**Proceedings for the 2015 PV Conference,
3-5 November 2015, Darmstadt, Germany**

APPLICATION OF OAIS TO THE PRESERVATION OF LINKED DATA

David Giaretta[1], Carlo Meghini[2] , Anna Fensel[3]

(1) Giaretta Associates, 2 High Street, Yetminster, Dorset DT9 6LF, UK (2) CNR Consiglio Nazionale delle Ricerche, Rome, Italy, (3) Semantic Technology Institute (STI) Innsbruck, University of Innsbruck, Austria.

Abstract

There is a demand that society reap the benefits of the investment made in creating the growing deluge of data with which we are faced. Data becomes more valuable and exploitable the more it is combined. Linked Data is one of the growing and most flexible ways of doing this. Yet this poses a problem. The very scale and diversity of the data is compounded by the scale, flexibility and diversity of the links, and there is a need to preserve some, if not all, of the data and the links since combining recent data with older data is a vital part of the overall process, not least for longitudinal studies. This paper summarises the state of the art of the understanding of the fundamental techniques of digital preservation using the concepts introduced by OAIS (ISO 14721), and then systematically discusses these techniques in the context of Linked Data current practices. It derives from the work performed in the PRELIDA (www.prelida.eu) project.

The question addressed is whether these techniques are applicable to Linked Data, whether current Linked Data practices involve new techniques which might be more broadly applicable and finally whether there are improvements to current linked data practices.

DIGITAL PRESERVATION AND LINKED DATA

OAIS defines the following fundamental concepts:

Long Term Preservation as *The act of maintaining information, Independently Understandable by a Designated Community, and with evidence supporting its Authenticity, over the Long Term.*

Independently Understandable: *A characteristic of information that is sufficiently complete to allow it to be interpreted, understood and used by the Designated Community without having to resort to special resources not widely available, including named individuals*.

However one can go a little further by looking in more detail at what “use” might mean. Preserving a simple JPEG image is relatively straightforward. To preserve it means that it can be displayed or printed in the future - it seems reasonable to say that there are few other options. More complex digital objects have many more options - perhaps too many to be able to state explicitly. Preservation aims, while not an OAIS concept, enable one to refine the definition of understandability and usability.

Preservation Aims

Understandability and usability are very general concepts. Being able to use some digitally encoded information could cover almost anything, including printing the 0’s and 1’s as wallpaper. Normally the implication is that the Designated Community can do things other than render (print) them - especially if it is scientific data. Even scientific data could be used in various ways. As an extreme example given a scientific dataset containing measurements of ocean temperature across the Pacific Ocean on 1st Oct 2014, one can think of uses ranging from (1) extracting the value of the temperature measured at some specific location, to (2) using the information to contribute to a study of global warming. In principle, depending on the intelligence/skills of the user, being able to do the first may allow the second to be achieved.

The repository could relatively easily provide Representation Information to achieve (1) but would need to provide much more Representation Information in order to guarantee the ability to achieve (2) - again depending on the definition of the Designated Community. Defining Preservation Aims helps to

guide the repository in terms of clarifying the amount of Representation Information to provide, and may make the work of the Designated Community easier. Examples of preservation aims for a dataset may include trying to enable the Designated Community to:

- process the dataset and generate the same data products as previously
- understand the dataset and use it in analysis tools
- combine the data with other data to calculate derived quantities

The Preservation Aims can also influence what data is selected for preservation. For example if the aim is to be able to combine two datasets then both datasets should be preserved - of course they may be regarded as a single digital object. For Linked Data one can look at a number of possibilities depending on the Preservation Aims.

LD Preservation Aim: Underlying data usability

Linked Data is often a means of publishing the underlying data, e.g. held in a scientific format such as HDF or in an SQL database, rather than simply publishing it in its native form. For example DBpedia is captured and stored in a database, not as RDF. The RDF is derived from the database using specific software which encodes some choices, for instance how to transform the original data values into URIs or literals, how to encode attributes as properties. Because of this the original data and their transformations into LD are not the same. Therefore one could choose to preserve the data and the software if we wish to preserve DBpedia more or less in isolation, but this may not be the same as preserving the RDF version of DBpedia.

