Digital Preservation: How to Be Trustworthy

Lina Bountouri1(✉), Patrick Gratz2, and Fulgencio Sanmartin1

1 Publications Office of the European Union, Luxembourg, Luxembourg
{Vasiliki.BOUNTOURI,fulgencio.sanmartin}@publications.europa.eu
2 infeurope, Luxembourg, Luxembourg
pgratz78@googlemail.com

Abstract. The Publications Office of the European Union has started a project for the long-term preservation of its digital publications to a new digital archival repository that contains legislative collections (such as the Official Journal of the European Union, treaties, international agreements, etc.), non-legislative collections (such as general and scientific publications), master data (such as descriptive, technical and provenance metadata specifications) and other data (such as datasets and websites). With the aim of safeguarding EU digital publications without any alteration during their life cycle, we have decided to follow standards ISO 14721:2012 (Open Archival Information System) to define the model of our digital preservation system and ISO 16363:2012 to verify the trustworthiness of the digital archival repository. In this context, we will deal with the following issue: how can we be sure that a digital object is the same as when it was created and has not been altered during its life cycle, both before and after its ingestion to the repository? In other words, how can our digital archival repository be trusted? The basic actions of the Publications Office towards this direction are: (a) to define a digital preservation plan; (b) to define and preserve representation information (master data and other specifications); (c) to define the designated community and its monitor; (d) to define and implement digital preservation strategies such as fixity, maintaining a read-only archive, keeping two copies in different physical data centres, etc.; (e) to define and implement provenance metadata; and (f) to have a technology watch of formats, standards and digital preservation strategies.

Keywords: Digital preservation · OAIS · Provenance · Authenticity · PREMIS

1 Introduction

One of the main problems that cultural heritage institutions, publishers and other organisations have to cope with is the efficient management of the digital resources they manage and/or produce in order to ensure their long-term digital preservation. Nowadays, a growing number of resources is digitised and/or digitally produced. How can archivists and information professionals ensure future access and long-term preservation to these resources, especially when software and hardware obsolescence have a direct effect on their management?
In order to tackle this issue, various institutions have recently started dealing with digital preservation/digital archiving. In line with [1], digital preservation is ‘the series of managed activities necessary to ensure continued access to digital materials for as long as necessary’, despite the problems that may occur because of any media failure or technological change. These activities both refer to the preservation of digitally born and to digitised resources.

It is of vital importance to separate the preservation of resources through digitisation and digital preservation, given that they are two distinct tasks. In the former, practitioners implement preservation strategies that hope to achieve the physical protection of analogue resources; while in digital preservation, practitioners try to ensure the continued access and usability of digital resources [2]. As a consequence, digital preservation issues differ from those that arise in the ‘traditional’ preservation field. This is mainly due to the subject of preservation, which in the first case is a digital resource while in the second case it is an analogue resource [3]. However, both processes may share some central ideas such as coping with the protection and long-term access to the resources, and both can be in danger, regardless of their analogue or digital substance.

The big question remains: how can information managers and archivists ensure future access to digital resources, especially when software and hardware obsolescence are having a direct impact on their management? In this paper we will present the main actions taken by the Publications Office of the European Union (OP) [4] towards the development of a trustworthy digital archival repository with the purpose of preserving its valuable digital publications over the long term. With this in mind, we will analyse the steps as well as the standards followed (such as the Open Archival Information System (OAIS) [5]) in order to create a trustworthy digital archival repository.

This paper is structured as follows: Sect. 2 presents the OP’s mandate for digital preservation and the implementation of its project for the creation of a trustworthy long-term digital archival repository; Sect. 3 presents all the action taken by the OP in order to support the long-term digital preservation goal as well as the future plans towards this direction; and Sect. 4 presents the lessons learnt from this effort and the future goals of the OP.

