**Report on Strategic Initiative to Provide Enhanced Intellectual Access to NYU-Curated Digital Collections**

## 0. Executive Summary

This report addresses Goal no. 4 of the NYU Division of Libraries' Strategic Plan 2013-2017,[1] namely, "Establish processes and support structures that ensure we can select, acquire, preserve, and provide access to the full spectrum of research materials," and specifically Initiative 4.3 (henceforth, "the Initiative"): "a plan to provide intellectual access to NYU-curated digital collections via the library's primary discovery-and-access interfaces."[2]

Since the Initiative's inception in July 2013, participants have identified and prioritized eligible collections, collected user stories, prototyped the "Ichabod" tool for metadata aggregation and normalization, mapped metadata elements to a local Nyucore schema, and harvested the processed metadata into the development instance of BobCat.

The Ichabod tool is based on Fedora, Hydra, Solr, and Blacklight. It was implemented using Agile methodology and involving developers from DLTS, KADD, and Web Services. The emerging code base, processes, and working relationships place NYU in a strong position to solve local discovery problems as well as innovate in the field of repository metadata management and enrichment.

Recommendations

- Continue aggregating digital collection metadata in Ichabod, and exporting them to BobCat (Dev)
- Implement harvesting mechanism from Ichabod into *production* instance of BobCat
- Ensure regular communications among collection curators, developers, metadata analysts, UX experts, and sponsors
- Build reconciliation and enrichment features directly in Ichabod to improve data quality and user experience
- Consider Ichabod as a discovery tool in its own right
- To enable sharing and reuse, release Ichabod metadata under a CC0 license
- In order to share tools and expertise more easily, consider enlisting NYU as a Hydra Partner

---

[1] http://library.nyu.edu/about/Strategic_Plan.pdf

[2] As stated in the Initiative charge, "The main deliverable will be a report (i.e., the 'plan to provide intellectual access'), but it is assumed that hands-on work and prototyping will be happening at the same time, in part because such activities are already taking place and will likely accelerate, and in part because they provide insights and reality checks on emerging functional requirements for discovery."

## I. Introduction and Problem Statement

The Initiative has benefited from two working groups: (1.) Collections and Functional Requirements; and (2.) Technical Specifications and Prototyping.  The first includes representatives of digital special collections (broadly construed) who inform functional requirements and submit user stories[3] that illustrate successful discovery and access of their materials. Thus far we have collected stories for geospatial data, numeric data, archived Web sites, Drupal sites, and categories of objects held in the Faculty Digital Repository (e.g., NGO reports), deemed high priority by Initiative sponsors but not yet easily findable through BobCat.

The role of the second group is to identify tools and procedures for aggregating, normalizing, enhancing, and disseminating metadata for the targeted collections. This group identified the Solr-Fedora-Hydra-Blacklight technology stack as ideal for prototyping and development, and implemented multiple proofs of concept (and even production-ready features) using this platform. See the "Ichabod Development" section, below, for more detail on current activities and recommended next steps.

The initial formation of both groups was tentative. We had identified several key stakeholders and advocates, and tried to position them for optimal impact in one group or the other, but the exact rosters and frequency of interaction continued to evolve. Over time, we realized that the Prototyping group would best organize itself as a Scrum agile project, now called "Ichabod." A great advantage of the agile approach is that it accelerates the feedback loop among stakeholders, while keeping the needs of users foremost in everyone's mind.[4,5]  This is discussed further in the Project Updates section, below.

The Collections group also evolved over time. A plenary meeting was held in January 2014, but after recognizing the diverse set of interests and needs represented there, Daniel Lovins and Carol Kassel proceeded to meet with collection managers in small sub-groups or one-on-one, which turned out to be more effective and efficient. They have met with representatives of the Spatial Data Repository, numeric data collections, Web archives, and NGO reports, and they have consulted with those managing "The Real Rosie the Riveter" and similar digital collections at NYU. Digital *archival* collections are also important for the Initiative, but we deferred working with them in Ichabod for two main reasons: (1.) EAD is a complex use case and more experience with Hydra was needed before attempting to model it; and (2.) there is another initiative currently in play to rebuild the Archives and Special Collections discovery portal and we wanted to avoid duplicating efforts. It will likely be useful to evaluate the relationship between these two separate but related projects in the future.

