

Repositories Unleashing Data: Ideas

Questions discussed by five breakout groups at RepoFringe, 4th August, 2015, Edinburgh. These are the comments and ideas that have been contributed. Written up by Pauline Ward, 13 August 2015.

Contents

1. Who, beyond the academic community, is currently using data from institutional repositories?	1
2. Where or who are the untapped audiences for the research data in institutional repositories?	2
3. How can repositories help researchers tell the story of their data?	3
4. How can repositories help researchers achieve impact with their data?	4
5. What should our vision be for the repository landscape of the future?	5

1. Who, beyond the academic community, is currently using data from institutional repositories?

- **SUCCESS STORIES: Examples.**
Software has more value for the non-academic community.
Environment website (Scotland – SEPA?) is a discoverable service.
British Library for theses.
Particularly with use of OCR for identifying text within them.
EUDAT.eu (repository of last resort) – faceted search.
SEPA: will look at repository data to help users.
Commercial research organisations – companies don't want to disclose where they got data from.
- **STRATEGY: Examples of effective approaches to generate this kind of outreach.**
When researchers move between research projects there should be a continuation strategy.
Faceted search (enables good searching through metadata).
Accessibility – finding things if they are in repositories with lots of traffic.
Need to consider which licences are used, perhaps more CC-BY.
Need to measure impact (licensing).
Need visible download metrics.
Need metrics for academics.
Need ability to search multiple databases at once.
Aggregated data centres might be better/more discoverable than institutional (silo-type) data repositories.
- **WISHLIST: What resources or university / funder policies do we need to help us achieve this?**
Data is out there, but how can you find it?

Data is out there, but we don't know how people re-use it.
We need mediation so the public can access it and the academic can get more impact.
We need impact going to both the academic and the institution (by knowing who is using their data).
Promoting – We need to let people know the data is there and is available.
How crawlable is content in repositories?
Support repo managers to support researchers to identify / reach their audience.

- **VISION: Future repository landscape.**
EPSRC Open Data requirement – impact of research -> who uses data.
How can we integrate data?
Lots more use of data by audiences beyond the professional academic community, and by undergrads.
When someone searches for a researcher's name, a site where they can browse that researcher's data is one of the top hits.

2. Where or who are the untapped audiences for the research data in institutional repositories?

Other scholars (interdisciplinarity).
Business.
Public.
Cultural heritage.
Schools.
Government.
Media.
Comedians -> mass appeal.

- **SUCCESS STORIES: Examples.**
Monash Music Archive.
Figshare.
Lost languages.
Music.
->Created Collections.
->Easy user interface.
->Framing to a non-academic audience.
Mining API for small businesses.
-> to find expertise.
- **STRATEGY: What should be the strategy of the repository community to help us and the researchers we support reach untapped audiences?**
Identifying opportunities for re-using and re-purposing data.
Public Engagement (as per the successful example of Figshare).
Institutional Repositories need open APIs and need to publicise them.
Quality control.
Peer review.

Workshops on how to do this.
 Teaching undergraduates.
 Increased visibility; workshops.
 Learn from other communities e.g. Art being more accessible to the public, likewise research data / outputs.
 Carrots for researchers.
 Facilitating depositors to link up with Public Engagement experts such as Beltane, and institutional Knowledge Exchange / commercialisation teams, and media/Comms teams.
 Draw on expertise of the Open Data Institute.
 Train / raise awareness / work with the people who provide Impact / Public Engagement training to researchers (e.g. Vitae?), so they will understand data sharing i.e. what repositories are, what kinds of licences may be applied to data, what kind of metadata is needed for services like Google to pick up datasets for interested Google users.
 Need to ask the researchers (individually, and their representatives).
 Find ways to facilitate data analysis, not just searching eg statistical analysis, and specialised structure/search/visualisation tailored to the researcher – need to ask researchers about this.
 Present the data in more fun / accessible ways – smartphone apps – need to consult with researchers and the public about this. Gather feedback on presentation / accessibility of data from as many potential users as possible.

- **WISHLIST: What resources or university / funder policies do we need to help us achieve this?**

Skills in Public Engagement.
 Tweet button on the repository.
 Major consultative exercises with researchers and with untapped audiences (public, government, charities, business, schools) to find ways to present data in more appealing and accessible ways. Resources to develop the software to do it.

- **VISION: Future repository landscape.**

Easily discoverable.
 Easy-to-use interface.
 Thinking of data as an asset.
 Lots of functionality to make data accessible / analysable in different ways.

3. How can repositories help researchers tell the story of their data?

- **SUCCESS STORIES: Examples of success stories.**

a. [JCB DataViewer](#): See [presentation](#) from Liz Williams, Exec Editor, Journal of Cell Biology

b. [LINCS](#): see [Publications](#) page for outputs and impact.

c. Dundee Data Repo: [Example Paper](#) and [associated data](#).

[added by Jason Swedlow, Univ of Dundee, jrswedlow@dundee.ac.uk, @jrswedlow]

A story of loss of data by keeping data on a website and then understanding the advantages of depositing data in a repository and making it available via website links.

Stories of re-use of data.

- **STRATEGY: What should be the strategy of the repository community to help us achieve this?**

Work with the researchers to tell the story(ies) of data.

