Designing Data Services for the Institutional Repository

Ron Jantz and Ryan Womack
Rutgers University Libraries
June 2, 2010
IASSIST 2010, Ithaca
About Us

- Ron Jantz, Digital Library Architect
- Ryan Womack, Data and Economics Librarian
- RUCore – Rutgers Community Repository
  - Institutional Repository for Rutgers, operational since 2006. Has expanded from initial faculty documents deposit module to include modules for Video, New Jersey materials, and Electronic Theses and Dissertations.
  - The goal of RUCore is to advance research and learning at Rutgers, to foster interdisciplinary collaboration, and to contribute to the development of new knowledge through the archiving, preservation, and presentation of digital resources.
  - **RUCore Open Source development**
    - OpenETD - open source module for management of dissertations.
    - OpenMIC – open source moving image collections cataloging.
    - OpenWMS – open source Workflow Management System - just about to be released!
- RUCore Data Working Group
  - Composed of subject librarians, technical services librarians and technologists. Members (in addition to the co-authors) are Isaiah Beard, Melissa Gasparotto, Qian Hu, Mary Beth Weber.
Project Background

• Support for data in RUCore identified as a critical need in Scholarly Communication, Cyberinfrastructure Working Group, Library Councils.

• Data Working Group formed in Fall 2009.

• Fall 2009 – discussion and decisions on basic capabilities, design, identification of pilot data projects.

• Spring 2010 – development of the prototype displayed here.

• Fall 2010 – survey of faculty on data needs, soft rollout of the prototype to selected clients, grant-seeking for further feature development.
Design Issues

• Who is the primary audience?
  – Our IR supports all Rutgers faculty and staff, so we must be ready to accept data according to their needs, not those that we define.
  – We want to support faculty in the research process as well as archive final results.
  – Open access to data is desirable, but will not be imposed.
  – The ability to protect restricted data is a requirement.

• What services are supported?
  – Our focus is on basic data access and download. Data is defined broadly, not only traditional numeric data. Can include text, audio, video, if the data is structured for the purpose of analysis.
  – The desirability of additional services (online query results, subsetting) was recognized. Since this will not work with all datasets, it is not part of the core services to be developed at the initial stage of the project.
Policies

• Some of our discussions have been devoted to development of appropriate policies for RUCore data.
• We are developing deposit and copyright agreements, modeled on those of ICPSR and others.
• What do we offer to the faculty depositor as part of the bargain?
• Policies will include the option to embargo data and limit access to authenticated users via an xacml policy.
System Architecture

- RUCore is Fedora based (v. 3.0).
- Until now, each object in the repository has been treated as a single item in a top level collection.
- Data is more complex – need to associate multiple data files, multiple formats, and related documentation and software within a single object.
- RUCore Data adds relationship mapping, additional object types, and a revised interface design to handle more complex relationships.
System Architecture

Workflow Management System

User Input (metadata)

Object Data Files

Object Ingest

Encapsulated Object

Object XML

Archive Objects

Research Objects

Fedora Repository (Main Server)

Collection & Preservation Services

ETDs

Videos

Data Portal

RUcore Portal
The Compound Data Object

Collection Object

hasMember

isMemberOf

View Object

hasMember

isMemberOf

hasMember

isMemberOf

Other Resources
On the Web

isRelatedTo

Aggregate Level

Data

hasPart

isPartOf

Numeric Data

Documents

Software

Supp. Materials

Simple Objects
with content models

Survey

Interview

Supp. Materials

hasPart

isPartOf

Numeric Data

Survey

Interview
Research Area

- Researchers will be given access to a “sandbox” where they can deposit files.
- A “light” version of metadata will be captured at this stage.
- Shibboleth and xacml allow inter-institutional collaborations on data.
- Once files are in the research area, we can begin to understand the structure of the data object and ask for additional documentation needed to make the data publishable.
Prototype

- The following slides are screenshots illustrating the prototype web site with tabbed search design.
RUcore Data is the data archive of RUcore, the Rutgers Community Repository. RUcore Data provides a means for researchers to share and collaborate on data related to their current research, and to make their final research data discoverable to the global scholarly community by publishing and permanently archiving the final data. RUcore Data holds numeric datasets and other structured data collections, along with associated documentation, software, and supplementary materials.

