonstration of implementation of a "Privacy by Design" approach, that include all the technical and organisational measures in order to protect data under European legislation such as GDPR[23] and Free Flow of Data[24], ensuring sovereignty and self-determination of scientific datasets.

- Services that implement innovative business models, that demonstrate how the planning and the cost of long-term data archiving is taken into account, allowing one to archive and preserve data in agreement with public organisations' procurement cycles, data management plans, and beyond individual researchers grant periods.

The following diagram illustrates the four levels of functionality, from the basic bit storage to complex data use and re-use.



	Providing insight into complex datasets. Research reproducibility. Sustainable business models
	Search, filter and access metadata to judge its relevance. Federated authentication and authorization. Automated indexing
	Preserving not just the bits but also the knowledge associated with the bits. Avoidance of vendor lock-in.
	Scaling to tens of Petabytes and beyond, very high ingestion rates, in an environmentally sustainable way.

Figure 1. ARCHIVER Levels

Responses to Challenges

LIBNOVA's existing software had been commercially successful in the cultural heritage sector (libraries and traditional archives) but was not scalable to be able to cope with petabytes of data, nor could it be claimed to the OAIS conformant or capable of preserving scientific information. In order to get to the point that it could meet the challenges outlined above a number of significant developments had to be made. The following sections describes the main areas of development.

Bit Preservation to Petabyte Scalability

A number of existing, largely cloud-inspired, hardware and software technologies were leveraged to reach the required scalability and responsiveness. Virtualisation and containerisation, orchestrated using Kubernetes, allow massive horizontal scalability by using compute and storage only as required, and scale as needed. This also allows the carbon footprint of the archive to be minimised by essentially allowing unnecessary hardware to be turned off when not required. Flexibility and expandability in terms of the hardware used is obtained through standardised interfaces (or connectors) which allow either a commercial cloud provider such as Amazon or the archive's own hardware to be used.

[23] https://gdpr-info.eu/
[24] https://digital-strategy.ec.europa.eu/en/policies/non-personal-data

LIBNOVA's existing techniques of internal, configurable, services to carry out the creation hashes, identify formats, and check for viruses, were modified to take advantage of the scalable hardware provision so that billions of objects can be processed in a reasonable time.

This cloud-related approach brings with it the configurability of the storage used, the scheduling of events, the workflow and many more capabilities, all at a massive scale.

Environmental Monitoring

As part of the sustainable business models that LABDRIVE has as an objective, the platform includes monitoring tools that allow users to have an almost immediate picture of the environmental impact of the preservation activities, mainly the storage and computing power involved. The images below show some of the basic information provided (overall monitoring in a time series, sources of energy and relative consumption, and geographical location of the data centers involved).

