

Infrastructure Considerations for Large Digital Libraries

A study to support the technical infrastructure decisions for the Digital Public Library of America

Geneva Henry

Developing a Prototype for the Digital Public Library of America

Council on Library and Information Resources / Digital Library Federation

2 February 2012

1. Introduction

Digital library initiatives over the past fifteen years have led to a wealth of scholarly materials being made available online. The impact of having these resources readily available has resulted in new business models and a cultural shift in how information is distributed. Mass digitization has been defined as the digitizing efforts that do not explicitly select materials, but that convert materials to a digitized format on an “industrial scale,” digitizing at high speed and applying OCR technologies to the texts to make them searchable (Coyle 2006). Coyle distinguishes mass digitization efforts from large-scale digital projects, with the latter defined as creating collections of resources rather than just large amounts of digitized print works. The purpose of this study is to examine the infrastructures of systems that manage large quantities of digital materials that one would think of as a digital library. This will help inform the design of the Digital Public Library of America (DPLA) system architecture. Though there are many smaller, specialized digital libraries, the large-scale and mass digitization libraries offer interesting challenges that come with scale and breadth of materials. The study will focus on non-commercial digital libraries since the infrastructures of commercial collections, though large, are not generally open for analysis.

Mass digitization libraries that many people are familiar with include Google Books, HathiTrust, and The Open Content Alliance from the Internet Archive. There are numerous large-scale digital libraries with varying architectures that can provide useful insights for architectural considerations when looking to develop the DPLA. This study has considered the infrastructures of a select number to understand the diverse approaches each has taken to manage the digital content. In addition to the mass digitization libraries, digital libraries that offer some interesting resource management approaches include Europeana, the National Science Digital Library (NSDL), DCC/Opening History, the California Digital Library (CDL), Networked Infrastructure for Nineteenth-Century Electronic Scholarship (NINES), and some sample U.S. State-wide digital libraries, including Florida Memory, Digital Library of Georgia, Digital Commonwealth (Massachusetts), and Seeking Michigan.

The study is guided by the following questions related to the digital infrastructures used for managing content:

- **Storage and Content Delivery:** *What are the storage approaches? What kinds of servers are needed? What databases and repository platforms are part of the infrastructure? How is content distributed? What are the assumptions/requirements for content formats?*
- **Metadata Approaches and Harvesting:** *What metadata formats are supported/required? How are content and its associated metadata managed (e.g. separated, tightly coupled)? What does the metadata look like (e.g. standards used, markup, mandatory fields)? How is metadata harvested?*

- **Search and Discovery:** *What discovery approaches are used to find content in the collections (e.g. full text search, image search, faceted browsing, type of search engine used)?*
- **Services and Applications:** *Are there special applications required to use the content (e.g. book turner, jpeg2000 viewers, etc.)? What services are provided?*
- **System Sustainability:** *What are the system sustainability practices and policies for the digital library?*

The goal of this study is to understand different approaches that are used to manage large digital libraries; it is not a comprehensive review and comparison of all the systems surveyed. The paper is organized into the following sections: Storage and Content Delivery, Metadata Approaches and Harvesting, Search and Discovery, Services and Applications, System Sustainability, and Summary.

2. Storage and Content Delivery

Managing large amounts of digital information requires a robust server and storage system that is both reliable and has strong performance. While smaller digital collections do not need to worry as much about the details of storage, larger digital libraries are likely to see a significantly increased amount of activity from users and will be expected to provide performance that allows resources to be accessed both quickly and reliably. Decisions about the storage hardware, server allocations, databases, and distribution approaches, along with bandwidth considerations, are key in establishing the digital library as a reliable resource that can be used by teachers, students, researchers and the general public. If a video can only be viewed after it is downloaded, access to content is unreliable, music does not play at an even speed, or content becomes corrupted – or lost – over time, the community will not view the digital library as a credible source of information.

2.1. Storage

Storage decisions for large digital libraries have the most impact when the content resides centrally rather than distributed among systems that have differing architectures for managing the content. Repositories that are content aggregators increasingly have looked to clustered storage solutions to provide reliable and robust performance. Disc speed, failover capabilities and automatic error correction functionality found in clustered storage solutions help to ensure high availability of the content and adequate performance to meet user needs. Depending on the type of content in the digital library and the approach for delivering it, higher speed storage may be needed to ensure adequate performance (e.g. streaming music); text-only content, however, may not require as high speed, but will still require failover and automatic error correction.