2 Digital Preservation in the Publications Office of the European Union

The OP has the legal mandate to manage and provide a long-term digital preservation service. The legal basis is defined by the following official documents:

- Decision 2009/496/EC on the organisation and operation of the OP, where it is stated that the part of the mandate of the OP is to preserve ‘the publications of the European Communities and the European Union’ [6]; and
- Council Regulation (EU) No 216/2013, which defines that with regard to the authentic electronic version of the Official Journal of the European Union, that the OP is responsible for ‘preserving and archiving the electronic files and handling them in line with future technological developments’ [7].
The OP has recently launched a project to migrate its existing archived digital publications to a new digital archival repository in order to allow their long-term digital preservation. This repository contains legislative collections (such as the Official Journal, treaties, international agreements, etc.), non-legislative collections (such as general and scientific publications), master data (such as descriptive, technical and provenance metadata specifications) and other data (such as datasets or websites). These publications are significant, in particular the authentic Official Journal, which has had legal value in its digital format since 2013 [7]. In this context, all efforts are targeted to safeguard these publications in the long term while preserving their integrity and authenticity.

3 Standards Compliance and Digital Preservation Policy

Based on the analysis of the bibliography and related projects in the digital preservation field, we have concluded that there are no perfect solutions or risk-free options. With the aim of safeguarding EU digital publications without any alteration during their life cycle, we have decided to follow the current internationally accepted standards: ISO 14721:2012 (OAIS) [5] to define the model of our digital preservation system and ISO 16363:2012 [8] to verify the trustworthiness of the digital archival repository.

The OAIS reference model is a conceptual framework for an archival system dedicated to preserve and maintain long-term access to digital information for a designated community, which is an identified group of users who should be able to understand the preserved information [9]. The OAIS helps archival and non-archival institutions become familiarised with the preservation procedures by providing the fundamental concepts for preservation and its related definitions. The objective is to avoid any confusion on the used digital preservation terminology. As a conceptual framework, the OAIS does not provide guidelines on policy issues, such as the standard metadata schema to use or the preservation strategy to implement. Nonetheless, it clearly states that ‘representation information’, including all kinds of information needed to interpret a data object (i.e. metadata schemas, Knowledge Organisation Systems (KOS), ontologies, documentation, style and format guidelines, etc.), should also be preserved so as to make the objects in the archival repository self-explanatory and self-contained. The OAIS can also be implemented and act as a solid basis for the certification of a digital archival repository, but it was not written to act as an audit or certification manual for archival repositories. For this purpose, the International Organisation for Standardisation has published the ISO 16363 reference standard, which provides analytical guidance for auditing an archival repository. The OP is in the process of implementing the OAIS in the interest of obtaining the ISO 16363 certification for its archival repository.

In this context, we will deal with the following issue: how can we be sure that a digital object is the same as when it was created and has not been altered during its life cycle, both before and after its ingestion to a digital archival repository? In other words, how can our digital archival repository be trusted? Both ISO 14721:2012 and ISO 16363:2012 emphasise the importance of the documentation, evidence and preservation of the semantics so as to achieve the aforementioned purposes. The more evidence is
provided for the preserved digital objects, the less the risk is of losing their integrity and authenticity.

With a view to support its digital preservation commitment, the OP has chosen a long-term digital archival repository, implemented mostly in Europe, which is called the Repository of Authentic Digital Records (RODA) [10]. RODA is a digital archival repository developed in Portugal in cooperation with the Portuguese national archive. This digital archival repository is open source and freely available to download. It is built on Fedora and can support the existing XML metadata schemas, such as the Encoded Archival Description (EAD) [11], the Metadata Encoding and Transmission Standard (METS) [12] and the Preservation Metadata: Implementation Strategies (PREMIS) [13]. In terms of preservation actions, the repository supports normalisation in ingest and other actions such as format conversion and checksum verification.

Apart from the implementation of the new software, the OP had to define related policies and take the following additional actions:

- definition of a digital preservation plan (DPP);
- definition and preservation of all possible representation information (master data and other specifications, such as XML syntax specifications);
- definition of the designated community of the long-term digital archival repository and its monitor;
- definition and implementation of fixity policies;
- definition and implementation of provenance metadata;
- technology watch of formats, standards and digital preservation strategies (i.e. migration, emulation and proactive digital preservation).

