Linked Data and Microservices at the Support of Customized Institutional Workflows*

Greg Jansen[1], Mark Conrad[2], Lyneise Williams[3], and Richard Marciano[1]

[1] University of Maryland, College Park MD 20742, USA
[2] Advanced Information Collaboratory (AIC), Keyser WV 26726, USA
[3] UNC Chapel Hill, Chapel Hill NC 27516, USA
jansen@umd.edu, conradsireland@gmail.com, lyneise@gmail.com, marciano@umd.edu

Abstract: This paper presents a highly innovative prototype infrastructure for linked archives. We propose a novel open-source approach to repository design, one that bridges the boundary between a disorderly world and an orderly inner sanctum. This design seeks to acquire data as early as possible from organizational workflows and active record systems, then employs microservices to extract, transform, and load (ETL) workflow-specific inputs into consistent and reconciled linked data graphs. We illustrate this approach using the digital assets of the National Park Service Mary McLeod Bethune Council House National Historic Site. We show the ability to create a flexible and reconfigurable interoperability layer that can bridge existing systems composed of a combination of independent proprietary, custom, and open-source components.

Keywords: Linked data, Infrastructure design, Mary McLeod Bethune Council House.

1 Background and Goals

This research was funded through a three-year collaborative agreement between the U.S. National Park Service Mary McCleod Bethune National Historic Site (MMBNHS) [1] and the research team at the School of Information Studies at the University of Maryland, College Park. Faced with an accumulation of digitized photographs and other media, the MMBNHS realized that current strategies for managing the digital versions of their archival collections could benefit from a rethinking of the underlying infrastructure, both in terms of long-term preservation and public access.

1.1 National Archives for Black Women's History Collections

Mary McLeod Bethune was a pioneering community organizer. The National Archives for Black Women's History (NABWH) Collections, document Mary McCleod Bethune's role in the federal government as a power broker [2]. Her success laid the path for Kamala Harris becoming Vice President of the United States of America. No other collection tells the story of Bethune's rise to power and her role as the most powerful Black woman of the first half of the 20th century. When she formally accepted the nomination for VP at the Democratic Convention in August 2020, Harris named

McCleod Bethune as one of the women who "inspired us to pick up the torch, and fight on."

Franklin D. Roosevelt and First Lady Eleanor Roosevelt appointed Bethune a presidential advisor of African American affairs in 1934. Bethune's position within the Roosevelt Administration would give her the leverage to form the Federal Council of Negro Affairs, which would become known as The Black Cabinet. Bethune achieved hearings for black concerns at the highest governmental levels, nurturing the principle that blacks were integral to the American body politic.

The collections also document the National Council of Negro Women and other African American women's organization. The collections include documentation of individuals – famous and not so famous - associated with these organizations.

1.2 Building Innovative Linked Data Infrastructure

Our goal is to preserve and manage the current and future digital assets of the Mary McLeod Bethune Council House National Historic Site, in order to enable increased access to these assets and better serve the African American community. The project has a public purpose of promoting greater public and private participation, and sharing the information, products and services to increase public awareness. The main objectives are to conduct a Digital Asset Management System needs analysis, create a model repository for the assets, and develop means of access both by NPS staff and members of the public. Our project leverages NCSA Brown Dog [3], Drastic Fedora [4], and Trellis LDP. Trellis is the software layer supporting Sir Tim Berners-Lee's Inrupt Solid technology.

2 System Design

In this paper we propose a different approach to repository design, one that bridges the boundary between a disorderly world and an orderly inner sanctum. This design seeks to acquire data as early as possible from organizational workflows and active record systems, then employs microservices to extract, transform, and load (ETL) workflow-specific inputs into consistent and reconciled linked data graphs. The results of the repeatable ETL processing are superimposed upon the source data, which is left in the user-supplied format and folder structure, fixed and undisturbed. This supports the archival principle of respect des fonds. We will demonstrate many benefits of this design through our prototype system.

2.1 Submission Workflows

When thinking about system design for cultural heritage, especially given the option of formal semantics afforded by linked data, we have a tendency to imagine an information architecture that is like a cathedral, an enclosure that expresses our desire for perfection in symmetry and elegant lines and that will stand the test of time. These virtues correspond in our minds to the need for fixity, valid metadata, content models,

standard ontologies, and durability, in short, a space where everything inside meets well-chosen data standards that are rigorous and timeless. In pursuit of such a pure data space, we risk creating a kind of inner sanctum, one in which all data must be valid prior to entry. However, the problem with such a vision and with the design mindset that may accompany it, is that the world outside of the software system is less than pure. It is more often messy, chaotic, and incomplete. Data presents itself in various forms at various times and it contains varying levels of user error and data corruption.

