Towards a Large Access-Oriented Digital Archive

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

Abstract

The ultimate goal of preserving digital records 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 application softwares and platforms with which those records were created. According to the Open Archival Information System (OAIS) model, Access to records is one of the main components of an archival system. The Access component allows the Consumer to search for records in the Archival Storage, and obtain relevant result sets. Although Access is at the end of OAIS data flow, the quality of service that it can provide to the consumers depends on both ingest and preservation planning. Indeed, the ETL (Extract-Transform-Load) process at ingest time will determine the richness and comprehensiveness of the created metadata, and the level of granularity or coarseness of the discovery and access process. Moreover, the planning and execution of preservation methods should take into account not only the pure transformation of file formats, but the quality of access and discovery of digital records. This paper examines the challenges to prepare and provide search and access to a large digital archives throughout the different phases of the OAIS functional flow, namely, ingest, and preservation. From the architecture perspective, the guiding principle of a large Access-Oriented Digital Archive should be to realize an Open Platform that facilitates the flow of data from ingest to access. The framework used to architect this platform layer should be extensible so that future search and access techniques could be easily inserted without major redesign of the system.

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 application softwares and technical platforms with which those records were created. According to the Open Archival Information System (OAIS) reference model [1], 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. As Access is at the end of the OAIS data flow, a lot of times it has been relegated to second rank by system analysts and designers. However, its importance cannot be neglected, since user experience for the Record Consumer depends on the very stages of ingest and preservation. Indeed, the extent and characteristics of the ETL (Extract Transform Load) process to be performed will determine the richness and comprehensiveness of the created metadata, and the level of granularity of the discovery and access process. Moreover, the planning and execution of preservation methods should take into account not only the pure transformation of file formats, but the quality of access and discovery of digital records.

Another challenge faced by an Access-Oriented Digital Archive is the need to reconsider the current business process to review records, determine their releasability, and the degree of record redaction. But, any adaptation of the process must resolve the duality between compliance with laws and regulations such as the Privacy Act and HIPAA (Health Insurance Portability and Accountability Act), and the desire to open large corpuses of digital records. Furthermore, an Access-Oriented Digital Archive will also need to adapt to the new ecosystem created by Web 2.0 paradigms and social networks. No longer will Access constitute a "passive" activity, but the Access experience will encompass searching, discovering, retrieving, and most importantly contributing to the digital archive. This contribution can take the form of enriching the metadata of the accessible digital records, or providing enhanced applications and services around the information extracted and analyzed from those records.

The rest of the paper is organized as follows. In Section 2, we discuss the motivation of having an access-oriented perspective for a digital archive. Section 3 discusses the multi-faceted challenges of an Access-Oriented Archive using the OAIS functional reference model. Finally, we summarize key ideas of the paper and potential future work in Section 4.

Background

In the traditional paradigm of an electronic archive, a consumer of electronic information initiates search requests or performs browsing some kind of asset catalog or "gallery". Only after this discovery and selection process that the consumer will issue requests for the retrieval or delivery of the desired assets. 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 non-proprietary collection of sensor data, such as earth science measurements, over a significant period of time. It is easy to foresee, that a custodian's ability to advertize the full extent of such content under management and to make it available would be circumscribed by 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. Therefore, on one hand, we have the reality of an increasing trend in the volume of digital objects being produced; on the other hand, there is also an increasing demand and/or desire to access that vast body of records. Unfortunately, current process does not seem to be able to reconcile this dual increase.

For the owners of digital archives and libraries, there is a clear motivation to continuously enlarge access to their digital object repositories as much as possible. This openness strives to maximize access along two dimensions R and U:

- The R-dimension is the number of records made available to the public;

- The U-dimension represents the number of users who are interested in researching digital records in the archive, and take action to search and retrieve those records.

Expanding the R-dimension can be motivated by government directives to have Open Government, so that the public can gain knowledge and transparency of governmental actions and policies. The result is accountability, which can in turn be a catalyst of government service improvement. In the US, recent directives concerning federal records management have reinforced this notion of openness, transparency, accountability and efficiency [19].

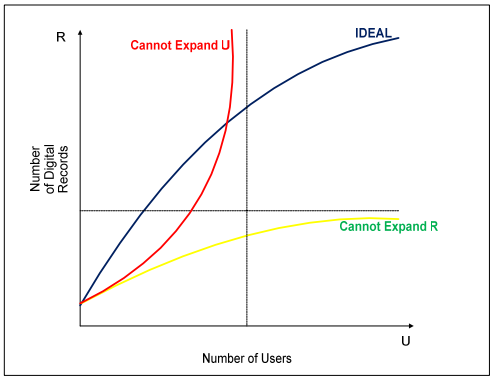

Figure 1. R-U Dimensions.

