

# Repository Comparison Report

---

June 2, 2015

Team Members: Ellen Ramsey, Anne Gaynor, Jill Heinze, Sherry Lake, Ray Lubinsky, Esther Onega

## Executive Summary

In November 2014, the University of Virginia Library's Libra Services Team submitted a Libra 2.0 Project Proposal to Library leadership for replacement of the Libra 1.x institutional repository.

The key factors indicating the need to replace Libra 1.x remain:

- 1) to improve information flow into and out of the repository (for depositors, for administration, and for analysis);
- 2) to eliminate the risks and inefficiencies of further development on a platform comprised of components that are now obsolete and unsupported.

The following report is a deliverable of the Libra 2.0 Project Proposal:

The Libra Services Team will compare advantages and features of other identified institutional repositories with in-house development of UVa-specific requirements. Usability testing will be conducted to evaluate fit with repository requirements as expressed in [the Libra 2.0 Project Proposal] document. A report comparing alternatives to building our own Fedora 4-based systems will be generated by March 2015.

As promised, the Libra 2.0 Project Team undertook this comparative study of current repository tools from reputable partnerships, peer institutions, and vendors. Software containers for open access to items including theses, dissertations, articles, books, book chapters, grey literature, and conference materials were compared as options for the "OA Repository" needs of the University. A separate comparison of "Data Repositories" for the University's research outputs was simultaneously undertaken.

Four broad categories of requirements were evaluated for each repository:

- User interface and user experience
- Statistics and reporting
- Infrastructure
- Metadata

Each system or product was shepherded through the evaluation process by a single team member, with a formal presentation of features, structure, and functionality presented in a live product walkthrough during a Libra 2.0 Team meeting. Team members gathered information about products by contacting users and administrators at peer institutions, program managers, in relevant online and in person community forums, and from vendor representatives.

# OA Repositories Comparison

## Evaluation Criteria

Four broad categories of requirements were evaluated for each candidate, each with detailed criteria for features, functionality, or structure:

### User interface and user experience

The team sought solutions that would enable depositors to maintain a high degree of involvement with their works, both during the deposit process and after, so that they can easily assess the impact of their scholarship. The team also sought solutions to give end users a better experience including giving more options for citation export and user-friendly system feedback.

- Depositor dashboard showing drafts in progress, completed submissions, and statistics (see below), with role-based access control: administrators will see more statistics, have more options, than the average depositor.
- Upload progress indicator.
- Expose object metadata in COinS for export to Mendeley/Zotero/RefWorks.
- Scientific notations, mathematical equations, and special character handling.
- System feedback and error handling.

### Statistics and reporting

In order to assess and demonstrate success, it is important that a repository solution provide tools for running detailed and consistent statistical reports. In keeping with OA community practice, a solution that provides some level of statistics to the end user is preferable.

- Total number of OA item deposits for a particular time period.
- Total number of visits to repository.
- Number of views and downloads per OA item. Statistics should include page views for embargoed files.
- Geographic metrics for views and downloads of objects (per Harvard model).
- Automated reporting based on statistics (e.g. automatic production of annual faculty effort report).
- Role-based levels of visibility and delivery mechanisms for statistics. Individual depositors should receive aggregated download and access statistics for their deposits via automated email reports and a user dashboard, while other statistics such as geographic metrics should be accessible to any site visitor.

### Infrastructure

As Libra ages, updates to its infrastructure are ever more critical to keep up with both user and administrative needs. The following criteria seek ways to give more authority to select users to upload batches of items as well as technical upgrades to enable the Library to actively engage the needs of the OA community.

- Allow embargoes.
- Allow records to include related URL(s).

