SITS[3]

The Scholarly Infrastructure Technical Summit

Meeting 3 @ OAI7 in Geneva (June 2011)

Following the two very successful Scholarly Infrastructure Technical Summit meetings in London and California, the third installment changed attendee backgrounds again by being positioned alongside the Open Archives Initiative (OAI) meeting taking place in Geneva over a few days in June 2011. This change of scenery meant that the meeting was a little light on heavily technical people and leaned towards advocates and political motivators looking to improve the Open Access world. This said, the overall involvement of the attendees in both institutional, national and international projects was very impressive, giving a global outlook perspective to the meeting.

Due to the schedule of OAI7 it was decided to split the SITs meeting over 2 days (afternoon and following morning). The morning session was also held at a local restaurant which overlooked the Rhône river in the center of Geneva. Due to the setup here the meeting was able to carry on pretty much as it started, if a little rushed towards the end (the restaurant wanted to set up for lunch).

As usual, proceedings started with a recap of the nominated subjects of previous meetings before attendees were invited to each pitch one or more ideas for nomination, conforming to the Open Agenda ethos. The topics which were raised and discussed during the meeting were:

Executive Summary

Due to the style of the conference this SITs meeting was co-located with, conversation at this particular meeting was mostly high level surrounding topics of concern for many of the delegates present. It is therefor possible to divide the subjects talked about up into 2 categories as listed below:

Areas to watch and requiring further development of ideas:

- **Researcher Identity, ORCid and disambiguation**
  There is clearly concern here over the adoption challenges of a common identification system for researchers across disciplines and the challenges faced in disambiguation.
- **Synchronisation of Resources**
  We generated a lot of use cases where synchronisation will be useful but there was not much technological knowledge of systems to enable each use (perhaps there is).
- **Preservation of Software, Sustainability of Software and Data**
  Although there are active projects surrounding software preservation. Scholarly infrastructure services don't yet have a strong enough requirement to start implementing or concerning themselves with these systems.

Active Research

- **Assessment of Research Impacts & Outputs**
  This is a very active area of discussion in the whole community. The altmetrics movement is gaining backing and while institutions may not be able to help with the metrics, they may need to consider the types of data that will be used here. Further, on a political note, institutions will need to start considering what this means to their assessment measures.
- **Better Harvesting of Better Usage Data**
  Follow on from the altmetrics discussion, there is some concern over the quality and comparability of current webometrics (such as download counts) which, so far, have not for assessment. Some work is needed here it was felt.
- **Linking Data to Publications for Discourse & Linked Data**
  Much of the conversation here focused on how to gather and expose data, there are many existing examples of data centric services from which lessons may be learnt.
- **Nano-Publications**
  A developing area. Delegates agreed that it may be a good idea to call for a special issue of a journal containing nano-publications and host a workshop around this.

Researcher Identity, ORCid and Disambiguation

A feeble attempt was made to not make this conversation about ORCid, but as it turned out ORCid is the central worry in this conversation. It was generally felt that ORCid was expected to be the project which is going to solve all the problems others have been fighting with for years. There are also a number of parallel efforts in this area (across many disciplines) and the impact of ORCid and rate of adoption may be affected by this and other factors. Being the OAI community, there is a lot of understanding about how long such technologies take to be adopted, there may need to be some sort of driver behind the adoption of such a service... providing it is suitable for use. With no-one highly knowledgeable on the progress of the ORCid project, it was hard to draw a solid conclusion from this. What was clear is that there is some caution around starting projects of a similar nature, even related to disambiguation, which ORCid isn't expected to solve on its own. There was wide agreement that ORCid is likely to succeed due to its backers.

What is of concern is how researchers, or institutions, claim their own or related research and connect it to the correct ORCid's. It was pointed out that ORCid is building a system which allows both of these methods of claiming research but the costs of doing so are unclear. Further to just claiming research was discussion on how an ORCid could be augmented with accurate information pertaining to the researcher, the general agreement was that the institution is likely to have the most up to data information pertaining to each of its staff members. This asks questions about collaborative updating of records and what happens when a researcher moves institution.