Commercial “cloud” storage solutions have become increasingly popular, but not generally for the large-scale digital libraries reviewed in this study. Depending on the transaction levels, cloud storage could be very expensive under the current

pricing models since the primary costs are driven by the amount of output, or use, of the data once it is in the cloud. If the repository is very active and users are downloading or streaming significant amounts of content, the cost of using current cloud storage vendors that provide a robust and reliable storage configuration may not be feasible. The current pricing structure for Amazon’s S3 service is tiered, with the first year of storage free for 5 GB of Amazon S3 storage, 20,000 Get requests, 2,000 Put requests, and 15GB of data transfer out each month (Amazon Web Services, LLC). Storage costs are very reasonable, but data transfers out of the cloud for an active site can be cost prohibitive. The data transfer pricing model, as of 12 December 2011, is shown in table 1. Amazon offers redundancy storage of the content for a reduced rate, providing a means of creating a back-up of the primary data store.

Pricing	
Data Transfer IN	
All data transfer in	\$0.000 per GB
Data Transfer OUT	
First 1 GB / month	\$0.000 per GB
Up to 10 TB / month	\$0.120 per GB
Next 40 TB / month	\$0.090 per GB
Next 100 TB / month	\$0.070 per GB
Next 350 TB / month	\$0.050 per GB
Next 524 TB / month	Contact Amazon
Next 4 PB / month	Contact Amazon
Greater than 5 PB / month	Contact Amazon

Table 1: Data Transfer Costs for Amazon S3, 12 December 2011 (Amazon Web Services, LLC)

HathiTrust is an example of a centralized, aggregated collection. Their storage solution is the Isilon clustered storage system that is highly scalable with the addition of new nodes to the storage cluster (HathiTrust). The Isilon’s OneFS operating system allows all storage nodes in the cluster to know the full file system layout, allowing each to act as a peer and service any incoming requests. Redundancy is controlled at the volume, directory and file levels, making this storage solution highly reliable (Isilon Systems 2011).

Large digital libraries may also choose to approach the collections from a federated perspective, relying on the original content providers to store their own content. Metadata is centralized to support resource discovery across the many repositories. When federating hundreds or thousands of collections, reliable and fast performance for searching the digital library is critical and the storage decisions can have a significant impact. The choice of metadata representations can have an impact on the storage solutions selected. Large federated collections, such as Europeana (Dekkers, Gradmann, and Molendijk 2011) and Opening History (Grainger Engineering Library, University of Illinois 2009), host the metadata for many federated collections, resulting in a large quantity of data that must be stored and searched efficiently (Dekkers, Gradmann, and Molendijk 2011). The structure,

such as a large flat set of metadata records, vs. relational or Resource Description Framework (RDF) representations, will lead to considerations of how the search will be performed and whether or not distributing the metadata across multiple storage servers can be readily supported with adequate performance.

2.2. Servers

Servers for the digital library are critical in providing a reliable platform for the digital library. Well-architected systems have configured a number of services that distribute the activity load over multiple servers, with services running on different servers and load balancers used to dynamically handle activity so that peak usage periods will not result in a server crash. Server hardware for large and mass digitization projects can vary widely. HathiTrust relies on commodity Intel-based servers running the Linux operating system. Most large digital libraries do not explicitly identify their server hardware or operating systems. Europeana maintains a Service Level Agreement (SLA) with Vancis, “a private company with firm roots in Academia and therefore with excellent and relatively cheap connectivity directly to the Amsterdam Internet Exchange (AMS-IX)” (Dekkers, Gradmann, and Molendijk 2011); those servers run the Linux operating system.

Redundant configurations and load balancing provide high availability and overall system reliability. The largest digital libraries are conscious of the need to provide this type of robust infrastructure, though the approach for configuring multiple servers varies by project. Europeana runs their Web servers, database servers and image servers on separate machines. The most commonly used web server is Apache.

Repositories that support audio and video resources may need to consider streaming servers as part of their server infrastructure. Supporting streaming of multimedia resources can be very advantageous since users will not be required to download these often-large files prior to listening to or viewing them. Because there are multiple streaming servers available on the market, platform compatibility for end users can be a key consideration. For example, Windows platforms and OSX platforms may have differing requirements for delivery applications. If users are required to have a plug-in prior to being able to stream the multimedia resources, they will need to have sufficient knowledge and/or support for installing the plug-in on their computer. In public settings or in many institutions, users are not given proper permissions to allow them to install applications needed to use online resources. If streaming servers are planned as part of the infrastructure, formatting the content to work with multiple streaming servers would be prudent. Permitting downloading of the resources, albeit less efficient, will allow users who cannot stream them to access them for listening/viewing on their computer.