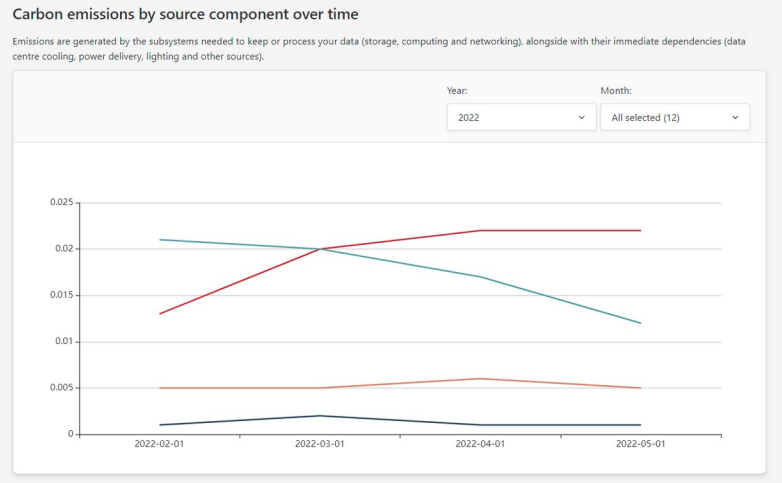


Figure 2. Overall monitoring in a time series

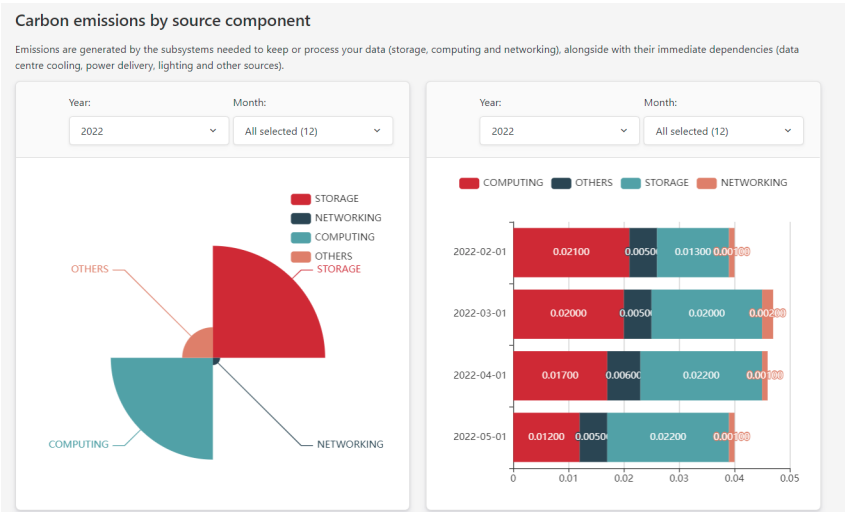


Figure 3. Energy sources and relative consumption

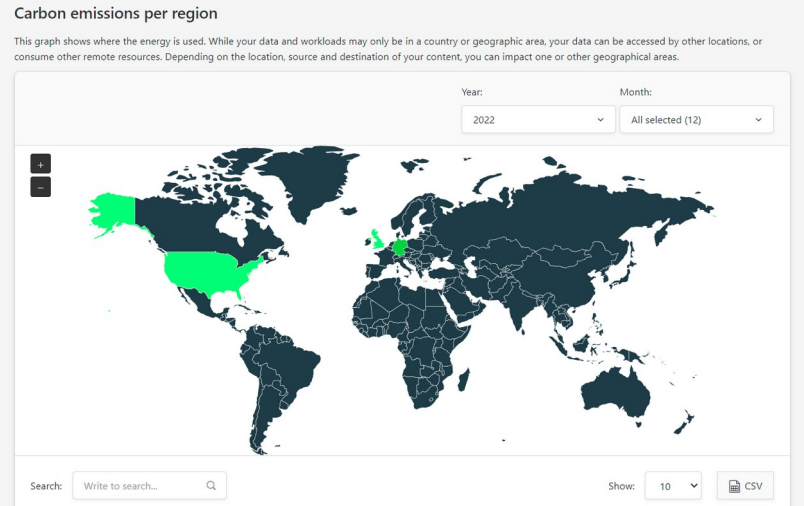


Figure 4. Geographical location of the data centers involved

OAIS Information Model Support

Commercial archive software systems all tend to support the normal requirements of document archives, including Dublin Core[25], and various descriptive standards such as ISAD(G), ISAAR, etc. LIBNOVA also provided a very flexible, configurable, extensible, metadata schema. This allowed a schema which contains the full OAIS Information Model to be associated with every digital object in the archive[26]. This means that any data object can have its complete OAIS Archival Information Package (AIP)[27]:

[25] Dublin Core specifications: https://www.dublincore.org/specifications/dublin-core/

[26] LABDRIVE support for OAIS Conformance https://docs.libnova.com/labdrive/concepts/oais-and-iso-16363/labdrive-support-for-oais-conformance

[27] Exporting Archival Information Packages: https://docs.libnova.com/labdrive/concepts/oais-and-iso-16363/labdrive-support-for-oais-conformance#exporting-archival-information-packages