### 3.1 Digital Preservation Plan

The DPP [14] is the set of documented strategies for preserving the collections of an archival repository and is a prerequisite for its trustworthiness. The OP’s DPP is a key instrument to describe and share how the OP fulfils its obligations in the domain of long-term digital preservation. It defines and documents the vision and strategy of the long-term digital preservation service that the OP is managing on behalf of the EU institutions. This policy document is an official commitment of the OP for the provision of this service and is in the process of receiving the official approval of the EU institutions, after having already been internally approved.

The DPP defines the legal basis on which the OP is based in order to provide the digital preservation service as well as all the important definitions that will make the implementation of the digital preservation policy accurate and complete. For example, it defines what an EU digital publication is: all information in digital format produced by the EU institutions, bodies or agencies, either directly or on their behalf by third parties, and made available to the public. Moreover, the DPP defines the vision, the mission, the strategy and the scope of the digital preservation service. Though the OP’s long-term digital preservation service should aim to cover all EU digital publications, the current scope of this DPP is narrower: it will cover all EU publications whose custody has been transmitted by the EU institutions to the OP to be preserved.
Defining the preservation policy is one of the most significant parts of the digital preservation life cycle. It is not only the important directives and guidelines that this document provides in order to cope with technological and organisational issues; it also helps to build the ‘preservation culture’ inside an organisation by defining its commitment towards a specific digital preservation policy [3]. In [15], the author reports that in order to build a strong foundation for a DPP, the documentation of policies, procedures and standards is one of the most important steps in the digital preservation process.

3.2 Designated Community

According to the OAIS [5], a designated community is ‘an identified group of potential consumers who should be able to understand a particular set of information.’ Defining the designated community for each set of information that has to be preserved is significant, since this will also indicate the content of the knowledge base and of the representation information that are needed in order for the designated community to be able to interpret the data. An analysis of them will be presented in Sect. 3.3.

In the context of the OP’s long-term digital archival repository, the designated community was defined based on the collection preserved, given that each collection has its own characteristics. In reference to this, there are different designated communities, one for each preserved collection. For example, for our legislative collection (**Official Journal of the European Union, EU and national case-law and pre-legislative documents**), the designated community consists of professionals linked to law (such as lawyers, academics of this discipline and re-users), EU and national public authorities, and EU professionals that may use the legislative collections as part of their work.

3.3 Representation Information

One of the main goals of OAIS implementers is ‘to preserve information for a designated community’ [5]. In order for a designated community to be able to understand the preserved information, a long-term digital archival repository also has to preserve and/or refer to representation information of the preserved information.

Representation information is the information that maps a data object (which is metadata and content) into more meaningful concepts. It is expressed in many ways, depending on the content and context of each preserved information set. For example, the representation information of an EAD metadata record can be, among others, the EAD XML schema onto which the generation of the EAD metadata record was based, the XML schema specifications onto which the EAD XML schema was based, or the KOS and the authorities from which values inside the EAD metadata record have been taken, etc. Representation information must enable or allow the recreation of the significant properties of the original object, meaning that the information should be able to recreate a copy of the original object [16].

Understanding what representation information is and managing/storing it at the same time can be a painful task, with many parameters to take into account. It can be very helpful to have a knowledge base, which is ‘a set of information, incorporated by a person or a system that allows that person or system to understand the received
information’ [17]. In the OP there is a knowledge base that is also a collection in our
digital archival repository, called master data, which helps the interpretation of the
archived content and its respective metadata. Some of our master data are the following.

- Controlled vocabularies, such as authority tables for countries, corporate bodies,
  legislative procedures, types of decisions, etc. These vocabularies are used both
  within the OP and within the EU institutions and agencies. The complete list can be
  found online and reused on the Metadata Registry website [18]. Therefore, every set
  of preserved information must be related to these tables in order to be understandable.
- Ontologies, such as the Common Data Model [19], which is implemented in the
digital dissemination repository of the OP for encoding and validating descriptive
metadata.
- Grammars and schemas, which are the schemas or document type definitions used
  in the OP to encode the content of documents (such as Formex for the Official Journal
  and case-law) [20], and the METS profile of the OP used to wrap metadata and content
during ingestion of data to the digital dissemination repository and to the digital
archival repository of the OP.

Master data mostly provide information to interpret the metadata. Nevertheless, this is
not enough to interpret it as well as the digital content stored in our long-term digital
archival repository. Some of the extra representation information that is needed is the
following.