LD Preservation Aim: RDF usability

Alternatively one may decide that the important thing is the RDF itself and its usability.

LD Preservation Aim: Services

On the other hand it may be more important to preserve the usability of the services such as the inference capabilities.

LD Preservation Aim: Time dependence

Another Preservation Aim may be to be able to see the Linked Data situation at some time in the past. As an example for DBpedia data (in RDF format, or Tables-CSV format) are archived and the user requests specific data (or the entire dataset) as it was at a specific date in the past, e.g., the RDF description of topic Olympic games at 1/1/2010. Conceptually the archive would keep AIPs containing snapshots of the information at specific times. In terms of practical implementations, we have two complementary approaches. One builds entirely on the web architecture, extending it in order to access past representations of a resource state. This avenue is pursued by the Memento project. The other relies on the development of ad hoc systems for the creation, maintenance and access to provenance information of any resource. Hyberlink and Diachron are two important projects following this approach.

Fundamental approaches

OAIS requirements

OAIS specifies what is required for conformance:

1. support the OAIS Information Model and
2. fulfil a number of mandatory responsibilities:
    a. Negotiate for and accept appropriate information from information Producers.
    b. Obtain sufficient control of the information provided to the level needed to ensure Long Term Preservation.
    c. 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.

d. 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.
e. 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.
f. 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.

All the above mandatory responsibilities apply to the preservation of LD. Addressing (b) we see that the OAIS must gain sufficient control over the information in order to ensure that it can preserve that information. Where the information is encoded as a single file then this is relatively easy - it can be copied to the repository. However for a distributed system such as we find for Linked Data, then we have a more difficult issue.

For Linked Data one could (1) maintain the distributed sources of information; alternatively one could (2) copy the some or all of the various components to the repository. We consider situation (2) where the information is kept within the OAIS - although we note that the OAIS may be distributed. If the various distributed sources of LD become part of the OAIS then we have case (1). As an intermediate case it would be possible to created a set of preserved LD datasets, each member of the cloud being an OAIS devoted to preserve a specific dataset to which the other members of the set somehow link. In both cases the obvious difficulty is the potential open-ended nature of Linked Data as one piece is linked to one or more other pieces. This linkage cannot be infinite since there are only a finite number of links in the world, but recognising that even a small part of this finite number could be very difficult. In the case that the linked objects are copied, and so the copies have new URIs, then an important part of the Provenance, and so important with respect to Authenticity, is the original URI and the time of collection. Therefore in what follows we refer to “the” archive although this may be distributed and we assume there are adequate resources to deal with the preservation of the Linked Data.

Fundamental techniques

The next sections look at the three basic techniques, which include the “emulate or migrate” as a sub-set. The three fundamental techniques are:

- Add Representation Information - this is very much broader than, but includes, emulators
- Transform - a more accurate description of "migration"
- hand over to the next in the chain of preservation

Each of these is described in detail in the next subsections. In each section a general description of the technique is given followed by a discussion of the applicability to Linked Data.

Maintaining AIPs

It is important to bear in mind that the creation of the AIPs is fundamental to the preservation of digitally encoded information, irrespective of any changes that may occur. Therefore in what follows it is essential to realise that all the information needed for an up to date AIP is maintained. This will not be repeated in the discussions in the following sections but is summarised here. For each of the techniques:

- **Add Representation Information**
  - In principle adding Representation Information should not require changes to the Preservation Description Information (PDI) or other parts of the AIP, apart from the Representation Information

- **Transform**
  - This requires great changes in the AIP. The Provenance Information needs to cover
    - the links to the original AIP
    - the details of the Transformation including the reasons why this new object is worthy to be regarded as sufficiently authentic
    - details of the Packaging Information which will probably change
    - any relevant changes to the Package Description
    - details of Access Aids which may be significantly changes
    - Representation Information for the new object
- **Handover**
  - The AIPs is maintained by the repository. If/when the AIPs are handed over to one or more other repository additional Provenance Information must be added about that handover from the first repository.