- Multimedia deposits may require the option to include a URL in an object record to point to a multimedia resource. For example, a multimedia file associated with a deposit may point to a streaming A/V resource in Avalon.
- A record could consist of only metadata with a pointer from the Libra record to item in a discipline-based or other vetted repository.
- Allow for links between Libra objects. Create a simple way for users to create links between objects in Libra. For example, allow for users to link an OA article to an OA dataset/product of research.
- Allow supplemental/individual files to have different release/embargo restrictions.
- Group/batch object deposit and management: Enable batch loading so superuser can add a group of items at one time.
- Implement a “Managed Works mode” so that an authorized user can manage a group of related items such as the contents of a journal.
- Copy-request button for embargoed items, using contact information authors provide for automatic reporting functions. The button sends an automatic request email from the researcher to the author, with no intervention by Libra staff.
- Create an API.
- Create an OAI-PMH endpoint (critical for SHARE participation and other metadata aggregation services).
- Compatibility or interoperability with Fedora/Hydra.
- Time to implementation/availability.
- Locally installed or remotely hosted.
- Maximum size of deposit item.
- Licensing cost.
- Estimated customization cost.

### Metadata

This development goes hand in hand with infrastructure. In order to realize developments in infrastructure, expansion and flexibility of metadata is key. Incorporating additional metadata fields and features also positions the Library to actively participate in research sharing and linking initiatives.

- Metadata structure to accommodate additional categories of materials, e.g. open educational resources, grey literature, etc.
- Allow for metadata/description fields for supplemental/individual files.
- Deposited items discoverable in Google and Google Scholar.
- Allow for hidden metadata fields (viewable to Libra administrators only) to track authors permissions sought and received, and to document reasonable efforts to contact authors of orphaned works.
- ORCID ID/ISNI (International Standard Name Identifiers) input and generation allows for unambiguous linking of research activities/output to a particular person.
- Allow depositors to include a DOI if they already have one for their deposit.
- Allow for DOI assignment should we choose to provide that service.
- Full-text searching.

## Comparison Highlights

Name	Sponsor	URLs	Description
Sufia	Penn State University	Cloud-based demo site: <a href="http://demo.curationexperts.com">demo.curationexperts.com</a>  UVa local test instance: <a href="http://sandbox.lib.virginia.edu:3000/">http://sandbox.lib.virginia.edu:3000/</a>  PSU ScholarSphere production instance: <a href="https://scholarsphere.psu.edu/">https://scholarsphere.psu.edu/</a>	Open Source, built on Hydra-Fedora stack.
<p><b>Pros:</b></p> <ul style="list-style-type: none"> <li>• Hydra-based.</li> <li>• Provides user dashboard for file management.</li> <li>• Allows role based access control.</li> <li>• Includes upload progress indicator.</li> <li>• Allows export to EndNote/Zotero/Mendeley.</li> <li>• Supports Unicode.</li> <li>• Allows batch uploading and editing.</li> <li>• Allows full-text indexing &amp; searching.</li> <li>• Capability to provide copy request button for embargoed items.</li> <li>• Allows records to include related URL(s).</li> <li>• Integration with Avalon multimedia streaming.</li> <li>• Local installation.</li> </ul> <p><b>Cons:</b></p> <ul style="list-style-type: none"> <li>• Global statistics not yet available.</li> <li>• Embargo periods available, but not yet implemented in production.</li> <li>• Requires a large amount of customization.</li> <li>• ORCID integration not yet functional.</li> </ul> <p><b>Costs:</b></p> <ul style="list-style-type: none"> <li>• No licensing cost.</li> <li>• Customization cost will be significant in terms of UVa programmer time, usability requirements gathering, testing, and implementation; administrators of Libra, SIS, and departments involvement in requirements gathering and evaluation of implemented solutions.</li> </ul> <p><b>Evaluation:</b> UVa's substantial involvement with the Hydra ecosystem and user community would be an advantage with this option. Customization costs and time to implementation would be significant. A careful balance of UVa-specific needs with out of the box Sufia functionality would be critical during the design and implementation phases. Balancing University, user, and Library desires for custom functionality against scalability, forward migration, and maintenance will be critical.</p>			

