SITS[4]

The Scholarly Infrastructure Technical Summit

Meeting 4 @ eResearch Australasia in Melbourne (November 2011)

Iteration four of the International Scholarly Infrastructure Technical Summit meeting following events in London and California and Geneva took place in Australia alongside the 2011 eResearch Conference (eXtreme eResearch).

The eResearch conference brings together CTOs and CIOs from various Universities whose job it is to provide, work with and support IT services for scientific research projects. This was a first for SITS, which previously had seen a mix of researchers and those working on the fringes of institutional IT infrastructure. The Australian focus on supporting research directly thus provided the 4th new platform for the SITS meetings, one closest to the infrastructure itself.

The meeting was kindly hosted at the conference venue and sponsors by the Australian National Data Service (ANDS) who also invited many key people to the meeting. This resulted in a rich set of topics being discussed, many of which are familiar to those following previous meetings but clearly with different issues and levels of concern expressed towards each.

This meeting was a one day meeting (on a Friday) and as usual, proceedings started with a introduction of attendees and a recap of a number of subjects from previous meetings. Attendees were then 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 (excluding the executive summary):

Executive Summary

With a high number of technical CIOs and CTOs present, it was accepted that technology is not the problem, things are achievable. General consensus in the room was that the biggest problem is providing technologies which are (1) Accessible, (2) Understandable (3) Usable and (4) Actually Used! Gathering, collating, carefully describing and then making project outputs (including data) available, does not result in people actually finding and then using those resources. Data citation is seen as a good forerunner in this area, as citations establish some level of trust (via proxy).

Methodologies and working practises became the logical place to start addressing some of these problems, however there seem to be other problems here as well. Agile methodologies are more successful in research, as they develop better communication channels, however the resulting products are often less maintainable as a result and funder's are not keen on such less well defined methods.

In terms of making outputs re-usable, the conversation here asked whether better information provenance would help. Re-usability seems to be an issue of trust. If a researcher does not trust information, and data can be re-collected, then they probably will. People networks are also important here as technology and provenance information may still not solve a problem which could simply be solved if the two people (producer and consumer) knew each other and trusted each other at that level.

Following on from this the topics shifted to large (cross organisation) project teams and their worth, are they too risky? An interesting discussion when considering the importance of connecting groups of people. The main conclusion of this conversation was that large projects should exist, however these should involve more stable and understood technologies and less random variables.
The conversation on building communities was continued when discussing the benefit of bringing developers together, something not done in Australia. Previously in the week the first Dev8D Australia (based upon the highly successful Dev8D UK) took place and bought together a number of developers who found the event useful as a platform to meet, learn and share knowledge with others in the community.

So with a lot of technical minded people, the main topic of the meeting focused on making technology, software, services and data usable and trusted. This seems to be the biggest problem facing data and the future of eResearch projects.

CERIF & RIF-CS

Following on directly from the "Linked Data - Applied" topic of the previous meeting in Geneva, it is clear there is a global concern over representation formats designed to describe items and relations between them. CERIF (in Europe) is being adopted as the standard for describing organisation, projects and people while RIF-CS focuses on collections and data. It is becoming clear that these two do not relate, thus could both be used... question is... for what?

It was commented that these standards have clearly been defined such that things can be described and related together however their use-fullness was called into question. Have they been engineered beyond dublin-core to do the same job in the most complicated way possible?

Neither RIF-CS or CERIF are going to let scientists delve into lots of relevant datasets to answer questions thus it is unfair to expect them to provide their data in such forms. The information both demand is valuable however a piece of the puzzle seems to be missing between these standards and the scientists. It was outlined that RIF-CS does not detail how to add community extensions, something which may or may not help discover this missing piece.

There is potential for RIF-CS to act as a training standard as it does make people think about what a collection is and what the metadata should be. A central body should not define the metadata as this will likely make no sense to those searching for the data. Semantic and free form data might be discoverable however this does not make it understandable (your ontology is only understood by 5 people). Discoverability and Understandability are two key problems which are not addressed by projects which get bogged down in inventing new models, something which could be changed?

As data is moved further from the source, its ability to be re-used decreases. Normalising data and describing it according to some well defined model is seen as one such way by which it is removed from the source. Some fields of research are good at aligning and sharing data, some should do better and some should not align their data due to the moving from source problem.

Is there a piece of the puzzle missing, or is this piece a different shape depending on the research area?

Semantic Suffering

Semantic Web and Linked Data are all the rage in scientific communities at the moment, providing a way to share "raw" data in a free form manner. In supporting such technologies, some are viewing semantic databases as a replacement for a traditional RDMS, a view clearly not shared by the majority of attendees in the meeting. Current Semantic Systems are like RDMS's of the 70s, they fall over a lot and not a lot of people understand them. Additionally they often expose too much information is ways which are not easy to understand. People like well defined, understandable models and it is clear that exposing a SPARQL endpoint is not classified as exposing a good API.