- Encoding languages specifications, such as the XML schema 1.0, the simple KOS
  (SKOS) specification, the Resource Description Framework Schema (RDFS) speci-
  fication, etc. (instead we provide a link to it in the namespace definitions of each
XML, SKOS or RDF file).
- Language dictionaries for the 24 languages of the European Union.
- The specifications of the PREMIS metadata schema [13], which is used to encode
provenance and preservation metadata (instead we provide a link to it in the name-
space definitions of each PREMIS file).
- Documentation, which can explain the various identifiers, structures and policies
  implemented such as the European Legislation Identifier [21] and the DPP.

Moreover, we refer to a format registry, such as PRONOM for having a stable reference
for the formats stored in the digital archival repository. The incorporation of additional
representation information as the one mentioned is in the future goals of the long-term
digital archival repository.

3.4 Fixity Policies

According to the PREMIS Data Dictionary [22], fixity is a property of a digital object
that indicates it has not changed between two points in time. Fixity checks have many
uses and they must always be encoded as part of the provenance/preservation metadata
since it can help to prove the authenticity and integrity of a digital object over time.
Fixity is often related to checksums of different algorithms, as it is associated with bit-
level integrity, though not exclusively: think of a workflow of files moving through many
temporary buffer folders; calculating the checksum for each step can delay the whole system, especially if it is real-time or big data ingestion. For these cases, other fixity methods could be acceptable, like file size, name or count. Checksums can be calculated at the beginning (reception) and at the end (archival) of systems.

In case a fixity check fails, the next step is to verify the other existing copies of a digital object. If the copy is satisfactory, it will substitute the incorrect object, and the long-term digital archival repository will register this event/action as well. If the copy is not satisfactory, then the digital object or part of it might be irrecoverable. If several (two or any other threshold defined) copies in the same physical support fail, then one option is to substitute the latter completely.

As part of our fixity policy, different fixities can be applied depending on the granularity of the digital objects collection, i.e. not only at the file level. There could also be fixity checks of a collection, controlling whether a particular file or even sub-collection (based on the language or the format of the collection) might go missing, but this only works for closed, well-defined collections. At the OP, we can think of a checksum of an Official Journal issue, including all its languages and formats. Granularity at different levels leads to another discussion: what to do in the case of updates on the digital objects? Depending on what is being modified (metadata or anonymisation of a page, for example), fixity to higher levels shall be recalculated.

Fixity also involves knowing the physical support in which the data are stored, and defining policies accordingly. It is well known that optical supports start to lose properties after 5 years, and most fail after 10 years. Magnetic supports keep data with electric power, which loses strength as time passes. Refreshing data (i.e. reading and rewriting the same bits) for magnetic supports and transferring for optical supports are the kind of periodical policies that should be taken into account. How often these actions will take place depends on many variables: the importance of the collection, when the content is extracted or if it will be migrated, and the financial means provided. A risk assessment will provide the answer to this question.

3.5 Provenance Metadata

In order to provide all the semantic information that is needed to support the digital preservation process, lately various models, standards and metadata schemas have been created and evolved. Digital preservation is an integral part of all the stages of resources management. Semantic information assigned to digitised/digitally born resources must be supplied, providing information on its content, structure, rights and technical characteristics. Providing explanations on the technical procedure of digitisation and digital creation will facilitate specialists involved with the digital preservation process to be aware of all the necessary information regarding the content, structure, rights and technical characteristics of the digital resource to be preserved [3].

One of the most significant parts of the semantic information needed to document the life cycle of a digital resource is the provenance metadata. Provenance metadata encode the custody of a digital object, in other words all the events that may have produced a change of any type during its life cycle [3]. As stated in [23], a provenance event is where ‘any event producing a change of the object has to be described and
documented at every stage in the life cycle to have, at any time, a sort of authenticity card for any object in the repository: the crucial point is to clearly state that the identity of an object resides not only in its internal structure and content but also—and maybe mostly—in its complex system of relationships, so that a change of the object refers not only to a change of the bits of the object but also to something around it and that anyway contributes to its identity, i.e. to its authenticity.’ Moreover, according to the OAIS standard [5], provenance metadata can support the authenticity of a digital object.