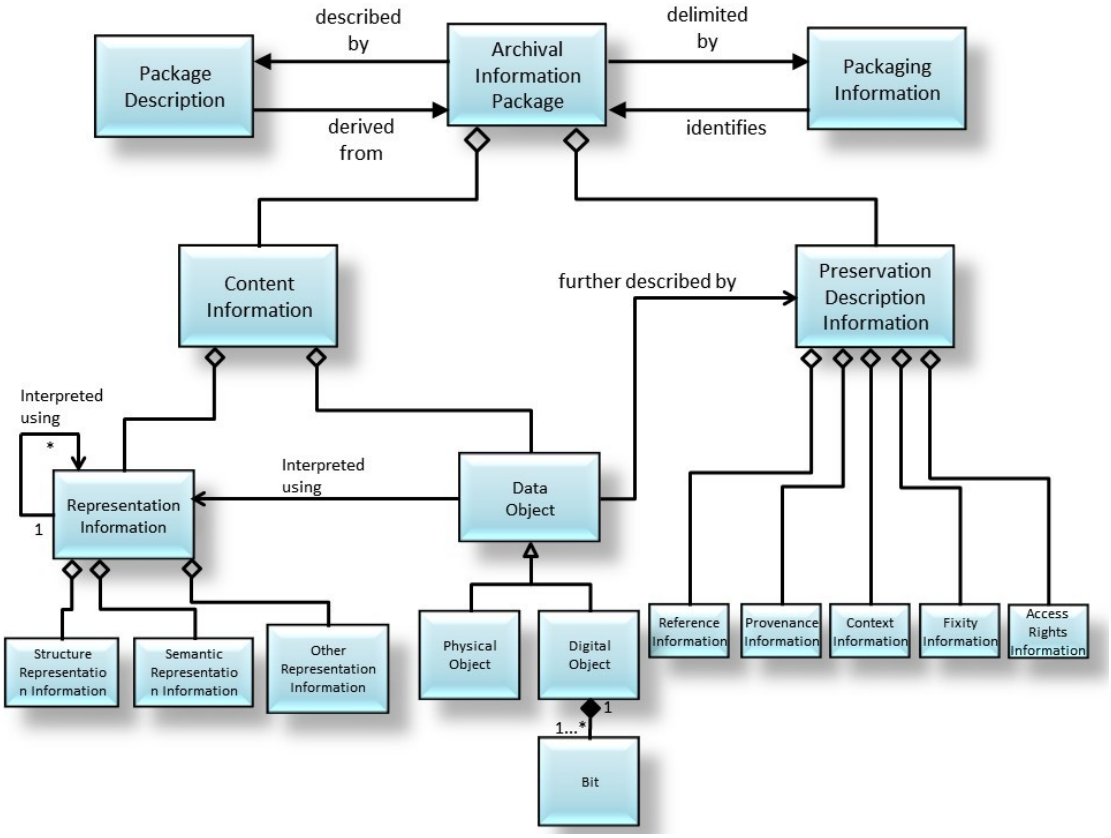


Figure 5. OAIS Information Model for AIP

The schema is shown below. The schema allows the various components to be either simple text or other objects in the archive or outside the archive.

OAIS Archival Information Package

This schema collects the mandatory pieces of information required for a complete OAIS AIP.

+ Add Descriptor ...

Search: Write to search...