2.3. Databases and Repository Platforms

Decisions regarding how the content and metadata are managed will often depend on the types of data stored and the search approach that is used. Repository platforms such as DSpace and CONTENTdm are popular choices for many US State

digital library collections and organizations that do not have a significant staff of programmers to support functional system development. An example of a CONTENTdm site is Seeking Michigan (SeekingMichigan.org 2008), the online archive for Michigan's digital cultural heritage resources.

The larger digital libraries reviewed in this study, outside of the State-level collections, have developed their own platforms for managing content. NSDL has developed the NCore platform for managing their collections. NCore uses the Fedora repository software and provides its own data model as well as a suite of tools to support NSDL (Krafft, Birkland, and Cramer 2008). The project created MPTStore as a means for robustly managing RDF triples (Cornell University 2006), noting that existing solutions for managing triple stores do not adequately support very large quantities of RDF data with the scalability, performance and reliability required by NSDL. SQL relational databases, both commercial and open source, are very common back-end databases for managing information in large digital repositories. HathiTrust uses MySQL which provides a good environment for handling the content and metadata formats used in that repository.

As projects grow and mature, the decisions about databases and platforms are likely to be revisited. An example is the Europeana project. Since Europeana does not manage the content objects, it has found that the use of the Lucene/Solr search engine provides an adequate database for the metadata it manages. Their current data model, the Europeana Semantic Elements (ESE) (Dekkers et al. 2011), provides a flat data model that works very well using Solr as the database. As more complex approaches are considered for a newer Europeana Data Model, however, there is a need to consider alternative database implementations. Under consideration are:

- Additional modifications to Solr to include caching and pre-computing search results to accommodate continued growth;
- Possibly moving to the noSQL document database and combining that with Solr; and
- A management solution for RDF triple stores. This, however, is of some concern because billions of triples would be needed to represent the content. Solutions in this space may have significant negative impacts on performance.

Deciding on a repository platform or data management infrastructures should be informed by the decisions made regarding content format, metadata, search/browse strategies, and the ability of the project to maintain the technology.

2.4. Content Distribution and Format Assumptions

High availability and overall system reliability are features that are important for most large digital libraries. It is important for users to know that the resource is reliable and stable. The decisions about how the content is managed may impact the approach taken to ensuring the content is reliably available when needed.

One means of providing reliable access is to mirror (i.e. replicate) the digital library in multiple locations so that if there is a problem at one location, the mirrored site can provide continued access to the resources and services. Furthermore, mirroring across multiple geographic locations can facilitate better, more reliable delivery of the resources at a global level. Examples of large archives that have mirror sites include the Internet Archive, with the production site in San Francisco and mirror at Bibliotheca Alexandrina in Egypt; Europeana, with mirrored sites at their host provider's data centers in Amsterdam and Almere in The Netherlands (unknown which site is the primary production site), and HathiTrust, with the production site at the University of Michigan and mirror at Indiana University's Indianapolis campus.

Configuring systems for high availability can also be accomplished by means other than mirroring. The projects/organizations mentioned in the previous paragraph are also attentive to load balancing their servers and distributing functions across multiple servers. High availability configurations will generally support scalability so that as traffic increases and content grows, systems will continue to meet the demands placed on them by users. NSDL supports high availability by using a Fedora-level transaction journaling system developed for the project. This allows for replication of transactions in real time to two "follower" systems, ensuring minimal downtime for all updates and failures (Krafft, Birkland, and Cramer 2008).

Backup and restore services are critical for ensuring the content can be recovered in the event of a catastrophic failure. As a first-level of digital preservation, this is one of the simplest yet fundamental actions that any repository must undertake. The determination of frequency of back-ups and the media chosen for backups (e.g. tape, disk) varies across projects. As discussed in the storage section above, storage solutions can have redundancy built into their architecture. In addition to the redundancy provided in its clustered storage, HathiTrust provides system-level backup and restore functionality with both file system backup and database backup. Tivoli Storage Manager software is used to support these backup services.

The decision of whether to host content centrally or to federate the content, whereby the content providers continue to provide access to the resources identified by the large digital library, can limit the reliability of the overall system. Large-scale federated digital libraries include Europeana, NINES, DCC/Opening History and several US State collections. Reliable configurations of large digital libraries will help to ensure resource discovery as well as the use of tools and services that the project/organization provides, but it will not be able to guarantee access to the resources when the user chooses to view or download a digital asset. Unless there are SLAs in place with the content providers that require the providers to maintain a high availability configuration, the large-scale digital library may or may not be able to successfully fulfill the user requests for access. Europeana, for example, has mirrored sites and highly-available, load-balanced servers with distributed functionality, but users will be taken to the content provider's site to

view the actual resource. If that site is down, or, worse, it has suffered a catastrophic failure and has lost content, it will be beyond Europeana's control to retrieve it.