*The Metadata Challenge*

---

[3] http://en.wikipedia.org/wiki/User_story

[4] http://blog.crisp.se/2012/10/25/henrikkniberg/agile-product-ownership-in-a-nutshell

[5] At one point (Sept. 2014), given the limited availability and work pattern of project members, we decided to switch from Scrum to Kanban http://en.wikipedia.org/wiki/Kanban_(development). This removed the immediate need for sprint planning, and gave more freedom to pick off user stories as members became available to tackle them. We continued to have regular scrum meetings, though.

In order to provide a centralized discovery portal for digital collections, irrespective of resource type, we wanted to be prepared to reconcile collections and standards as diverse as these:

- Books and journals: MARC, MODS, RDA, LCSH, NAF
- Special collections: EAD, DACS
- Museum and art objects: CCO, VRA, CDWA
- Geospatial data: CSDGM, ISO 19115
- Article indexing: Dublin Core, MODS, various I&A thesauri
- Full-text corpora
- Web sites: HTML, Schema.org, MODS, Locally defined metadata.

Though mitigated by metadata crosswalks and string normalization routines, this mixing of heterogeneous content and descriptive standards threatens the kind of precision and recall that library systems traditionally provide.  Moreover, as publishing moves increasingly to the Web, metadata is increasingly assigned by non-catalogers, which means that names and subjects may have been applied inconsistently or as free-text.  This can render indexes, facets, and filters unreliable or unusable for discovery. Therefore, we are exploring the use of  natural language processing (NLP) to perform entity recognition and extraction, and remediation tools like Open Refine to reconcile named entities with linked-data vocabularies.  This approach is described further in the "Recommendations" section, below.


## II. Project Updates

The following is a list of collections for which we are clarifying priorities, collecting user stories, and finding ways to make discoverable.

### A. Collections

1. Available in Ichabod and BobCat Dev

- Archive of Contemporary Composers' Websites
- Spatial Data Repository
- ESRI
- NYU Press Open Access Books
- Voices of the Food Revolution
- The Real Rosie the Riveter
- Asian NGOs Reports

2. Identified but Not yet available in Ichabod or BobCat Dev

- Additional ArchiveIt web archives (i.e., beyond Contemporary Composers, above)[6]
- CDL web archives[7]
- The Masses
- The Liberator

---

[6] E.g., University archives, African American themed websites, Hemispheric Institute (Already discoverable in BobCat), RG39: Student organizations and publications (http://archive-it.org/collections/3771)

[7] There is a large set of websites Tamiment is archiving with the CDL web archiving service tool: http://webarchives.cdlib.org/institutions/NYUL

- The Fight against War and Fascism
- Second Avenue Yiddish Theater Digital Archive
- Witness to the Early American Experience
- Numeric data sets
- Indian Ocean postcards[8]

## B. Ichabod development

The process we've been using to develop Ichabod is Scrum, a form of Agile development that focuses on constant communication and reprioritization of tasks. A basic idea of this philosophy is that "you don't know what you want until you see what you don't want." The focus is thus on strategies such as iteration: get a feature working, review, refine, repeat. It involves a continuous flow of creation and refinement, with the goal of always having something that can be deployed to production.

In Scrum, the goal of the product owner (that is, the person or persons with the most knowledge of end users' desires) is to create user stories that can translate into application features. The user stories (i.e., the plain English description of what the users need the program to do as part of their job function) help the developers understand how patrons may interact with features of the application. User stories are split into features, which lead to testing and eventually to working code. As a result, every piece of functionality occupies a persistent niche in the minds of product owners and end users. If a piece of functionality cannot be described in plain English, it is probably not yet sufficiently understood by product owners or actionable by developers.

