Starting from scratch - building the perfect digital repository

Piotr Hebal, Violeta Ilik, and Kristi Holmes

Galter Health Sciences Library, Feinberg School of Medicine, Northwestern University

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. 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**. These factors played a significant role in determining the best platform for our needs with special attention to interoperability and strength of the technology roadmap for the future. These factors are especially 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]

BACKGROUND

The Galter Health Sciences Library team, as a member of the Northwestern University Clinical and Translational Sciences Institute (NUCATS), is establishing a digital repository to enable open representation of diverse scholarly outputs and outcomes by our scholars. 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 provide a digital home for traditional and non-traditional scholarly outputs in the Galter Digital Repository (GDR). Non-traditional outputs (defined for this purpose) are items produced during the scholarly process but which are often not discoverable or made available for reuse through the traditional scholarly publishing workflow, including measurement devices, patient education materials, curriculum materials, conference materials, community engagement materials, and so on. Open access and availability to the products and outcomes of research are increasingly required by funders and can serve as an important way to demonstrate return on investment to partners and our communities. For these reasons, the GDR serves as an important lynchpin in the evaluation and continuous improvement activities of NUCATS and other projects at FSM. FSM also has a rich digital heritage which we will continue to build through the GDR, as well.

After taking into account the possibilities of the different frameworks that provide digital repositories architecture we selected the Fedora open source architecture. From 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]

GV Black Collection

Our first test collection 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. 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), Library of Congress Subject Headings (LCSH), Subject Names, and Subject Geographic Names to enable users to select keywords and subjects from a controlled vocabulary. We also expanded the possible "Resource type" options by adding publication types from VIVO-ISF Ontology and the Local Northwestern Ontology to accommodate all the publication types from the National Library of Medicine. This will allow us to seamlessly move data between systems: Repository and VIVO.

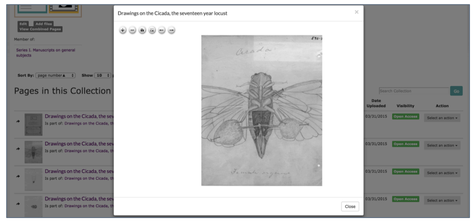
IIIF & OpenSeadragon

All images in the repository are viewable through our IIIF server. Furthermore, we serve IIIF presentation metadata for all of our collections and files. For the front-end pager we use the actively maintained OpenSeadragon, included with Sufia. In addition to paging, it supports features such as zooming, panning and browsing. Other pagers can be easily integrated as long as they support IIIF.

Inclusion of IIIF allows us to group series of files into entities such as books and then display them using OpenSeadragon. Users have the ability to mark any Collection as 'Multi-Part'. A 'Multi-Part Collection' contains individual files and a link to a combined PDF file.

Each file in such a collection can have a page number that can be anything alpha-numeric as found in the source document, backed by a sort-number indicating the order pages are to be displayed. Only files with page numbers are considered part of the collection and are viewable in OpenSeadragon pager.

All IIIF related information are indexed in Solr along with other RDF metadata for fast retrieval. The IIIF server respects the authorization scheme used in the hydra-access-controls gem. The IIIF service uses permission information stored in Fedora, indexed in Solr and it serves images only to properly authorized users.

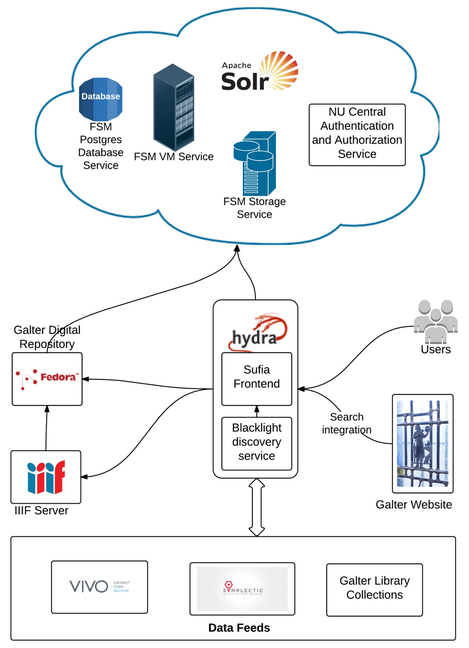


System Architecture and Customizations

Our repository runs Fedora 4.2 and is front-ended by heavily customized version of Sufia 6.0. We have integrated the authentication with LDAP and groups used for authorization are also sourced from LDAP. Furthermore, certain metadata fields such as 'Creator' only allow entries that are verifiable in LDAP to consistently identify individuals. All our software runs on CentOS 6.5 and we use Postgres as a relational database.

We have the ability to batch various feeds including EAD to create hierarchical collections. Rails stack is proxied by Passenger and Apache and includes various parts of the Hydra stack, custom RDF vocabularies, IIIF server and background workers for Resque.

Our code is open to the public on Github: https://github.com/galterlibrary/digital-repository/tree/master/app. The main goal of customization is to make sure that we are able to keep up with the upstream. If changes to the upstream code are needed pull requests should be opened with the upstream. All other changes are done by monkey-patching appropriate objects.



Repository & FSM Databases

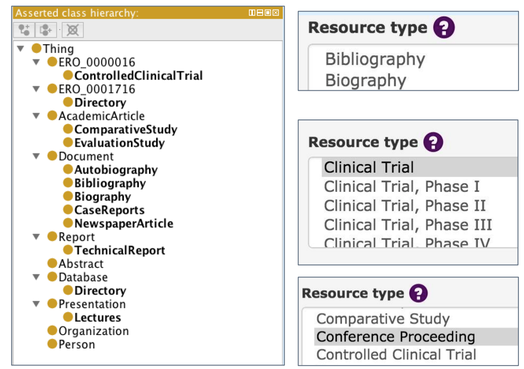
Current: LDAP integration

We 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 which is the same data source for the Northwestern VIVO instance.

Future: Symplectic, VIVO, ORCID, Shibboleth, FASIS

VIVO-ISF Ontology and Local Northwestern Ontology Extensions to represent National Library of Medicine Publication Types

Feinberg School of Medicine researchers need to be able to correctly represent their publications, therefore we created an extension to the VIVO-ISF Ontology to represent most of the National Library of Medicine publication types. This allows for granularity and correct representation of types of scholarly outputs by FSM researchers.



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