#	Name	Description	Data Type	IECode
1	Structure Representation Information	The Representation Information that imparts information about the arrangement of and the organization of the parts or elements of the Data Object	Link Fields	structureRepInfo
2	Semantic Representation Information	The Representation Information that further describes the meaning of the Data Object, and its parts or elements, beyond that provided by the Structure Representation Information.	Link Fields	semanticRepInfo
3	Other Representation Information	A type of Representation Information which cannot easily be classified as Structure Representation Information or Semantic Representation Information, for example software. It is a type of Information Object.	Link Fields	otherRepInfo
4	Provenance	The DATA object for the information that documents the history of the Data Object. This information tells the origin or source of the Data Object, any changes that may have taken place since it was originated, and who has had custody of it since it was originated. The Archive is responsible for creating and preserving Provenance Information from the point of Ingest; however, earlier Provenance Information should be provided by the Producer. Provenance Information adds to the evidence to support Authenticity. This Data Object and its own Representation Information form a Provenance Information Object.	Link Fields	provenanceOais
5	Context	The DATA object for a set of information that is the original target of preservation. This Data Object and its own Representation Information form a Context Information Object.	Link Fields	contextOais
6	Reference	The DATA object for the information that is used as an identifier for the Data Object. It also includes identifiers that allow outside systems to refer unambiguously to a particular Object. This Data Object and its own Representation Information form a Reference Information Object.	Link Fields	referenceOais
7	Fixity	The DATA object for the information which documents the mechanisms that ensure that the object has not been altered in an undocumented manner. This Data Object and its own Representation Information form a Fixity Information Object.	Link Fields	fixityOais
8	Access Rights	The DATA object for the information that identifies the access restrictions pertaining to the object, including the legal framework, licensing terms, and access control. It contains the access and distribution conditions stated within the Submission Agreement, related to both preservation (by the OAIS) and final usage (by the Consumer). It also includes the specifications for the application of rights enforcement measures. This Data Object and its own Representation Information form a Access Rights Information Object.	Link Fields	accessRightsOais
9	Package Description	The information intended for use by Access Aids. It is a type of Information Object.	Link Fields	packageDescriptionOais

Figure 6. LABDRIVE metadata schema for OAIS Information Model

The full OAIS Information Model (as updated) is supported as illustrated in the following diagram showing an example OAIS Archival Information Package. Note that the example shows many arrows to an element which describes ASCII, because many of the examples were simple ASCII text in English.

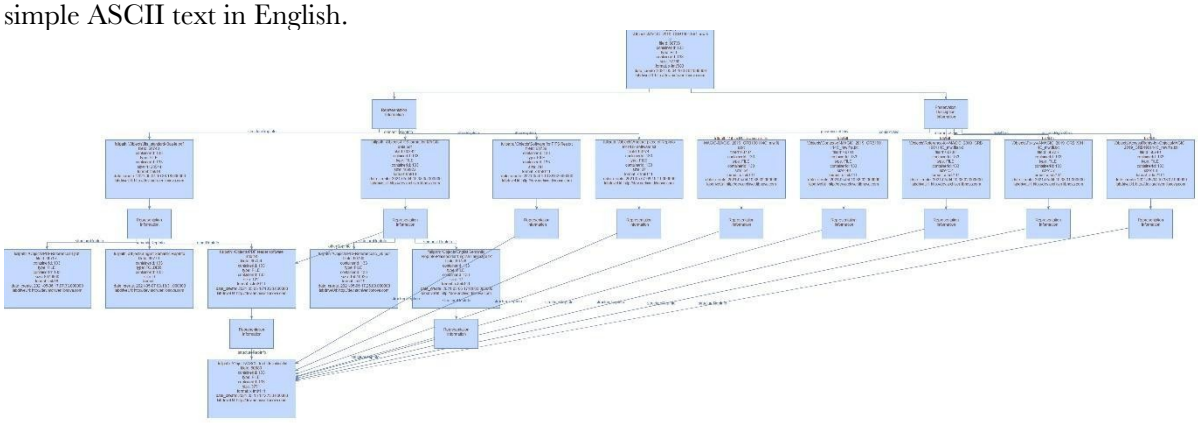


Figure 7. Example OAIS AIP in LABDRIVE

In order to export the full Archival Information Package, Bagit, for example, may be used because it is basically a container which can be configured in a way which can be consistent with an OAIS AIP as the following example shows.

```
|   bag-info.txt
|   bagit.txt
|   manifest-md5.txt
|   tagmanifest-md5.txt
|
└──data
       fetch.txt
       oais-aip-manifest.txt
```

