Institutional and research repositories: Characteristics, relationships and roles | Feb. 26, 2015

Judith Coffey Russell
Dean of University Libraries
University of Florida

Michael Witt (@mwittin)
Head of the Distributed Data Curation Center (D2C2)
and Associate Professor of Library Science
Purdue University

James Toon (@jamestoorn)
Research Information System Manager
University of Edinburgh

Alicia Wise (@wisealic)
Director of Universal Access
Elsevier
Public Access

A shared view on a shared challenge!

Judy Russell & Alicia Wise
February 2015
# Challenges

<table>
<thead>
<tr>
<th>Funder</th>
<th>Institution</th>
<th>Publisher</th>
</tr>
</thead>
<tbody>
<tr>
<td>Researcher success</td>
<td>Researcher success</td>
<td>Researcher success</td>
</tr>
<tr>
<td>Research funding</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Compliance</td>
<td>Compliance</td>
<td>Sustainable business models</td>
</tr>
<tr>
<td>Compliance tracking and reporting</td>
<td>Compliance tracking and reporting</td>
<td>Usage tracking and reporting</td>
</tr>
</tbody>
</table>

Our interests are complementary; more work needed to investigate roles and responsibilities and ways to avoid duplication of effort.
Why Chorus?

- **Portfolio Analysis of University of Florida Articles by Publisher 2013**

  - **x-axis**: Relative Field Weighted Citation Impact
  - **y-axis**: Citation Share
  - **Bubble Size**: Number of University of Florida published articles 2013

  - **Elsevier**: Orange
  - **Springer**: Red
  - **Wiley**: Blue
  - **T&F**: Light blue
  - **Sage**: Green
  - **Wolters Kluwer**: Gray
  - **AIP**: Light purple
  - **OUP**: Purple
  - **IEEE**: Red/purple
  - **ACS**: Green/purple
  - **IOPP**: Black
  - **APS**: Black/red
  - **PLOS**: Black/green

  - **x-axis**: so the field weighted impact factor of each publisher divided by the average across all publishers, so a RFWCI of 1.0 means you have an average quality.
  - **y-axis**: you have the share of all citations that the University of Florida authors receive.

- **Why Chorus?**
Why University of Florida?

Scholarly Output

Outputs in Top Percentiles

Field-Weighted Citation Impact

Source: SciVal
Project = collaboration through linking
maximize visibility, impact, and dissemination

Why link?
• **Maximize Research Impact** – citation metrics and alt metrics are used to assess the impact of research so it is essential that these data are aggregated and shared rather than fragmented across platforms

• **Connect Researchers to the Best Available Version** – dynamic ingest of publisher holdings information and metadata into the university IR allows local management with delivery of the best (published) version, and also addresses the concern of some authors about reliance on a pre-publication manuscript for public access

• **Present the Article in Context** – enable researchers worldwide at institutions with licensed access to Elsevier journals to read the article in the context of the other articles in that issue of the journal, and with access to supplementary and linked datasets and provide alternative means of access to the article for others.

• **Realize Cost Savings and Efficiency** – this collaboration eliminates the burden of self-archiving on researchers; facilitates better compliance, impact tracking and reporting; increases the visibility of and access to the scholarship of UF authors; and establishes a new, and valued, library service to the campus and beyond, with minimal initial investment and low ongoing costs

Access
Denials
## Collaborative Open Access Project

<table>
<thead>
<tr>
<th>Phase</th>
<th>Overview</th>
<th>Status</th>
</tr>
</thead>
</table>
| 1 Bilateral Project (underway) | Populate U of Florida Sobek repository with metadata for articles published by its researchers with Elsevier + Test embed/linking services | • Project team formed  
• First meeting held  
• Technical discussions underway about Elsevier API & metadata |
| 2 Multilateral Project (under consideration) | Engage other stakeholders (e.g., CHORUS and its participating publishers, one or more agencies) in a pilot project to  
 a) investigate roles to:  
• Provide access to content  
• Track and report on compliance  
• Provide services to the others  
 b) recommend ways to minimize duplication of effort and burdens on researchers | • Initial discussion by CHORUS board is positive |
SobekCM Open Source Repository Software

- Open Source Software
- Employs industry standard practices, tools, and components
- Deployed at massive-scale
- Embargo support for restricting access by IP range and for authenticated users
- Fully integrated digital curation and digital production workflow tools; online submission, editing, tracking, reporting, and processing tools
- U of Florida uses SobekCM to host collections and items: rare books, manuscripts, antique maps, children's literature, newspapers, theses and dissertations, data sets, scientific reports, and more
Phase 1 – Bilateral Project between Florida and Elsevier
Metadata feed: How does it work?

