Electronic lab notebooks and data repositories

Complementary responses to the scientific data problem

ACS Dallas session on
Research data and electronic lab notebooks

17 March 2014
Rory Macneil, with thanks to Sunny Yang, Robin Rice, George Hamilton and Gary Ferguson
What we’re going to cover

• The scientific data problem

• ELNs and data repositories: Complementary approaches to solving the problem

• $1 + 1 = 3$? An example of ELN-repository integration:
  - RSpace and Edinburgh DataShare

• Looking ahead
Research workflow 1

Enter data  Document research  Publish results

And the data gets lost
The scientific data problem: How to capture data and make it available for scrutiny and reuse?

- Ease of data entry drives uptake, but adding structure to data enhances value

- Different types of data, different fields of research, different styles of documentation
Whose problem?

- Funders
- Communities
- Researchers
- PIs/Labs
- Institutions
The scientific problem rephrased:
How to add structure to data?

– Who decides the parameters?
  • Researcher
  • PI
  • System (ELN/Repository)
  • Community (Dial-a-Molecule, Henry Rzepa et al)

– Who adds it and how?
  • Researcher
  • Curator
  • Machine

– When is it added?
  • Pre-documentation (in system)
  • During documentation (in ELN)
  • Post-documentation (in Repository)

– How much scope for variation?
How do ELNs and data repositories deal with this problem?
What is an electronic lab notebook?

Software that enables researchers to manage and present research data
ELNs: The first generation
1990s origins for Big Pharma

- Windows based
- For pcs
- Domain and project specific
- Complex
- Required lots of training
- Very expensive
ELNs: The second generation
2000s for academic researchers

– Generic tools adapted by individuals
  • Evernote, Dropbox
  • Web-based, platform agnostic
  • Generic, flexible
  • Very easy to use, very cheap

– Lab-oriented ELNs
  • eCAT, Lab Archives, Ruro
  • Web-based, platform agnostic
  • Generic, flexible
  • Easy to use, cheap
  • Sharing and groups
ELNs: We need a third generation

2011/12 Wisconsin pilot

“We need an ELN that can be rolled out across the university”

It has to:

• Be easy to use
• Be platform agnostic
• Support intra- and inter-group collaboration
• Have enterprise capabilities
• Support data publishing and archiving

and . . .

Still be affordable!!!!!!!!!!”
Institution as customer: Brave new world

- ELN must provide flexibility and breadth across disciplines
- University provides
  - Funding
  - Support
    - IT
    - Training
- Driving
  - Mass uptake
  - Consolidation of providers
Who and what is driving demand for ELNs?

- Researchers
  - Utility and convenience of paper lab book + online capabilities
  - On multiple devices
  - File management/integration

- Groups/PIs
  - Controlled sharing
  - Collaboration
  - Group management
  - File management/integration

- Institutions: Data librarians, IT, commercialisation offices
  - Enterprise features: Scalable deployment, Single Sign On
  - IP protection: audit trail, signing
  - Publishing
  - Archiving
  - Repository integration
  - File management/integration
Data repositories

An information repository is an easy way to deploy a secondary tier of data storage that can comprise multiple, networked data storage technologies running on diverse operating systems, where data that no longer needs to be in primary storage is protected, classified according to captured metadata, processed, de-duplicated, and then purged, automatically, based on data service level objectives and requirements.
Three types of data repositories

• Domain specific
  – Life Sciences: Dryad, NCBI Databases

• Institutional

• Generic
  – Figshare
What’s driving demand for data repositories?

• Domain Specific
  – Who?
    • Research communities
  – Why?
    • Capture data for review and reuse
    • Improve quality and reliability of data
    • Standardize research techniques to improve productivity

• Institutional
  – Who:
    • Funders → Data librarians
  – Why?
    • Maintain data
    • Make data available for reuse
ELNs and data repositories: what do they do?