Figure 8. Example BAGIT structure

An example of a ***fetch.txt*** file is as follows:

```
http://acme.libnova.com/download/file/962847 - data/MAGIC_2019_GRB190114C_mw.fits
http://acme.libnova.com/download/file/962850 - data/accessRightsOais/AccessRights-for-Obj
http://acme.libnova.com/download/file/962855 - data/semanticRepInfo/English Semantics
http://acme.libnova.com/download/file/962858 - data/structureRepInfo/ASCII_text_definitic
http://acme.libnova.com/download/file/962854 - data/fixityOais/Fixity-of-MAGIC_2019_GRB19
http://acme.libnova.com/download/file/962848 - data/referenceOais/Reference-for-MAGIC_201
http://acme.libnova.com/download/file/962851 - data/contextOais/Context-of-MAGIC_2019_GRB
http://acme.libnova.com/download/file/962846 - data/provenanceOais/Provenance for MAGIC-M
http://acme.libnova.com/download/file/962849 - data/otherRepInfo/Software for FITS files.
http://acme.libnova.com/download/file/962852 - data/semanticRepInfo/FITS format for MAGIC
http://acme.libnova.com/download/file/962857 - data/otherRepInfo/PDF reader software info
http://acme.libnova.com/download/file/962859 - data/semanticRepInfo/Placeholder-English-I
http://acme.libnova.com/download/file/962862 - data/structureRepInfo/PDFReference15_v6.pd
http://acme.libnova.com/download/file/962853 - data/structureRepInfo/fits_standard40aa-le
http://acme.libnova.com/download/file/962861 - data/structureRepInfo/PDFReference14.pdf
```

Figure 9. Example fetch.txt for OAIS AIP in LABDRIVE

This identifies all the component files (one must have the correct permissions to allow the links to be downloaded). The file "***oais-aip-manifest.txt***" defines the structure of the AIP, specifically identifying which file is the Content Data Object, which is its Context etc. as the following example shows.

```
ContentDataObject: data/MAGIC_2019_GRB190114C_mw.fits
structureRepInfo: data/structureRepInfo/fits_standard40aa-le.pdf
semanticRepInfo: data/semanticRepInfo/FITS format for MAGIC data.pdf
otherRepInfo: data/otherRepInfo/Software for FITS files.txt
otherRepInfo: data/otherRepInfo/Another piece of RepInfo for PDF software.txt
provenanceOais: data/provenanceOais/Provenance for MAGIC-MAGIC_2019_GRB190114C_mw.fits.t
contextOais: data/contextOais/Context-of-MAGIC_2019_GRB190114C_mw.fits.txt
referenceOais: data/referenceOais/Reference-for-MAGIC_2019_GRB190114C_mw.fits.txt
fixityOais: data/fixityOais/Fixity-of-MAGIC_2019_GRB190114C_mw.fit.txt
accessRightsOais: data/accessRightsOais/AccessRights-for-ObjectsMAGIC_2019_GRB190114C_mw

DataObject: data/accessRightsOais/AccessRights-for-ObjectsMAGIC_2019_GRB190114C_mw.fits.
structureRepInfo: data/structureRepInfo/ASCII_text_definition.txt

DataObject: data/fixityOais/Fixity-of-MAGIC_2019_GRB190114C_mw.fit.txt
structureRepInfo: data/structureRepInfo/ASCII_text_definition.txt

DataObject: data/referenceOais/Reference-for-MAGIC_2019_GRB190114C_mw.fits.txt
structureRepInfo: data/structureRepInfo/ASCII_text_definition.txt
```

Figure 10. Example oais-aip-manifest.txt file showing AIP components

It is worth noting that e-ARK package structures are also supported but these are not, indeed cannot, be conformant with the OAIS Information Model because there is no place, for example, for Semantic Representation Information, which is essential even for something as simple as a text table; in this case the Semantic Representation Information would include the table column names and units.