Submissions as physical transfers: Members of the team have been involved in several previous repository projects that attempted to bar the gate to disorderly information. However, the inevitable result of enforcing a strict order inside of the repository is that there are more complex and cumbersome ingest workflows created outside the system to meet its demands for perfect consistency. We see this in baroque and failure-prone ingest pipelines that deliver highly processed submission packages to repositories. These ingest pipelines process submission packages, such as ZIP files, that are much like miniature repositories themselves, with their own internal folder structures, naming conventions, and metadata. These pipelines create a pristine package for physical transfer into the storage platform. The complexity involved in staging the ingest event, complete with boxing and unboxing the content, is a bit punishing.

NPS Workflows: The digital collections at the MMCBHS are growing all of the time through several workflows. *Archival description* is created through the Interior Department Catalog Management System (ICMS), which also provides archives and museum management functions. This information is periodically exported as **XML files**. Vendor-supplied *digitization packages* contain sets of **TIFF image files** that follow a file-naming convention that mirrors collection structures. MD5 digests for each scanned page file and item-level Dublin Core descriptions are also supplied in **Excel files**. *Authority records* are kept in curated lists in the ICMS and may be exported as **XML files**.

Submissions as units of work: In the prototype software uploading starts when the files and staff are available to begin the work of a submission, even if a batch of files is incomplete or has known issues. Such a submission is marked as a work in progress until errors have been detected and fixed and staff can mark it as complete. Early upload, even by drag and drop of files onto a web page, moves digital assets off of vulnerable storage as soon as possible. Instead of expecting every submission to conform to a limited set of platform content models, the submission structure will mirror the content stream produced by the organization. These are vendor digitization batches and museum catalog exports at the NPS, but other organizations might capture the outputs from scientific instruments, entire email accounts, or drive transfers. In most cases a microservice will be created to interpret the submission files, identify errors, and extract linked data. However, the raw data can be captured, preserved, and potentially accessed even before a more specific microservice has been created. For preservation and historical research, the ability to access the original file format may be desirable. The platform anticipates that there will be new streams of data that need to be captured first and only fully processed later.

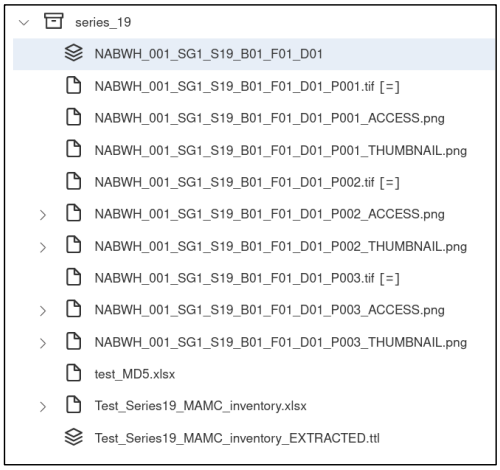

Fig. 1. Uploaded images and Excel files with extracted objects.

2.2 Reproducible Linked Data Graphs

Having arranged submission files according to how they are produced or managed outside of the system, we run microservices that extract data from the workflow-specific file structures. Each microservice captures the knowledge required to transform a particular set of files into a linked data graph that follows a platform-wide content model and metadata structure. This linked data is stored in named graphs that are associated with the files from which they were derived, making it straightforward to trace any assertions to a submission file, a necessary step towards trustworthy archives. We can regenerate the named graphs in an automated way any time submission files are added, replaced, corrected, destroyed, or enhanced. We can also regenerate the graph when the microservice that performs linked data extraction is improved. Because the extracted linked data is isolated in its own graph, it cannot pollute the rest of linked data in the repository. Problems can be detected and reported out to users for corrective measures. For example, a staff member can receive an email from a microservice that prompts them to add the missing Dublin Core record to a vendor supplied Excel file. Another microservice, responsible for building paged documents conforming to the Portland Common Data Model (PCDM) [5], might inform staff that a page file is missing from the sequence of file names that is expected.