If you are a Rutgers researcher with data you would like to share with collaborators or make publicly available, please contact us.

RUcore Data Policies and Documents
Search Results

Displaying 3 of 5 results

ACRL Academic Library Statistics


Name Molyneux, Robert E., Association of College and Research Libraries

Filesize 100 (KB)

Description This is a .zip archive containing the complete datasets associated with the ACRL Academic Library Statistics, 1978/79-1987/88. Data from 1978-1984, 1986, and 1988 are available in multiple formats.

Persistent URL http://ndl.rutgers.edu/1782.2 /lefty64.rucore00000000090.Dataset.000002139
Select an identity provider

The Service you are trying to reach requires that you authenticate with your home institution, please select it from the list below.

Choose from a list:

- Choose your institution — Select — Remember for session

or

Search by keyword:

Search

Need assistance? Send mail to shib@njedgecn.net with description.

NJEDge.Net
New Jersey's Higher Education Network
Designing Data Services for the IR

RUcore
Rutgers Community Repository

Search Results

Displaying 1 - 5 of 5 results

ACRL Academic Library Statistics


Name Molyneux, Robert E. (Author); Association of College and Research Libraries

Description The 1986 data is available in Excel, ascii, csv, and ssd formats.

ACRL Academic Library Statistics


Name Molyneux, Robert E.; Association of College and Research Libraries

Description The 1986 data is available in Excel, ascii, csv, and ssd formats.
Designing Data Services for the IR

Name: Meynens, Robert E. (Author), Association of College and Research Libraries

Filesize: 10 (KB)

Description: This is the SAS program for processing the SAS data file containing the 1988 data from the ACRL Academic Library Statistics.

Genre: Software

Persistent URL: http://cdl.rutgers.edu/1782.2
lefty44.rucore000000001111.Daset.000002191


Organization Name: Rutgers University
Issues - Metadata

- The metadata scheme will identify the genre of research (survey, experiment) and the style of data (time series, microdata, instrumental), in addition to identifiers for subject, geography, time period, and file formats. Advanced search will be built around these identifiers.
- Metadata will need to flex to adjust to different disciplines, but retain simplicity.
- We must structure the data in a way that serves the researchers' needs, not imposes our own idea of structure. If they say it is data, it is data [3-D models].
Issues - Standards

• DDI - We made the decision not to implement DDI 3. Due to the heterogeneous nature of the data we are likely to receive, including lots of science data, and the lack of staff for intensive processing of the data, DDI 3 will not provide great benefits and would be labor-intensive.

• We need to maintain compatibility with the rest of the repository metadata (METS and MODS). We will be able to provide a DDI2 Lite compatible version of the core metadata.
Issues – Formats

• In all cases, we will preserve the original submitted files
• For numeric data, we will generate a .csv version
• We will support common formats from software such as R, Excel, SAS, SPSS, Stata, and commit to migrating files forward from version to version
• Beyond this list, we will make decisions based on the number of files submitted in the format.
Issues – Intervention

• The research area allows for inspection of submitted files and coaching towards the goal of a complete, well-documented, and well-structured dataset.

• We envision subject librarians and RUCore developers will work on reviewing data during its move from research area to archive.

• Working within datafiles to reformat and produce variable level information will be limited. Researchers should provide variable descriptions when information cannot be automatically extracted from files.

• Post-archiving interventions will be limited (retractions, issues with software versioning)
Questions? Comments?

Contact Ryan Womack (rwomack@rci.rutgers.edu) or Ron Jantz (rjantz@rci.rutgers.edu)