It is a challenge for large federated systems to keep abreast of system availability, especially for projects that have been funded by grants and have no commitment to ongoing maintenance and availability following the end of the grant. Organizations may also move their collections, not notifying systems that aggregate their metadata, and the reference to it may return an error to the user. DCC/Opening History (Grainger Engineering Library, University of Illinois 2009) maintains a federated system that relies on many grant-funded projects to maintain their content after the grant is over. The staff there work to keep up on the links, but with the many collections that are included, it is a challenging task. Digital Commonwealth, the federated repository for the state of Massachusetts (Digital Commonwealth 2007), describes itself as a portal, relying on minimal collection of information centrally and redirecting users to the content provider sites. This helps to minimize the resources required to keep Digital Commonwealth running, but risks the unavailability of digital assets for users who need to access them.

Centralized repositories such as HathiTrust have more control over the content and the availability of the overall system. Challenges, however, for centralized repositories include the ability to grow content more rapidly and the willingness of content providers to give their content over to the centralized digital library. Concerns about rights, attribution and the ability to draw-in users to the provider's larger site or facility are some of the reasons content providers are hesitant to contribute their materials to a centralized large digital library. Central repositories also require content to be provided in specific, supported formats. If contributors do not have sufficient staffing or knowledge to comply with those standards, the burden falls on the large digital library to do the work or decide not to accept the content. Thus, staffing requirements are generally much higher for centralized digital library systems.

Whether centralized or federated, the content format for large digital libraries may or may not be defined by the large digital library. The flexibility of formats that can be included impacts the participation of content providers. NINES is an example of a large digital library whereby initially content format was assumed to be transcribed text with XML markup, most notably Text Encoding Initiative (TEI) (nines.org 2012). As repository platforms became more pervasive, the organization realized that they needed greater flexibility in their guidelines regarding both content and metadata formats, thus they modified their requirements to better accommodate content in repositories such as Fedora, DSpace and CONTENTdm. This has allowed them to scale to a much larger federated collection.

3. Metadata Approaches and Harvesting

Whether centralized or federated, large digital libraries rely on standardized metadata for the resources in their collections. Federated digital libraries harvest

the metadata from their contributing collections and some also harvest full-text of digital text to support resource discovery. Metadata formats, how the metadata is managed in relationship to the content, and harvesting approaches are reviewed in this section.

3.1. Metadata Formats

The Dublin Core (DC) metadata standard is widely used by large digital libraries, though most will also add some additional elements to support the services they provided. Large digital collections that use a common repository platform, like many of the large US State digital libraries, do less customization. For aggregators who harvest metadata from content providers, the metadata requirements that must be met are usually published and the providers are often the ones responsible for ensuring that their metadata meets the specifications. Larger organizations will take the metadata from the providers, along with a mapping, then get the metadata into the needed format for the repository.

As the quantity of digital objects grows to hundreds of thousands of items, organizations sometimes find that they need a new data model to represent the content that will support enhanced services and improved performance. NINES initially planned to manage all content centrally, with protocols for submission, an assumption that content would be marked up in TEI XML, and a METS wrapper would be used to describe and link the resources. After a few years, however, they realized that the collections in NINES could grow much more significantly using a federated approach and adopting a more widely-used metadata standard that would capture content stored in scholarly repositories. NINES now requires content providers to submit their metadata in DC “flavor” of RDF (Nowviskie and McGann 2005). While the burden is on the provider to conform to the metadata requirements, this nonetheless increases the ability of a provider to participate in NINES.

Europeana has used a data model called the Europeana Semantic Elements (ESE) until now. ESE is described as “Dublin Core plus a few project-specific element” (Dekkers et al. 2011). The data is “flat,” enabling Solr to serve as the repository for the metadata as well as the search engine. While this is convenient, there are improvements that the team would like to realize. They are moving to the Europeana Data Model (EDM) where contributors supply their metadata with a mapping file that maps it to EDM. During the mapping ingestion process, enhancements will occur to enrich the metadata such as named entity recognition, linking to Geonames or Virtual International Authority File (VIAF) records, normalization of date values etc. All enriched and normalized fields are stored in separate fields, or aggregations, next to the original record. Moving towards an RDF approach, the enhancements will support links to related content, lending support to a Linked Open Data environment (Heath and Bizer 2011).

