Sharing is Good: Getting the Most Out of Your Library's Data.

Bethan Ruddock, MA, MCLIP. Jisc.

Introduction

As individuals, we make decisions daily about what data to share. Sharing our data can bring us many rewards, from the personal boost of likes on Facebook and Instagram, to financial rewards and discounts for allowing companies to track your shopping profile.

Data sharing is not without its perils. For individuals, over- or misguided sharing may result in anything from embarrassment to identity fraud, and learning appropriate levels of data sharing is an important 21st century literacy.

For organisations and institutions, decisions about what data to share are even more fraught. Organisations must comply with regulations to keep certain data confidential (such as the EU data protection directive (European Commission, 2014), while complying with other regulations to make other data open, (such as the open data policies of the US (the White House, 2015) and UK (data.gov.uk, 2015) governments, and the Wellcome Trust (Wellcome Trust, 2015)).

It’s not just legal regulations that organisations have to consider when deciding what data to share. Data can confer a significant competitive advantage, which applies whether organisations are competing for customers, students, or funding grants. For organisations need to make carefully informed decisions about data sharing, they need to channel Kipling’s Elephant’s Child, and always repeatedly ask:

“What and Why and When / And How and Where and Who.” (Kipling, 2010, 83)

Why: Share Data?

It may seem that the first of these questions to ask is ‘why?’. Why share your data? In deciding why to share data, beyond what is mandated, organisations need to consider how sharing will affect them and others. Things they might want to consider include:

Who owns this data? Do we have the right to share it?
Is it moral and ethical for us to share this data?
Will sharing this data benefit us as an organisation?
Will sharing this data benefit society as a whole?
Are there any potential negative implications of sharing this data?
Are there any potential negative implications of not sharing this data?
You will see that it is very difficult to consider these questions without the what, where, when, how, and who. You need to decide what data you are thinking about sharing and who you want to share it with before you can start to make a good case for why you should (or shouldn’t) share it. As with any project, it’s a good idea to start thinking about data sharing by identifying your needs and research question:

What do we want to achieve?
Can we achieve that by sharing some of our data?

Once you know what you want to achieve, you can start to investigate how you can work towards it.

**HOW: ACTIVE AND PASSIVE DATA SHARING**

One of the first decisions to be made is the choice between passive and active data sharing.

Passive data sharing is what many people will think of when you talk about opening up data or data sharing. If you put an open data licence (Open Definition, 2015) on your data, then make it accessible and available for download and reuse without any further interaction from you (beyond basic maintenance) or any specific question or end case in mind, that can be defined as passive data sharing.

Passive data sharing may be the right thing for you if:

- You have limited resources
- You have no specific research question or desired outcome
- Your research question is ‘what happens if...’
- Your desired outcome is ‘open our data...’

If you have more specific results in mind, then it is likely that you want to do some active data sharing. That’s where you actually have a purpose in mind: you decide to share specific data with specific people for a specific purpose. Active data sharing means that you are more involved with the process, and actively select and curate data, with an end goal in mind. This can be more resource intensive than passive data sharing, but the rewards can be commensurately higher.

Active data sharing may be the right thing for you if:

- You have some resource/expertise to work with your data
- You have a specific research question (‘can our data tell us if...’)
- Your desired outcome is concrete and measurable (‘can our data help us achieve...’)

One of the most famous quotes of the Open Data movement is

"The best thing to do with your data will be thought of by someone else" (Pollock, 2015)
For active data sharing it could be more appropriate to say

"The best thing to do with your data will happen when it meets someone else’s data."

This is the principle behind Big Data: that large data sets can do things that single data sets can’t. An excellent example is systematic reviews and meta-analyses, used so often now in evidence-based medicine (Cochrane, 2014). Data from different sources can allow us to see patterns and draw conclusions that we just can’t find in the single sources (Higgins and Green, 2011, 1.2.2), allowing researchers to synthesise and make connections. This is an important principle to consider when you’re deciding whether and why to share data: will this sharing produce insights and information that wouldn’t be possible for you to discover from your own data? Consider how this new information might benefit your sector and society as a whole. If this new information is also available to competitors, your competitive edge then becomes what you do with the information, not merely possession of it.

How: Can you get your Data Out?

"Garbage in, garbage out". It’s a commonly known saying that is usually used to refer to mean that if the data you input is bad quality or nonsense, that’s all you will get out. But it can be extended to think about whether you can export your data at all. No matter how good the quality of your data is, if you can’t get it out of the system you’re using in a reusable format, then it has no reuse value. For the purposes of data sharing, it is effectively garbage.