Curating an enhanced archival context: The separation of file management from linked data processing holds great promise for enhancing historical context. It allows a file to describe just one or multiple archival items or topics. Submitted files can contribute facts about the holdings, such as EAD or Dublin Core metadata, but they can also provide facts about the archival context that surrounds the holdings, including authority lists and further descriptions of entities like organizations and people. Workflows for managing archival context can be created in the same way as archival submissions, perhaps processing Excel files that staff create to describe historical individuals and organizations.

An example document: *Figure 1* shows a view from the NPS prototype of a document produced through the existing digitization workflow and microservices. It shows three uploaded TIFF images and two Excel files, containing batch MD5 checksums and Dublin Core metadata. Around these uploaded files several new objects have been created, including a Portland Common Data Model paged document representation (second line in the figure - shaded) that was implied in the names of the uploaded files. A PREMIS-based subgraph for each file records fixity, both before and after file upload. Additionally, there are two access copies for each image and a batch metadata object that contains the Dublin Core re-encoded as linked data (RDF turtle format).

2.3 The Big Graph

As the graphs of linked data from various submissions, stemming from a variety of different workflows, are added, curated and refined over time, they are also aggregated in a triple store index (technically an Apache Fuseki quad store in this case). The triple store presents a unified linked data landscape that has many uses. Data that is curated in many different ways is united in the same big graph. For instance, the NPS descriptive hierarchy from the Department of Interior Collection Management System (ICMS) – a commercial off-the-shelf software system-- is united in this structure with paged documents from the vendor digitization process. More than anywhere else in this system, this big graph represents the data according to a platform-wide content model. This consistent and conformant overview of all the platform data enables publishing, exports, reporting, and search indexing. For instance, we can leverage this unified graph to publish collections or certain archival series in the NPS NPGallery Digital Asset Management System. We can also use it to create a search index and a finding aid-based website for archival access by historians and community researchers.

2.4 Distributed Services Architecture

The microservices that support ingest processing are just one part of an ecosystem of microservices, storage clusters, and web servers that together create the functions of digital asset management [6]. In the prototype for the MMBNHS we have incorporated important architectural features from our previous repository research. In the DRASTIC Fedora Project we demonstrated the capacity and performance benefits of horizontal scaling through stateless servers and distributed databases. In that research we showed that this approach can reliably keep pace with a performance load of around 500 upload requests per second [4]. Now we have added the industry-proven Kafka event streaming platform in order to run microservices and coordinate workflow actions. With these features we can maintain acceptable performance under any user load and meet arbitrary data storage demands [7].

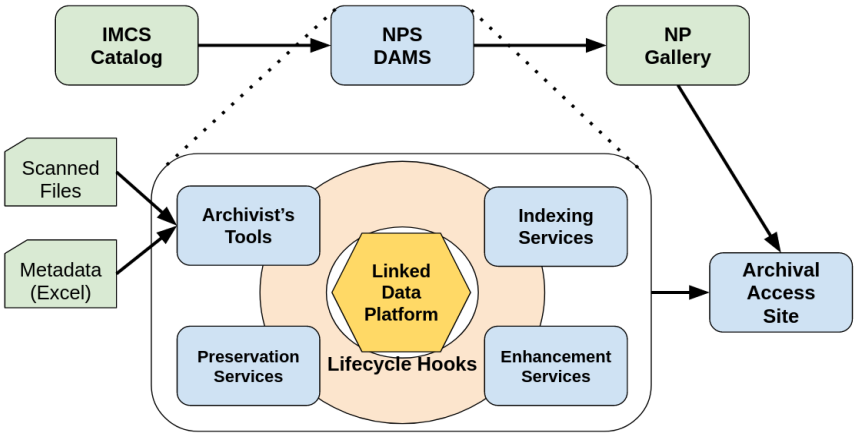

Fig. 2. Digital Asset Management functions as microservices in blue, as orchestrated by lifecycle hooks (event streaming) in light yellow around the persistence services provided by the linked data platform in gold. Context of the U.S. National Parks Service is shown in green.

The open-source technology in the prototype was chosen in order to make the system open to extension. In particular, Apache Kafka (a community distributed event streaming platform capable of handling trillions of events a day [8]) and the HTTP application programming interface provided by Trellis LDP are programming language neutral. Microservices for submission processing, format migration, or computational enhancement may be written in Python, Java, Ruby, or any programming language. New microservices can support additional content streams from organizational workflows or computational treatments such as optical character recognition or other types of media and image processing.