The ideal strategy is to expand both R and U as depicted by the "ideal" curve in Figure 1. Meanwhile, the other two curves exhibit weak patterns in cases where one cannot expand the number of users due to the lack of resources or "marketing" (red curve), or due to the inefficiency of releasing records to the growing user demand (yellow curve).

In previous work [12], we have proposed the Archive Place model to expand the accessibility of digital records along the U-dimension. The strength of the model relies on third-party collaboration and cooperation in order to achieve exponential expansion of user population of archival content in a Web 2.0 paradigm. In this regard, the Archivist of the United States of America has emphasized the active participation of access to digital records via social networks, which have become so popular and part of our daily life:

"Access to records in this century means digital access. For many people, if it is not online, it doesn't exist. The use of social media to increase access is the new norm. NARA [National Archives and Records Administration] has been going after innovative tools and projects that increase digital access to our records, including projects that invite public participation. We are developing a Citizen Archivist Dashboard that will encourage the public to pitch in via social media tools on a number of our projects." [23]

Lastly, the technological aspect of access via multiple channels of communications cannot be dismissed. Indeed, with the proliferation of mobile devices such as smart phones, iPads, iPods, and electronic tablets, systems providing access to digital repositories must include mobile apps, if the goal is to maximize access.

In summary, since the ultimate goal of a digital archive is to provide to the users access or rather enhanced access services of its digital objects, it would be beneficial for system owners and consumers to consider the system from the access perspectives, hence the term of access-oriented digital archive.

Challenges

In this section, we will examine the challenges that an access-oriented archive has to face. Using simple computer system model, a digital archive system can be viewed as having three parameters:

- Production Rate with which digital records are produced and planned to be ingested into the digital archive system. It has been an almost accepted fact that there is no slow down in this rate, especially with the recent talk about Big Data.
- Service Rate which measures the throughput of the system processing until the records can be accessed. This "service" phase can encompass the ingest processing of records, the appraisal, and maybe redaction of privacy data.
- Consumption Rate which models the demand of the public to access records.

The optimal situation of is to achieve a balance of all three rates: Production, Service and Consumption, so that to avoid backlog and/or starvation.

Ingest Challenge

The Production Rate creates Ingest challenge. Indeed, with the proliferation of IT systems, the increase of federal agencies using computer applications to conduct business, it is widely accepted that the volume of electronic records will keep increasing. Furthermore, the collection and creation of data vital to some missions has generated Big Data that are used for analytics and ad-hoc query to answer business intelligence questions. The mere transfer of Big Data from the producer's location to the large digital archive will soon become a prohibitively expensive practice, as pointed in [21].

Added to the volume of digital records, record formats and types also present a challenge. If archiving is for archiving sake, then the challenge is minimal. But, in the case of an access-oriented archive, where access is paramount, some processing of the ingested records will be required so that those records can be accessed intelligently by future users. Using the model of the Computer System as described above, the Service Rate may present a bottleneck to the whole process. As mentioned above, "Service" is not only restricted to the performance of the computer system, but may also include the business process and current practices to ingest and validate digital records. In order to increase the throughput, serious thinking should be done to maximize parallelism, that is eliminate all bottlenecks in the ingest system, as well as the business ingest process. This bottleneck elimination may require some simplification of the workflow, so that automation via reduction of human intervention can be achieved. As noted in [20], archivists will have to perform a balancing act between emphasizing on provenance and record grouping metadata as in traditional archival science, and just providing the search of record content in a Google-like fashion.

Preservation Challenge

The core functionality of an OAIS system is digital preservation. Even with this classical concept of preservation, access still plays an important role during the process. According to the Digital Preservation Coalition, it appears that access is the driving factor of digital preservation process: "Digital Preservation can be understood as the series of managed activities necessary to ensure continued access to digital materials for as long as necessary. It combines policies, strategies and actions to ensure access to reformatted and born digital content regardless of the challenges of media failure and technological change. The goal of digital preservation is the accurate rendering of authenticated content over time" [8]. Moreover, the report on audio preservation by the Columbia University Libraries noted that "[b]est practices in digital preservation continue to evolve and may encompass processes that are performed on content prior to or at the point of ingest into a digital repository as well as processes performed on preserved files post-ingest over time" [6, 7]. Preservation for access applies not only for audio objects, but also for other types of digital records. In fact, the archivists may have to determine the format for storing digital objects at the very ingest phase in order to provide adequate search and access in the future. In some sense, this is related to the hard archival problem of defining what the record is. Preserving email records is an example: should we store emails in large aggregate .PST files, or should individual email messages be preserved in archival storage?