Elsevier has developed a policy to help support repositories enable users to search scholarly information written in that research institution. We do this by supplying a metadata feed through an API which enables these websites to show this information to all users.

**Process**
- The institution can get API access through our developers portal
- After a simple online click-through agreement they will receive an API key
- Metadata is available from Scopus or ScienceDirect
- Metadata can be (easily) integrated into the repository

**Details**
- You can store the following metadata:
  - Journal/book title, Article title, Authors, publication date, journal / issue, page numbers, author keywords, abstracts.
- Search results/records need to link back to the article or chapter on ScienceDirect.

www.developers.elsevier.com
Embedded functionality: how does it work?

Functionality which enables the final full text version of the published article to be shown within the repository

Technical details

- **Full text is embedded on repository page** using Embed tag / I-frame / object tag
- **The repository specifies the size** at which you want the document to be displayed
- Provides **the possibility to display other content** next to the article, for example the navigation pane for your repository
- **Opening PDF in new browser** window

What do users access?

- **All users** will be able to access full text of open access articles
- **Users with recognized IP addresses** will see full text of subscription article
- **Users who are not entitled based on IP address** (guests and remote) will see one-page preview of subscription articles
- **Remote users can login to our API** for access to the full text article based on their entitlements
- **Guest user** options too

Helps promote research published by your university

Encourage citations as researchers want to cite the final published article

Enhance the user’s experience
Piloting Repository services

Expanding access to publications by University of Florida authors through the university’s institutional repository using publisher supplied data and links

**Structured Metadata:**

**Helps the repository:**
- Get indexed in search engine results
- Track research output

**What is being piloted?**
- Integration of the Elsevier API enabling:
  - Retrieval of metadata for Elsevier published articles written by its authors from all titles whether open access, subscribed, non-subscribed.
  - Provides clean and structured metadata which includes the abstract
- Gathering requirements and how to include information relevant to green OA (e.g. embargo periods and DOI links).

**Embed services:**

**Help the repository:**
- Display the best available version to users
- Provide access from their repository

**What is being piloted?**
- Simple tool integrated into the repository which enables the platform to present final full text article views to entitled users served via ScienceDirect.
Phase 2 – Multilateral Project (hopefully!) to include CHORUS and its member publishers
How Does CHORUS Help?

- Identification
- Discovery
- Compliance
- Preservation
- Access

Publisher
Agency
Librarian
Public
VP Research
Researcher
Project Goal - draft

- Clear, simple, aligned guidance for researchers about how they can publish and easily comply with mandates in ways that are hassle-free, easy for their institution to report, and sustainable for their journal.
- Explore roles and responsibilities
- Facilitate interoperability between systems without duplication of cost/effort
- Develop reports that are helpful to all stakeholders
CHORUS + SHARE

• CHORUS aspires to be inclusive and interoperate with scholarly repositories and other systems providing access to scholarly articles
• Agreed to work jointly on persistent identifiers, metrics, and notification system
• Continuing to explore more areas of collaboration
• Member of SHARE workflow working group

Potential to deepen/strengthen this collaboration
How do Institutional Repositories fit in?

- Store information
- Provide access
- Track research output
- Facilitate compliance reporting
- Manage access and entitlements
- Support copyright compliance

That is a lot of different tasks!
How do Publishers fit in – Elsevier’s view

• Support scholarly sharing and want to make it easier for researchers

• Believe that hosting platforms and publishers should work together to make this as seamless as possible

• Find ways to work with repositories to make copyright and compliance easier and less expensive

We are ready to help!