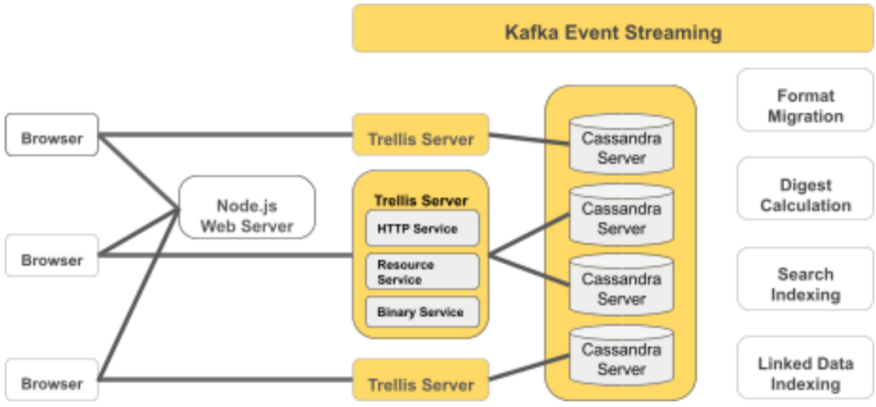

Fig. 3. Horizontal-scaling services at work in the prototype, including Trellis LDP, Apache Cassandra, Apache Kafka, and some pluggable microservices.

3 Conclusion

This project represents the culmination of several threads of research in digital archives and linked data (based on prior funding from NSF and IMLS in particular). The open-source software prototype has demonstrated the versality and potential for workflow processing for MMBNHS catalog data and digitization packages, while not a fully functional digital asset management system or preservation system. We designed, implemented, and tested an innovative approach to linked data infrastructure that leverages open-source and massively scalable systems including Trellis LDP, NoSQL Apache Cassandra / Fuseki / Kafka. We demonstrated the ability to create a flexible and reconfigurable interoperability layer that can bridge existing systems composed of a combination of independent proprietary, custom, and open-source components.

References

1. "Mary McCleod Bethune Council House National Historic Site", website. https://www.nps.gov/mamc/index.htm
2. "National Archives for Black Women's History", website. See: https://www.nps.gov/mamc/learn/historyculture/nabwh_collections.htm
3. S. Padhy, J. Alameda, R. Kooper, R. Liu, S.P. Satheesan, I. Zharnitsky, G. Jansen, M. Dietze, P. Kumar, J. Lee, R. Marciano, L. Marini, B. Minsker, C. Navarro, M. Slavenas, W. Sullivan, K. McHenry, "An Architecture for Automatic Deployment of Brown Dog Services At Scale into Diverse Computing Infrastructures", July 2016 XSEDE16: Proceedings of the ACM XSEDE16 Conference on Diversity, Big Data and Science at Scale.
4. Jansen G., Coburn A., Soroka A., Thomas W., Marciano R., "DRAS-TIC Fedora: Evenly Distributing the Past", Invited for submission to a Special Journal Issue "Selected Papers from Open Repositories 2018", July 4, 2019 at MDPI Open Access. See: https://www.mdpi.com/2304-6775/7/3/50.
5. Wilcox, David. "A linked data approach to digital newspapers with Fedora and PCDM." (2016). See: http://origin-www.ifla.org/files/assets/newspapers/2017_Iceland/2017-wilcox-en.pdf
6. M. Villamizar *et al*., "Evaluating the monolithic and the microservice architecture pattern to deploy web applications in the cloud," *2015 10th Computing Colombian Conference (10CCC)*, 2015, pp. 583-590, doi: 10.1109/ColumbianCC.2015.7333476.
7. Jansen, Gregory, Aaron Coburn, Adam Soroka, and Richard Marciano. "Using Data Partitions and Stateless Servers to Scale Up Fedora Repositories." In *IEEE BigData*, pp. 3098-3102. 2019.
8. P. Le Noac'h, A. Costan and L. Bougé, "A performance evaluation of Apache Kafka in support of big data streaming applications," *2017 IEEE International Conference on Big Data (Big Data)*, 2017, pp. 4803-4806, doi: 10.1109/BigData.2017.8258548.