OAIS Mandatory Responsibilities

As noted previously, these responsibilities apply to the organisation and not the software. However, one can describe what the software solution should support in order to enable the archive to meet its responsibilities.

1. **Negotiate for and accept appropriate information from information producers.**
    - LABDRIVE is able to check the Submission Information Packages (SIPs) to ensure that they are what is expected and have not been corrupted, having been defined to ensure the AIPs can be created. It can also allow the archive staff to add additional metadata to the packages.
    - LABDRIVE has automated workflows including automated collection of "metadata" of the various types defined by OAIS. Additionally, if the SIPs include, for example, Provenance Information, then there can be adequate Representation Information for the way it is encoded. The PREMIS standard is widely used in some domains; in this case the Representation Information would be the PREMIS standard as well as the specific vocabulary used. Other domains use other Provenance encodings, even "home-grown" systems, all of which would require their own Representation Information. LABDRIVE can support all these.
2. **Obtain sufficient control of the information provided to the level needed to ensure Long Term Preservation.**
    - LABDRIVE can preserve the proof of control e.g., to make copies. It also has extensive support for restrictions on access with configurable authentication and authorization capabilities.
3. **Determine, either by itself or in conjunction with other parties, which communities should become the Designated Community and, therefore, should be able to understand the information provided, thereby defining its Knowledge Base.**
    - LABDRIVE allows one to identify which Designated Community applies to which object being preserved by adding appropriate schema elements.
4. **Ensure that the information to be preserved is Independently Understandable to the Designated Community. In particular, the Designated Community should be able to understand the information without needing special resources such as the assistance of the experts who produced the information.**
    - LABDRIVE allows the repository to maintain as much of the Representation Information Network (RIN) as required, including identifying types of Representation Information and links between them, and allow staff to add Representation Information as required. The platform offers functionality to create/collect, and link Representation Information required Designated

Community, including those which are human-actionable as well as machine-actionable.

5. **Follow documented policies and procedures which ensure that the information is preserved against all reasonable contingencies, including the demise of the Archive, ensuring that it is never deleted unless allowed as part of an approved strategy. There should be no ad-hoc deletions.**
    - LABDRIVE:
        - can be configured to keep as many backup copies, distributed geographically and over different technologies, as desired, with periodic fixity checks. Deletion policies can be configured, with multiple authorizations required.
        - maintains all the Information Objects' types, defined by OAIS, of metadata, with interfaces to add, edit, import, export or search it.
        - supports the handover of all the information, in particular complete AIPs, to another repository in such a way that the other repository can extract the components of the AIPs are required.
    - Day-to-day administration is well supported as well as decision support for Management in terms of Preservation Strategies to configure the Archive, taking into account costs, both in terms of financial resources as well as environmental burden, and risks.
6. **Make the preserved information available to the Designated Community and enable the information to be disseminated as copies of, or as traceable to, the original submitted Data Objects with evidence supporting its Authenticity.**
    - LABDRIVE is able to construct DIPs in a flexible way to support changing demands of all types of consumers, as well as members of the Designated Community. In addition to specific interfaces such as Web GUIs, a general API to query and access holdings allows users to create their own applications.
    - Provenance Information, to support claims of Authenticity, can be provided, ranging from the origins and previous custodians of the preserved objects as well as detailed events within the repository.

OAIS Conformance Summary

Full OAIS conformance cannot be achieved by software alone but requires processes and procedures; the way in which LABDRIVE can provide detailed templates for these.

Putting these together all the requirements for OAIS conformance, namely support for the OAIS Information Model and the mandatory responsibilities, are met and so LABDRIVE is OAIS conformant.

