Towards an Archive Place for Disseminating Digital Records

Quyen L. Nguyen - National Archives and Records Administration, College Park, MD, USA

Abstract

The ultimate goal of a digital records archiving is to make them accessible to the public and authorized parties for centuries. Such access should be done in a way that is independent of technical platforms with which those records were created. According to the Open Archival Information System (OAIS) reference model, Access to records is one of the main components of an archival system. The Access component allows the Consumer to search for digital records in the Archival Storage, and selectively retrieves them based on the returned result sets. In order to provide enhanced e-discovery and retrieval services, we are proposing the conceptual model of Archive Place, which would allow record producers to deposit data, and consumers to discover, and also develop applications to display and manipulate digital objects stored and preserved in the archive. Advantages of this model are to leverage emerging technologies offered by Web 2.0, whose main characteristics are semantic web, social media, and collective user participation. Different options for exposing digital records will be described. We will also show how key architectural elements based on service-oriented architecture, and web technologies can support the Archive Place's realization.

Introduction

The ultimate goal of a digital records archiving is to make them accessible to the public and authorized parties for centuries. Such access should be done in a way that is independent of technical platforms with which those records were created. According to the Open Archival Information System (OAIS) reference model [3], Access to records is one of the main components of an archival system. The Access component allows the Consumer to search for the digital records in the Archival Storage, and retrieves them based on the returned result sets.

The design and implementation of the Access component have to face a double challenge. On one hand, digital records volume in need for archiving from various information domains will grow with an unprecedented rate. On the other hand, the representation of the data produced by software applications from diverse specialized domains will have a wide range of formats from Microsoft Office documents, relational database files, geospatial images, to multimedia objects. In order to meet this daunting challenge from the Access perspective, we are proposing the conceptual model of *Archive Place*, which would enable the exploitation of Web 2.0, where users and web developers can contribute indirectly to develop applications based on the digital records stored and preserved in the digital archive. The proposed concept is based on emerging trends in today's information technology. Archive Place's paradigm is inspired from Web 2.0 principles of openness and collaboration, while its supporting infrastructure relies on Cloud Computing to achieve efficient sharing of computer resources.

With Web 2.0, the Archive Place model will be able to extend the Access component beyond the portal supported by a single institution. Ultimately, Consumers of digital records can reap the benefits as they get innovative, enhanced and rich experience for discovering and accessing digital records thanks to the contribution of a community of consumers.

Different options for exposing digital records will be described. In our proposal, Archive Place is supported by the cloud-based infrastructure presented in [11]. Furthermore, Archive Place has to implement the following architectural elements:

- Globally unique identification of the digital objects,
- Preservation service with checksum for record authenticity,
- Extensible metadata structure that would help digital records developers to have sufficient information to implement enhanced services.

Thus, the main contribution of this paper is to formulate the novel concept of Archive Place where record producers and consumers can share computing resources of a digital archive in order to promote and enrich access to preserve digital records.

Motivation

In the traditional paradigm of an electronic archive, a human consumer of electronic information initiates search requests or performs browsing some kind of asset catalog or "gallery". For instance, the gallery of the ARC (Archival Record Catalog) web application offers to the public popular topics and collections of records such as War, Civil Rights, etc. Only after this discovery and selection process, that the user will issue requests or the retrieval or delivery of the desired assets. One notable limitation of this usage pattern lies in the fact that the records custodians bear the primary responsibility of systematizing, advertising and presenting the collections of digital objects. This "owner as a curator" challenge is compounded if a collection is large or complex. One example of a large and complex collection would be the entire volume of the publically available electronic records of a national government over the lifespan of a nation. Another example would be a time-longitudinal collection of sensor data, such as earth science measurements, over a long period of time. It is easy to foresee, that a custodian's ability to advertise the full extent of such content under management and to make it available would depend on the availability of funding and conflicting business priorities. As a result, only fractional portions of the total holdings may end up being available for public access in a timely fashion. Moreover, due to budget limitation, a record custodian organization may not have the opportunity to invest sufficiently to design and implement the Access component, hence falling short of providing an attractive and accept application to users.