• Provide public access
• Metadata feeds
• Embed functionality
• Other piloted services
• Partner w/ libraries on preservation services
Michael Witt
Head of the Distributed Data Curation Center (D2C2) and Associate Professor of Library Science
Purdue University
http://www.lib.purdue.edu/research/witt

An Institutional Data Repository Service at Purdue
Data = Evidence

http://epicgraphic.com/data-cake
Some Examples of Data

- Tables and spreadsheets
- Sensor and instrument output
- Software source code and simulations
- Images, video, and audio
- Observation logs
- Workflows
- Interview transcripts
- Survey instruments and results
- Etc.
Campus Collaboration

Purdue University Research Repository (PURR)

The PURR service is a collaborative effort of the Purdue University Libraries, Executive Vice President for Research and Partnerships, and Information Technology at Purdue. PURR is a designated university core research facility.

Designated community:
Purdue University faculty, staff, and graduate student researchers; their collaborators; and the current and future consumers of their data.
Library Strategic Plan

Data is written into the three pillars of our strategic plan:

• Learning
  “...information literacy defined broadly to include digital information literacy, science literacy, data literacy, health literacy, etc...”

• Scholarly Communication
  “Lead in data-related scholarship and initiatives”

• Global Challenges
  “We will lead in international initiatives in information literacy and e-science and ... contribute to international information literacy, learning spaces, data management, and scholarly communication initiatives.”

Publish Datasets with DOIs

Use PURR to publish datasets with Digital Object Identifiers (DOI) that make it easier for people to cite your data and give you credit. Purdue is a founding member of DataCite, the international agency that registers DOIs for data.

http://purr.purdue.edu
FACT:

MANY FUNDING AGENCIES REQUIRE DATA MANAGEMENT PLANS WITH GRANT PROPOSALS.

purr.purdue.edu

PURR IS YOUR SOLUTION FOR:

DATA MANAGEMENT PLANS
READY-MADE BOILERPLATE TEXT TO PUT IN YOUR PROPOSALS, TUTORIALS AND BEST PRACTICES, SUPPORT FOR DEVELOPING A GOOD DATA MANAGEMENT PLAN

COLLABORATION
PURR PROVIDES A FREE HUBZERO™ PROJECT SPACE FOR PURDUE RESEARCHERS AND THEIR COLLABORATORS TO WORK TOGETHER ON RESEARCH AND SHARE DATA AND TOOLS ONLINE

PUBLISHING YOUR DATA
PUBLISH YOUR RESEARCH DATA WITH DIGITAL OBJECT IDENTIFIERS THAT OTHER SCHOLARS CAN USE TO EASILY FIND AND CITE YOUR DATASETS

ARCHIVING YOUR DATA
YOUR DATA WILL BE PRESERVED AND MADE ACCESSIBLE FOR LONG-TERM ACCESS IN A SECURE, TRUSTWORTHY DIGITAL REPOSITORY

OVER 500 GRANT PROPOSALS FROM PURDUE HAVE INCLUDED PURR IN THEIR DATA MANAGEMENT PLANS SINCE 2011

PURR IS A PURDUE RESEARCH CORE FACILITY DEVELOPED BY THE OFFICE OF THE VICE PRESIDENT FOR RESEARCH, PURDUE LIBRARIES, AND INFORMATION TECHNOLOGY AT PURDUE

To learn more, visit PURR at purr.purdue.edu or contact Courtney Earl Matthews purr@purdue.edu 765-496-2770
Data Management Plans

- Boilerplate text
- Example DMPs
- DMP Self-Assessment
- **DMPTool**
- Workshops
- Tutorials
- Reference and consultation with subject-specialist librarian and/or data services specialist

[https://purr.purdue.edu/dmp](https://purr.purdue.edu/dmp)
Create a Project

PURRR project tutorial video:
http://www.youtube.com/watch?v=q5xGO_oF9uQ
## Storage Allocation