It must be noted that preservation is also important for what may collectively be termed "metadata". For example, a PREMIS file which records the Provenance of an object must itself be preserved. In particular one must ensure that the PREMIS file has not been changed, so its Fixity must be recorded as with any scientific object. Similarly, the Access Rights to that PREMIS file must ensure that it cannot be changed by unauthorised people. Also, the Representation Information for the PREMIS file must be available, for example the version of the PREMIS format (Structure Representation Information), and the semantics of the controlled vocabulary used (Semantic Representation Information), which may be quite

different from the Library of Congress vocabulary. LABDRIVE enables the PREMIS file to be preserved in the same, OAIS conformant, way as a scientific data file.

In the same way, each piece of Representation Information can have its own Representation Information (as well as Provenance, etc.) and so a full Representation Information Network can be constructed.

Preservation Options

LABDRIVE supports the three basic options for preservation which are described in OAIS as follows[28]. The Digital Object of the Information Object may be:

1. Kept by the archive unchanged or
2. Kept by the archive but may be changed i.e., Transformed or
3. Handed over to another archive

Each of these are supported by LABDRIVE as follows:

- In case (1) the archive must keep the bits unchanged, using multiple copies and regular checks by recalculation of the hashes. As time passes additional Representation Information can be added, using a variety of GUI and command line methods in LABDRIVE to ensure the Information is independently understandable.
- In case (2) the archive may use LABDRIVE to Transform the Data Object – using a variety of built-in functions and external applications. The Transformational Information Properties, associated with the Data Object, can be checked, for example using procedures in Jupyter Notebooks. The new Data Object must have appropriate Representation Information and Provenance.
- In case (3) a complete Archival Information Package can be exported, as described above.

ISO 16363 Conformance

At each stage of the Archiver project evaluations against OAIS, ISO 16363, CoreTrustSeal, and FAIR were demanded, and these evolved from self-evaluations at each stage to actual evaluations with a real archive, including detailed templates for contracts, procedures and processes, in a full third-party ISO 16363 formal audit.

The performance evaluations tested the network response, ingestion rate, scalability and single sign on with federated identifiers. At every stage issues were discovered which required work arounds, not with the core software but with the standardized services to which the core connected.

OAIS Information Model conformance, for preservation, reproducibility and re-purposing was tested against complex digitally encoded information in the domains of astronomy, high energy physics with data, software and documentation, bioinformatics and laser-neutron science.

ISO 16363 certification is described in the LABDRIVE documentation[29].

[28] Preservation Activities https://docs.libnova.com/labdrive/data-curation-and-preservation-1/oais-based-information-preservation-curation-and-exploitation/preservation-activities

[29] ISO 16363 certification guide https://docs.libnova.com/labdrive/concepts/oais-and-iso-16363/iso-16363-certification-guide

Examples of Preservation and Re-usability

LABDRIVE has the ability to associate extensive Representation Information Networks with any object within an archive. It can also associate extensive Provenance, captured in a variety of objects, for example, lists of events, PREMIS files, or entries in the headers of data files, with the objects. This allows LABDRIVE to support the reproducibility of research, because a user can obtain and understand the steps involved in the research.[30] In particular, even complex software used in any of these steps can be preserved.[31]

The following provides some specific examples of the use of LABDRIVE to preserve information.

Scientific Information

One of the key distinctions, stated in a broad-brush way, between what may be referred to as rendered information, such as images and documents, compared to scientific information, is that for the latter the Semantics of the elements must be available and cannot be left to the guesses of human observers. For example, if a simple text table, with rows and columns of numbers, could be printed in 50 years' time then this would be regarded as successful preservation, if the table were regarded as a document. On the other hand, if that table represents scientific data, then without the meaning and units of the columns, that table would be unusable as scientific information. The authenticity of scientific information is important, for example, in areas such as climate change; authenticity is also important for rendered objects. s

The provenance of the scientific data from the original data processing centre, where it may have been kept for processing and immediate use may have been maintained in one of many different ways, some specific to that data centre. LABDRIVE can take that Provenance in whatever way it is encoded and preserve it using the OAIS conformant ways described above.