Add Representation Information

Representation Information is defined by OAIS as: *The information that maps a Data Object into more meaningful concepts*. This is expanded as follows: *Since a key purpose of an OAIS is to preserve information for a Designated Community, the OAIS must understand the Knowledge Base of its Designated Community to understand the minimum Representation Information that must be maintained. The OAIS should then make a decision between maintaining the minimum Representation Information needed for its Designated Community, or maintaining a larger amount of Representation Information that may allow understanding by a larger Consumer community with a less specialized Knowledge Base, which would be the equivalent of extending the definition of the Designated Community. Over time, evolution of the Designated Community's Knowledge Base may require updates to the Representation Information to ensure continued understanding*.

Each piece of Representation Information is encoded, often as a digital object, and that digital object may require its own Representation Information in order for the Designated Community to be able to understand and use it. The same applies to each of those pieces of Representation Information, producing a network of connections, which is called a Representation (Information) Network (RIN). As described above, one of the key threats to preservation is that the digital objects will not be understandable and usable. Another threat is that software or hardware is no longer available.

One, perhaps the only, way to overcome this threat without changing the object is to add Representation Information. In the case of hardware being unavailable the Representation Information would be an emulator. A repository would need to be clear on its definition of the Designated Community, which in turn determines the Representation Information required. As the Knowledge base of that Designated Community changes, additional Representation Information must be added. There are services which can assist the efforts required to have adequate Representation Information including

- Network of Registries of Representation Information - to share such things as shemas and ontologies
- Orchestration service - to help to share information about changes

Examples of relevant efforts include the SCIDIP-ES project http://www.scidip-es.eu, the software from which can be tested at http://int-platform.digitalpreserve.info.

We can examine this in terms of several different potential decisions of Designated Communities and potential evolution of its Knowledge Base. Note of course that we cannot predict these future changes, and moreover OAIS does not demand that the all the possible Representation Information be collected at once. However these "thought experiments" should clarify what information we should be prepared to add to the collection of Representation Information required to ensure the Designated Community can understand and use the LD of interest.

LD Representation Information

Designated Community: users of Linked Data

One can think about the stack of information and processes which such a user currently has support for, noting at each point the potential threats/ changes which may need to be countered. Let us assume that the user has a link - a (HTTP) URI - to some Linked Data and consider the various steps.

- the HTTP URI must be resolved so that the RDF can be accessed
  - this requires the ability to recognise the HTTP URI as a string which can be resolved, and then resolving it to a particular address; this address must be able to then provide the correct sequence of bytes in some way
    - currently the DNS system (perhaps after a Persistent Identifier look-up) resolves the HTTP URI to an IP address. The Internet infrastructure and TCP/IP directs the packets to that address. The recipient must recognise the byte sequence and send the appropriate response using an acceptable protocol via TCP/IP back to the requesting machine.
    - in the future, assuming the internet and DNS are available, the various hosts may no longer exist or the HTTP URIs may no longer be resolvable given the rate of decay of URLs.
- On receipt of the requested RDF, local software parses it and identifies schema, ontologies and other files that may be needed. These can be retrieved in a similar way to the initial file. Each retrieved file is parsed and may itself point to further files.
- Having parsed and gathered the information together, software, not necessarily the same as the parser, is used to satisfy queries from the user

The required pieces of information, including files and software, are Representation Information; the lookup system and network infrastructure **could** also be regarded as Representation Information in the sense that in order to use the RDF files one needs to resolve the URIs. This could be the "normal" network infrastructure including DNS; an alternative would be to create a local version of the resolution system. The amount of Representation Information collected depends upon the Knowledge Base of the Designated Community, which may change over time. As an extreme example if at some point in the far future the internet as we know it is about to replaced by something very different then those responsible for preserving LD could decide that additional Representation Information such as the definition of the network protocols used by LD such as TCP/IP, HTTP and various RFCs should be collected, and perhaps local implementations be kept available. As a concrete example one could image that some time in the future when the current network infrastructure, such as the use of TCP/IP, is to be replaced, then the archive could add information about TCP/IP as additional Representation Information as one of the potential preservation techniques. We can see here the Representation Information Network (RIN) is made up of