<table>
<thead>
<tr>
<th>Option:</th>
<th>For what:</th>
<th>Space Available:</th>
<th>For how long:</th>
<th>For who:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Default/trial projects</td>
<td>Just trying things out, or don’t need much space</td>
<td>10 GB</td>
<td>3 years</td>
<td>Any Purdue faculty, staff, and graduate student</td>
</tr>
<tr>
<td>Default data publications</td>
<td>Great for small publications</td>
<td>1 GB</td>
<td>10 years</td>
<td>Any Purdue faculty, staff, and graduate student</td>
</tr>
<tr>
<td>Supported projects</td>
<td>Funded projects with PIs from Purdue</td>
<td>100 GB</td>
<td>Life of grant</td>
<td>Purdue faculty with a verifiable grant or account number</td>
</tr>
<tr>
<td>Supported data publications</td>
<td>Publishing work done on a funded project</td>
<td>10 GB</td>
<td>10 years</td>
<td>Purdue faculty with a verifiable grant or account number</td>
</tr>
</tbody>
</table>

### Estimated cost of additional space *

<table>
<thead>
<tr>
<th>Option:</th>
<th>Space Available:</th>
<th>For how long:</th>
<th>For who:</th>
<th>Cost:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extra project space</td>
<td>As needed</td>
<td>Per year</td>
<td>Any Purdue faculty, staff, and graduate student</td>
<td>$2.10 per GB per year **</td>
</tr>
<tr>
<td>Extra publication space</td>
<td>As needed</td>
<td>10 years</td>
<td>Any Purdue faculty, staff, and graduate student</td>
<td>$14.30 per GB</td>
</tr>
</tbody>
</table>

[https://purr.purdue.edu/about/pricing](https://purr.purdue.edu/about/pricing)
Data Publication & Archiving

PURR publication tutorial video:
http://www.youtube.com/watch?v=jYBcsfiRhio
Organization and Staffing for PURR

• **Executive Committee**: Dean of Libraries, Vice President for Research, Chief Information Officer

• **Steering Committee**: 2 from libraries, 2 from IT, 2 from research office and sponsored programs, 3 domain faculty researchers

• **Personnel**: Project Director (.50), Technologists (3.85), HUBzero Liaison (.35), Metadata Specialist (.20), Digital Archivist (.25), Digital Data Repository Specialist (1.0), Data Curator (1.0)
Assessment

2013 was first full year in operation, to date:
• 1,405 data management plans
• 163 grant awards
• 1,294 registered researchers
• 468 research projects
• 101 citations

DMP analysis (n=111 NSF proposals from Purdue, Jan-Jun 2013)
• 49% PURR
• 29% Local computer or server
• 14% Disciplinary repository (e.g., ICPSR, Protein Data Bank, nanoHUB, NEES)
• 8% No data or not applicable
James Toon
Research Information System Manager
University of Edinburgh
@jamestoon

Pure at the University of Edinburgh
Overview

• Background
• Pure support model
• Summary of approach
• Key challenges
• Success factors
University of Edinburgh context

• Large, research-led Russell Group institution
  – 32,000 students (11,000 postgrad)
  – Over 6,000 academic staff
  – Ranked 4th by research power in UK (REF 2014)
• Early-adopter for open access, repository since 2003 (Dspace)
• Pure as CRIS (Current Research Information System), in place since Sept. 2011
• Pure used as main institutional repository system for research outcomes since launch
What is a CRIS system?

This is a reasonable ‘starting point’...

A current research information system (CRIS) is a database or other information system to store and manage data about research conducted at an institution.

Current research information system - Wikipedia, the free ...
en.wikipedia.org/wiki/Current_research_information_system

...But far too simplistic really.
Why we have a CRIS system

Demands for research information have increased substantially in the past few years.

• Research Excellence Framework
• Information for external reporting and compliance (RCUK, HESA, open access)
• Information for strategic management (bibliometrics, collaborations and benchmarking)
• Information for researchers (Web pages, CV in support of proposals or for annual review, digital footprint)
Pure support model

- Business owners - Governance and Strategic Planning (University Secretary’s Group)
- Dedicated Pure support team – 2 FTE
- OA and publications support from University Library Scholarly Communications team – 4 FTE
- Distributed local ownership across 3 Colleges and 22 Schools - approximately 6 FTE and growing
- Service contracts with local IS infrastructure and applications support
Architecture

Pure Intermediate data (built in-house)

Managed via

Pure backend system – the admin interface

Edinburgh Research Explorer

Published via

Web service layer

Content synchronised via HR, Finance systems on nightly basis
### What information are we tracking @ Edinburgh?

