Digital Archiving 3.0

Christophe Guéret <christophe.gueret@dans.knaw.nl>

Royal Netherlands Academy of Sciences (KNAW), Data Archiving and Networked Services (DANS)

Den Haag, The Netherlands

Abstract

Whereas physical archives used to be a place to catalog and keep content safe, their digital counterparts do not have to worry about the risks of altering the items by manipulating them. Digital archives can afford making the content of their safe accessible to the outside world and, as such, are turning into data repositories. In addition to classical tasks of identification and preservation of content, digital archives have to deal with new specific challenges such as the obsolescence of data formats, the control of the dissemination of the digital copies and the access of the content from applications. In this position paper we argue that Semantic Web technologies are a good way to approach this problems and constitute (part of) the future of digital archives.

From archives to repositories

Archives are typically seen as a collection of documents racked in shelves and to which an identifier is associated. A classification system provides a mapping from keywords to the document number to enable searching for content. The documents can be physically accessed under some conditions and are rarely so to preserve them. Digital archives share a lot of this, replacing documents by e-contents (Image, Spreadsheets, Videos, …), shelves by disk-based storage and classification system by indexes. But rather than just mimicking their physical counterpart, digital archives have the opportunity to turn themselves into persistent storage solution by leveraging the latest advances in Web based data publication. In a time where Open Data, data sharing, and data preservation are gaining importance digital archive have the potential to become a key element of the next generation of Web: a place where data is securely persisted and made available to the outside world.

The elements of a data repository

Dataverse[1] and Figshare[2] are two example of data repository allowing users to dispose data (in a wide definition, non restricted to any particular format) and have it archived. This data is also made accessible. In the following, we abstract a set of features from these two systems and propose Semantic Web driven adaptations to it.

Identification of data items

Following the Linked Data principles, resources on the Web are assigned HTTP URIs as persistent identifiers. This is the resource for which a description is provided using the Resource Description Framework (RDF). The "resources" of an digital archive are the data items are the items that are persisted in it but, more exactly, what is being described is the meta data about the resource rather than the resource itself. Let's consider, for example, the file with the identifier "urn:nbn:nl:ui:13-kkx-n5k" stored in the archival system "EASY" from DANS[3]. This identifier can be replaced by an URI like "http://dans.knaw.nl/id/nbn-nl-ui-13-kkx-n5k". The main advantage of using URIs over URNs is that URIs are associated with their dereferencing mechanism. An URI can get a machine directly to the content on the Web whereas a URN needs first to be resolved into a URL by an external service (in the case of EASY, this service is found at http://www.persistent-identifier.nl).

The propose URI would not make the difference between the location of the archived content and the content itself. This distinction can be added in several way, one of them being the usage of suffixes to the URI. That is make the metadata available at "<URI>/about" and the data as "<URI>/data". Such a scheme is employed by Geonames to refer to an entity (suffix "about") and entities surrounding it (suffix "nearby").

Meta-data

(Meta)data is expressed in RDF as triples combining a subject, a predicate and an object. The meta-data associated to a given resource will use its URI as a subject and then a set of property/value for every meta-data element. The properties are not unique, several values can be given to the same

[1] http://thedata.org/

[2] http://figshare.com

[3] https://easy.dans.knaw.nl

property. This is similar to the result format of a “GetRecord” OAI-PMH query[4] and the vocabularies recommended and most often used in both approaches are the same, namely Dublin Core. The main difference here is that the meta-data of a record is directly accessible from its URI. Using HTTP URI as identifiers for resources enables the implementation of RESTful services: a GET on the URI returns the description of the resource.

The application of Linked Data principles for the description of the metadata make it possible to easily link items within and across archives. An object in a statement can be the URI of another persisted resource coming from the same archive or a different one. Cross-referencing resources across archives is not only facilitated but even becomes a core part of the design of the system.

Format migration

One of the specific challenge digital archiving has to deal with is the obsolescence of data format. Paper documents become unreadable when a language is lost, digital documents face the same fate when the format they are encoded in is no more supported by any software. Technologies are, unfortunately, evolving very fast and archives have to deal with that. One of the solution to provide smooth migration between formats is to use the content negotiation feature of HTTP [1]. Requests can be transparently redirected to a readable representation of the document depending on what the client have been asking. The usage of this content negotiation feature, along with the redirection of queries is also recommended by the best practices around Linked Data to deal with the ambiguity between the identifier of a resource and that of the document containing the description about the resource [2]. Format migration can thus be dealt with in a way that is compatible with the recommendations concerning the publication of Linked Data.