When considering purchasing or implementing a new system, you should always consider how data is stored, and how it can be output. Questions to ask your systems maintainer or supplier include:

Can the data be exported?
What export formats are available?
Are these export formats standard formats?
Are these export formats transformable and reusable?
How does the exported data need to be stored?
Can you export the data yourself?
If the provider must export the data, is there a charge for this?

Questions to ask yourself include:

Do you know what to do with it once you’ve got it?
Do you have the expertise to work with it or transform it?
Do you know what questions to ask of it?
And how to understand/interpret the answers?
Do you need any additional expertise? (eg programmatic, statistical)

Active data sharing can often help you to answer these questions, as providers of shared services, or other contributors, are often able to help give advice on data extract, transformation, and provide expertise in various areas.
**What: Uses for shared library data**

At Jisc, we provide digital solutions for UK education and research, including many that are built on various types of shared data, including bibliographic, census, geospatial, statistical, and learning object. What follows is an overview of a selection of Jisc projects and services that use shared library, archive, and bibliographic data.

**WHERE: COPAC**

Copac [http://copac.ac.uk/] started in 1996 as a shared OPAC for a specific group of research libraries, and now takes in bibliographic data from nearly 90 UK academic, specialist, and research libraries. This data is combined into a union catalogue which gets thousands of searches daily from academics and researchers worldwide. Copac is a physical union catalogue, which means that data is loaded into a single shared database, rather than remotely searched.

**How: Data and Deduplication**

Data was traditionally submitted to Copac in MARC format, but in recent years we’ve started to accept data in other formats. The most common format is XML, but we’ve also had data sent to us in Excel spreadsheets. Most major library systems can export MARC data, but as Copac has started to include smaller and more specialist libraries, they are using smaller and more specialist library systems, which don’t always have a MARC export. For data to be acceptable for inclusion into Copac, it needs to be consistent and transformable. This applies to MARC data as well as non-MARC data. All data is checked for consistency, and transformation rules are created and passed to the programmer.

![Figure 1: example of mapping XML data to MARC fields](image)

4
Much of the MARC data that comes to Copac is consistent internally, but inconsistent with MARC data that is supplied by other libraries. This is partly because the MARC guidelines allow room for local practice, and partly because of the long history of the changes in MARC mean that acceptable practices for record creation have changed over time.

One of the key features of Copac for users is deduplication, where records for the same item from different libraries are merged into a single record (Sanders, 2011). For this to be completed with reasonable accuracy, key bibliographic information about the item must be in the same MARC fields, so these can be compared. However, even where the data is in the same fields, differences in cataloguing practices mean that it is not possible to deduplicate records on Copac with 100% accuracy. We always err on the side of non-merging, and certain fields are considered to be ‘dealbreakers’ for merging. For example an item catalogued with the date of c.2009 will not be merged with an item with a date of 2008, no matter how well the other details match.
Once the data is consistent MARC, it is transformed into MODS (XML for MARC) which is then made available through the Copac website for searching, but also for remote searching and harvesting through an API. Data which comes in originally as MARC data goes into the RLUK database [http://www.rluk.ac.uk/services/], a subscription-only database for shared cataloguing.

Why: Contribute Bibliographic Data to Copac?

Exporting this data to Copac isn’t always a trivial task, with the demands placed on the organisation varying according to their library systems and the amount of expertise and support they have available. Neither is it a one-off task: to maintain database currency, we ask contributors to send details of new, updated, and deleted records. Contributing data to Copac is an ongoing commitment. So why do libraries do it? These quotes from Copac contributors show some of the benefits that libraries feel from sharing their data through Copac:

"[Being on Copac] provides a link to Higher Education and other research libraries, and that’s good because [specialist library] are isolated in many ways. We mention it to visitors – gets us a bit of kudos! … We’re thrilled to have got on." (Specialist library)

"Making our resources identifiable to new audiences and other professionals." (Specialist library)
"If our OPAC is down, can still be searched via Copac. Allows [us] to identify possible partners for collaboration. Our library is sometimes closed for functions and it is useful to direct researchers who have travelled to London to other London libraries that hold what they want." (Specialist library)

"Very useful for academics and post-graduates – and feeling of supporting the scholarly community." (University library)

"We’re not the most approachable library; [specialist area] libraries aren’t, by the very nature of things, so this is another branch out into the wider world, which is so important for us … Copac reaches that wider audience. Cataloguing is important to me because otherwise the books just gather dust. I don’t want that: I want them used and read. We can’t really justify ourselves to the [governing body] because they want a different readership, or usership – I mean they’re after tourists specifically. We’ll never get thousands of tourists or visitors through our door, because we’ve got specialised book collections. So we’ve got to reach out ‘over the hills and far away’, which is what you [Copac] help us do." (Specialist library)