Need plans, and communication, and where to go.

Should have business cards with DOIs.

Use cases – Discovery

- Lost data (and the importance of preservation).

Learn from other services.

Familiarity of system and ease of use.

Self-deposit must be encouraged.

Do not have to have all data; record and link.

Repo managers can use blogs and twitter to promote the content in their repositories.

- **WISHLIST: What resources or university / funder policies do we need to help us achieve this?**

Funders to provide case studies or finance.

Credit for data.

Data Journals.

More money and staff.

Metrics.

Jisc to collect Success User Stories.

- **VISION: Future repository landscape.**

Ability to visualise the data within the repository.

Ability to leave feedback, and links.

Ability to attach a wiki page to a deposit.

Ability to redeposit dataset with additional value.

Impact story i.e. summaries.

The Story needs more than data – need to link to projects.

Allow users to follow datasets.

Alert people that new datasets have been deposited.

Integration with colleagues and staff lists.

Researchers have easy-to-use information on the reaction to their research in media , social media.

4. How can repositories help researchers achieve impact with their data?

Does 'Impact' still have negative connotations in the UK? i.e. associated with the REF?

Financial impact cf. impact on society.

What is impact: change, or benefit?

Is there value in predicting impact?

- **SUCCESS STORIES: Examples of success stories.**

DOIs for papers e.g. DataCite.

Visualisations.

Visualisations of elements of data.

Data: targeted but open.

The Impact Story.

Paid-for service.

- **STRATEGY: What should be the objectives of the repository community in this respect?**

We need to stop using acronyms.

Who are the 'non-academics' most likely to benefit from the data and how will they find it? -> twitter, searching...

Facilitate searching across multiple (all?) institutional repositories.

University library searches (like Ebsco's Searcher) should highlight repository holdings in other institutions. Undergrads use these search interfaces.

Visibility outside of academia – industry and charitable sector – and ease of re-use for same.

Flickr-like metadata.

Searching across other data repositories.

Discussion of social media and citations – these don't necessarily indicate impact.

Lay-person summaries (will also help researchers from other disciplines).

We need better descriptors of datasets for other potential users.

Repo managers need good awareness of what kind of (measures of) impact researchers are incentivised to / want to achieve.

Engagement with researchers.

- **WISHLIST: What resources or university / funder policies do we need to help us achieve this?**

We need a 'widget' to provide a good visualisation of the data.

Training for repo managers in what Impact is / can be, so we can design repositories and our processes to facilitate that.

Training for repo managers in Public Engagement, so we can design our repositories and our processes to facilitate that.

Universities should treat publication of a dataset as an opportunity for communicating impact to the public i.e. have the comms team monitor the deposits in your data repository for news-worthy datasets.

5. What should our vision be for the repository landscape of the future?

Re-use of data: Lack of evidence for desire for re-use. -> Does the appetite truly exist? See Open Source code for comparison.

Dataset quality: Do people trust others' datasets? Is machine data regarded more highly? Impartial? Would more information on methods / reagents help?

Provide a vehicle for sharing data in a long-term and self-sustaining way.

Data as VoR? [Voice of the Researcher?]

Public are generally aware that university research data is openly available, and re-use (including re-analysis, re-packaging, mashing up) by non-professional researchers is seen as natural and a good thing.

- **What are the opportunities for mining and analysing the data?**
Metadata standards: - interoperability; - finding and using are very important.
Licensing – needs to be open to possibility of future data mining.
- **What is the role for commercial publishers?**
Data Journals – does this risk us getting into the same situation as with traditional publishers?
Currently set standard e.g. peer review of data.
In future, will they be so influential? Depends on evaluation criteria e.g. for REF 2020.
Need to demonstrate added value if they are going to continue to be profitable.
- **How can we ensure the public gains the greatest possible benefit?**
Creating an evidence-curious new generation through schools.
Accessible language.
Attractive, easy-to-use interfaces. But don't patronise – they're smarter than we think!
Lay summaries of all outputs.
Data plugs in to easy-to-use visualisation.
Maximise discoverability.
Increase awareness that university researchers now share their data, and where it can be found (whether that's the institutional repository, or Jisc's UK service, which needs a snappier name, since I can't remember it, and I'm on one of the consultative groups).
Draw on expertise of Open Data Institute and e.g. Wellcome Trust in Public Engagement, expertise of others in Knowledge Exchange, to identify potential users and reach out to them.
Work with researchers to find out what they want from repositories (and from universities / funders?!) in terms of helping them deliver greatest possible benefit to the public.
- **Do we need to beware the commodification of data?**
Only when we've made it happen.
Data is an output from research; is it a commodity already?
Why would commodification be a problem?
Reduction of variety?
Commodification -> standards -> greater usability.
- **What will the relationship be between papers and data?**
Will be stronger, because Research Councils require data sharing when it underpins a publication; but until funders give data parity of esteem with articles the relationship will be "look but don't touch".
As close a relationship as possible – this way scholarly communications can truly generate re-use.

Data paper = human readable overview of how data was generated. Required context to re-use data in the future and/or replicate / validate analyses based on the dataset.

Increased blurring of the lines. Researchers publishing more data which stands alone, or is a follow-up to a paper.