It was pointed out that Australia already has a bibliographic catalogue (of book authors) and there is a current funded project which aims to allow institutions to freely populate a national index with whatever data they want. Currently this works by each institution running some disambiguation processes and sending the results to a national service. The return of this process is a list of potential author IDs related to this data from which the institution selects the correct IDs and stores them alongside the existing record locally. At the highest level this is a binding for the home institution and re-usable for everyone else, simple but effective. The key benefit is that the institution does not have to change their internal identifier scheme. In the UK something similar is already practised with the identifiers produced by Thompson and Scopus and ORCid is likely to become yet another provider for this type of data.

It was clear from conversation that there is some concern over publicly revealing an identifier which can be used to obtain too much information relating to the researcher.

Further, what happens when one of the identity providers disappears. The approach in Australia was to focus on the national library (who already provide this service for book authors) on providing identifiers for researchers as well.

To drive the institutions to populate the registry, this data is then linked together with the CRIS systems which enables an entire CV of research to be constructed. A process known as building bridges between data, a lightweight approach which I personally quiet like.

Again there is some question of why we would need national or international IDs if institutions are driven to mint and maintain (my new favorite phrase for linked data URIs "mint and maintain") IDs themselves. Finding the identifier and having the rational is the most important part of adoption of service. Trust is the last important aspect, what people trust... may we get used.

Even with IDs created, disambiguating existing research may be difficult and it unclear what work needs to done here currently and who is responsible for doing it.

Lastly there is a clear need to have IRIs to represent researchers so they can be used in the linked data world.

Definitely a hot topic which might in the coming years see a lot of work and a potential solution...we hope.

Assessment of Research Impacts & Outputs

A couple of years ago this topic had very few active researchers, however now seems to be gaining traction somewhat. Also conveniently this was also my PhD area :)

This area, known as altmetrics "alternative metrics for the assessment of research" observes the increased rate of research as well as the changing nature of research, realising that current techniques are not suitable for its analysis.

Firstly the altmetrics website is well worth a read, along with the proceedings of the first altmetrics workshop, held in Germany at the ACM Web Science Conference 2011 the week before this SITS meeting.

It is clear that there is a change is culture required is we are to take into account more outputs than just the final publication when measuring the quality of research. This was also something bought up during the OpenScience session at OAI7 itself. Here it was asked how science can become more open, with datasets, code and processes being shared as well as the publications. It was concluded that in order to encourage the controlled release of these types of outputs by researchers requires a change in the way in which researchers are assessed, thus providing a drive to release this type of data. It is this point you get into a big of a Chicken-and-Egg situation where altmetrics for this requires more data, and more data requires altmetrics...

Altmetrics have to also be very carefully considered, they should be designed to measure and direct research, not to cause major collateral damage along the way and corrupt themselves. This was particularly the case in an experiment where all journals were rated, resulting in a local increase in the number of publications in journals in the highest rated category... this is a consequence which should be expected.

In the UK, it was proposed that the next REF exercise be based entirely upon mathematical bibliometric techniques. Due to the major differences in publication techniques between the various disciplines, it was decided that this was not an appropriate technique to use on its own. What is clear is that impact is too laggy, publications can't be accurately judged until they are at least 3 years old.

Interestingly there is some feeling that Australia and the UK get blinded by these assessment exercises. There is some question as to weather this is the researchers or assessors fault (e.g. are the assessors just trying to process lots of data and the researchers trying to give them the idea result).
Services like OAI-PMH provide a good basis on which to begin to gather data in order to

produce a set of altmetrics however there is concern (from the USA) on how to get more content into the repository. Typically in the UK and AUS, there has been an increase in content ingested into repositories surrounding the REF exercices, so there is a very positive side to them here :P

Confusingly there seems to be the observed gap between the metrics used for assessment and those people would actually like.

- Who is the most productive?
- Which group is the most productive?
- Who is struggling?
- Where can I help?

These are all questions which data formats like CERIF aren't directly helping to solve and perhaps a greater focus needs to be applied here.

At this point the conversation took an interesting turn. From focusing on repositories, the mention of OAI-PMH started people thinking about other sources for data. We have seen systems which harvest from Scopus, WoS and Google Scholar in order to populate a repository. Can these be used to help gather more data for altmetrics?

Further, if we require expanded citation indexes (beyond publications) are there any tools which can extract citations from various document formats. It was mentioned that the freecite project might be able to help here.