Security

The possibility of duplication of digital content and the increase of its accessibility through APIs call for more security for access and traceability of content. To this end, several tools can be deployed:

- **Encryption**: asymmetric encryption is a first way to ensure the authenticity of the content delivered. The archive can sign the content with a private key and make their public key, needed to decrypt the signature, publicly available.
- **Provenance description**: the W3C recently published a set of recommendations concerning the description of the provenance of digital resources[5]. This vocabulary, along with its data

[4] http://www.openarchives.org/OAI/openarchivesprotocol.html#GetRecord

[5] http://www.w3.org/TR/prov-o/

model and associated best practices, indicate how archives can describe what is the origin of the content they are serving. Because it is a W3C recommendation, several tools will be able to consume this information and put it into use.

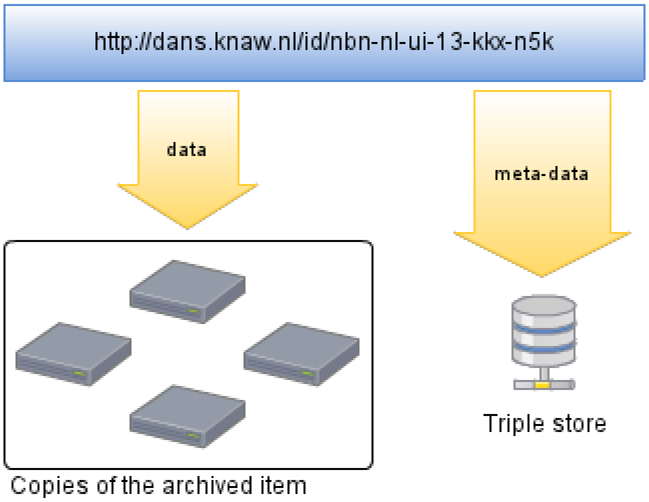
- **OAuth**: this protocol has become very popular as a way to control the access to APIs by applications. The archive provide its authenticated users with a panel where they can control the list of applications having access to their data. Applications making a new request will be guided through a user-controlled process leading to the granting of an API token. This token is then requested in every communication with the API of the archive.
- **WebID**: this solution, formerly known as FOAF+SSL, combines the usage of the TLS secure connection protocol with users URIs to implement a login process. Instead of providing a name and a password, users provide a certificate that authenticate them.

The implementation of these techniques can give a way for digital archives to give access to their content while monitoring and/or filtering its usage.

Persistence

Lastly, the archived items and their meta-data have to be persisted. Whereas the meta-data expressed in RDF can be stored in a triple-store, a data based optimised for this type of data, and mirrored, there is no expectation made concerning the format of the items. It thus makes sense to apply two different policies, ensuring database replication for the first and the dispatching of copies for the later[6]. One of the advantage in publishing the meta-data of the items in a triple-store is that, through the flexibility of SPARQL, users can express complex queries over the content of the archive. In order to enable this kind of search feature, the meta-data will have to be provided with sufficient elements to properly describe the data item it is associated to.

[6] http://www.lockss.org/


http://dans.knaw.nl/id/nbn-nl-ui-13-kkx-n5k
data
meta-data
Triple store
Copies of the archived item

The data items and their associated meta-data are two different type of data to be persisted, with two different strategies

Conclusion

This short paper aimed at positioning digital archives in the evolution of the Web from a document platform into a open data platform. In this upcoming new Web generation, accessible to both machines and human, digital archive have the opportunity to become the data repositories that provide meta-data rich, persistent, storage solution for Web accessible data. Sketching out the architecture of such a repository we highlighted the proximity with Semantic Web technologies and the potential that resides in using them. The natural following step for this proposal is the implementation of a prototype Semantic Web powered digital archive.

References

[1] David S. H. Rosenthal, Thomas Lipkis, Thomas Robertson, Seth Morabito: Transparent Format Migration of Preserved Web Content. D-Lib Magazine 11(1) (2005)

[2] W3C Interest Group, "Cool URIs for the Semantic Web" note, http://www.w3.org/TR/cooluris/ (2008)