A long-term digital archival repository is responsible for generating provenance metadata starting from the ingestion of the digital object to the repository; however, provenance metadata can be provided in earlier stages of the life cycle of the digital object, such as by its producer and in different information systems than the long-term digital archival repository. It is advisable that the documentation of provenance metadata starts at the early stage of the creation of a digital object and that these metadata are also implemented inside the production systems and not only inside long-term digital archival repositories. Preservation metadata are a prerequisite for ensuring the authenticity and integrity of a digital object, and they encode the preservation actions taken on a digital object. Preservation actions are specific activities that are parts of preservation strategies, such as digitisation, integrity checks of digital objects, and policies such as migration and emulation. As previously mentioned, digital resources, similarly to the actions taken on analogue resources, have to be treated as a fragile object on which preservation rules and actions should apply.

In the OP, the implementation of provenance metadata is planned so that it starts in the early stages of a digital object’s life cycle. In detail, before its archiving in the long-term digital archival repository, a digital object passes through two additional information systems. The first system deals with the reception of metadata and digital content. During the reception workflow, the OP receives digital objects and metadata, while at the same time takes action on both of them, such as generating new formats of a digital object and transforming the metadata to richer structures. The second system oversees their storage and dissemination. In this system, different actions are taking place in parallel on the digital content and its respective metadata, such as modifications and deletions, etc. In this context, provenance metadata will be attributed during the reception and storage workflows so as to encode all the actions aforementioned. It is important to note that in the long-term digital archival repository, all provenance and preservation metadata are already encoded based on the PREMIS Data Dictionary [22]. The following is an indicative example with all the events encoded during ingestion:

- ingest start: the ingest process has started;
- unpacking: extraction of objects from package in the Submission Information Package (SIP) format;
- well-formedness check: checked that the received SIP is well formed and complete and that no unexpected files were included;
- well-formedness check: checked whether the descriptive metadata are included in the SIP and if these metadata are valid according to the established policy;
- message digest calculation: creation of base PREMIS objects with file original name and file fixity information (SHA-256);
format identification: identification of the object’s file formats and versions using Siegfried;
authorisation check: user permissions have been checked to ensure that they have sufficient authorisation to store the Archival Information Package (AIP) under the desired node of the classification scheme;
accession: addition of the package to the inventory—after this point, the responsibility for the digital content’s preservation is passed on to the repository;
ingest end: the ingestion process has ended;
replication: the replication of AIPs, its events and agents to RODA.

3.6 Technology Watch

The requirements of digital preservation and the action towards its long-term implementation involve continuous updating. This is due to the fact the technological obsolescence occurs so drastically: formats are updated to newer versions, software can be continuously updated and conformance may be in question. With the aim of being continuously updated on the new trends and evolutions on the digital preservation field, the OP team responsible for long-term preservation is focusing on the following resources and actions:

- participations to scientific conferences, scientific committees of the related discipline and seminars (such as the International Conference on Digital Preservation and the PREMIS Working Group for events);
- consultancy contract;
- list of events, news, etc.

4 Conclusion

Useful outcomes have been delivered during the long-term archiving projects of the OP. One of the most significant conclusions is that even if an information system for digital preservation/archiving is highly sophisticated and reliable, it always needs to be accompanied by an accurate and well-defined digital preservation policy. Moreover, implementers have to take into account the international standards of the field, which must be correctly implemented during the whole process of digital preservation. In addition, traditional archival theories and values, such as preserving ‘the custody of the archive’, are also implemented in the digital preservation field, in which they have equal importance.

Regarding the encoding of provenance metadata, it is important to note that there is no perfect solution: there is not one model or schema that covers all the documentation that is needed for preservation descriptive information. The OAIS provides the outline that must be followed when developing a long-term digital archival repository as well as guidelines on what kind of semantic information is needed for long-term preservation [3]. PREMIS focuses on encoding the preservation actions taking place before and during the ingestion of a digital object into an archival repository, while others, such as PROV-O and OPM, are instead focusing on encoding the provenance history.
Combining different schemas and models for encoding the most possible provenance information could be an adequate solution for ensuring the authenticity and integrity of a digital object.