It was discussed how the majority of data services (including data.gov.uk) all store and manage data in RDMS or other existing systems. Triples are an exchange format and method to index all this data in a standard, sharable way. The key here is not the understandability of the data, but of the exchange format itself.

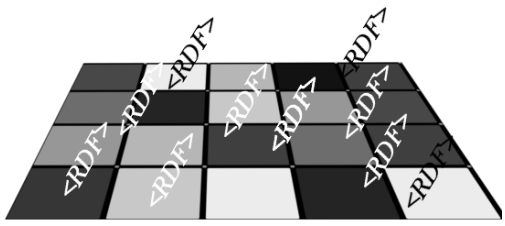
Semantic services and the advanced stuff such as inference, rules and use of ontologies is still being researched in the area of computer science (as it should be).

Linked Data is seeing a lot of traction and we see this happening with the work at doi.org on referencing data and using content negotiation*, also something happening in repository software's (EPrints). Semantic Services are not as clearly defined as the Linked Data standards, however are shared semantic services the way forward or should inference and rules around this be domain specific?

If sustainability is key to future success, then are these technologies ready. There is no denying they will mature and be the next big thing, however what is the path between here and there.

On the topic of ontology's, it was asked where the VIVO ontology was and who was responsible for accepting changes (people issues). As it happened, all the right people were at the meeting who could answer these questions and actions were noted that VIVO would help support the continued development and sustainability of their ontology's and semantic services.

* The doi.org URI is used to reference both the metadata about the data and data itself, what happens if your data IS rdf, how do you get the dataset and how do you get the rdf metadata record about the dataset?



Agile Methodologies

Due to the delay in project funding (from point of specification) following a waterfall model according to this initial specification will often result in a product which doesn't actually represent what was intended. This is either due to the initial specification being wrong or needing to change due to the elapsed time. Additionally due to academics operating a lot of projects, the timing does not suit a dedicated effort. There also exists a great amount of "It will be alright on the night" which is not supported by a waterfall project or development model.

One of the key areas of concern focused on documentation and the demands of each stakeholder involved in the project. It was realised that project management and funder's require documentation which can hinder the project actually succeeding. Conversely documentation for an agile project is often suitable, but incomplete... is complete documentation ACTUALLY necessary? A suggested way to make a project succeed and to keep the right people happy was to lie (which is why this meeting conforms to Chatham House rules).

Many of the meeting attendees also had different definitions of "agile" and it was agreed that any method can work if the correct people skills are utilised most effectively. Additionally it was pointed out that unless the researcher is involved in building the infrastructure, there is no point continuing as each product needs an owner.

Documentation can be just a proxy to close engagement, hence the question about the necessity for it all to be produced. While document templates are good, there should exist more flexibility in what is deemed appropriate documentation to accompany each project, possibly making the entire process more streamlined.

The idea of clandestine development was mentioned, where any compelling prototypes get commissioned. This is a potential way to change the project bidding process which may raise the success rate of projects.

Also it was noted that perhaps projects should be split into parts, with management and paperwork only involved during the set up phase. Project continuation (with more money) is then only granted for the purposes of documenting projects which have produced something solid or need money for the change management process in order to achieve first full deployment.

It was concluded that funding is not the problem, engaging and managing risk is. Flexability can make a project a success. Perception is the problem.

APIs and PUSH-PULL

At the center of most systems sits some kind of repository that traditionally has expected people to actually contribute and upload data manually. Protocols such as SWORD have help somewhat to reduce the deposit demand on real users and allow agent based deposit. The Tardis project has had great success with machine based deposit using techniques similar to SWORD. Similar to SWORD as HTTP is not suitable for transferring of really large files and the source cannot be changed easily. Ironically it is the long tail (smaller science) which has more support in this area as SWORD and HTTP are appropriate here.

Concern was also raised that data is transferred using non guaranteed protocols, there is an opportunity to use message queues in this situation, along with tokens to connect data. Such systems are already in widespread use within the commercial sector.

Like many sources of data, machines which produce the data are manufactured by companies who will want a customer to invest in their own proprietary technologies. This means that data endpoints will need to fit to these devices and not the other way round, so should DROWS (reverse SWORD) be looked at as a library of plug-ins which can PULL data from such devices?

The idea of creating a library or information source for DROWS clients was a good one and people would like to see clients which pull data from services such as dropbox, google docs and other sources of publications and data.

There is clearly work needed in building pipelines between the machine which produces the data and the repository, Somebody HAS to write the glue libraries and these libraries should be advertised and shared widely, something that ANDS is good at already.

Use to Scientists

We are producing exa-bytes of data, and people are hosting this data. The problem is not based around managing this amount of data (that can be solved with money), the problem is establishing trust in the data such that it can be re-used. How can the quality of data be assured such that scientists see value in the data? Currently it is often less effort to re-create data than to establish trust.

