Storage and infrastructure discussion group

Discussion group participants included Kevin Ashley, DCC (facilitator), John Kaye, Jisc (notes) and research data management support staff from a number of universities.

Q: How easy it is to determine and provide capacity?

Most found it difficult estimating data volumes. Researchers often didn't know how much data they held and much was on unmanaged devices. Several had done questionnaires, but responses weren't always useful. Needs are complex and are sometimes changing due to the use of new technology e.g. the move to audiovisual content in arts and social science. Some haven't bothered to estimate and are using feedback from allocating out a moderate amount (500GB) to get a better feel for it.

- Know tip of iceberg, but lots of hidden data. You hope this is going to be small, have an idea about big stuff.
- Questionnaire - academic response: "what a stupid question" Needs are challenging but people should be able to estimate at least. Simple question, how big is the problem? Most people don't have a clue.
- DCC DAF should tell quantities and who owns it, should give sort of a clue. Sometimes lower and upper limits are useless, ranges are too wide. People don't know.
- 4PB stored on server systems, but don't know what's under the desk. That level of data needs own infrastructure.
- If people don't have enough from provision then it's a good way of gaining requirements through complaints.
- DMP entry on amounts is flaky/patchy - researchers handle on amounts is patchy. Some folks have an idea. Those who can estimate accurately don't generally have a lot of data. Longitudinal studies have a good idea.
- As yet people haven't been asked what needs to be archived and shared. What % of data generated.
- Transition in AH and SS from text to media (audio/vids) → massive increase in data use. Tools for minutes per media can estimate accurately.
- Some people can tell us can give us a lower bound and estimate on top.
- This isn't different from other uni resource allocation decisions.
- How many folks IT procurement? 3 or so. Rest of group have to deal with issues, but get other folks to procure and do deals.
- Haven't bothered to try, bought expandable storage 400TB can go to PB, both active and archiving for now. Get people using it for free and you get an idea of what people use.
- Give people 500GB you can see what people use it for/need

Q: To what extent can a single storage infrastructure provide all your RDM needs?

Consensus seemed to be that a single solution can't solve everything. Needs differ greatly between active data storage and archiving. Different tiers of storage are needed to offer greater/lesser resilience, speed of retrieval, costs etc. The more complex needs of handling sensitive data and meeting stringent standards such as ISO 27001 were covered in detail too.

- Deep archive/access/active etc. etc. can you have one solution?
- In engineering 200-300TB would do, 80% data needs archive, capture from instruments, but must not change. Needed a landscape of storage. Archive piece could be relatively cheap. Wouldn't mind delay in retrieving storage.

- Agree, once beyond a certain scale you need cheap offline storage.
- Tender for 3 tier Active, Data made available, Archive + safe offsite copy.
- Cost differences between high and low - factor of 10
- Disk means many things £1k enterprise £300 mid-tier £100 cheap £20 tape
- 3 tiers have some advantages, but it is only suitable for large scale.
- Can do it on an ad-hoc basis 1TB per PI and come and talk to us if they run out.
- Sensitive data - have to be kept in infrastructure that is seen to be separate. Health/Rights? Not separate storage, but separate systems with access controls
- Very few unis can host patient data, have to work with NHS trust. Most uni data centres don't have 27001, which is a minimum and you may well need safe haven infrastructure. Gap that could be filled with shared data centre and getting standards.
- Some unis are looking at safe haven - need to create a way for other unis to stop creating a club BRISSkit would like to make applications available over a secure network over a secure data centre, not just BRISSkit other collaborations if there was a shared infrastructure
- ESRC admin data centre - secure pods that you can bid to have as uni. Safeshare initiative. Jisc tech needs 27001 and will be putting service in place for everyone not just ESRC.
- Shared Data Centre. Kings and South London Maudsley working together to bring an N3 connection and would be available to others.
- RDS dashboard would be good to see out of a shared data centre

Q. What approaches are being taken to improve the discoverability of data and monitor access?

Several ways to improve discoverability are being pursued, including data catalogues, CRIS systems, use of the national registry, data centres, statements in research papers etc. Monitoring access is more of a challenge and responses here were more ideas than definite solutions. Statistics are being collected about downloads and page views, but it can be difficult to understand usage.

- Monitoring access - how do we have any idea about what people are doing with open data? Institutions need to be clear about what it has measured, exclude robots etc. and this will allow comparison, not perfect, but needs a common approach.
- DataCite stats, how often do dois get resolved? There is an indication of a level of interest. Is that enough esp. if people read metadata and not download? That analytic is still useful to highlight problems in description.
- Concern about keeping data for longer and retention policies? Do EPSRC set research policy, might change preservation policy around 10 years, but could still access over 10 years.
- Full text log files and ip address analysis.
- Expect the long tail to end up whether deleted. Auto delete or not.
- ESRC researchers every piece on request. How to manage paper that says contact us. Strong relation in citations where it's completely open. OK if name department and not actually a researcher.
- RDA has 2 levels here's the link and here's the people. See what the difference is.
- DataCite not indexed by google scholar. Implications for discovery. National data registry/aggregation and highly search engine friendly. RDA visibility went up through google searches. How to present basic metadata in a google friendly way.
- Pure or datacite for discovery. You want to keep track of stuff you have some sort of ownership of. Stats from national data centres. AHDS example. You need to know where your data is just in case
- Discovery service. Data centre push info back to you (via the service?)
- CRIS procurement not about RDM and there's no scope to harvest in some institutions.

Q: What shared services are desirable or feasible?

The main services people were looking for were shared data centres / repositories and discovery services. Some institutions didn't want to develop their own catalogue and would like a central service to avoid the overhead. Ways for smaller institutions to buy-in to services like figshare were desirable too.

- Data centres
- Repositories
- Shared discovery services and metadata schema (need to help people discover stuff)
- DataCite monitor usage
- Repository infrastructure - commonalities between collected data into disciplinary data centres.
- Strong evidence between strong domain specific data centres to usage. Gather evidence. 10 years+ could the archives take on long term preservation
- Institution doesn't want to have to run a catalogue at all, advantage of central service is for collaborative work across institutions. Registry alert people to new data and collaborations. Uni admins can query.
- Overhead for smaller unis and can use an instance of registry
- Guidelines for external repo software and stuff that helps you make decisions. Where should you use figshare for institutions=. Advice on what point should you buy or just use.
- UWE can't afford institutional account with figshare
- institutional figshare - you can have your own schema