Given these challenges and the desire to expand the access capability, a digital *Archive Place* concept is proposed, that draws inspiration from the well-known virtual marketplaces on the Internet such as Amazon, eBay, and Craigslist. While these marketplaces provide a virtual environment for merchants and

consumers to sell and buy goods, an Archive Place could serve as a content clearing house and allow web developers not directly affiliated with the electronic repository to develop applications based on the digital records stored and preserved in the digital archive.

Archival Place Concept

Overview

The concept of Archival Place is the realization of the reference OAIS model in a Web 2.0 world. Therefore, an Archival Place does implement the primary functionalities of Ingest, Access, Archival Storage, Data Management and Preservation of the classical OAIS model, where Record Producers transfer data into the system and Consumers search and retrieve data stored and preserved in the archive.

The essential extension brought by Archive Place lies in the interactions of Producers and Consumers with the system and their respective roles. In addition the customary user interface via the Ingest component, Record Producers will be allowed to interact directly with the Data Management component for directly manipulating the digital objects' metadata, such as metadata enrichment, curation, or other maintenance. Such maintenance is necessary for Producers, who are also record owners, to make decision about when and how to open their digital records to the public. The question of "How" arises in the case of Archive Place, which offers more than one mode of digital objects' dissemination. Traditionally, an OAIS system has a portal with search and browse application that allows researchers to discover and retrieve digital objects of interest. With Archive Place, the architecture is engineered to enable multiple modes of use by data consumers of the holdings. The design must lend itself to holdings consumption via most of the available access modes: mashup participation on the third-party web pages, automated crawling, bulk downloads by third parties upon request, as well as the classical access of data in the stores by "link clicking" users. The novelty and the challenge of this approach is the support of plethora of access modes, rather than a more traditional focus on select few. In addition to Web 2.0 features and links to Social Networks implemented on the single Access portal, the approach would open a "network of doors" to the public. A Consumer can be an end-user or a third-party website. In other words, the end-consumer of digital records can interact directly with the portal of an Archive Place, or indirectly via a third-party website. The latter may offer customizable enhanced services to access digital records in the Archive Place. Pictorially speaking, an Archive Place can be depicted as the main tree with accessible branches and distribution outlets. Theoretically, access of digital objects by users could be expanded in an exponentially fashion. If successful, the Archive Place can engender a chain reaction of record access and enhanced access. For instance, the chain reaction effect can happen after April 2nd, 2012, when the US National Archives and Records Administration (NARA) releases the 1940 US Census Data [20]. There is already a spreading interest as posted on websites such as Ancestry.com and The1940Census.com. Being a community of 1940 Census project including Archives.com, familysearch.com, findmypast.com, etc., the latter has made the call to invite third-party to collaborate in the indexing of the digitized images of census forms [21].

Another perspective is that the Archive Place concept has impacts on the Record Lifecycle and Roles of the actors interacting with the system. According to the strict OAIS model, Consumers access digital holdings via the Access component. In the case of an Archive Place, a third-party website or an automated system consuming Archive Place holdings would function as an end-user along with human consumers. Since Consumers enhance the digital records, the Archive Place can offer to the Producers the option to allow the re-ingest (by adding more descriptive metadata, and rendering services).

The lifecycle of digital records in Archive Place changed from the original OAIS model which had more of the encapsulated "enterprise" vision in mind. In a sense, the Archive Place represents an evolution of the original model in the Web 2.0 ecosystem. Content Server, as discussed in Section 4, consisting of the Archival Storage and Data Management (metadata repository) components, is a key architectural element in the implementation of Archive Place. Archival Storage within the Content Server in an internal element and has no interfaces at Archive Place boundary, thus simplifying holdings security concerns.

There are several options to expand the accessibility of the digital objects in an Archive Place. One possibility is to publish a sitemap which will enable crawling by robots launched by third-party web sites to perform indexing and content searching. Another option is to deliver content wholesale, and let consumers cache digital objects in batch along with their aggregate metadata. The third option is to expose digital records via well-designed RESTful APIs so that dynamic applications such as web mashups can access those objects and integrate them within specialized or customized applications at runtime.