There are several aspects raised as being crucial to good data preservation and re-use enablers:

- Provenance
- Attribution
- Clear Licencing
- Perceived quality, achieved through being complete and understandable.
- Distance from origin (the closer the data is to source the better)

It was enquired as to weather the OAIS model was broken when looking at data preservation, preservation and data-reuse start on the desktop and the OAIS model may disconnect resources from their origin.

Standards are another way to communicate data quality, but very few standards or automated systems exist which analyse data quality at the point of creation. As an example, a number of years ago the unit for measuring salt levels changed and using an old data set it was discovered that the Great Barrier Reef would die within 5 years (not true if the salt reading had been in the correct units).

It was stated that data sharing data only verifies the sampling, not the science. The only way to verify the science is to sample again! This questions how much we really need data re-use if a discovery has been proven several times. It was also stated that it is often impossible to connect oneself to the original data as no-one knows the exact conditions.

Time sensitive data is obviously the exception to all this, here data cannot be re-sampled, disciplines using such data have evolved to accept historical data and process it with care and due diligence. It is clear that solutions can only be engineered to a point, the use of data will never be consistent for all disciplines and trust may never be established. It is these conditions which lead to good scientific progress and should not be stopped.

Are Large Projects (cross-org) too risky and less worthwhile

Following on from the discussion on agile methodologies, concern was raised over the worth and value for money of large cross organisation projects. It was stated that bigger projects are not able to fail quietly and are not agile, changing direction to be successful is very hard when there are so many peoples interests to account for.

The problem originates right from the bidding process. Large collaborative bids are actively encouraged as a means to get funding, however once funded, big projects have little or no idea how to actually achieve what they want to do. It might be better therefor to fund based on the best proposals, not the size and involvement and encourage organic collaboration via townhall meetings and conferences which also emphasises the importance of dissemination. If several smaller projects are funded in the same area, why not get them to swap their data half way through to force collaboration and more generic problem solving.

Another key problem with big projects is that too many variables are unknown and are based around developing or using immature software. There is a potential for big projects to work when the number of unknown variables is strictly limited and well established technologies are used at the heart of the project. Take companies like Facebook, Google and Apple, all three of which changed the world with one simple, well implemented idea.

At the end of the day Institutions want to fund their staff, however changing the model might lead to some interesting developments in the community. For example what if an institutions software was chosen by a funding body (like ANDS) to be the de-facto software for a certain purpose and because of this you got no further money to develop it, instead this money was given to the other institutions to help them deploy and use it? Are support contracts a possible way to gain this money back and extend the life of the software? (a concept similar to the EPrints model)

Obviously if a number of Institutions decide on adopting the same software, then the deployment challenges become less as there is a bigger (more local) community with the same problem. However, there are bad points to this, a lack of different implementations and distributed knowledge means that replacing any endangered or beleaguered solution is much harder.

It is clear that racing to the carrot (most users) is not something which works in Universities, however there should be more incentive for successful projects and not just big ones.

Developer Communities

Understand and developing technologies around cutting edge science is not easy, so way are developers, who gain expertise in this area, seen as people employed on short term contracts to do highly specialised work before being forgotten. This is a clear call to encourage and maintain expertise with in the academic community and prevent developers leaving for industry jobs. Currently career pathways for developers are non existent. It was mentioned at the conference in one of the keynotes that one professor employed a key developer on professorial wage due to the importance of their role.

There are still big questions surrounding who it is that needs convincing to the importance of developers at institutions. Part of the problem may be that institutions don't have an issue with training experts in any area, it's what they do, hence why turnover is so high. Interestingly, developing technical experts has been in the eResearch road-map since 2006, yet this years Dev8D conference was the first attempt at pulling the developers together.

Stability is a key issue (more than salary level) and underwriting multi-year contracts is seeing success at some institutions. "Dangling a carrot in front the the Donkey only works for a limited time, eventually the Donkey will kick you".

There is definitive agreement that specialised and highly skilled developers should not be regarded in the same category as printers and photocopiers. These developers are close to research units who provide interesting problems which keep developers interested.

A start would certainly be to establish a career pathway for developers, broadband-ding salaries would be a start.

Dev8D Australia was deemed a success and there is definitely room for it to grow into a highly valued event such as Dev8D UK.

Identities

Without fail, every meeting at some point ends up talking about this. In the case of this meeting this topic was nominated right at the end of the day in order to find out if anyone had anything new to say, or any new concerns.

Will researchers care? This main concern certainly has carried throughout the meeting, will researchers be engaged, do they need to be. It was put forward that if ORCID is successful, researchers will never know about it, it will become transparent to the process. ORCID will be used as a means for less form filling.

It was acknowledged that ORCID is very heavily focused on certain areas and does not give IDs to organisations or fictional characters. There are several other systems which are being developed, TROVE and (PARTI?), all of which will have to be worked with as well.

Conclusion was that all these systems are always 6 months away and we are going to have to work with them anyway.. shrug.