Starting from scratch – building the perfect digital repository

Piotr Hebal, Galter Health Sciences Library, piotr.hebal@northwestern.edu
Violeta Ilik, Galter Helath Sciences Library, violeta.ilik@northwestern.edu
Kristi Holmes, Galter Health Sciences Library, kristi.holmes@northwestern.edu

Session Type (select ion in bold)
- ☐ Panel
- ☐ **Presentation**

Abstract

By establishing a digital repository for the Feinberg School of Medicine (FSM) (Northwestern University, Chicago campus), we anticipate gaining the ability to create, share, and preserve attractive, functional, and citable digital collections and exhibits. Galter Health Sciences Library did not have a repository as of November 2014. In just a few months we formed a small team that was charged at looking to select the most suitable open source platform for our digital repository software. We followed the National Library of Medicine master evaluation criteria by looking at various factors that included: **functionality**, **scalability, extensibility**, **interoperability**, **ease of deployment**, **system security**, **system, physical environment**, **platform support, demonstrated successful deployments**, **system support**, **strength of development community**, **stability of development organization**, and s**trength of technology roadmap for the future**. These factors are important for our case considering the desire to connect the digital repository with platforms that produce VIVO-compatible structured linked data. VIVO is a linked data platform that serves as a researchers' hub and which provides the names of researchers from academic institutions along with their research output, affiliation, research overview, service, background, researcher's identities, teaching, and much more.

Conference Themes
(selection in bold)

☐ **Supporting Open Scholarship, Open Science, and Cultural Heritage**
☐ Managing Research (and Open) Data
☐ Integrating with External Systems
☐ Re-using Repository Content
☐ Exploring Metrics and Assessment
☐ Managing Rights
☐ Developing and Training Staff
☐ **Building the Perfect Repository**

Keywords
Open Science, Open scholarship, Semantic web

Audience
Repository managers, developers, data producers, librarians

Background

Our proposal is in line with the conference theme, more specifically with **"**Supporting Open Scholarship, Open Science, and Cultural Heritage Online" and "Building the perfect repository."

Presentation content

By establishing a digital repository for the Feinberg School of Medicine (FSM) (Northwestern University, Chicago campus), we anticipate gaining the ability to create, share, and preserve attractive, functional, and citable digital collections and exhibits. Galter Health Sciences Library did not have a repository as of November 2014. In just a few months we formed a small team that was charged at looking to select the most suitable open source platform for our digital repository software. We followed the National Library of Medicine master evaluation criteria by looking at various factors that included: **functionality**, **scalability, extensibility**, **interoperability**, **ease of deployment**, **system security**, **system, physical environment**, **platform support, demonstrated successful deployments**, **system support**, **strength of development community**, **stability of development organization**, and

strength of technology roadmap for the future. All these factors played a key role in determining the best platform for our needs with special attention to interoperability and strength of technology roadmap for the future. These factors are important for our case considering the desire to connect the digital repository with platforms that produce VIVO-compatible structured linked data. VIVO is a linked data platform that serves as a researchers' hub and which provides the names of researchers from academic institutions along with their research output, affiliation, research overview, service, background, researcher's identities, teaching, and much more. VIVO's semantic approach to research networking has been widely adopted and the VIVO data standard is a recommendation and best practice for representation of information about research and researchers across the 62-member Clinical and Translational Science Award (CTSA) Consortium. CTSA Hubs are encouraged to "implement research networking tool(s) institution-wide that utilize RDF triples and an ontology compatible with the VIVO ontology… [and] people profiles at institutions should be publicly available … as Linked Open Data."[1]

Considering all of these factors and considering the important role the library plays as an information hub on campus, the Galter Health Sciences Library team, as a member of the Northwestern University Clinical and Translational Sciences Institute (NUCATS) looks to establish a digital repository that will enable open representation of FSM and NUCATS scholarly outputs and outcomes. Open access principles can help guide dissemination strategies for the broad range of products and outcomes of research from the diverse biomedical workforce. Our goal is to house traditional and non-traditional scholarly outputs in the Galter Digital Repository (GDR). Non-traditional outputs are defined for this purpose as items produced during the scholarly process but which are often not discoverable or made available for reuse, including measurement devices, patient education materials, curriculum materials, conference materials, community engagement materials, etc. Open access to the products and outcomes of research are increasingly required by funders and therefore we plan to place all visualizations, workforce development materials, published articles, and non-traditional outputs openly available via the GDR.

Taking into account the possibilities of the different frameworks that provide digital repositories architecture we selected the Fedora open source architecture. According to the Fedora's DuraSpace wiki page Fedora's "flexibility enables it to integrate with many types of enterprise and web-based systems, offering scalability and durability. It also provides the ability to express rich sets of relationships among digital resources and to query the repository using the semantic web's SPARQL query language." [2]

Our first collection that served as a test was the collection of photographs, manuscripts, letters, and addresses (speeches) by/about Greene Vardiman Black, the father of modern dentistry. The collection was previously digitized and described with the help of Encoded Archival Description (EAD). EAD is an XML standard for encoding archival finding aids, maintained by the Technical Subcommittee for Encoded Archival Description of the Society of American Archivists, in partnership with the Library of Congress.[3] We cross-walked the existing EAD metadata in order to display it in our repository stack Fedora/Sufia/Blacklight. To provide for rich metadata we added the Medical Subject Headings terms (MESH) to enable users to select keywords and subjects from a controlled vocabulary. We also extended the data streams to include additional fields to accommodate all the publication types present in National Library of Medicine. This will allow us to seamlessly move data between systems: GDR and VIVO. We also integrated the Lightweight Directory Access Protocol (LDAP) into the workflow enabling users to select names of creators and contributors from a controlled list of names.

Two other collections have been submitted to the Digital Projects Working Group, established to help the technical team with outreach and support. The faculty members and the owners of the collections provided the digitized copies and the appropriate metadata. The Digital Projects Working Group will continue to actively solicit and identify potential candidates for inclusion in the GDR. The Galter Library Team has great support from the FSM, NUCATS, and the Center for Data Science and Informatics. They all provide marketing, outreach and technical support to the library team, poising us for success.

Conclusion

Commitment to open source software is the primary driving force in establishing a repository based on Fedora framework. As we look to integrate the repository with existing projects and databases from the FSM, NUCATS, and CDSI we are encouraged by the stability and strength of the open source community that is behind Fedora and VIVO and look forward to leveraging standard data architectures and workflows to unveil a more rich picture of the outputs and activities within and across our organization.

References

[1] The Clinical and Translational Science Awards (CTSA) Consortium, 2013. Research Networking, https://www.ctsacentral.org/best%20practices/research%20networking

[2] DuraSpace, 2013. Fedora, http://www.duraspace.org/about_fedora

[3] Society of American Archivists, 2015. Encoding Archival Description (EAD). http://www2.archivists.org/groups/technical-subcommittee-on-encoded-archival-description-ead/encoded-archival-description-ead