The Representation Information, including semantics and software (as Other Representation Information) will also have their own Representation Information to ensure that it can be used in future to understand, use, and re-use, the scientific data.

Software and Processing Preservation

LABDRIVE supports reprocessing through virtual machines, of all types, code in Jupyter notebooks, and source code and build systems for very flexible, tailorable, processing. The software objects themselves can be preserved as with any other digital object with LABDRIVE, with as extensive a Representation Information Network as is required, which could include anything from software build systems to CPU instruction sets.[32]

Of particular importance for software which has been containerized, to deal with very large amounts of data, if the issue is the unavailability of the original computer hardware, preservation of container engine and all the container software, could be achieved using a hardware simulator such as QEMU, so that the container engine itself may be left unchanged. Otherwise, the system would rely on the preservation (including replacement) of the container engine, as would be done for any other piece of software.

---

[30] Reproducing research: https://docs.libnova.com/labdrive/data-curation-and-preservation-1/oais-based-information-preservation-curation-and-exploitation/reproducing-research

[31] Preserving complex software: https://docs.libnova.com/labdrive/data-curation-and-preservation-1/oais-based-information-preservation-curation-and-exploitation/preservation-activities/adding-representation-information/other-representation-information/software-as-part-of-the-rin/preserving-complex-software

[32] Software as part of the RIN: https://docs.libnova.com/labdrive/data-curation-and-preservation-1/oais-based-information-preservation-curation-and-exploitation/preservation-activities/adding-representation-information/other-representation-information/software-as-part-of-the-rin

LABDRIVE and FAIR

The FAIR Guiding Principles (Wilkinson et. al., 2016) for scientific data management and stewardship are being used by many repositories to show that they have ensured that their data is valuable by being easier to find through unique identifiers and easier to combine and integrate. The principles provide a checklist when managing scientific and other data to help making decisions which will enable the data to be more useful.

LABDRIVE's support for FAIRness for an archive has been evaluated in detail and had been shown to fully support the FAIR Principles.[33]

Acknowledgements

We would like to acknowledge the ARCHIVER project under Grant agreement ID: 824516, and our colleagues in the LIBNOVA consortium, namely CIC – IFCA, University of Barcelona, Amazon Web Services (AWS), Voxility and Bidaidea.

[33] LABDRIVE support for FAIRness:
https://docs.libnova.com/labdrive/concepts/oais-and-iso-16363/labdrive-support-for-fairness

References

ICA ISAD(G): General International Standard Archival Description second edition or later https://www.ica.org/en/isadg-general-international-standard-archival-description-second-edition

International Standard Archival Authority Record (Corporate bodies, Persons, Families) (ISAAR) - 2nd edition, 2003 https://www.ica.org/en/isaar-cpf-international-standard-archival-authority-record-corporate-bodies-persons-and-families-2nd

International Standard For Describing Institutions with Archival Holdings (ISDIAH) - 1st edition, March 2008 https://www.ica.org/en/isdiah-international-standard-describing-institutions-archival-holdings

International Standard for Describing Functions (ISDF) - 1st edition, May 2007 https://www.ica.org/en/isdf-international-standard-describing-functions

Information and documentation – Records management Part 1: General, ISO 15489 (2001)

Information and documentation -- Principles and functional requirements for records in electronic office environments ISO 16175-1:2010, ISO 16175-2:2011 and ISO 16175-3:2010 available from https://www.iso.org/search.html?q=16175

Model Requirements for Records Systems, MoReq2010 ® see https://www.moreq.info/files/moreq2010_vol1_v1_1_en.pdf

Reference Model for an Open Archival Information System (OAIS) Draft Recommended Practice CCSDS 650.0-P-2.1 (Pink Book) Issue 2.1 October 2020. Retrieved from https://public.ccsds.org/Lists/CCSDS%206500P21/650x0021.pdf

Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18