- distributed files each of which will have its own RIN. Note that each file could be generated on the fly by the server which receives the request.
  - schema
  - other ontologies
  - imported files
  - explanations of the meaning of the various symbols in the ontologies, supplementing the information provided by the ontologies i.e. the linkages between the symbols.
  - structural information such as Unicode definitions
- software written in various languages, and each of which relies on RINs including various libraries, operating systems and hardware
  - parsers
  - query resolvers

In general terms one could:

- collect the files, to whatever required depth, and use appropriate indirection to ensure that embedded locator strings (e.g. the HTTP URIs) still find the appropriate files; the Memento system is an example of this, adding in timestamping.

- collect the software, either source code or compiled files, together with the required libraries etc. or different underlying software - emulators.

Migration

As an alternative to keeping the bytes unchanged and adding Representation Information, it may be advantageous to change the bytes - and of course add a whole new set of Representation Information associated with those bytes. The advantages may be because of cost savings or for more straightforward perhaps wider, usability. The issues which is arise include:

- the choice of the particular transformation to use
- whether the new object can be claimed to be an authentic representation of the original. Note that it becomes impossible to verify the authenticity of the "new" object simply by checking that the fixity (e.g. checksums or digests) match the original (unless the transformation is reversible in which case one can reverse the process before calculating the digest).

Experience indicates that, except in the rare cases that the transformation is reversible, there will be loss of some information and so a key question is: has enough information been retained? OAIS introduced two related terms related to these questions:

- **Information Property Description:** The description of the Information Property. It is a description of a part of the information content of a Content Information object that is highlighted for a particular purpose.
- **Information Property** is *that part of the Content Information as described by the Information Property Description. The detailed expression, or value, of that part of the information content is conveyed by the appropriate parts of the Content Data Object and its Representation Information*.
- **Transformational Information Property** *as an Information Property the preservation of the value of which is regarded as being necessary but not sufficient to verify that any Non-Reversible Transformation has adequately preserved information content. This could be important as contributing to evidence about Authenticity. Such an Information Property is dependent upon specific Representation Information, including Semantic Information, to denote how it is encoded and what it means. (The term 'significant property', which has various definitions in the literature, is sometimes used in a way that is consistent with its being a Transformational Information Property)*.

The choice of Information Properties and Transformational Information Properties (TIFs) are made according the judgement of people - perhaps the various stakeholders such as data creators, funders or repository managers. The judgement that the TIFs have acceptable values is the responsibility of a person, for example the repository manager. The reputation of that person is then an important factor in the consumers' judgement of the authenticity of the "new" object.

LD Migration

Assume the Linked Data we are interested in is in the form of RDF/XML serialised as a file on a server. There are several possible transformations - for example to various serialisations, as files or as byte sequences in a database: RDF/XML, Turtle, N-Triples, JSON, etc. In the future there may be other serialisations. These transformations are not reversible in that going in a cycle e.g. from an RDF/XML file to a Turtle file and then back to an RDF/XML file again, the latter will be very similar, but not identical, to the original e.g. there may be differences in statement order. This lack of reversibility will not affect the "understandability and usability" of the RDF **but** means that authenticity cannot be verified by checking at the bit-level such as comparisons of checksums or digests, and hence Transformational Information Properties must be identified. Transformational Information Properties might include:

- the ability to link between named things
- the ability to describe the meaning of those links at some level
- the ability to resolve queries about the things

- it may be required to provide information about the changes over time of each object/link

The parser would need to be changed to deal with the serialisation chosen. It might implement one of the proposed LD APIs. The links within each object may be transformed or else the Representation Information for the links must specify how to find the location of the linked object.

Hand-over

If the repository is unable to preserve the information, perhaps because there are insufficient resources allocated, then it should hand over the information, together with any information needed to make up the appropriate Archival Information Package e.g. Representation Information, Provenance, Access Rights etc.

LD Hand-over

Handing over LD should present no additional challenges except that there may be a significant number of distributed objects. It is worth re-iterating the importance of updating and maintaining the Provenance, in particular the URIs - including the new URIs, the URIs used by the previous archives, and the original URIs.