ELN
- Data entry
- Data organization
- Documentation
- Metadata creation
- Data and metadata export

Repository
- Data import
- Data preservation
- Data sharing
- Data reuse
ELNs + Data repository enables new workflow

ELN
Enter data and document research

Data repository
Store data and metadata

Publication
Publish results

Data is captured, structured and made available for reuse
Research workflow 1

Enter data → Document research → Publish results

And the data gets lost
ELNs + Data repository enables new workflow

Data is captured, structured and made available for reuse
An example of integration: RSpace and Edinburgh DataShare
RSpace

• Conceived in response to Wisconsin RFP and trial

• Developed with Wisconsin by Research Space 2012 - 2013
## Researcher experience

<table>
<thead>
<tr>
<th>Feature</th>
<th>√</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sketching</td>
<td></td>
</tr>
<tr>
<td>Image annotation</td>
<td>√</td>
</tr>
<tr>
<td>Chemical structures</td>
<td>√</td>
</tr>
<tr>
<td>Notebook</td>
<td>√</td>
</tr>
<tr>
<td>Forms</td>
<td>√</td>
</tr>
<tr>
<td>Templating</td>
<td>√</td>
</tr>
<tr>
<td>PDF export</td>
<td>√</td>
</tr>
<tr>
<td>Export to Word</td>
<td>√</td>
</tr>
<tr>
<td>File gallery</td>
<td>√</td>
</tr>
<tr>
<td>Journal view</td>
<td>√</td>
</tr>
<tr>
<td>Tablet friendly</td>
<td>√</td>
</tr>
<tr>
<td>Clean design</td>
<td>√</td>
</tr>
<tr>
<td>Performance</td>
<td>√</td>
</tr>
<tr>
<td>Round trip editing</td>
<td>√</td>
</tr>
<tr>
<td>Offline access</td>
<td>√</td>
</tr>
<tr>
<td>Sample management</td>
<td>√</td>
</tr>
</tbody>
</table>
## PI/Lab support

<table>
<thead>
<tr>
<th>Feature</th>
<th>Available</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sharing</td>
<td>✓</td>
</tr>
<tr>
<td>Messaging</td>
<td>✓</td>
</tr>
<tr>
<td>Lab set up enabled</td>
<td>✓</td>
</tr>
<tr>
<td>Group management</td>
<td>✓</td>
</tr>
<tr>
<td>Inter-group collaboration</td>
<td>✓</td>
</tr>
</tbody>
</table>
Institutional requirements
(IT, data librarians, commercialisation)

<table>
<thead>
<tr>
<th>Requirement</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single sign on</td>
<td>✓</td>
</tr>
<tr>
<td>Sys Admin/RSpace admin</td>
<td>✓</td>
</tr>
<tr>
<td>Group set up</td>
<td>✓</td>
</tr>
<tr>
<td>IP support</td>
<td>✓</td>
</tr>
<tr>
<td>Export to XML</td>
<td>✓</td>
</tr>
<tr>
<td>Archiving</td>
<td>✓</td>
</tr>
<tr>
<td>Repository integration</td>
<td>✓</td>
</tr>
</tbody>
</table>
RSpace design advantages

• Easy data entry
• Easy and flexible data structuring
• Multiple ways of getting data out (and back in)
  – Export PDF
  – Export Zip (XML)
  – Re-import, preserving structure
  – Archive (with metadata)
• Inter – institutional support
  • Re-import, preserving structure
What is Edinburgh DataShare?

Edinburgh DataShare is a free-at-point-of-use data repository service which allows University researchers to upload, share, and license their data resources for online discovery and re-use by others.
Whence Edinburgh DataShare?

- The service was built as an output of the DISC-UK DataShare project, which explored pathways for academics to share their research data over the Internet at the Universities of Edinburgh, Oxford and Southampton (2007-2009, Jisc Repositories and Preservation Programme).
Why Edinburgh Datashare?

“7. Research data management plans must ensure that research data are available for access and re-use where appropriate and under appropriate safeguards.”
“9. Research data of future historical interest, and all research data that represent records of the University, including data that substantiate research findings, will be offered and assessed for deposit and retention in an appropriate national or international data service or domain repository, or a University repository.”
“10. Exclusive rights to reuse or publish research data should not be handed over to commercial publishers or agents without retaining the rights to make the data openly available for re-use, unless this is a condition of funding.”
Does DSpace work for data?

- Communities, collections, data items, files
- Metadata subset from DCMI “dcterms” vocabulary, RDF-compliant – aids discovery
- Persistent identifier (handle) assigned
- Suggested citation provided; viewable download statistics
- Single-sign on with EASE
- Embargo option – delayed publication
Quality controls

- New users can register with EASE login but are given permission to add to a collection.
- Data Library staff or collection administrator approves each item before publication.
- Dataset/data item must include documentation as well as data files.
- Metadata field to point to related publication (“is referenced by”).
RSpace – Edinburgh Datashare integration: Overview

• Background

• People
  – Edinburgh: Robin Rice, George Hamilton
  – Research Space: Sunny Yang, Gary Ferguson