Content in HathiTrust is described using a METS file for preservation, structural and technical metadata. PREMIS preservation metadata is updated whenever actions

occur on an object. In addition, MARC21 bibliographic records are held for all of the content. Maintaining traditional cataloging records for digital library content is not as common with other large digital libraries, though creating bibliographic records from digital metadata would not be difficult.

Metadata in NSDL consists of DC elements plus some additional NSDL qualified elements. Additionally, the project stores aggregation objects and agent objects. Relationships among the objects in the digital library are expressed as RDF triples. Support is provided for nested aggregations, enabling richer searches and relationships between digital objects.

3.2. Management of Metadata with Content

Metadata management with their corresponding digital content objects can either be loosely or tightly coupled. Proponents of loose coupling argue that separating the metadata from the objects supports better overall scalability for the digital library and better performance for searching and browsing. The addition of new storage to accommodate large data objects where the data will need to change locations can occur frequently yet seamlessly to end users when the data and content are not tightly coupled. Platforms such as DSpace support this type of architecture for metadata.

Advocates for tighter coupling of the metadata with the content object note that the metadata is an essential component of the digital resource and should always accompany it as it moves. There should be no risk associated with losing the metadata due to its separation from the resource it is describing. It is also critical for the digital preservation of the resource to have the metadata bundled with the object.

Large digital libraries are starting to do both approaches: structured metadata, such as DC, loosely coupled with the content and a descriptive metadata file, such as METS, included with the digital objects.

Aggregated metadata objects are maintained by both NSDL and Europeana, with NSDL also supporting aggregated content. HathiTrust maintains a METS file with the digital content and also maintains catalog records for each resource in its integrated library system, Aleph. NINES keeps the metadata tightly coupled, with RDF embedded with the content. While there is no clear right or wrong approach, it is worth noting that there are advantages and drawbacks with each approach. Understanding the content, how the metadata will be used, who will be responsible for creating the required metadata, and how long-term preservation will be done are important considerations when making decisions about the best approach for metadata that can be sustained by the project.

3.3. Harvesting and Content Ingestion

Harvesting metadata is a way to gather the descriptive information about items in distributed collections for a federated digital library. Having the metadata

centralized can enable common support for functions such as discovery services, timelines, tag clouds, or geospatial visualization that can be used with all the federated collections, even though the content remains distributed. There are differing approaches to harvesting metadata, but many rely on having contributing collections make their data available to the large digital library in a known format. Sophisticated digital libraries such as the DCC/Opening History digital library at University of Illinois will accept multiple formats, but most all harvesters have clear guidelines that either the contributing collection must conform to or that staff at the harvesting repository will create based on the metadata that is provided.

As large digital libraries have discovered, harvesting metadata from varying collections can be a big challenge. While guidelines are provided, content providers may not follow them exactly, leading to additional work that needs to be done by the metadata aggregator. For digital libraries that have lean staffing, this additional work may not be possible and the collection will simply not be included in the larger digital library. Collections using DC metadata can be harvested using a DC harvester such as OAI-PMH. The DCC/Opening History beta sprint for DPLA demonstrates the effectiveness of this, but there are also challenges with harvesting metadata from collections who use non-DC formats (e.g. RDF and TEI markup).

The challenge with standards such as TEI and RDF lies in their flexibility, which has advantages for describing many facets of a work but makes it difficult to do standard processing. DCC/Opening History has successfully shared records with collections that have embedded TEI headers for the object metadata where there is a clearly defined XML Schema Definition (XSD). Embedding MODS and METS in sharing metadata has also been successful. As Tim Cole has noted, the challenge with TEI is that there is not a canonical XSD for it. The modularity of TEI has providers to define an XSD for a given module. There does exist a “tei-all.xsd” to use as a default, but it does not represent the breadth of TEI usage. TEI headers are generally the key element for exchanging metadata files in TEI marked-up collections and providers will create their own XSD for the headers. A very similar challenge exists with RDF. As with TEI, there is no canonical XSD to support RDF; indeed, there is resistance within the RDF community to developing an XSD since it would be almost impossible to capture all of the nuances in the RDF data model. The DCC/Opening History team has, however, found some workarounds to be able to share RDF metadata (Cole, Habing, and Palmer 2011).