Another aspect of a digital preservation system is to sustain continuous evolution to adapt to changes in technologies. In [11], we stressed on using the Service-Oriented Architecture (SOA) paradigm to build a flexible service platform for a long-term and evolvable digital preservation system; the platform allows the assembly of "tool-services", which can be easily replaced or augmented, into useful preservation services. Bradley views the preservation issue in terms of "digital sustainability", which consists of activities to "facilitate access" in the future; this would also require a flexible infrastructure and "continuous maintenance" [9, 10].

Access Challenge

It seems to be redundant to mention about the challenges of access when we are discussing the concept of an access-oriented digital archive. But, an understanding of how access should be provided to the users in today's environment will help the pre-access processes, namely ingest and preservation. There are three aspects of access challenges:

- First is the personalization of digital record research. It may sound trivial to say that different types of users have different needs or rather approaches in searching for records. A seasoned archivist is most likely interested in obtaining record groups of interest, or will pay more attention to the record provenance. On the contrary, a student writing a class paper on a historic event will probably utilize a Google-like search to find documents or digital objects related to the topic being studied.
- With the advent of social media such as Facebook, Twitter, Flicker, etc., today's users expect to have a more active interaction with digital records. They would like to tag them, share them with friends in the same social network group, or discuss about them in an open forum.
- Access is not limited to a browser running on a desktop, due to the growing popularity of mobile devices. Therefore, the system has to support various channels of communications, including traditional web browsing on the desktop and low bandwidth mobile apps on smart phones and electronic tablets. This definitely will impact the design approach and maintenance consideration of the access website.

So, the question is what needs to be done in the face of these access requirements. Any solution should include careful consideration of data processing during the ingest and preservation phases, as well as the design of the access application to accommodate user expectations of a rich Web 2.0 experience. In the case of still picture preservation, some current practice requires at least two formats: a digital high-resolution format, and a lower resolution JPEG format; the former may be used for reproduction to a hard copy photo, while the latter is destined for web users to avoid annoying slow download time.

The white paper [22] promotes a novel vision for managing records stored in different clouds. The idea is to limit datta movement since this presents a transport challenge, and to process the data in place. In this case, a digital record will have three Access states: owner/producer, NARA, and public. While Legal Custody is still needed to transfer legal ownership of digital records from one entity (producer, federal agency) to another (custodian, NARA), Physical Custody may lose its actual semantic of a location change of the in-progress accessioned records. The lifecycle of a digital record may remain the same, with some expected simplification of the accessioning business process.

One way to perform in-place processing is to use the Manager-Agent pattern, as depicted in Figure 2. A centralized Manager component can dynamically create an Agent in the Cloud where records are being stored using APIs published by Cloud service providers. Then, this Agent can be scheduled to perform the dedicated service, such as ingest processing, preservation transformation, or content indexing. Thanks to the elasticity and virtualization of cloud computing, adequate resources can be allocated to those services, depending on the data volume, desired throughput, and processing demand. The author believes that today's technology makes the vision described in [22] feasible. In fact, there has been research done in the area of Service Composition with services spanning over multiple clouds [18].

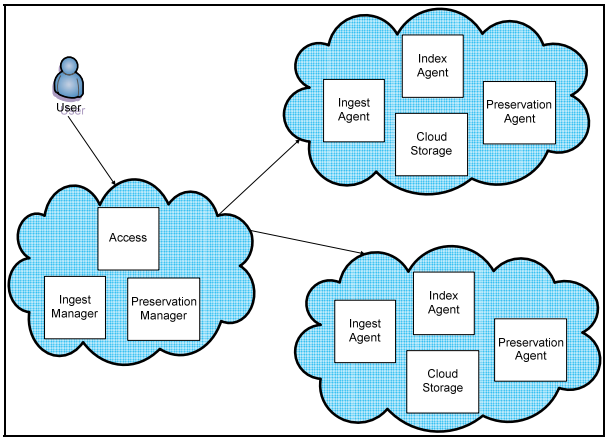

Figure 2. Intercloud System.

With this "Intercloud Digital Archival System", the Access subsystem may function like a web search engine such as Google. A user will interact with the Access subsystem to perform searching for objects that she is authorized to search. Most likely, the index repository is located within the same cloud as the Access subsystem. However, when the user wants to retrieve the digital object itself, her retrieval request will go directly to the Cloud Storage belonging to some cloud containing the object.