"Prestige. Being part of the research community and having what you know are very valuable holdings exposed to the research community … Can see lots of advantages to being part of the community. Exposure. Potential for collaboration in the future – in collection management? Part of UKRR [UK Research Reserve]. Envisage eventually being able to compare holdings using a central tool." (University library)

**WHERE: COPAC COLLECTION MANAGEMENT TOOLS**

Over the past few years we’ve been putting the Copac data to another use, driven by user demand. As identified in the last quote from a Copac user, the Copac data has significant potential to be used to collaborative collection management. The White Rose Consortium, a group of UK university libraries, had an idea for a collaborative collection management tool, and asked us to work with them to build it. The libraries had the idea for the benefit they wanted, and Copac had the critical mass of shared data to achieve it.

Copac Collection Management Tools (CCM Tools) are allowing libraries to use the bibliographic and holdings data they’ve shared through Copac to take very practical steps to inform and assess their collection management policy. Use cases for the CCM Tools include:

- Identify last copies.
- Assess material for retention, digitisation, and conservation.
- Visualise collection strengths.
- Discover rare and unique holdings.
- Develop new collections.
- Explore subject collections across the UK

More details, and case studies demonstrating how CCM Tools are being used to address these questions, can be found on the CCM website [http://ccm.copac.ac.uk/]
You can search CCM Tools in a number of ways. One of the most common is by uploading files of ISBNs or local record numbers, exported from an LMS, by which libraries can assess the rarity of their holdings. The CCM Tools allow you to see a list of each item, with details of other libraries that hold that item. This is made possible by the deduplication of the Copac records.

Figure 5: partial result list from a CCM Tools ISBN batch search

CCM Tools visualisations allow librarians to quickly scan the data to see trends.

Figure 6: graph from CCM Tools showing the relative rarity of items
Figure 7: graph from CCM Tools showing which libraries hold most items in a results set

Figure 8: map from CCM Tools showing distribution of holding libraries

This data is available to download as graph data, map data, MODS XML, and pseudo-MARC (not “proper” MARC data, but a small subset of content that gives collection information in the right format for import into systems which require MARC import. This lets libraries incorporate a snapshot of collection comparison information into
their own records). Users can then transform and manipulate this data to produce customised reports and meet institution-specific needs.

**Why: use CCM Tools?**

The CCM Tools are currently in a closed beta, transitioning into service in summer 2015. Beta users have been enthusiastic about the benefits of CCM Tools:

“The tools’ ease of use has been impressive….we can already see the benefits these tools can provide: a better understanding of our collections in comparison with other HE institutions; the ability to highlight and promote research strengths to our user community, and support for our collection management decisions.”

“We were keen to explore how our holdings compared with other UK institutions. By running the collection through the [CCM Tools] we immediately saw that our instincts about the strength and scarcity of the material held in the collection seemed to be true.”

“The CCM Tool was found to be quick and easy to use. In both trials the search results and visualisation of the search results were available within a few minutes and the export of the graph data only took about one minute. Overall, the actual use of the Tool took just over an hour of staff time, for each of the trials. Without the Tool, and to use similar withdrawal criteria as above, the online Copac database would have to have been searched for each of the 6,836 items initially submitted. Based on previous experience of an average of 3 minutes a search, this would have taken about 342 hours just to do the search.”

**WHERE: THE ARCHIVES HUB**

The Archives Hub [http://archiveshub.ac.uk](http://archiveshub.ac.uk) is also a shared catalogue, for descriptions of archive collections (or ‘finding aids’). Contributions to the Archives Hub come from a variety of repositories across the UK, from small specialist collections to huge research institutions. For some of them, the Archives Hub is the only online window into their collections; for others, contributing to the Hub is a single component of a wider online presence.

**How: EAD**

The Archives Hub uses EAD (Encoded Archival Description) an XML format for encoding archival finding aids [http://www.loc.gov/ead/index.html](http://www.loc.gov/ead/index.html). EAD is ISAD(G) compliant; that is, it is compatible with the General International Standard for Archival Description. EAD provides a software- and hardware-independent way of storing data, so it is interoperable, and not tied to any particular system. Interoperable does not, however, mean that the data that comes to the Archives Hub is all consistent.
EAD comes to the Archives Hub from a variety of sources. Most commonly, contributors create descriptions (either transcripts of existing finding aids or original cataloguing) in an online tool we provide, the EAD Editor.