In terms of political motivation, having a Research Manager work with a Repository Manager can help bring together knowledge from the CRIS system and presentable evidence from repository.

Summary

- The practical implementations can and do change researcher practise.
- Assessment can majorly change behaviour.
- What is assessed needs to be expanded.
- Assessment needs to be more recent.
- Libraries can provide unique value via data is already there.
- Building the system is not the problem, data quality is important when tieing stuff together.
- The need for an open source tool to splice citations, make ONE awesome!
- Change in conditions of Scopus and WoS allow us to republish data obtained via the API

Future Event: Citation Summit - Oxford (UK) on September 3rd and 4th 2011.

Linking Data to Publications for Discourse

Following on from the previous topic, this seems like the most logical place to go. In order to access more aspects of research requires the creation of more first class objects. The complication comes where there exist a series of complex and often free form relations between these first class objects.

e.g. One publication can be related to many datasets, one dataset can be responsible for many publications.

Although this sounds like a database problem, there are many more scenarios you can imagine without even needing to talk to researchers to get their view.

One important realisation is that the metadata defined by the domain will be the most useful when it comes to researchers re-using the resources being shared. There is little point trying to limit the model.

It was observed that systems such as the Human Gnome Database is already viewed as an extremely valuable dataset already, can we replicate this success factor in other diciplines?

The key will come when we have a mechanism for giving credit for publishing data.

Although we talked about altmetrics being a driver for data release, there are many areas which aren't funded and examined in this way. So there needs to be a solution which encourages (makes it easy) for everyone to publish data. What are the other value adds?

People have started giving data to Google and MS Azure in order to utilise the capabilities provided by these services. Keep the barrier of entry low and a capability list high. This is a concept backed up by the Opendap initiative for marine data, Pangaea (earth and environmental science) and also ADEA for dental data. Interestingly all of these examples are in areas where the main form of communication is not scholarly journals. Humanities is another such area, where monographs are more common and there exists orphaned tools and data. Is there a potential here to examine one of these areas for ideas around how altmetrics can apply to non-publication data?

As for collecting all this data, there is no problem having 100s of disparate places where it is put as long as you can create new objects by linking between existing ones. In terms of data capture, systems engineers should "**make the right thing easy**". **Many scientific fields have metadata in the workflow, we should be looking to capture it**.

Preservation of Software, Sustainability of Software and Data

What struck me here was the discussion around how carefully software development is done, e.g. using a version control system and hopefully good documentation. However the sustainability of the final product is still lacking support and good practice. With the amount of data being produced increasing there is a need to preserve the tools which can interpret/process this data... or is there?

Interestingly CERN (our local large data creator) see the data as the primary resource, they then try and tie version control together with data. Software which can use the data is a secondary concern along with hard written notes. It is interesting that they view some (non-essential) software and data processors on the same level as written notes. They are more interested in preserving the algorithms and resulting scientific theories rather than the exact process used to obtain this result. If you wish to preserve the data (which seems to be the next stage) then some work may be required to translate proprietary data into open data.

In terms of preserving software, it was bought to the floor that a much more short term problem exists. Here projects, or degree modules, which produce software don't have suitable methods for preserving this software in the short term such that it can be picked up and extended by subsequent projects. It was mentioned that one institution now provides all 3rd and 4th year project students with a virtual machine on which to carry out their coursework dissertation implementation. Those deemed valuable are then snap-shotted and preserved ready for future reference.

Obviously, commodification of preservation as a science would help solve the problem but a lot more research on different solutions and standards is required to get to this point. In the different areas it is the domain researchers which will define the requirements and currently a lot seem to be very short term, there is a general understanding that in the future everything will be better and faster and the software will be superceeded. Preservation of software is not seen as valuable enough in the area, people want to re-process data for new results, not re-run existing work.

Preservation of data standards can be thought of differently if you think that no one asks about image formats, there are many of them but they are understandable. Can the same level of understanding be applied to spreadsheets and the cells in spreadsheets.

Software preservation is a new research area with a couple of European projects investigating the problem and the possibilities. I think there is some concern over the applicability of the area and how much effort should be put into long term vs. short term software preservation.