• Process
  – Understanding DataShare (DSpace + SWORD + Datashare)
  – Backend integration
  – METS (metadata encoding transaction standard) development
  – Interface decisions
RSpace – Edinburgh Datashare integration: Backend platform

– Edinburgh DataShare has three interfaces/APIs
  • Web-UI
  • Python
  • SWORD (simple Java based web-service which supports repository deposits)

– RSpace uses the SWORD Interface

– The SWORD server accepts a file for deposition if a METS description file is provided
Four part METS implementation in RSpace – Datashare integration

- RSpace uses the standard METS header
- DMD -- field definitions are based on Dublin Core
  - Four required fields in Edinburgh DataShare -- contributor, publisher, title, and data creator -- must be completed as part of the deposit through RSpace
  - Additional optional fields can be filled in later by DataShare administrator:
    - FUNDER, SPATIAL_COVERAGE, TIME_PERIOD, DATA_CREATOR, AVAILABLE_DATE, DESCRIPTION_ABSTRACT, DESCRIPTION_TOC, LANGUAGE, RELATION_VERSION_OF, RELATION_REFERENCED_BY, SUPERCEDES, RIGHT, SOURCE, SUBJECT_KEYWORDS, SUBJECT_CLASSIFICATION, ALTERNATIVE_TITLE
- All zipped files and their mime-types (e.g. application/pdf, text/html) are included
- A structure map describes the full structure and relationships between the above three elements
RSpace – Edinburgh Datashare integration: Workflow

- **Front end trigger**
  - An RSpace user selects files/folders/notebooks to be deposited from RSpace, and starts the deposit process

- **Backend to support the user workflow**
  - RSpace extracts the associated data and resources from its database and file-store
  - These are turned into xml files
  - METS is used to describe the zip file and each selected file
  - The xml, resource, and METS files are zipped into a zip file for archiving
  - The DSpace SWORD client deposits the zip file to DataShare after an authentication and validation procedure
  - File deposited in Collection associated with Depositor
User workflow 1

My records in RSpace
User workflow 2a
Record(s) for export

continue with ADI western.

Wash strips once in 10 ml PBS then 3 x 5 min in about 10 ml PBS.

Make up secondary anti-mouse HRP antibody as 1:1000. Incubate with rotation, at room temp, 1 hr.

Wash once in 10 ml PBS as before, then 3 x 5 min in TBS.

Take out developers from fridge before starting to wash, mix solution A and B 1:50 after last wash. Place strips on saran wrap, align their height to fit master, Develop in dark. Place on screen with master. Develop 30', 1 min, 2' and take photo of strips with master under light.

Lane 1: uninfected sample 8:330; Lane 2: infected 1:330; Lane 3: 1:600, Lane 4: 1:900

Conclusion: Can definitely detect protein induction after infection! Lots of background as always obviously, but can definitely see a nice specific band.
User workflow 2b

Record(s) for export

Date

2013-05-20

Objective

To determine the ED50 value of Drug A on Cancer cell line XYZ

Method

1. A dilution series of Drug A was made based on the calculations below.
2. 100 µl of each dilution series was then plated in quadruplicate.
3. Plates were incubated for 72 hrs and cell viability assessed using an XTT assay.

Plate layout:

Calculations:

Drug A Desired concentration 100 µM
Desired volume: 4 ml

Mix 0.007 ml of Stock Drug A + 3.943 ml of media to generate the top dilution.
User workflow 2c

Selecting record(s) for export

<table>
<thead>
<tr>
<th>Options</th>
<th>Type</th>
<th>Owner</th>
<th>Name</th>
<th>Last Modified</th>
</tr>
</thead>
<tbody>
<tr>
<td>✔️</td>
<td>gFerguson</td>
<td>Sensitivity of Cancer cell line XYZ to Drug A</td>
<td>2014-03-13 11:29</td>
<td></td>
</tr>
<tr>
<td>□</td>
<td>gFerguson</td>
<td>bam mutant RT-PCR</td>
<td>2014-03-13 11:29</td>
<td></td>
</tr>
<tr>
<td>□</td>
<td>gFerguson</td>
<td>SJHF3 ELISA</td>
<td>2014-03-13 11:29</td>
<td></td>
</tr>
<tr>
<td>✔️</td>
<td>gFerguson</td>
<td>Notebook</td>
<td>2014-03-13 11:29</td>
<td></td>
</tr>
<tr>
<td>□</td>
<td>gFerguson</td>
<td>Extract Protocol for ChIP</td>
<td>2014-03-13 11:29</td>
<td></td>
</tr>
</tbody>
</table>
## User workflow 2d
### Selecting record(s) for export