Using a crawling approach to harvest metadata records is one way of overcome variances in formats. In their DPLA beta sprint submission, the California Digital Library (CDL) used Apache Nutch to crawl sites to form its search index (California Digital Library 2011). This approach can have significant performance issues, especially if it is not configured for a high-performance computing environment, and presents challenges with reconciling metadata fields. It does, however, enable collections with otherwise incompatible formats. A combination of harvesting and crawling could be very powerful in gathering metadata for access to a variety of collections.

As noted in section 3.1, with the move to EDM, Europeana has taken on the metadata ingestion work where they collect the metadata and mapping files from the content providers, then enhance the metadata when they ingest it. HathiTrust ingests its content through a batch process using the Google Object-Oriented Validation Environment (GROOVE). Originally developed for the Google Books project, GROOVE is a means of batch ingesting works from other collections as well. HathiTrust shares DC metadata for its content to OAI-PMH harvesters.

4. Search and Discovery

Search technologies have improved significantly over the past decade. For text search, the search engine that is most often used by large repositories is Lucene, an open source information retrieval platform (The Apache Software Foundation). Solr is a scalable search engine that uses the Lucene library and is implemented in many digital repositories to support full-text search, hit highlighting, faceted search, dynamic clustering, database integration, and rich document (e.g., Word, PDF) handling (The Apache Software Foundation).

Europeana supports simple and advanced searches, but it only searches metadata, not full text. The Lucene/Solr search engine has worked well for the ESE data model. The flat catalog list simplifies searching, but the move to EDM will enable richer searching and browsing. This will likely lead to modifications in the overall Europeana system for supporting searching and browsing. Other interesting approaches used by large digital libraries include HathiTrust's support for Z39.50 searches of their bibliographic records in OCLC and their ILS, Aleph, and the use of Apache Nutch by both NSDL and NINES. Though NINES is a federated collection, it uses Nutch to also crawl the content in addition to the metadata so that there is a full-text index to support searches in NINES.

Search technologies continue to evolve and improve. The types of discovery the digital libraries need, the varying content formats, and overall performance considerations will shape the decisions regarding the appropriate discovery technologies that should be included in the system.

5. Services and Applications

There is increased attention in large digital libraries to modular development around services. A true Service Oriented Architecture (SOA) approach supports scalability and the addition/subtraction/substitution of technologies over time. HathiTrust and NSDL are two examples of large digital libraries that have embraced this approach in their system development efforts. Defining the functional components of the digital library in terms of services enables changes to be made in isolated or semi-isolated parts of the code with little impact to the other software components. New capabilities and improved technologies can more easily be integrated into the overall system if designers follow SOA principles while

developing the system. Services, once defined and developed can be reused or modified, supporting flexibility and a modular architecture.

Efforts at defining services specifically for education and libraries have been started, though there is no complete registry for all services that are needed by a digital library. Reusing services already defined for other SOA projects, such as the planned services for project Bamboo (Project Bamboo 2012), will enable flexibility and sustainability. The JISC e-Framework initiative maintains a registry of useful reusable services that would likely be of use and provide a shared development community (eFramework Partners 2010). The Digital Library Federation (DLF) explored the establishment of a services framework for digital libraries in 2006 (Lavoie, Henry, and Dempsey 2006), laying a foundation for establishing a services model for developing digital library systems. The micro-services being developed by the CDL are an example of using finer-grained services as the defining building blocks for modular development (Abrams, Kunze, and Loy 2010). All of these efforts can inform the development approaches for a large digital library that can be modified as new and/or improved functionality is available.

In addition to an overall SOA architecture, large digital libraries often provide value-added applications and services. NSDL provides a WordPress MultiUser blog, a MediaWiki and Shibboleth user authentication. HathiTrust offers access to its content through a page turner application and offers a Collection Builder interface. Europeana provides map and timeline views of its resources. NINES applications and services include:

- Juxta, a bibliographical collation system;
- IVANHOE, a multi-player game of literary interpretation;
- Collex, a tool for collecting and annotating digital objects and for publishing interlinked online exhibits; includes support for folksonomic tagging; and
- XML-to-RDF stylesheets for TEI encoded documents.

This is just a sampling of some of the useful applications and services being provided by a handful of large digital libraries. They each provide many more that help to meet their users' needs in working with digital resources. Streaming services for audio and video are another consideration for inclusion when the digital libraries support large audio and video files. JPEG2000 viewers can enhance image viewing for digital libraries with image collections. With the proper systems architecture, new services can be added as needed to ensure that the digital library provides value to its users.