Benefits

There has been a known problem that a vast wealth of digital objects is stored in a "deep Web" inaccessible to public search engines and hence public users, as discussed by Bergman [1]. Note that besides providing access of digital objects to external websites, the Archive Place, as envisioned here, also incorporates Web 2.0 features such as links to social media networks, tagging, comments, and RSS Feed. But, the Archive Place strategy goes beyond just standing up a Web 2.0 portal by creating a kind of ecosystem for storing new data or developing new features on archived data. Since it is based on the Web 2.0 principles of "Web as a platform" and "collective intelligence" through user participation [14], the Archive Place concept can offer the following benefits in terms of providing discovery and access to the digital objects in the archive.

Reachability. First, by opening up the access of digital objects in its archival storage to third-party search engines and web sites, an Archive Place has expanded exponentially the reachability of objects to researchers via multiple e-discovery channels. An end-user can search for publicized digital records by either going directly to the Archive Place website, or by using third-party websites. If the external website is a search engine such as Google, Bing, Ask or Yahoo, then the discovery phase of digital objects can be performed there. The popularity of public commercial search websites among the web-surfer population would definitely promote the reachability of the Archive Place. In the case where the third-party has cached and processed original digital objects

extracted from the Archive Place in order to provide enhanced and customized features such as Ancestry.com or Footnote.com, end-users will have the extra capability to retrieve and view the cached copies.

Collective resource utilization. By allowing third-party to work on its digital objects, the effective resource utilization for dissemination is not limited to the computing resources of the Archive Place. Indexing does require CPU cycles and storage capacity, as well as processing search requests. According to Figure 1, an end-user can search for a digital object not only by accessing the Archive Place website but also by going to third-party ones. This achievement can be viewed as a powerful World Wide Web level of load balancing and distribution.

Enhanced services. We believe that the Archive Place can leverage the power of "collective intelligence" by providing an open environment for developers to access its digital objects and develop innovative and enhanced functionalities for viewing, accessing and manipulating the records. Such manipulation should be understood as acting with the records without modifying their original content and context. An example of this is the zooming in and out of an old digitized document.

Sitemap

An effective means of providing a readily consumable sitemap is an adoption of the Open Archives Initiatives (OAI) [13], Protocol for Metadata Harvesting (OAI-PMH). In the OAI terminology**,** Archive Place serves primarily as a *Data Provider,* the entity administering the data-carrying infrastructure supporting OAI-PMH. It relies widely on the third party consumers to function as Service Providers that consume metadata harvested via the OAI-PMH to build tailored value-added services. The combination of publishing both sitemap and OAI-MPH services will help expanding discovery of digital objects and/or their metadata via existing and future Internet search engines.

Table 1 below shows a sitemap with three galleries containing pictures of the "New Century", the "Great War" and the "Great Depression". This sitemap example is encoded using sitemap XML schema and protocol [18]. An Archive Place has to generate sitemaps listing the digital objects which it wants to be crawled and indexed by external search engines. Such activity can be done by informing interested search engines and maybe forging some partnership, as it is not guaranteed that all sitemap listings will be crawled and indexed.

Table 1. SiteMap Example.

```
<?xml version="1.0" encoding="UTF-8"?>
<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9">
  <url>
    <loc>http://www.domain.org/</loc>
    <lastmod>2012-01-01</lastmod>
    <changefreq>weekly</changefreq>
    <priority>0.8</priority>
  </url>
  <url>
    <loc>http://www.domain.org/galleries/newcentury</loc>
    <changefreq>weekly</changefreq>
  </url>
  <url>
    <loc>http://www.domain.org/galleries/greatwar</loc>
    <lastmod>2012-01-02</lastmod>
    <changefreq>weekly</changefreq>
  </url>
  <url>
    <loc>http://www.domain.org/galleries/greatdepression</loc>
    <lastmod>2011-12-23T18:00:15+00:00</lastmod>
    <priority>0.3</priority>
  </url>
</urlset>
```

Corpus Retrieval

A number of benefits can be realized by an Archive Place implementation where a portion of records reside on a partner site, transferred there as a wholesale data corpus. In this case crowd-sourcing by the user community patronizing the partner site may add significant value to the initial curation of the holdings.

The corpus of holdings may be transferred from the original point either via a network connection or via physical media transfer.

In the paradigm of the electronic transfer of a large body of data Service Level Agreements (SLA) may be negotiated between the data originator and the data consumer. The stability and size of the connection bandwidth are critical factors. At the data provider interface, Content Servers focus significantly on optimization of the transfer throughput. Transfers of very large files as well as transfers of very large numbers of small files are a challenge. In both cases it is important to be able to do transfer checkpointing. Consider the issue of large file for which normal FTP can break. The need for a method of transfer that has checkpointing and is resilient to data transfer interruptions/resumptions is obvious.