The Scrum approach is designed to be flexible and iterative. It is based on the recognition that "trying to formulate an ideal product will probably result in no product at all." Above all, it recognizes the necessity of communicating at every phase of the project so that, from the highest to the lowest levels, all parties know what is going on and why.

At the software level, Ichabod is an implementation of Hydra, which is a stack of interweaving open-source technologies that provide a framework for managing a repository of digital content. The technologies it includes are Fedora, Solr, Blacklight and the Hydra piece itself. Fedora is the repository application Hydra was developed around. We had not previously used Fedora at NYU; however, because of its integration within this stack, we have adopted it as the nexus for the disparate data sources and the tool for augmenting metadata on these ingested objects in the repository. It is worth noting that we are not using Fedora as an object repository. This implementation is currently for metadata management and augmentation only, with repository functionality delegated to other systems. Solr is a fast, Lucene-based indexing and searching tool. Blacklight is a front-end Solr discovery tool conveniently developed by the same group of library technologists as Hydra. The Hydra "gem" (i.e., module of Ruby code)[9] allows for the management of Fedora objects through a Web interface and also integrates the access control layer inherent in Fedora.  These pieces fit together into what is called a "Hydra head," a single application sitting on top of this web of technologies. This "Hydra head" is neatly packaged as a Ruby on Rails engine, which means it fits into the stack that Web Services already supports. With

---

[8] The Indian Ocean Collection includes 50+ books; 1,000+ postcards (front and back); and 230 maps. Digitization is ongoing, beginning with books, and continuing with maps, all of which will be represented in Aleph. Postcards will not be represented in Aleph and will potentially benefit from Ichabod functionality.

[9] http://guides.rubygems.org/what-is-a-gem/

this integration with Web Services' existing applications, we get shared assets, single-sign-on and automated deployment for no additional effort.

Another major aspect of our technology stack is a robust testing framework. Because we start thinking about our functionality at the highest level, we write feature requests in Cucumber, a technology that allows for integration testing by being written in plain English (but is syntactically structured for computer readability). The beauty of this approach is that product owners with no technical background can ask how a feature works and answer their own questions by reading the Cucumber test.

The final pieces to this puzzle are continuous integration and deployment. We use GitHub for software version management. We can share ideas, review code, report issues, and cross-check code compatibility before deploying and can easily roll back or branch off into different areas of development. Jenkins is our deployment application, which allows us to maintain a hands-off approach to the minutiae of pushing code into various environments, updating databases, and other tasks best left to machines. When our application is fully tested, we can have the confidence to commit our code to GitHub, which will trigger a job on Jenkins, which in turn automatically runs our test suite and deploys code updates only if all tests have passed.

### C. Metadata workflow development

The Initiative involves multiple collections that have been described with different metadata standards and sometimes with no standards at all. Building effective indexes and user interfaces for such an aggregation is a major challenge. There are tools available to help, though, including those for metadata ingest and mapping, batch and individual metadata editing, schema transformation, Natural Language Processing (NLP), and linked open data.

As a step toward addressing this need, the Initiative has been implementing a local Metadata Application Profile, Nyucore, as a "Hub and Spoke" metadata format to aid in crosswalking and normalization. The idea of "Hub and Spoke" is to have a central format that is used for data storage and management, with all ingest from and dissemination to other formats handled by mappings to and from this central format. Generally, "Hub and Spoke" formats are lossy, though Nyucore is designed to be extensible and to evolve to meet the needs of both upstream (e.g., the Faculty Digital Archive) and downstream (e.g., BobCat) systems.

The idea of an Nyucore Metadata Vocabulary goes back to January 2011 and work done for the Union Catalog for Digital Projects. The strategic goal of getting heterogeneous metadata from various digital library projects into the Libraries' discovery layer is not a new goal, and has been a significant challenge for years due to the MARC-centric nature and non-intuitive normalization management environment presented by Primo. An intervening metadata management system such as Ichabod, along with a pre-normalization metadata format such as Nyucore, provides a staging area to collect, normalize, enrich, and augment metadata *before* it is consumed by Primo and incorporated into the Libraries' primary discovery system.