6. System Sustainability

While the overall infrastructure of the digital library is important for providing a robust management system for the digital content, it is equally important to know that the system will continue to operate reliably into the future. Thinking about overall sustainability, not just of the hardware and software, but of the entire

organization early in the process can help to create a successful digital library that users can trust. Decisions about whether to have a federated collection or to manage the content centrally may be driven by a realistic assessment of whether or not there is sufficient commitment to the level of staffing needed to keep the system functioning at all levels.

At the systems level, sustainable systems must be easily maintained and need to scale easily to meet growing traffic and content. For federated systems, there is a challenge with overall sustainability since the content does not reside centrally and is, therefore, out of the control of the library people rely on for discovery and enhanced services. As content contributors become less reliable, the reliability of the overall system is degraded, calling into question whether or not the larger digital library can be considered to be sustainable.

Sustainability can be a key factor in users' trust of the system. As resources are used and referenced, users need to know that those resources will be available to them over time. Therefore, it is important for large digital libraries to publish their efforts and policies that demonstrate they are a viable, trustworthy resource on which users can depend.

HathiTrust is an example of a digital library that has given serious consideration to its long-term sustainability. At the system level, they have implemented a modular architecture where discrete functional services are integrated for effective and efficient operation. This supports quick resolution of specific issues and supports distributed software development. Open standards and open systems are used to enable partners to develop new services and components for repository functionality, thereby leveraging the expertise and contributions of a larger community. HathiTrust is Trustworthy Repository Audit and Certification (TRAC) certified, demonstrating that it meets strict standards for trustworthiness. Good software development practices are in place such as the use of a concurrent versioning system (CVS) repository to support code versioning for software consistency. The University of Michigan Library security policies are followed to ensure that the system meets security guideline. HathiTrust has identified its approach to providing long-term preservation and curatorial services for content. There is a plan in place for disaster recovery. All of these, along with other policies and practices, provide evidence of a thoughtful, sustainable digital library.

7. Summary

The decision to establish a large digital library such as DPLA leads necessarily to a complex set of interwoven considerations. Decisions in one area will impact decisions in other areas. The focus of this study has been on understanding the infrastructure elements of a few large digital libraries that offer diversity in their approaches and can serve as models for DPLA as it embarks on developing a large-scale digital library for US cultural heritage digital assets. While a daunting task, there is much that can be learned from the experiences of those that have preceded

this effort, borrowing the proven successes and learning from the challenges that have already been faced.

In sum, the following are considerations to be mindful of during planning and development for DPLA. Scalability is critical to support long-term growth of the system, thus the architecture decisions should support this. Building modular, following SOA principles, will enable flexibility, code reusability, and stronger support for system sustainability. Understanding who the target audience is for DPLA and what their needs are when interacting with the digital library are critical. Talking with users and documenting how they would use the system is important to ensure that a useful system is implemented. It is important to decide on a realistic sustainability plan and to publish the policies and guidelines that will help to enforce that plan.

As implementation begins, good project planning will keep the project on track and in scope. Work should be planned at a level that can ensure success. Developing documented usage scenarios to guide decisions about the most important functions can help with the scope and architecture. The most important functions, especially those that impact the overall architecture, need to be prioritized over the “nice-to-have” features that can be implemented after the core system has stabilized. While decisions about the infrastructure can be made based on research and discussion alone, a better approach will be to establish a sandbox environment to experiment with differing technologies and architectures. This will also help with launching beta code in the open source “release early, release often” ethos.

Lastly, an area where DPLA has already demonstrated leadership is in open communication and inclusion in its efforts. With proper management, this will help to keep the community informed and invite participation from those whose expertise can contribute to the common good that DPLA strives to provide.