Name	Sponsor	URLs	Description
DSpace	DuraSpace	<p>Partner documentation site:  <a href="https://wiki.duraspace.org/display/DSDOC/All+Documentation">https://wiki.duraspace.org/display/DSDOC/All+Documentation</a></p> <p>University of Minnesota's production instance:  <a href="http://conservancy.umn.edu/">http://conservancy.umn.edu/</a></p>	Open Source (or hosted solution available via DSpaceDirect)
<p><b>Pros:</b></p> <ul style="list-style-type: none"> <li>• Provides user dashboard for file management.</li> <li>• Allows role based access control.</li> <li>• Includes upload progress indicator.</li> <li>• Provides ability to generate monthly reports.</li> <li>• Allows embargo periods.</li> <li>• Accommodates multimedia deposits.</li> <li>• Allows customization.</li> <li>• Allows input of existing ORCID ID.</li> <li>• Allows full-text indexing &amp; searching.</li> <li>• OAI-PMH ready.</li> <li>• Enables retrieval of SHERPA/RoMEO information.</li> <li>• Has API.</li> <li>• Possible to use controlled vocabulary for keywords, but also can be user generated.</li> <li>• Allows local installation or remote hosting.</li> </ul> <p><b>Cons:</b></p> <ul style="list-style-type: none"> <li>• Not Hydra-compatible</li> <li>• Group/batch object deposit not available at user level</li> <li>• Does not allow export to EndNote/Zotero/Mendeley (possible to add?)</li> <li>• Limited metadata capability</li> <li>• No sub-collections, so a collection could be a whole school or department, or it could be a single journal with all its issues as items.</li> <li>• Requires a large amount of customization</li> </ul> <p><b>Cost:</b></p> <ul style="list-style-type: none"> <li>• Customization costs and time to implementation would be significant with a locally built and locally hosted instance, comparable to costs for Hydra/Sufia development or possibly more since local expertise in DSpace is not extant within the Library.</li> <li>• For a hosted <a href="#">DSpaceDirect solution</a>, initial configuration is at least \$5k, with a yearly \$750 fee. Annual subscription costs start at \$12k/year.</li> </ul> <p><b>Evaluation:</b> The hosted, configured service through DSpaceDirect is attractive in</p>			

terms of speed of implementation and freeing up programming resources for other Library projects, but flexibility of the hosted repository is limited in the standard configuration. Peer institutions of UVa have not typically opted for this solution because it is best scaled for smaller institutions with more limited repository footprints. If the Library decides to host and build a DSpace repository rather than opting for DSpaceDirect services, substantial time and development will be needed to stand up an operational service meeting our minimum OA repository requirements. The Library would need to make a substantial commitment to the DSpace community, build Library programming experience with DSpace, and contribute significant programming resources to make a locally hosted service and interface as polished as Minnesota's.

Name	Sponsor	URLs	Description
Digital Commons	Bepress	<p>Vendor information site:  <a href="http://blog.digitalcommons.bepress.com/">http://blog.digitalcommons.bepress.com/</a></p> <p>Clemson TigerPrints production instance:  <a href="http://tigerprints.clemson.edu/">http://tigerprints.clemson.edu/</a></p>	Vendor hosted, annual subscription required.
<p><b>Pros:</b></p> <ul style="list-style-type: none"> <li>• Provides user dashboard for file management.</li> <li>• Allows role based access control.</li> <li>• Includes upload progress indicator.</li> <li>• Allows export to EndNote/Zotero/Mendeley.</li> <li>• Supports Unicode.</li> <li>• Provides excellent customer support.</li> <li>• Provides statistical reporting, including geographic metrics.</li> <li>• Allows embargo periods.</li> <li>• Allows batch uploading and editing.</li> <li>• Allows full-text indexing &amp; searching.</li> <li>• Allows records to include related URL(s).</li> </ul> <p><b>Cons:</b></p> <ul style="list-style-type: none"> <li>• Not Fedora-compatible.</li> <li>• Remotely hosted.</li> <li>• Allows limited customization.</li> <li>• Metadata limited to Dublin Core.</li> <li>• No ORCID capability.</li> <li>• Items do not automatically publish.</li> <li>• “Display tool” rather than “archival tool.”</li> </ul>			

**Cost:** Annual subscription fee.

**Evaluation:** Vendor hosted is the only option for this product. Peer institutions of UVA have not typically opted for this solution because it is best scaled for smaller institutions with more limited repository footprints. Customization is limited to a standard set of options implemented upon request to the vendor, local customization beyond preconfigured choices or product wide updates rolled out by the vendor is not available.