An additional advantage of an intermediary metadata management system is that metadata experts and collection curators have a central place to edit metadata and enrich descriptions. Hydra is designed to provide Web forms for managing metadata, and Ichabod takes care to manage metadata provenance so that local edits are separated from metadata derived programmatically from source systems. This ensures that cataloger- and/or curator-supplied enrichments are retained even when source data is re-ingested.

This same approach can be used for machine-assisted enrichments, such as metadata that is algorithmically derived from source data. Just as Ichabod can keep manual edits separate from the source data, it can also maintain the distinct provenance of metadata enrichments from automated processes. There are two main categories of machine-assisted enrichments, both of which fit loosely under the heading of Natural Language Processing (NLP):

- **Named Entity Recognition** involves scanning chunks of free text, such as abstracts, titles, folder lists, or even full-text documents for any strings resembling proper nouns, including People, Places, Topics, Events, and Corporate Bodies.

- **Metadata Reconciliation** is the process of adding structure to metadata values such as unstructured strings, uncontrolled author headings, user-supplied subject tags, or collection-specific vocabulary entries. When the values are known to match particular types of entities, reconciliation tools can try to match them against controlled vocabularies like LC's Name Authority File, LC Subject Headings, Getty's Art & Architecture Thesaurus, or the GeoNames list of Place Names.

These processes can also work in concert. Named Entities extracted from article abstracts, for example, can be reconciled against controlled vocabularies providing the basis for machine-assisted subject analysis.

The initial implementation of Nyucore in Ichabod is flat, consisting entirely of literal strings. However, the application profile has been built on the principles of linked open data, allowing the vocabulary to be easily extended to provide classes of "authority control" objects, which the Hydra and Fedora system design supports. Just as Aleph supports using authority records to control the form of personal names and subjects that appear in bibliographic records, Ichabod can build out more complex models of digital library resources in which creators are represented as links to objects of class "Agent" and subjects are represented as objects of class "Concept." Even Collections and Linguistic Systems can be modeled as complex data types. Managing some metadata via authority objects is in line with traditional cataloging systems, fits with emerging best practices for managing library metadata as linked data, and brings Nyucore closer to other emerging specifications, such as BIBFRAME, the Europeana Data Model, and the DPLA's Metadata Application Profile.

This complex modeling provides a number of distinct advantages. Most notably, metadata managed via authority control objects are less likely to contain errors that are introduced by data entry. Additionally, alternate forms of name can be included in indexes. These alternate labels can also be exposed to Primo, so that digital materials use the same indexing and faceting as MARC-based resources coming from Aleph and elsewhere.  As an added bonus, managing preferred and non-preferred labels for subjects and authors using standard controlled vocabularies will ensure that Ichabod-based materials appear consistently alongside Aleph-based materials in Primo v. 4's new "Browse" functionality.

In the case of "collection" complex objects, managing these metadata as separate objects in Ichabod will allow for the development of "collection editing" pages, where collection curators can manage aggregation-level metadata such as collection title, collection abstract, and rights information. This data will propagate down to descriptions of individual resources in the collection, ensuring that information is represented consistently across resources. If NYU users find the native Hydra User Interface useful as a complement to the primary, Primo-based discovery interface, we may consider developing user-facing

collection pages. These pages could either be built directly in the Hydra / Blacklight interfaces, or using tools such as SpotLight[10] and the soon-to-be-released ArcLight.[11]

Finally, modeling these metadata elements as complex objects lends itself to more automated methods of Natural Language Processing. In early phases of Ichabod development, NLP has been performed using Open Refine[12], formerly Google Refine. This methodology is useful in developing a proof of concept and provides the benefit of having a cataloger operating the NLP workflow and providing quality assurance, but it does not scale well to large collections with a lot of free text. Managing subjects, authors, and genres internally as complex objects will allow Ichabod's ingest routines to employ probabilistic and "fuzzy" string matching techniques to suggest relevant metadata values. Tools for this kind of matching, such as Boston Public Library's Bplgeo[13], are already appearing in the Hydra community. Data sources for NLP, such as DBPedia and LCSH, are shown at the bottom of Diagram I in Appendix A of this report.