Bibliography

- Abrams, S., J. Kunze, and D. Loy. "An emergent micro-services approach to digital curation infrastructure." *International Journal of Digital Curation* 5, no. 1 (2010).
- Amazon Web Services, LLC. "Amazon Simple Storage Service (Amazon S3)", n.d. <http://aws.amazon.com/s3/#pricing>.
- Apache Foundation. "FrontPage - Nutch Wiki." *Nutch Wiki*, November 27, 2011. <http://wiki.apache.org/nutch/>.
- California Digital Library. "CDL's DPLA Vertical Search Demo." *DPLA Vertical Search Demo*, 2011. <http://crawlspac.cdlib.org/>.
- Cole, Tim, Thomas G. Habing, and Carole Palmer. "Opening History Email Clarification", December 9, 2011.
- Cornell University. "MPTStore 0.9.1 Documentation." *MPTStore 0.9.1*, 2006. <http://mptstore.sourceforge.net/>.
- Coyle, K. "Mass digitization of books." *The Journal of Academic Librarianship* 32, no. 6 (2006): 641–645.
- Dekkers, Makx, Stefan Gradmann, and Jan Molendijk. "D3.4 Final Technical & Logical Architecture and future work recommendations". Europeana group, October 5, 2011. http://www.version1.europeana.eu/c/document_library/get_file?uuid=d0327b50-2e86-45bd-81c1-7bff4b9a449b&groupId=10602.
- Dekkers, Makx, Stefan Gradmann, Carlo Meghini, Catherine Lupovici, Go Sugimoto, Robina Clayphan, Julie Verleyen, et al. "Europeana Sematic Elements Specifications v3.4", March 31, 2011.
- Digital Commonwealth. "Digital Commonwealth." *Digital Commonwealth: Massachusetts Collections Online*, 2007. <http://www.digitalcommonwealth.org/>.
- Digital Library of Georgia. "About the Digital Library of Georgia." *Digital Library of Georgia: Sharing Georgia's History and Culture Online*, January 27, 2012. <http://dlg.galileo.usg.edu/AboutDLG/>.
- eFramework Partners. "e-Framework for Education and Research." *Service Genre Registry*, April 9, 2010. <http://www.e-framework.org/Default.aspx?tabid=987>.
- Grainger Engineering Library, University of Illinois. "Opening History." *IMLS DCC*, 2009. <http://imlsdcc.grainger.uiuc.edu/history/>.

- HathiTrust. “Technological Profile | www.hathitrust.org.” *HathiTrust Digital Library: Technological Profile*, n.d. <http://www.hathitrust.org/technology>.
- Heath, Tom, and Christian Bizer. *Linked Data: Evolving the Web into a Global Data Space*. Vol. 1. 1st ed. Synthesis Lectures on the Semantic Web: Theory and Technology 1. Morgan & Claypool Publishers, 2011. <http://linkeddatabook.com/editions/1.0/>.
- Isilon Systems. “OneFS Operating System | Isilon Systems.” *Isilon Systems*, 2011. <http://www.isilon.com/onefs-operating-system>.
- Krafft, D.B., A. Birkland, and E.J. Cramer. “Ncore: architecture and implementation of a flexible, collaborative digital library.” In *Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries*, 313–322. ACM Press, 2008. <http://arxiv.org/abs/0803.1500v1>.
- Lavoie, Brian, Geneva Henry, and Lorcan Dempsey. “A Service Framework for Libraries.” *D-Lib Magazine* 12, no. 7/8 (August 2006). <http://www.dlib.org/dlib/july06/lavoie/07lavoie.html>.
- nines.org. “N I N E S.” *NINES: Nineteenth-century Scholarship Online*, 2012. <http://www.nines.org/>.
- Nowviskie, Bethany, and Jerome McGann. “NINES: A Federated Model for Integrating Digital Scholarship”. NINES.org, September 2005. <http://www.nines.org/about/wp-content/uploads/2011/12/9swhitepaper.pdf>.
- Project Bamboo. “Project Bamboo Technology Wiki - Home.” Wiki. *Project Bamboo Technology Wiki*, February 1, 2012. <https://wiki.projectbamboo.org/display/BTECH/Technology+Wiki+-+Home>.
- Schmitz, D. “The Seamless Cyberinfrastructure: The Challenges of Studying Users of Mass Digitization and Institutional Repositories.” *Council on Library and Information Resources* (2008). www.clir.org/pubs/archives/schmitz.pdf.
- SeekingMichigan.org. “Seeking Michigan.” *Seeking Michigan*, 2012 2008. <http://seekingmichigan.org/>.
- State Library and Archives of Florida. “Florida Memory Project.” *Florida Memory*, 2012. <http://www.floridamemory.com/>.
- The Apache Software Foundation. “Welcome to Apache Lucene!”, n.d. <http://lucene.apache.org/>.
- — — . “Welcome to Solr”, n.d. <http://lucene.apache.org/solr/>.

Vancis BV. “Vancis BV - a subsidiary of SARA | Vancis.” *Vancis Advanced ICT Services*, n.d. <http://www.vancis.nl/en/>.

W3C. “RDF - Semantic Web Standards.” *W3C Semantic Web*, March 7, 2010. <http://www.w3.org/RDF/>.

Wikipedia. “Apache Solr - Wikipedia, the free encyclopedia”, n.d. http://en.wikipedia.org/wiki/Apache_Solr.

— — —. “Lucene - Wikipedia, the free encyclopedia”, n.d. <http://en.wikipedia.org/wiki/Lucene>.