References

1. Digital Preservation Coalition: Preservation management of digital materials: the handbook. Digital Preservation Coalition, October 2008. [http://www.dpconline.org/docs/digital-preservation-handbook/299-digital-preservation-handbook/file]. Accessed 04 Sept 2017
2. Deegan, M., Tanner, S.: Digital Preservation. Facet Publishing, London (2006). ISBN 9781856044851
3. Bountouri, L.: Archives in the Digital Age: Standards, Policies and Tools. Chandos Publishing, Cambridge (2017). ISBN 9781843347774
4. The Publications Office of the European Union: The Publications Office of the European Union, 11 September 2017. [https://publications.europa.eu/en/home]. Accessed 08 Sept 2017
5. International Organization for Standardization: ISO 14721, 2012: space data and information transfer systems, open archival information system (OAIS), reference model. International Organization for Standardization for Standardization (2012). [https://www.iso.org/standard/57284.html]. Accessed 28 Mar 2017
6. European Union: Decision of the organisation and operation of the Publications Office of the European Union. Official Journal of the European Union — Series L, vol. 168, pp. 41–47, 30 June 2009
7. Council of the European Union: Council Regulation (EU) No. 216/2013 of 7 March 2013 on the electronic publication of the Official Journal of the European Union. Official Journal of the European Union — Series L, vol. 69, 13 March 2013
8. ISO: ISO 16363, 2012: space data and information transfer systems, audit and certification of trustworthy digital repositories (2012). [https://www.iso.org/standard/56510.html]. Accessed 27 Mar 2017
9. The OCLC/RLG Working Group on Preservation Metadata: Preservation metadata and the OAIS information model: a metadata framework to support the preservation of digital objects, June 2002. [http://www.oclc.org/content/dam/research/activities/pmwg/pm_framework.pdf]. Accessed 27 Mar 2017
10. RODA: Welcome to RODA: an open-source digital repository designed for preservation, 08 September 2017. [https://demo.roda-community.org]. Accessed 09 Sept 2017
11. The Library of Congress: Encoded archival description: official site, 31 July 2017. [https://www.loc.gov/ead/]. Accessed 08 Sept 2017
12. The Library of Congress: Metadata encoding and transmission standard, 18 August 2017. [http://www.loc.gov/standards/mets/]. Accessed 08 Sept 2017
13. PREMIS Editorial Committee: PREMIS data dictionary for preservation metadata, 31 August 2017. [http://www.loc.gov/standards/premis/]. Accessed 08 Sept 2017
14. The Publications Office of the European Union: Digital Preservation Plan (2017)
15. McGovern, N.Y.: A digital decade: where have we been and where are we going in digital preservation? RLG DigiNews (2006)
16. OAIS 7: Representation Information. Alan’s notes and thoughts on digital preservation, 24 January 2008. [https://alanake.wordpress.com/2008/01/24/oa-oa7-representation-information/]. Accessed 08 Sept 2017
17. Rusbridge, C.: Representation Information: what is it and why is it important? 06 July 2007. http://www.dcc.ac.uk/news/representation-information-what-it-and-why-it-important. Accessed 08 Sept 2017
18. The Publications Office of the European Union: Metadata Registry (MDR) (2017). http://publications.europa.eu/mdr/index.html. Accessed 08 Sept 2017
19. The Publications Office of the European Union: Common Data Model (CDM), 21 July 2017. Accessed 08 Sept 2017
20. The Publications Office of the European Union: Formex: formalized exchange of electronic publications (n.d.). http://formex.publications.europa.eu/. Accessed 08 Sept 2017
21. The Publications Office of the European Union: ELI technical specifications, 21 September 2016. http://publications.europa.eu/mdr/eli/index.html. Accessed 08 Sept 2017
22. PREMIS Editorial Committee: PREMIS data dictionary for preservation metadata, November 2015. http://www.loc.gov/standards/premis/v3/premis-3-0-final.pdf. Accessed 08 Sept 2017
23. Factor, M., et al.: Authenticity and provenance in long term digital preservation: modeling and implementation in preservation aware storage. In: Proceedings of the First Workshop in Theory and Practice of Provenance (2009)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.