A lower-tech solution to bulk data transfers is to transfer the corpus on physical media. It must be noted, that while this transfer mechanism does not require the network site-to-site connectivity, all the system functions at the both interface boundaries of the origination site and the receiving site remain the same, as do the functions that are implemented further down the data stream in the architecture. In other words, while no data is flowing electronically between the sites, the transfer media, playing the role of the connecting data pipe, must still be created at the origination point and, then, either the media containing the data must be assimilated into the receiving system or the data on it.

Providing specific collections of digital objects third-party websites by forging partnership has been practiced by the National Archives and Records Administration (NARA). In one instance, Ancestry.com has processed millions of digital records and images under the physical and legal custody of NARA, and made them searchable via Ancestry.com's website with a presentation geared towards end-users interested in researching family history [5]. The collections of records range from census data, immigration manifest, and military records such as the "Consolidated Lists of Civil War Draft Registration Records from 1863 to 1865.

Another instance illustrating the option of providing access to a whole set of digital objects is NARA's partnership with Fold3, which used to be Footnote [6]. Indeed, public users can go to Fold3's website endowed with Web 2.0 functionalities to enjoy rich experience with historical records. Examples of those digital

records are "Papers of the Continental Congress (1774-1789), Mathew B. Brady Collection of Civil War Photographs, Census Records, War Department Collection of Confederate Records, Records of the American Expeditionary Forces of World War I, etc.

The power of this wholesale Corpus Retrieval is that the end-consumer of digital records preserved by the Archive Place can experience not only Web 2.0 features implemented on its portal, but also various other innovative Web 2.0 sites. For example, a subscriber of Fold3 can tag a record with "I'm related", or share his/her own experience to enrich the metadata of a related digital record [6].

Dynamic Access

It is almost a given that the Archive Place portal will incorporate known Web 2.0 applications such as bloggers, social media networks, Internet emails. The screenshot in Figure 3 illustrates the integration of Web 2.0 functionalities and links implemented on NARA's Online Public Access application. These links would allow public users to share interesting records, comment on them, tag them, cite them in documents using booksmarks, or discuss on community forums, etc.

As in the previous option, a further step is to leverage the power and innovation of crowd-sourcing by providing access of the digital objects to external services and applications at runtime. As such, an Archive Place will also provide dynamic access functions as a content node in a Content Delivery Network (CDN). To a consumer, a content server within Archive Place may function as just one of the nodes among a number on the CDN backbone. Archive Place owners then must design for all the attendant bandwidth, performance and availability concerns. From the system engineering point of view, nodes within an Archive Place have to satisfy the desired availability and performance quality of service requirements of consumers in order to enable dynamic access. Such access is provided to third-party websites via a set of simple RESTFul APIs, which work as Google Maps APIs [7]. With the dynamic access to digital objects stored in Archive Place, a mashup application can in turn deliver to end-users Web 2.0 experience without having to worry about preservation issues such as format obsolescence; the latter should fall on the responsibility of the Archive Place.

System Architecture

The main system infrastructure supporting an Archive Place is the cloud-oriented system used for Long Term Digital Preservation as a Service (LDPaas) [11], which consists of a set of Content Servers running on top of a virtualization middleware, and a layer of atomic and composite services designed according to the Service-Oriented Architecture (SOA) paradigm. Two additional data elements are critical to the realization are: Persistent Identifiers, and an extensible Metadata Structure for digital assets. Regardless of the chosen set of access modes, the Archive Place has to assure the inclusion of following aspects in its architecture:

a) Provide globally unique identification of the digital objects in line with the concept of Web-Oriented Architecture (WOA). WOA axioms rely on the institution of universally unique URIs for the resources. In this paradigm a digital record is a resource in the domain of Semantic Web.
b) Ensure at least bit-level preservation with Integrity Seal applied for authenticity of the digital records.
c) Provide complete metadata structure that would help digital records developers to have enough information about the records to implement enhanced services.