## III. Recommendations

**A**. Our short-term recommendation is to continue aggregating curated digital collections into Primo, i.e., the BobCat "Books & More" tab.  This approach leverages Ichabod's ingest, aggregation, and editing functionality which can then serve as a staging and normalization area before adding new digital collection metadata to BobCat.

Most of the collections short-listed for the Initiative are uncontroversial. There is an ongoing debate, though, on the extent to which the Faculty Digital Archive should be considered curated, and therefore appropriate for harvesting into Ichabod and BobCat.  The main concerns are that the resources have not always been consciously "selected" by the Library, and so they would not be considered curated. Certain FDA collections have been formally selected, e.g. NGO reports, and numeric data sets.  We propose that these, at least, be made discoverable through BobCat.

There is also concern, however, about the quality of metadata assigned to FDA objects. In order for institutional repositories like the FDA to get adopted by a community like NYU, the barrier for metadata creation must be low. This is to say, a key reason for maintaining such repositories is to empower researchers to deposit and describe their own work, and thereby leverage their disciplinary expertise and self-supplied metadata for the common good. At the same time, institutional repositories help manage the flood of digital publications that professional catalogers cannot be expected to describe.   We see an important role for catalogers, though, in providing metadata analysis, building metadata crosswalks, developing methods for entity recognition and extraction, selecting appropriate core ontologies, and reconciling access points to controlled vocabularies.

**B**. Our longer-term recommendation is to build advanced enrichment and reconciliation features within Ichabod. There are significant opportunity costs to our current normalization workflows in Primo. The forced Web interface to the Primo Back Office along with a restrictive Oracle license makes it impossible to edit normalization rules programmatically. The labor-intensive web forms and opaque database structure slow down application maintenance and development. A compelling goal, therefore, is to normalize

---

[10] https://github.com/sul-dlss/spotlight

[11] https://groups.google.com/forum/#!topic/blacklight-development/v5R42LYxsv8

[12] http://openrefine.org/

[13] https://github.com/projecthydra-labs/Bplgeo

data outside of Primo based on Nyucore, followed by bulk ingest and 1:1 mapping to Primo's internal PNX format, after which we can still rely on Primo, if desired, for services like deduping, FRBRization, and real-time availability.

**C**. We also recommend looking at Ichabod as a discovery tool in its own right, i.e., rather than as just an aggregation, staging, and normalization tool for BobCat. Given that Hydra and Blacklight are bundled into Ichabod, the core components are in place for a self-contained and flexible discovery system. Indeed, Hydra excels at combining a single metadata repository with multiple user interfaces, each optimized for a particular collection or content type, such as GeoHydra for Geospatial Metadata and Avalon Media Systems for A/V materials. [14,15]

**D**. We see value in the sharing and reuse of metadata to improve resource discovery and workflow efficiency. Therefore we recommend that Ichabod metadata, to whatever extent possible, be released under a CC0 license.[16] This could help with projects such as Arabic Collections Online (ACO), where the digital content is open access and we want anyone to be able to reuse our records.

**E**. We recommend that NYU become a Hydra Partner.[17] There is no cost involved, but establishing an explicit relationship with the Hydra community would allow us to collaborate more effectively on common challenges and opportunities.


## IV. Conclusions

Regardless of how Hydra gets deployed at NYU, we believe there are larger lessons to be learned from the Initiative, including the importance of test-driven and behavior-driven development, user-centered design, robust automated testing, iterative prototyping, and software that fits the environment our Web developers already support. The Ichabod team is gaining valuable experience in these areas while also pursuing the first large-scale collaboration between DLTS, KARMS, and Web Services. The same cross-section of departments is now using the same approach for development on the Finding Aid Portal Bridge project. The Ichabod project contributes to the Library's Strategic Plan beyond Initiative 4.3. It advances "User Experience" (Goal 1), by inviting collection curators to submit user stories and review prototypes; "Digital Infrastructure" (Goal 3) by fostering collaborative software development and metadata best practices; and "Organizational Culture" (Goal 5) by emphasizing entrepreneurship, collaboration, and open communications. We recommend building on recent progress by continuing to invest in