![Image of ResearchSpace interface]

<table>
<thead>
<tr>
<th>Options</th>
<th>Type</th>
<th>Owner</th>
<th>Name</th>
<th>Archiving</th>
<th>Data Share</th>
<th>Last Modified</th>
</tr>
</thead>
<tbody>
<tr>
<td>✔</td>
<td>0</td>
<td>gFerguson</td>
<td>Sensitivity of Cancer cell line XYZ to Drug A</td>
<td>✔</td>
<td>✔</td>
<td>2014-03-13 11:29</td>
</tr>
<tr>
<td>□</td>
<td>0</td>
<td>gFerguson</td>
<td>bam mutant PCR</td>
<td>✔</td>
<td>✔</td>
<td>2014-03-13 11:29</td>
</tr>
<tr>
<td>□</td>
<td>0</td>
<td>gFerguson</td>
<td>SJHF3 ELISA</td>
<td>✔</td>
<td>✔</td>
<td>2014-03-13 11:29</td>
</tr>
<tr>
<td>✔</td>
<td>0</td>
<td>gFerguson</td>
<td>Notebook</td>
<td>✔</td>
<td>✔</td>
<td>2014-03-13 11:29</td>
</tr>
<tr>
<td>□</td>
<td>0</td>
<td>gFerguson</td>
<td>Extract Protocol for ChIP</td>
<td>✔</td>
<td>✔</td>
<td>2014-03-13 11:29</td>
</tr>
</tbody>
</table>
User workflow 3
Deposit dialog
User workflow 4
Completed deposit dialog
User workflow 5
Depositing dialog
User workflow 6

DataShare agreement message
User workflow 7

DataShare agreement email

datalib@ed.ac.uk
13/03/2014 11:56

to: gary@researchspace.com

DataShare: Depositor Agreement Required

Thank you for deposit of Gary’s Notebook via the Edinburgh DataShare SWORD interface. In order to curate, reproduce and distribute your submission worldwide, you must agree to the terms of the Depositor Agreement [view: http://www.ed.ac.uk/schools-departments/information-services/services/research-support/data-library/data-repository/depositor-agreement] by clicking on the link below:

http://devel.edina.ac.uk/9228/deposit-agree?item=54821

Many thanks!

Edinburgh DataShare Repository Team

The University of Edinburgh is a charitable body, registered in Scotland, with registration number SC005336.
You submitted: Garys Notebook

To collection: ResearchSpace Test

Your submission has been accepted and archived in DataShare, and it has been assigned the following identifier:
http://delta.edina.ac.uk:9220/handle/123456789/54510

Please use this identifier when citing your submission.

You can find a copy of the Depositor Agreement at: http://www.ed.ac.uk/schools-departments/information-services/services/research-support/data-library/data-repository/depositor-agreement

Many thanks!

Edinburgh DataShare Repository Team

---
The University of Edinburgh is a charitable body, registered in Scotland, with registration number SC005336.
User workflow 9
Submission screen in DataShare
Looking ahead 1: Integrating ELNs with other digital resources

– Other institutional repositories

– Domain specific repositories
  • PubChem, ChemSpider

– Archives
  • Institutional
  • Commercial
    – Chemical Semantics
    – CDD Vault
Future workflow 1

ELN
Enter data and document research

Data repositories
Store data and metadata

Publications
Publish data and results
Looking ahead 2:
ELN as interactive publishing platform

- Institutional repositories
- Domain repositories
- Archives
- Datasets (Publications)
- Data Metadata

ELN
ELN transformed from point tool to infrastructural glue at the heart of the research process
ELN also becomes the backbone of data management at the University

Integration with

- Data storage
- Data repository
- Data archive
Scientific data: whose problem?
Which model(s) will win?

Those that best satisfy the core requirements of the key constituencies

- Easy data entry and usability
  - Researchers, Labs/PIs
- Easy integration with write up/publication
  - Researchers, Labs/PIs
- Enhanced collaboration and better management of data and research
  - PIs/labs, Institutions
- Comprehensive and easy data capture and availability for reuse
  - Institutions/Communities
- Flexibility to support different workflows and outputs
  - Researchers, Labs/PIs, Communities

- And by the way, as Wisconsin said:

It has to be affordable!
Thank you

Rory Macneil
rmacneil@researchspace.com