Content Servers

The openness of the Archive Place to public access accentuates the need for the encapsulated internal design with well defined external interfaces. The effectiveness of the independent web developers in developing applications based on the holdings of the digital archive relies on accessibility and intelligibility of the services and interfaces presented by the archive.

A Content Server in LDPaaS system consists of services to store digital assets, provide search capabilities and access to the assets, and allow the building and deployment of enhanced archival applications. As a point of reference to the OAIS model, a Content Server encompasses the Data Management and the Archival Storage OAIS components. Most importantly, Content Server is a self-sufficient entity that manages a body of assets, and provides to applications and external services a unified and standard interface to operate on the assets. One key aspect of its self-sufficiency is that a Content Server manages both the assets and their metadata. The Content Server pattern is based on the following layers:

- Physical Layer comprising hardware machine servers and storage.
- Virtualization Layer using a middleware that facilitates the sharing of the underlying physical computing resources.
- Atomic Service Layer comprising of basic services.
- Composite Service Layer containing the services constructed from the atomic services using service orchestration method.

A customer of LDPaaS can select to subscribe for one ore more Content Servers; each Content Server can be dedicated to a category of digital assets, and has its own set of Levels of Service (LoS). In our previous work of LDPaaS, we have defined main services of Ingest, Preservation, Discovery, Access, and Content Server with possible levels of service [11].

Federator

The principal role of a Federator is to manage the Content Servers, including activating and de-activating Content Servers. The Federator has also the function of routing transaction requests to the target Content Server, including e-discovery and retrieval of digital assets.

Related Work

Web 2.0 has gained momentum not only among Web surfers, but in the very community of digital libraries and archives, as a new paradigm for web-based applications to promote usage of digital objects and enrich their associated metadata [2, 9, 18]. Maslov et al. [9] advocated relinquishing control over geospatial data collections stored in institutional repository in favor for a cooperation and decentralization model; instead the interface was integrated with external APIs such as Yahoo! instead of being built in-house notwithstanding based on open standards. Given the significance of Web 2.0 in digital library, Sastry et al. developed

an end-user interface with Web 2.0 features for rich user interactions, blogs, wikis, RSS feeds, bookmarking, information sharing via social networks, and tagging on top the backend DSpace repository. Cheng et al. proposed a similar approach in [2]. Similarly, the new web application OPA (Online Public Access) from US NARA offers links to social networks so that public users can share records of interest with social media friends. Moreover, a user can choose to register to the OPA website and perform record tagging.

DigitalPreservationEurope (DPE), a consortium of national libraries and research institutions in Europe, reported their study and project vis-à-vis Web 2.0 [4]. By exploiting the widespread adoption of Web 2.0 among public users, DPE has made Web 2.0 sites such as Slideshare, Youtube, Facebook, etc. as part of their toolset to raise awareness and share information related to digital preservation.

Palmer [16] promoted the notion of Archives 2.0, as a new generation of performing record archiving. Inspired by the famous Web 2.0 paradigm, she calls for "openness" and "user-centered" characteristics for an archive system. Palmer also suggested a radical change of mindset in the archiving business, including the role of a Producer who also is a Consumer in the new model. Our discussion about roles of the actors of an Archive Place is similar to Palmer's view. However, no system architecture and system design was proposed in [16] for achieving the new Archives 2.0 vision.

As far as the authors are aware, the formulation of the Archive Place concept and its supporting architecture based on Cloud Computing has not been discussed. Indeed, Archive Place is related to Archives 2.0 since both are inspired by the openness and user collaboration over the Web medium. However, Archive Place is not limited to an institution, but is a cloud-based system that allows subscribing organizations to deposit digital records for preservation [11]. Furthermore, as web-based companies and Web 2.0 users can access publicized objects, we could view Archive Place as an environment built on Cloud Computing and virtualization for exchanging data, enhanced metadata, and enhanced access-related services. In other words, what is proposed in the Archive Place concept is dual and synergetic collaboration at Web 2.0 application and data level as well as the computing resources level.

