Open Horizons: Archiving Perspectives for Services and Frameworks

Stefan Bürer, Historisches Museum Basel (Switzerland)

Abstract

Sooner or later archives storing electronic data will most likely face the task how to archive resources from the internet. These will usually come in as an URI (Uniform Resource Identifier, pointing at a webpage or some other resource in the internet.

References to external documents are not something fundamentally new to archives. Traditionally these references point at a document kept safely in another archive like a National Library. In contrast, information in the internet underlies a continual process of change, pages will be updated, merged or they disappear suddenly, whole domains will raise and fall without leaving a useful trace for reference.

The obvious way of treating these resources is to download the concerning page. Despite the copyright issues, what if there are other links on the page? Instead of downloading whole sites or starting to archive the internet, it is advisable to treat these dynamic resources as an information type at its own. At first, it is necessary to examine the structure of this kind of resource.

At first glance: the URI

An appropriate way of capturing the resource is to store its address, the URI. This seems feasible for simple URI's like: **http://www.anywhere.xyz/anything.htm**. This URI could point to a simple, static document, but this doesn't have to due to the capability of web servers of rewriting addresses.

Today many websites will use databases for storing and composing on the fly the requested resource. They use elaborated session management techniques for richer user experience. An URI of this type will often look like: **http://www.database.web/dynamic/databasedriven/site.xsp?sesID=e840acb2fe6813057e18c7248fd7010b&docid=268791&bypass=yes**.

An URI of this type is not suitable for archiving purposes because it contains session information which will expire after finishing the session, it depends on a specific application and is incomprehensible for human being. A slight change in the application environment will result in totally different URI.

The most suitable type of URI for archiving is the modern, search-engine friendly URI commonly called "clean URI", like **http://www.framework.web/documents/myfunnyvalentinescore**. They hide quite well from the details of a specific implementation of a website and designate well the content of a resource. They are typically based on an application framework, which can support archiving of resources in other useful ways, as will be shown later.

Below the line URI's are very unsafe candidates for archiving purposes due to the many factors influencing their composition. More reliable data is needed to overcome these limitations.

Behind the curtain: Metadata

Beside the URI a webpage can contain Metadata which refer to the context and the content of a webpage and which should conform to the requirements for archiving of dynamic resources in the web.

One of the most common and well known standard of Metadata is the Dublin Core from the „Dublin Core Metadata Initiative" (DCMI). This descriptive set of metadata encodes straightforward in 15 properties the content of a resource in the internet:

```
<meta name="DC.Title" content="myColex">
<meta name="DC.Creator" content="Stefan Bürer">
<meta name="DC.Subject" content="Collection
  Documentation System">
….
<meta name="keywords" content="Open Source,
Collection Documentation, cultural objects,
```

Administrative Metadata specific for archiving purposes provides the standard "PREMIS Data Dictionary for Preservation Metadata". Based on XML, the Extensible Markup Language, it is divided into a main container and the four entities: Object, Event, Agent and Rights, each with a detailed and elaborate structure. Both, DC and PREMIS are extensions of METS, a standard for encoding descriptive, administrative, and structural metadata.

On a more abstract layer, but flexible and extensible are the concepts of the so called "Semantic Web". This initiative has established several frameworks to facilitate the interaction of people and computers so to make for instance the results of a search more meaningful for human beings.

One of the central concepts of the semantic web is the Resource Description Framework (RDF), which specifies methods and syntax for describing metadata of a resource. The Web Ontology Language (OWL) also uses RDF as a syntax and allows the formulation of ontologies and hence the representation of knowledge. RDF uses XML for serialization and has for example following structure (without data):

```
<rdf:Description>
 <ex:editor>
  <rdf:Description>
   <ex:homePage>
    <rdf:Description>
    </rdf:Description>
   </ex:homePage>
  </rdf:Description>
 </ex:editor>
</rdf:Description>
```

The extensibility makes RDF very suitable for archives, because it allows combining both administrative and descriptive Metadata. With the triple form of expressions (subject, predicate and object) it is possible to formulate complex statement about resources and to provide specific information for archiving purposes.

Consuming Metadata

Metadata provide a convenient way to cope with the variability of the Internet and to overcome the limitations of storing just URI's. If the access over an URI fails, metadata will facilitate the retrieval of the original resource even if the URI changed.

But this needs a more elaborate model for storing references to resources than just keeping the URI embedded in a document. It will be necessary to set up a repository of references in the documentation system, preferably realized as a service. This will facilitate the reuse of this functionality for different and future purposes.

Providing Resources and Services

Maintainers of collections and archives who publish their resources in the internet will find in the concepts of the semantic web an emerging and vital field for enriching their services. A good starting point is to adopt "clean URI's", so the addresses have a chance of surviving application changes. A supplementary benefit is the better support by search engines. Delivering Metadata together with the resources makes these resources more accessible over time. This could be in one of the forms mentioned above, like the Dublin Core, METS, PREMIS, or RDF, using an existing ontology or formulating an ontology on the base of OWL.

The Service Oriented Architecture (SOA) approach for implementing information systems seems very promising in this context. This approach will not only allow integrating information from different sources but will also enable archiving duties, for instance by realizing a particular archiving service. This service would respond to a request for a certain archived item with a rebuild of the originally referenced resource.

Of course the request for an archived item has to conform to an established standard of metadata as mentioned above, so it can be interpreted correctly by the archiving service. Under this premise there will be no major difference if the request is coming from the system itself or from another system, hence an archiving service is a cornerstone for the archiving of complex resources in a ever changing environment like the Internet.