Better Harvesting of Better Usage Data

Normalisation Standard Practise for Processing
Re-Downloads Comparable Metrics

This area really leads on from altmetrics. Many people and organisations are already process and revealing web based metrics such as download statistics. There is some concern about how this is done however. There seems to be no industry standard way to normalise the data collection (e.g. removing bots and re-downloads) and then processing this data so it can be compared with other similar services.

There is some question as to the continued value of anonymous statistics. With people now carrying and using so many different devices, there is concern that the only way to track re-downloads is to force the user to go through some sort of identification process. Question is will this block current processes, it is certainly another stage and will an opt in system be suitable?

If you are to force users to identify themselves, then there needs to be some form of “reward scheme” for this. The example of mendelay, which finds similar publications, was raised and maybe a similar thing can be done once enough data and links are available.

The idea of “social searching” was raised and since the meeting I am now interested in what google+ are doing around this area and if repository can track the google cookie and ask users for permission to added/retrieve data from their google+ “cloud”.

Summary

- In the statistics being gathered, area people using the same processors and measures?
- There is some existing work to look at, popirus2 and counter
- Better expectation and information about services is key to users understanding their privacy.
- There is a key difference between anonymity and privacy.
- Incentivisation is key to the success of this area.

Synchronisation of Resources

Distributed Enrichment and synchronisation of the resulting resources is the key use case

During OAI7 itself there had been chat about extending OAI-PMH to support synchronisation of metadata records, but the main point here seemed to be that there was no way to notify a system about a record deletion operation which has taken place. This was pitched as OAI-PMH v1.X.

Although proposals were pitched about developing OAI-PMH v2.0 with synchronisation being the main focus, there was some apprehension about OAI-PMH being the best place to start here.

With the capabilities being provided by linked data it is becoming unclear when you do and don't need synchronisation and the following use cases were constructed:

- Search engines being up to date with page content.
- Other indexes (specifically linked data stores) which need a complete, and local, copy in order to perform complex operation.
- Centralised backup/copy (or partial copy) of (volatile) resources
- Local updating of remotely controlled vocabularies
- Distributed enrichment of single resources. (e.g. lots of people analysing one CERN result)

Many of these use cases could also potentially utilize many different techniques in their solutions, from simple http HEAD operations, through OAI-PMH delete (atom pub delete maybe), to more complex custom systems.

There was some interesting discussion surrounding the perceived operations which can be carried out on different sizes of objects. Perhaps separating the small, medium and large objects will allow people to understand the different capabilities and enable current policy to retained for objects of sizes we know how to handle.

An interesting question was ask about whether we are moving away from a collection based world. For the sake of the human brain, and the way it partitions things I hope not here.

There is certainly use cases for synchronisation, perhaps though the demand is not that high...

Linked Data - Applied

Perhaps one of the subjects with the largest amount of currently relevant work, it was surprising that this did not come up earlier in the conversation. Or parhaps everyone is happy with the progress made and is using linked data successfully. In order to make the conversation a bit lighter, this topic was introduced specifically with the **applied** word in mind.

Many parties are now contributing and consuming linked data, a couple which come to mind at Southampton are http://data.southampton.ac.uk/ which features the open data map application (click a bus stop for coolness).

In the repository world many are now contributing linked data, from the repositories theselves to the standards based around them. The CERIF data model is heading towards a technique by which more data can be exposed, not as linked data yet, but it will enable distributed processing. Other systems (including VIVO?) are also making various objects addressable in many disparate areas.

The key to the whole linked data movement is the 5th star (see below) and very people have yet empowered the linked data web with links to and from external resources to their own systems.

★	Available on the web (whatever format), but with an open licence
★★	Available as machine-readable structured data (e.g. excel instead of image scan)
★★★	as (2) plus non-proprietary format (e.g. CSV instead of excel)
★★★★	Use open standards from W3C (RDF and SPARQL) to identify things, so that people can point at your stuff
★★★★★	All the above, plus: Link your data to other people's data to provide context

Source: http://www.w3.org/DesignIssues/LinkedData.html

In the scholarly community there are world there are also more standards than just CERIF in existinace and UKOLN is looking at the relationship between CERIF and RIF-CS (Registry Information Format - Collection Service). There is also ISO 2146:2010 to consider which provides a high level information model and means of representing parties, organisations and activities for an information which can possibly be mapped to CERIF.