Figure 9: example of the EAD Editor interface

```xml
<ead id="g03184sk/jr">
  <archdesc level="fonds">
    <did>
      <unitid countrycode="GB" repositorycode="3184" identifier="SK/JK">g03184sk/jr</unitid>
    </did>
    <physdesc>
      <extent>853 linear metres.</extent>
    </physdesc>
    <repository>University of Liverpool Archives</repository>
    <langmaterial>
      <language languagecode="eng">English</language>
    </langmaterial>
  </archdesc>
  <creation>
    Created by Jane Ronson using the Archives Hub EAD Editor
    <date>2014-09-22</date>
  </creation>
  <profiledesc>
    <titlestmt>
      <title proper>Personal papers of Stanley Kubrick</title proper>
    </titlestmt>
  </profiledesc>
</ead>
```

Figure 10: example of EAD exported from the EAD Editor
We also receive EAD that has been exported from archival management systems as EAD, or as another form of XML which has then been converted to EAD.

EAD, like many XML standards, allows for slightly differing implementations, with optional attributes and elements. Combined with differences in cataloguing practices and the unique nature of archival collections, this means that there is a great deal of variation across descriptions sent to the Archives Hub. This variation is dealt with in three main ways:

1. Processing scripts. These allow us to programatically make certain changes to standardise the data, and produce the particular form of EAD we require for the Archives Hub. These changes are never to the actual descriptions, but only to the structure of EAD - for instance, we add attributes to a particular field which allow us to generate a link on the Archives Hub display to email the repository. Scripts can only be applied in certain circumstances, where we can see patterns that provide a hook for us to write a transformation script, and the required outputs must be consistent.

2. Manual edits. Where the change that we require is in the content of the description rather than the structure of the EAD, we ask the contributor to edit the content (or for permission for us to make the edit on their behalf). This is usually used for descriptions that do not meet our requirements for mandatory fields, those fields that we consider absolutely necessary for a successful online interaction with a finding aid [http://archiveshub.ac.uk/mandatoryfields/]

3. No action. Where discrepancies do not affect the validity of the EAD document, the Archives Hub functionality, or the usability of the finding aid for Archives Hub users, we do not take any action to change them. Standardising data to a valid and usable level already consumes a significant amount of time and resources, and we do not persist when there is no direct immediate material benefit. While we consider data standards and consistency to be highly important, our priority must be to provide a good, usable service.

Why: Contribute to the Archives Hub?

As with Copac, exposure of collections is considered an important benefit from sharing data through the Archives Hub, as are the impact on the UK research landscape, and contributing to the organisation’s mission, and the wider archival and cultural heritage landscape.

"[Contributing to the Archives Hub] increases the number of collections catalogued online, leading to more users, leading to more impressive stats for fundraising. On this basis, Archives Hub is highly important in the strategic development of our service."
(Museum archive)

"[The Archives Hub’s] existence means for the first time we can have an archival catalogue that exposes our collections, used in conjunction with the digital archive on the repository, means for the first time internationally important collections will be discoverable and accessible." (Higher Education archive)
"Good to be able to show it to management/stakeholders as proof of work done. It looks professional and the exciting projects done by Hub (like sharing data with Archives Portal Europe) sound good to management ears. And it's free." (Higher Education archive)

"I think it's good for our collections to be presented in context of other UK collections." (Higher Education archive)

"I think people who have not used archives before may be nervous and put off because they don't have the chance to familiarise themselves with the catalogue and plan their visit in advance. Being able to browse archive catalogues online is akin to going onto a shop website before entering the intimidating show room!" (Museum archive)

**WHAT: LIBRARY AND LEARNING ANALYTICS**

The Library Analytics and Metrics Project (LAMP) started as an investigatory project in 2013, exploring how to enable libraries to capitalise on the many types of data they capture in day-to-day activities, using this to support the improvement and development of new services and demonstrate value and impact in new ways across the institution. Developed in response to community demand, the LAMP aimed to bring together usage, attainment, and descriptive data from university libraries across the UK, and provide access to this through an attractive dashboard.

Since early 2015, LAMP has been integrated into Jisc’s Effective Learning Analytics project, which is tackling the challenge that: “universities and colleges don’t have enough useful data about students and how they are learning. What they have they don’t analyse and interpret. They are missing opportunities to use technology to provide feedback to students. They need to support staff who could be using analytics and a standard set of tools and technologies to monitor and intervene.” (Jisc, 2015).