**Systematic collection via PURE**

1. **People**
2. **Research publications**
3. **Research activities and awards**
4. **Press clippings**
5. **Bibliometric data**
6. **Projects and funding**
7. **Research data**
8. **Impact**

And part of future plans:

1. **Equipment and facilities**
What information are we tracking @ Edinburgh

Systematic collection via PURE

i. People

ii. Research publications

iii. Research activities and awards

iv. Press clippings

v. Bibliometric data

vi. Projects and funding

vii. Research data

viii. Impact

And part of future plans

i. Equipment and facilities
Workflows

Established for:
• Staff inclusion
• Publication/Activity data
• Press clipping/Press release
• Funding/Project data
• Open Access (i.e., RCUK/REF2020)

Emerging flows being developed:
• Research data
• Impact
Content added to Pure is available publicly via the Edinburgh Research Explorer.
Publications with full text (open access) content available via Edinburgh Research Explorer, together with information on rights as required.

Able to link related funding and other content types.
Scale of collection

Growth in content significant since 2011, and no sign of slowing

- People – 4,110 current staff
- Research Outputs – 89,755 (26% Full Text)
- Activities – 14,512
- Press Clippings – 5,233
- Projects - 13,976
- Impacts - 665
Open Access Implementation Programme - 2013

Growth in content over time since launch (total content vs content with full text attachments)
- With document attachments
- Total Items
- % Documents

Total number of research output records:
- August 09
- September 09
- October 09
- November 09
- December 09
- January 10
- February 10
- March 10
- April 10
- May 10
- June 10
- July 10
- August 10
- September 10
- October 10
- November 10
- December 10
- January 11
- February 11
- March 11
- April 11
- May 11
- June 11
- July 11
- August 11
- September 11
- October 11
- November 11
- December 11
- January 12
- February 12
- March 12
- April 12
- May 12
- June 12
- July 12
- August 12
- September 12
- October 12
- November 12
- December 12
- January 13
- February 13
- March 13
- April 13
- May 13
- June 13
- July 13
- August 13
- September 13
- October 13
- November 13
- December 13
- January 14
- February 14
- March 14
- April 14
- May 14
- June 14
- July 14
- August 14
- September 14
- October 14
- November 14
- December 14
- January 15
- February 15

Date

Percentage records with full text attachments

The University of Edinburgh

Webinar
Library Connect

#LCwebinar
James Toon
Edinburgh communities

Challenges facing Edinburgh research information management communities

• Approaches are very different across different subjects
• Research Outcomes Management as a profession
• University data governance – managing the institutional data model
• Coordinating messages across Pure/Library/Research data management teams
• Introduction of new content types
• Managing expectations of academic staff
Whose data is it anyway?

Ownership issues need careful consideration

- Offering to maintain publication lists for academics as a service. Works, but can cause disengagement
- Over emphasis on data quality tends to transfer data ownership from academic to library (or local admin) and can cause academic disengagement
- Synchronising information from corporate systems makes their data my problem
Issues of data quality

Keeping on top of maintenance chores is a BIG job, made more complex by having a CRIS.

For example, on top of dealing with management of outcomes data we have;

- 50398 external organisations
- 144976 external persons
- 10320 journals
- 3768 publishers
Key factors supporting adoption

• Key driver is University research administration – allowed a real clarity of purpose

• Having the REF was a great focus for implementation and recent UK policy developments (Finch Report/RCUK/REF2020) have reinforced this

• Research Information Management now a clear academic issue. Sanctions proposed for non-compliance with policy may directly affect their funding.
Has Pure been a success?

- Successfully delivered against the REF2014 requirement
- Data in Pure now regularly used for review of research activity (strategic/appraisal)
- Increasing number of schools using Pure data to develop local web presence
- Pure embedded as corporate system
- Growing dependency on Pure data for compliance reporting and other data submissions
- REF2020...
- Research outcomes management now established at the University of Edinburgh
Questions & Thank You!

Judith Coffey Russell
Dean of University Libraries
University of Florida

Michael Witt (@mwittin)
Head of the Distributed Data Curation Center (D2C2)
and Associate Professor of Library Science
Purdue University

James Toon (@jamestoon)
Research Information System Manager
University of Edinburgh

Alicia Wise (@wisealici)
Director of Universal Access
Elsevier