Security Challenge

Adding to these challenges is the unavoidable security challenge, which poses a duality with the desire to maximize access along both R and U dimensions. How can the system comply with laws and regulations, while still satisfying the high release rate? Such laws and regulations such as the Privacy Act or HIPAA have been in place to govern the usage and disclosure of data while still protecting individual privacy. As of today, there is no automated tool that can guarantee a perfect redaction of privacy data. Thus, without the 100% assurance, using exclusively an automated tool to accelerate the redaction process will not be sufficient, and remain an outstanding issue.

Conclusion

This paper presents the view of a digital archive from the access perspectives, based on the motivation that all archival activities performed on a digital archive have the ultimate goal of providing long term and continuous access to the digital objects preserved in the archival storage. With the advent of Big Data, some reengineering of OAIS functions and business processes may be necessary to evolve and adapt archiving to the growing challenge of larger and larger volume of data.

We also have described how an Access subsystem can be designed to work in the case where digital objects are stored in various Cloud Storages.

Disclaimer

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

References

[1] 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.

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

[3] 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.

[4] Fay Chang et al. Bigtable: A Distributed Storage System for Structured Data. OSDI '06 Proceedings of the 7th USENIX Symposium on Operating Systems Design and Implementation.

[5] Jie Tao, Daniel Franz, Holger Marten, and Achim Streit An Implementation Approach for Inter-Cloud Service Combination. International Journal on Advances in Software, vol 5 no 1 & 2, year 2012.

[6] Digital Audio Working Group. Collaborative Digitization Program (2006). Digital Audio Best Practices (Version 2.1). Aurora, Colorado. p. 4.

[7] Columbia University Libraries (2010). "Preserving Historic Audio Content: Developing Infrastructures and Practices for Digital Conversion. Final Report to the Andrew W. Mellon Foundation", p. 5.

[8] Digital Preservation Coalition (2008). "Introduction: Definitions and Concepts". Digital Preservation Handbook. York, UK.

[9] Bradley, K. (Summer 2007). Defining digital sustainability. Library Trends v. 56 no 1 p. 148-163.

[10] Reagan Moore. Towards a Theory of Digital Preservation. International Journal of Digital Curation (2008), ISSN: 1746-8256.

[11] Quyen Nguyen. Towards a Design Approach for an Effective System Evolution of a Large Electronic Archive Information System. 3rd International Workshop on a Research Agenda for Maintenance and Evolution of Service-Oriented System, Edmonton, CA, September, 2009.

[12] Quyen Nguyen. Towards an Archive Place for Disseminating Digital Records. Archiving 2012, Copenhagen, Denmark. Jun 2012, p. 141-146; ISBN / ISSN: 978-0-89208-300-8.

[13] Stephen Chapman, Paul Conway, and Anne R. Kennay. Digital Imaging and Preservation Microfilm: The Future of the Hybrid Approach for the Preservation of Brittle Books.

[14] Kia Ng et. Al. Preservation of Interactive Multimedia Performances. Int. J. Metadata, Semantics and Ontologies, 2008.

[15] Robert A. Schrier. Digital Librarianship & Social Media: the Digital Library as Conversation Facilitator. D-Lib Magazine, July/August 2011, Volume 17, Number 7/8.

[16] Michelle Springer et al. For the Common Good: the Library of Congress Flickr Pilot Project, October 2008. Available: http://www.loc.gov/rr/print/flickr_report_final.pdf.

[17] National Archives and Records Administration. Open Government Plan 2.0. Available: http://www.archives.gov/open/open-government-plan-2.0.pdf.

[18] J. Octavio Gutierrez-Garcia and Kwang-Mong Sim. Agent-Based Service Composition in the Cloud Computing. Grid and Distributed Computing, Control and Automation. Springer Berlin Heidelberg, pp 1-10. 2010.

[19] Managing Government Records Directive Memorandum, August 24, 2012. Available: http://www.whitehouse.gov/sites/default/files/omb/memoranda/2012/m-12-18.pdf.

[20] Kate Theimer. Web 2.0 Tools and Strategies for Archives and Local History Collections. Neal-Schuman Publishers, Inc.: 1 edition, December 31, 2009.

[21] Karen D. Schwartz. How Agencies are Scaling Mountain of Data. FedTech, October 31, 2012. Available: http://www.fedtechmagazine.com/article/2012/10/how-agencies-are-scaling-mountains-data.

[22] Archival Data-at-Rest, September 21, 2012. Available: http://www.archives.gov/era/acera/presentations/.

[23] Alex Howard. The Future of Social Media at the National Archives. O'Reilly Radar. November 18, 2011. Available: http://radar.oreilly.com/2011/11/national-archives-social-media.html.

Author Biography

Quyen L., Nguyen is a System Architect in Information Services, Strategic Systems Management 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 security architecture.