---

[14] https://github.com/sul-dlss/geohydra

[15] http://www.avalonmediasystem.org/project

[16] Cf. announcement from University of Illinois to release locally created metadata under CC0 an OCLC-derived metadata under ODC-BY 1. (Cf. http://catalogdata.library.illinois.edu/)

[17] "Partners sign a formal one page agreement agreeing to support the formal, legal Memorandum of Understanding signed by the Steering Group in April 2012, and each Partner is asked at the time of joining the group to write a brief letter of intent indicating why they want to become a Hydra Partner and what they intend to contribute to the project. (Note that we post these letters publicly in the wiki, to share among the community.) The one-page agreement can be found here (with the full MOU attached). Partners are also strongly encouraged to file a corporate Contributor License Agreement to ensure that the project can accept code and other intellectual contributions from the partner institution." See https://wiki.duraspace.org/display/hydra/Hydra+Community+Framework

appropriate staff positions, training, and technologies and continuing to foster collaboration among collection managers, usability experts, developers, and metadata analysts.

Fortunately, this Initiative is backed by the Library's Strategic Plan, which implies a high degree of institutional buy-in, transcending the immediate interests of any particular person or department. We must stay vigilant, though, given the Library's many competing priorities and resource constraints. It is especially important that key stakeholders continue to be represented on the project teams. We see two specific risks in this area:

First, the capacity of Web Services is currently reduced as they try to fill an open developer position; at the same time, the number of projects in their portfolio has increased. These factors mean they have less time available for Ichabod. Web Services has been a critical partner on the Initiative not only because of their excellent work on strategy and implementation, but also because of an early decision to adopt their preferred platforms and workflows such as Rails, Jenkins, and specific server configurations. Without their continued active involvement in scrums, sprints, and planning sessions, there is a risk that those investments may feel (or become) more of a legacy and an obstacle rather than the forward-looking decision that we believe it to have been.

Second, the communication between the prototyping group and collections curators has so far been informal and irregular, and we need to take care to ensure that all key stakeholders remain included in the decision-making process. The main locus of communication has been the "user story," which product owners, collection managers, and developers submit and refine within the PivotalTracker tool.[18] This is an effective approach, but we will need input and review from a broader range of stakeholders in order to ensure that the Initiative remains viable. Similarly, we need to deepen our consultation with the Library's UX department. The agile approach of Ichabod is predicated on user-centered design, and we risk diverging from user needs if collections managers or UX experts are not firmly embedded in the feedback loop.

By coordinating the efforts of collection curators, software developers, and metadata analysts, under the protective umbrella of the Strategic Plan, we have identified appropriate methodologies for continuous improvements to our discovery environment. Moreover, we have built a scalable platform for aggregation, enrichment, and discovery of underserved digital collections.

---

[18] https://www.pivotaltracker.com/n/projects/1025368

## IV. Appendix

**Significant Dates**
- July 19, 2013: Kick-off meeting
- Nov 4, 2013: First Commit to Github
- Dec. 12, 2013: Initial Technical Specs and Prototyping Meeting
- Jan. 22, 2014: Initial Collections and Functional Requirements Meeting
- Feb. 10, 2014: Initial Development of SDR user stories
- Mar. 5, 2014: Initial Review of indexing and prototyping with sponsors
- Apr. 17, 2014: Scrum kick-off meetings
- June, 2014: Ichabod application deployed to development
- Sept. 2014: Ichabod application deployed to production
- Nov. 13, 2014: Initial harvest of Ichabod records into BobCat Dev
- Dec. 23, 2014: Proposed date for submitting report and recommendations

**Participants**
Initiative Lead:

- Daniel Lovins *

Collections group

- Nadaleen Tempelman-Kluit
- Charlotte Priddle
- Chela Weber
- Tim Johnson
- Michael Hughes