USE CASES

DBpedia

Wikipedia provides the "raw" data used by DBpedia, from which DBpedia creates RDF. DBpedia stores different versions of the entire dataset as RDF or CSV dumps, as a versioning mechanism. The DBpedia contains more than 27 million links to other Linked Datasets. The preservation strategy is to keep these RDF and/or CSV dump files. The external LD to which these are linked are not part of the preserved information, nor are the querying or other software. The evolution of terms is not tracked. In terms of the basic preservation techniques described in section 3, this is a very minimal approach to preservation.

Improvements

- Add Representation Information

Europeana

Europeana functions as a metadata aggregator: its partner institutions or projects send it (descriptive) metadata about their digitized objects to enable centralized search functions. The datasets include links to the websites of providers, where users can get access to the digitized objects themselves. As this metadata is stored by Europeana, Europeana has no specific requirement for specific metadata preservation policies on the provider's side. This is less true for the problem of link rot on providers' websites. Often providers do not use (or do not send) persistent web identifiers, which results in broken links between Europeana and provider's object pages, when these get different web addresses.

Also Europeana has embarked on enriching this data, linking for example to GEMET, Geonames and DBpedia. While sets GEMET are very stable, DBpedia is much more dynamic, and not monotonic (i.e., DBpedia facts may sometimes be retracted during updates, while others are added). Europeana download dumps of external sets to store a part of it in its main databases, so the Europeana services would not be disrupted, should the external datasets undergo massive changes. There are several complexities in the Europeana case. There are several potential sets of preservation aims.

Preservation Aim: Preserve the RDF

Versions of the RDF can be stored by Europeana.

Potential Improvements

The Europeana RDF so that it is usable would require

- the schema - those specific to Europeana as well as any imported
- all the raw data on which it depends - see the next Preservation Aim

- associated software

Preservation Aim: Preserve the "raw" data

By "raw" data is meant the data about which the metadata is harvested - because the data that Europeana uses changes. This implies some level of link rot.

- Dumps of this remote data are being collected. They would presumably be used within a preserved European because if the published links may no longer meaningful in the context of updated third-party sets. Europeana could re-publish its "cached" version of the third-party data. But in a Linked Data setting it would be extremely confusing for users, if such re-publication shows statements that have become very different, or even incompatible with the original source.
- In any case Europeana generates its internal identifiers from the identifiers sent by its providers, which are not always persistent.

Potential improvements

- Add Representation Information (as described above) e.g.
    - information about schema involved
    - information about the link between the original URIs and the re-directed identifiers
    - Add versioning information (a special type of Representation Information)
        - Currently there is no versioning at all in the data that Europeana re-published. Europeana hopes to make progress at some point in the future, by providing information on incremental modification using the tested means of an OAI-PMH server for RDF/XML representation of the object records stored by Europeana.

Preservation Aim: Preserve the services

To keep the Europeana services running one could undertake the 2 preservation aims above

CONCLUSIONS

Systematic study of the fundamentals of general digital preservation indicates that these are applicable to Linked Data, although of course there are aspects that are specific to Linked Data, for example the specific types of Representation Information that must be collected.
None of the cutting edge activities in preserving Linked Data seem to indicate any missing general concepts. On the contrary, the general preservation aspects do suggest a more systematic approach to preserving Linked Data are useful. On the hand there may be specific tools which are needed. Some exist already, including:

- a system of Registries of Representation Information (see the SCIDIP-ES services http://int-platform.digitalpreserve.info/dashboard/registry/)
- a system to collect information about changes e.g. changes in schema or software (see the SCIDIP-ES service http://int-platform.digitalpreserve.info/dashboard/orchestration-service/

Other services and tools are still to be developed. The document D4.3 Consolidated Roadmap (PRELIDA, 2014) discusses additional work that is needed.

REFERENCES

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., & Ives, Z. (2007). DBpedia: A Nucleus for a Web of Open Data. The Semantic Web, 722-735.

OAIS (2012) available from http://public.ccsds.org/publications/archive/650x0m2.pdf

PRELIDA (2014), see http://www.prelida.eu/