The mapping between these ontology's is seen as a necessary stage to linking data together. Systems such as VIVO (and the VITRO tool) is enabling the building of a series of open data compatible endpoint for various first class objects. Using this system it is also possible to combine many ontologies to make a more general ontology to provide interoperability across a number of systems. Drawing on other external systems is still hard however and the amount of normalisation required to get things into a Vivo ontology is quiet a high.

Interestly the Australia National Data Service (ANDS) have a VITRO for RIF-CS which might also be worth looking at as a reference when comparing this project to CERIF and other standards.
VIVO as a system still has many limitations, it does not de-duplicate information across domains, making collaboration discovery difficult.

More lightweight approaches to connecting objects (by linking data) are certainly having some traction, such as that used by Mendelay to connect researchers and publications. It was mentioned that sameAs.org, which is meant to be a further simple service can be confusing when considering temporal issues surrounding an object. I still believe sameAs.org could be an extremely effective solution if used in a lightweight (but accurate) manner.

More data will provide more opportunities, but more lightweight approaches to connecting and collecting this data together are required. How about a facebook group for the funders of research where they gather "stuff". It turned out that we are not really bothered about how the data is represented, as long as we know it is available and re-usable in some form.

Question is what can we do now?

"Go forth and make links!" (Wendy Hall)

There were several other projects and activities listed during this discussion which I list here:
CERIF, ACCPF, RIF-CS, Repository Data, CRIS data (CRIS pool), RIM-Info

Digital Preservation Architectures in the Real World

Q: Any else relying on central IT for their most critical preservation role... bit stream storage?

If your answer is yes to this then you are not alone and there are further questions to ask...

Q: Are you 100% certain that their backup solution is full proof and that the backup tapes/disks have the correct and up to date copy of the data on them?
Q: Are you able to fully restore from these backups?

As we start to collect more and more data, it is clear that the successful validation for the purposes of answering these questions is also getting harder (even more latency is introduced). There are a great many practical matters to consider when managing a tape based preservation environment. There also seems to be a large disconnect between a digital repository or CMS and the underlying environment, these systems are simple not aware of this infrastructure and thus are not able to manage and report upon it.

What is hoped is that whoever provisioned the original hardware has a plan to migrate to newer storage platforms in the future with minimal disruption.

There are 2 main issues here, one is of cost while the other is around perception on how much data an institution can capably handle. It is suspected that as with cost, each institutions panic point when it comes to the amount of data will vary. It was suggested that it would be nice to know what these figures are, what constitutes a large dataset to an institution? What are the differences in these bodies?

Even those with systems as policies in place currently are not 100% confident these are both correct.

There was a call here to continue the Preservation, Archiving Special Interest Group (PASIG) which focuses on these issues. Perhaps this point could become a key one for the meeting.

Nano-Publications

Due to the fact there is no wikipedia page for nano-publications, there we defined as follows:

> "Nano Publications simply tell you what is being asserted. (e.g. Europe is a continent)"

Related Reading: The anatomy of a nano-publication : P Groth et al. - Information Services and Use, 2010 - IOS Press

Nano publications are seen as a possible approach which generalises well across disciplines. Again the problem is how to get authors to think about nano-publications. Once again there is possible work to be done both with tools and motivation of the community.

There is already a problem with people producing well behaved documents, e.g. a navigable contents, author data and key words for searching embedded nicely etc... If people did start to care about the quality of the documents they produce would this help in creating new nano-publications.

Microsoft are working on a series of Word Add-ins and among this is one which can heal create a nano-publication. It does this by informing the user on the way Word interprets a sentence and then attempts to form these concepts and statements into a nano-publication.

Many other services have been developed to identify people, places, times and other objects but the main problem lies in inferring the predicates related to these first class concepts (although the field of sentiment analysis is getting better). In focused disciples this might be easier, but the generalisation is likely to be difficult.

Certainly another area to keep an eye on for future development, particular around the area of citations where is great interest in establishing simply if the citation is a positive or negative one.

For general amusement, here are some nano-publication type statements:

- ^* causes Cancer
- Nano-Publication DoesNot Work

On a more serious note there was general agreement that it might be a nice idea to publish a special issue of a journal involving non-publication, potentially alongside a workshop on the same topic.