Prototyping/Ichabod group

- Barnaby Alter *
- Hannan Butt
- Scot Dalton
- Esha Datta
- Carolina Garcia
- Corey Harper *
- Carol Kassel *
- Alberto Ortiz Flores
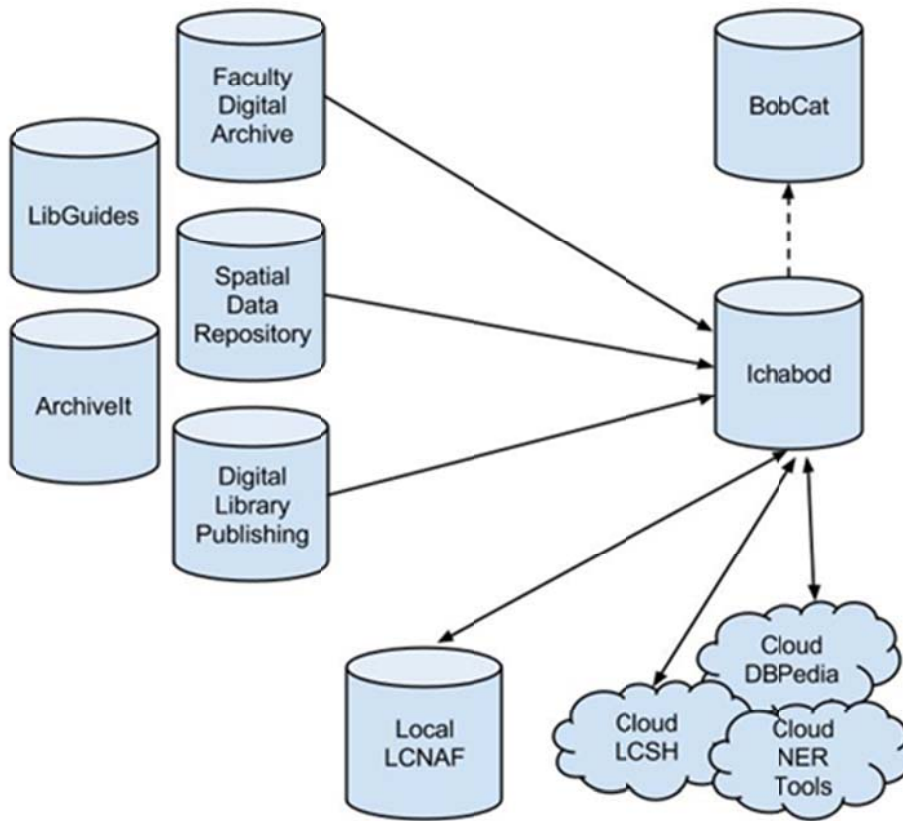- Joseph Pawletko
- Ekaterina Pechekhonova

Sponsors

- Roddy Austin
- David Millman
- Marty Kurth
- Michael Stoller

* = Contributed one or more sections to the Report

**Diagrams**
Diagram 1. Some components of aggregation, remediation and discovery

## Resources Cited

- Hydra Community Framework: https://wiki.duraspace.org/display/hydra/Hydra+Community+Framework
- Strat43 Charge (draft) https://docs.google.com/a/nyu.edu/document/d/12UAbiGhLPuk3eUGD8NGQZxHj2V0-pZs799Pd3YE2hxI/  (distributed 7/19/13; last viewed 10/15/14)
- Strat43 Assumptions and Directions (draft) https://docs.google.com/a/nyu.edu/document/d/1ndp4GVveGKXqzzOB70Rj9dTvxEJvuQ6BvagPcwa4vfw/edit#heading=h.4bzz4v9bhh57 (distributed 1/14/14; last viewed 10/10/14)
- NYU Libraries Strategic Plan 2013-2017 http://library.nyu.edu/about/Strategic_Plan.pdf (viewed 10/11/14)
- PivotalTracker  https://www.pivotaltracker.com/n/projects/1025368
  - Union Catalog of Digital Projects (restricted access): https://itwikis.nyu.edu/pages/viewpage.action?pageId=50725221