**WHY: Questions and Benefits**

One of the key aspects of the LAMP pilot phase was to identify business cases, use cases, and user stories: identifying the research questions and benefits that libraries want to get out of sharing their data. The use cases to be prioritised for the prototype were:

- Merge data from multiple systems: library, student registry, IT services
- Contribute to the institutional analytics mission
- Avoid data and reporting silos i.e., data sets, spreadsheets and reports
- Compelling visualisations
- Map e-resource usage to actual users
- Key usage indicators by discipline
- Identify usage levels i.e., high or low
• Analyse library gate data
• Examine events by specific user groupings i.e., overseas, distance

From these high-level use cases, we granularised to user stories, prioritised as:

• Collect comprehensive and granular eresource data
• Merge data from the student registry
• Identify how different faculties use the library and the collection
• Identify least-used resources/where usage is less/more than expected
• Produce compelling visualisations
• Provide a dashboard with key indicators by department
• Correlate feedback (e.g. NSS), enquiries and collection strength
• Correlate reading lists with actual usage
• Examine events by specific demographics/groupings

These were then granularised further into job stories, taking the format

When…
I want to…
So I can…

We started with a completely open remit, without limiting the uses to those that could be addressed by the data.

One of the key desirables for LAMP was the ability to benchmark with other, anonymised, institutions. Libraries agreed to share valuable data around student achievement and retainment, so that they could compare their performance with other institutions; and high-level insights could be discovered about the situation across the UK higher education sector. We produced a ‘dummy’ benchmark for Phase 1, and hope to continue developing this in the future.

How and What: the Data

Six libraries initially agreed to send us data about their students and how they used the library. For each student we got:

• demographic information
  o e.g. age, country of residence,
• learning data
  o e.g. course and grade achieved
• library data
  o e.g. number of loans and visits to the library

What data exactly we received depended on what the university was collecting, in what format, and whether the library was able to extract data that was held on non-library systems and send it to us. All data was pseudonymised, standardised, and small data sets removed, meaning that no-one except an authorised representative of the contributing institution would be able to identify an individual. Future work will take this a step further, as outlined in the draft Code of Practice for Learning
Analytics (Sclater, 2015), which covers responsibility, transparency and consent, privacy, validity, access, action, minimising adverse impacts, and stewardship of data.

The LAMP data was extremely variable. While the data that comes in to Copac and the Archives Hub does vary in quality and format, we get certain core fields to certain standards - and we know what we’re looking for, and what we can do with it. The data is transformable, interoperable, and reusable.

The data for LAMP so far came in in spreadsheets, with no standardisation or consistency. The data wasn’t reusable or interoperable. It took a lot of work for us to even make the data useable. This is not to fault the libraries who sent us this data. We had not given them any guidelines for the data we wanted to receive: part of what we exploring was simply what it was possible for libraries to send us. The systems that they were exporting from were not, in the main, designed to export data in a format that would be interoperable and able to be combined with data from other systems. It is an excellent practice to consider interoperability and future-proofing when purchasing or designing a system, but we realise that constraints mean this is not always possible, and that many legacy systems simply do not have this capability.

![Figure 11 - example of varying data for a single LAMP field](https://library3.hud.ac.uk/blogs/lidp/)

LAMP was heavily influenced by the Library Impact Data Project [https://library3.hud.ac.uk/blogs/lidp/], which discovered a statistically significant correlation (note: not causation) between library usage and student attainment (Stone and Ramsden, 2013). Their data processing had been an entirely manual process. We did a lot of manual mapping and standardisation for LAMP, but with the increased scale of LAMP, and the resources we had available, this would not be sustainable.
As LAMP is now part of the Effective Learning Analytics project, the data will be being harvested by the analytics solution provider, and stored in one big data warehouse, along with other Learning Analytics data. A standard format for export from the data warehouse has been agreed: xAPI (or the experience API, or the Tin Can API), which uses subject-action-object triples, e.g. ‘Bethan attended the SLA Conference’. Data will be exported to the data warehouse from different systems in different formats, but it should come out clean, consistent, and reuseable.

**What and How: Next Steps**

If you are interested in data sharing, here are some questions that can help you get moving:

**What:**
- business needs do you have that you think could be answered with your data?
- benefits will you get from conjunction with other people’s data?
- format is your data currently in?
- format do you need your data to be in?
- legal/moral/ethical implications are there of sharing?

**Why:**
- do you want to share data?
- do you want to share this particular data with these particular people?

**When:**
- does this data need to be available by?
- is the right time to start sharing?
- will your data be ready for sharing?

**How:**
- can you get your data out of your systems?
- can you prepare your data for sharing?

**Where:**
- is your data?
- will your data be available once shared?

**Who:**
- can currently access your data?
- owns the rights to your data?
- can give permission for sharing?
- do you want to share your data with?
- has the expertise to work with your data?
- has done similar work?
- can help?

Answers to all of these questions should help you to understand the benefits and processes, and provide you with the framework for a business case about data sharing.
Endnotes