Name	Sponsor	URLs	Description
Eprints	Eprints	Partner documentation site: <a href="http://www.eprints.org/uk/">http://www.eprints.org/uk/</a>  University of Southampton EdShare production instance: <a href="http://www.edshare.soton.ac.uk/">http://www.edshare.soton.ac.uk/</a>	
<p><b>Pros:</b></p> <ul style="list-style-type: none"><li>• Provides user dashboard for file management.</li><li>• Allows role based access control.</li><li>• Includes upload progress indicator.</li><li>• Provides statistical reporting, not geographic metrics.</li><li>• Allows embargo periods.</li><li>• Provides copy request button for embargoed items</li><li>• Allows records to include related URL(s).</li><li>• Has API.</li><li>• Allows local installation or remote hosting.</li></ul> <p><b>Cons:</b></p> <ul style="list-style-type: none"><li>• "Editor review" required before items appear on public website.</li><li>• Allows export to EndNote, but not Zotero or Mendeley.</li><li>• Perl-based, probably not Hydra-compatible.</li><li>• Not scalable.</li></ul> <p><b>Cost:</b> Possible setup costs, significant customization costs.</p> <p><b>Evaluation:</b> This repository is built on a Perl-based platform and is not sufficiently scalable for UVA's needs at this time.</p>			



## Data Repositories Comparison

### Evaluation Criteria

The four broad categories of requirements listed above for OA repositories were also applied to the evaluation for Data repository candidates, with the same detailed criteria for features, functionality, or structure as for OA candidates:

- User interface and user experience
- Statistics and reporting
- Infrastructure
- Metadata

Additional criteria for data repositories were pulled from existing Libra 1.x functionality determined to be needed in a replacement system:

- Allows for deposit of at least two files for a "package". One for a readme, one for a data file. (Libra checks to make sure at least one of each is included in the deposit.)
- Can embargo (set release date)
- Minimum metadata: author, title, license (agreement & requirement), date created, abstract
- Other metadata: keywords, notes, associated URLs (one for dataset, one for article), date collected (single or range), sponsor info

### Comparison Highlights

Name	Sponsor	URLs	Description
Dataverse	Harvard University	Harvard's "anyone can deposit" site: <a href="https://dataverse.harvard.edu">https://dataverse.harvard.edu</a>  JHU has their own locally installed instance (not tied to other Dataverses): <a href="http://archive.data.jhu.edu/dvn/">http://archive.data.jhu.edu/dvn/</a>	Free to install and use, significant user community (Open Source)
<b>Pros:</b> <ul style="list-style-type: none"><li>• Depositor dashboard showing drafts in progress.</li><li>• Allows records to include related URL(s).</li><li>• Provides copy-request button for embargoed items.</li><li>• Allows versions of data deposits.</li><li>• Provides ability to point to data that resides somewhere else.</li><li>• Provides copy request button.</li><li>• Downloads come in multiple formats of tabular data (processed on back end into R, etc.).</li><li>• Allows user to add files to an existing dataset.</li><li>• Allows embargoes, but definition may not match ours.</li><li>• Good metadata, based on DDI (<a href="#">Data Document Initiative</a>). Can also make</li></ul>			

metadata “templates” based on domain or type of file.

- Has API.
- Allows local installation or remote hosting.
- Dataverse user community is active and dynamic.

**Cons:**

- Compatibility with Hydra of concern--records are stored in database tables as opposed to a repository system like Fedora.
- Does not include upload progress indicator.
- Limited statistics available (views, citations, and share statistics coming “soon”).