Conclusion

This paper promoted the concept of an Archive Place where Producers can archive and publicize digital records and Consumers can access and enhance those records in a variety of ways at a large scale. In that context, the main role of an Archive Place is to maintain an open and secure cloud environment with guaranteed preservation capabilities. The concept is realized thanks to the flexibility and extensibility of the architecture with the Content Server pattern, virtualization, Service-Oriented design and service composition. As Archive Place is based on current technologies, the true challenge of Archive Place is the engineering and integration of these technologies to provide a service in a large scale. On the business side, the success of an Archive Place depends on how much consumers and web developers are attracted to Archive Place to access its digital objects and build services around those objects. The ultimate deciding factor will then be the interest of the content within the digital objects. Therefore, we believe that national archives and libraries would be in a solid position to become successful Archive Places thanks to the wealth of their digital objects.

Disclaimer

The content of this paper is the personal opinion of the authors and does not necessarily reflect any position of the U.S. Government or the National Archives and Records Administration.

References

[1] Bergman M. The Deep Web: Surfacing Hidden Value. *Journal of Electronic Publishing, Volume 7, Issue 1, 2001*.

[1] Bergman M. The Deep Web: Surfacing Hidden Value. *Journal of Electronic Publishing, Volume 7, Issue 1, 2001.*

[2] Cheng J., Zhao J., Huan J, Tian J., Wang X. The Study on Management and Service of Digital Archives in Web 2.0. *2010 International Conference on Networking and Digital Society*.

[3] The Consultative Committee for Space Data Systems. *Reference Model for an Open Archival Information System,* 2002. Available: http://public.ccsds.org/publications/archive/650x0b1.pdf.

[4] Digital Preservation Europe. Report on Social Networking Presence for DPE, WePreserve and Digital Preservation/Curation, 2009. Available: http://www.digitalpreservationeurope.eu/publications/reports/D7.3ReportOnSocialNetworkingPresenceforDPE.pdf

[5] Family meets history. http://www.ancestry.com/nara_new.

[6] Fold3. http://www.fold3.com/nara/.

[7] Google Maps API Family. http://code.google.com/apis/maps/index.html.

[8] Lifecycle Data Requirements Guide (LCDRG). http://www.archives.gov/research/arc/lifecycle-data-requirements.pdf.

[9] Maslov A., Mikeal A., Weimer K., Leggett J. "Cooperation or Control? Web 2.0 and the Digital Library". *Journal of Digital Information, Vol 10, No 1 (2009)*.

[10] Menasce D., Ewing J., Gomaa H., Malek S., Sousa J.P. SASSY: A Framework for Utility-Based Service-Oriented Design in SASSY. *Proceedings 1st Joint WOSP/SIPEW International Conferencer Performance Engineering, ACM Press, 2010, pp. 27-36, 2010.*

[11] Nguyen, Q., Lake, A. Content Server System Architecture for Providing Differentiated Levels of Service in a Digital Preservation Cloud. *IEEE Cloud 2011.*

[12] Nguyen, Q. L., Lake, A., Huber M. Evolvable and scalable system of content servers for a large digital preservation archives. *4th Annual Systems Conference, 2010*.

[13] NISO Metadata for Images in XML Schema. http://www.loc.gov/standards/mix/.

[14] Open Archives Initiative. http://www.openarchives.org.

[15] O'Reilly T. What is Web 2.0: Design Patterns and Business Models for the Next Generation of Software. *Communications Strategies (2007), Issue 65, Publisher SSRN, pp 17-37*.

[16] Palmer J. Archives 2.0: If we build it, will they come? *Ariadne, Issue 60, July 2009*. Available: http://www.ariadne.ac.uk/issue60/palmer/.

[17] PREMIS – Preservation Metadata. http://www.loc.gov/standards/premis/.

[18] Sastry H.G., Reddy L. C. Significance of Web 2.0 in Digital Libraries. *International Journal on Computer Science and Engineering, Vol. 02, No. 06, 2010, pp. 2208-2211.*

[19] Sitemaps.org. http://www.sitemaps.org/.

[20] 1940 Census Data. http://www.archives.gov/research/census/1940/.

[21] The 1940Census.com. https://the1940census.com.

Author Biography

Quyen L., Nguyen is currently working in the Systems Engineering Division of the ERA Program Management Office at the U.S. National Archives and Records Administration (NARA). Before joining NARA, he has worked for telecommunications software companies. His experience is in developing software systems for large scale deployment. He has a BS in Computer and Information Science and Applied Mathematics from the University of Delaware, and a MS in Computer Science from the University of California at Berkeley. His research interests are in system architectures for digital archives and intrusion tolerance architectures.