**Cost:**

- No licensing cost.
- Customization cost will be significant in terms of usability requirements gathering, testing, and implementation; administrators of Libra and researcher/departmental involvement in requirements gathering and evaluation of implemented solutions.
- If integration via API with Libra 2.0 interface is indicated, customization cost will be significant in terms of UVa programmer time.

**Evaluation:** Dataverse is a known and trusted solution for open access data archiving and access needs in social sciences disciplines, and could be branded either as part of the Libra 2.0 suite of services or promoted as a UVa-branded instance of Dataverse, as JHU does with their instance. The primary focus on data, polished user interface, ability to host multiple versions of products of research including completed, citable versions as well as dynamically growing data sets, set it apart from any other data repository option available in production at this time. It is essentially a complete turnkey solution unless the Library opts to integrate Dataverse display of deposited data into Libra 2.0’s interface. A dedicated Dataverse administrator would need to be identified from existing Library talent.

Name	Sponsor	URLs	Description
Sufia (data aspects)	Penn State University	Cloud-based demo site: <a href="http://demo.curationexperts.com">demo.curationexperts.com</a>  UVa local test instance: <a href="http://sandbox.lib.virginia.edu:3000/">http://sandbox.lib.virginia.edu:3000/</a>  PSU ScholarSphere instance: <a href="https://scholarsphere.psu.edu/">https://scholarsphere.psu.edu/</a>	Open Source, built on Hydra-Fedora stack.
<b>Pros:</b> <ul style="list-style-type: none"><li>• Compatible with Hydra.</li></ul>			

- Allow items to be in multiple collections.
- Allows access levels to be set by item inside of a collection.
- Allows versions of data deposits.
- Includes upload progress indicator.
- Allows records to include related URL(s).
- Allows group/batch object deposit and management.
- Requires local installation.

**Cons:**

- Data is currently file-based, not package based, though Portland Community Data Model (PCDM) and Hydra:Works changes are coming to Sufia.
- Same structure as other Sufia objects, very few special data aspects.
- Limited statistics available.
- Does not allow embargo periods.
- Does not allow for ORCID ID input.

**Costs:**

- No licensing cost.
- Customization cost will be significant in terms of UVa programmer time, usability requirements gathering, testing, and implementation; administrators of Libra and departments' involvement in requirements gathering and evaluation of implemented solutions.

**Evaluation:** UVa's substantial involvement with the Hydra ecosystem and user community would be an advantage with this option. Customization costs and time to implementation would be significant with compared to turnkey solutions. A careful balance of UVa-specific needs with out of the box Sufia functionality would be critical during the design and implementation phases. Development in the Hydra community has centered on OA solutions to date.

Name	Sponsor	URLs	Description
DSpace	DuraSpace	DSpace Partner documentation site: <a href="https://wiki.duraspace.org/display/DSDOC/All+Documentation">https://wiki.duraspace.org/display/DSDOC/All+Documentation</a>  Production instance (DRUM): <a href="http://conservancy.umn.edu/handle/11299/166578">http://conservancy.umn.edu/handle/11299/166578</a>	Open Source (DSpaceDirect is not an option for data repository).
<b>Evaluation:</b> See DSpace OA evaluation comments above regarding a locally hosted service and interface similar to Minnesota's. The DSpaceDirect product is not configured for products of research/data needs at this time.			

## Data Repositories Not Considered:

### CDL-DASH

Reason: DASH is a California Digital Library data sharing platform created for the UC schools. Unless institution is part of the UC system , all deposits go to DataONE (see below).

### Eprints

Reason: Eprints is mainly used in the UK and only recently entered into the research data environment. It was too new to evaluate.

### SEAD

Reason: Still in development. SEAD is not a single repository, but a virtual archive (federated over multiple institutional repositories). So far it contains metadata from IUScholarsWorks and IDEALS (two DSpace repositories).

### DataONE (ONEShare)

Reason: This is a discipline repository for primarily environmental data. It is not an institutional repository option, but an open repository that anyone can use.

### PURR

Reason: Based on hub-zero platform with its only instance at Purdue University.