# Research data management requirements interview thematic analysis report

## iridium project output

| | | | |
|---|---|---|---|
| **Project Information** | | | |
| **Project Title** | *iridium - Institutional Research Data Management at Newcastle University* | | |
| **Project Hashtag** | *#iridiummrd* | | |
| **Start Date** | 1 October 2011 | End Date | 31 March 2013 |
| **Lead Institution** | Newcastle University | | |
| **Project Director** | Janet Wheeler | | |
| **Project Manager** | Lindsay Wood | | |
| **Contact email** | lindsay.wood@ncl.ac.uk | | |
| **Project Web URL** | http://research.ncl.ac.uk/iridium/ | | |
| **Programme Name** | *JISCMRD02* | | |
| **Programme Manager** | Simon Hodson | | |
| **Document Information** | | | |
| **Author(s)** | Phil Heslop | | |
| **Project Role(s)** | Senior Computing Officer (Digital Institute) | | |
| **Date** | 05 July 2012 | Filename | *iridium*_interview_thematic_analysis_5_7_2012_v1_PH.docx |
| **URL** | http://research.ncl.ac.uk/iridium/outputs/ | | |
| **Access** | DRAFT for external review | | |
| **Document History** | | | |
| **Version** | **Date** | **Comments** | |
| v1 | 05 July 2012 | *iridium*_interview_thematic_analysis_5_7_2012_v1_PH.docx by PH | |
| | | | |

# 1. Deductive Thematic Analysis

One of the key aims for a project of this scope is to gather information about the wide range of uses of data within the university. To some extent this can be done systematically through the existing data infrastructures that already exist (such as file servers etc.). However this does not capture the whole picture, just as important are the perceptions of the users of data. How are people actually thinking about their data usage? How much does this correlate with existing services and facilities? And, most importantly, can we generate a useful set of recommendations by taking onboard insights from users?

## Data Collection - Interviews

One method of gathering user information is through semi-structured interviews. Participants are directed by the interviewer to talk about their data, whilst having the scope to expand on their answers in an unrestricted way. This produces a rich output that reflects real opinions; however the downside is that the output is also very large. There are however qualitative techniques that allow the most salient points to come to the fore – to distil this large corpus of user data into something manageable and most importantly actionable.

The interviews were conducted across a wide range of users within the university – from PhD students to Professors and Deans. In all, more than thirty interviews will take place, in this initial report we will look at ten interviews. The interviews were conducted by post graduate students and members of staff.

## Thematic Analysis – Deductive Stage

One such tool is thematic analysis, which as its name suggests is a method for identifying overarching themes from a corpus of data. This is not done in a mechanical way, but rather a researcher would examine the data and extract the themes based on their frequency but also their interestingness. Normally, the generation of themes is inductive, themes emerge from the data. However, in the case where the corpus of data is large, an initial deductive phase can occur. The deductive phase begins with predetermined themes, to be used in an initial classification. The themes should be vague at this stage, as they will be made more concrete later. For this study we chose four main themes:

- Perception
- Purpose
- People
- Process

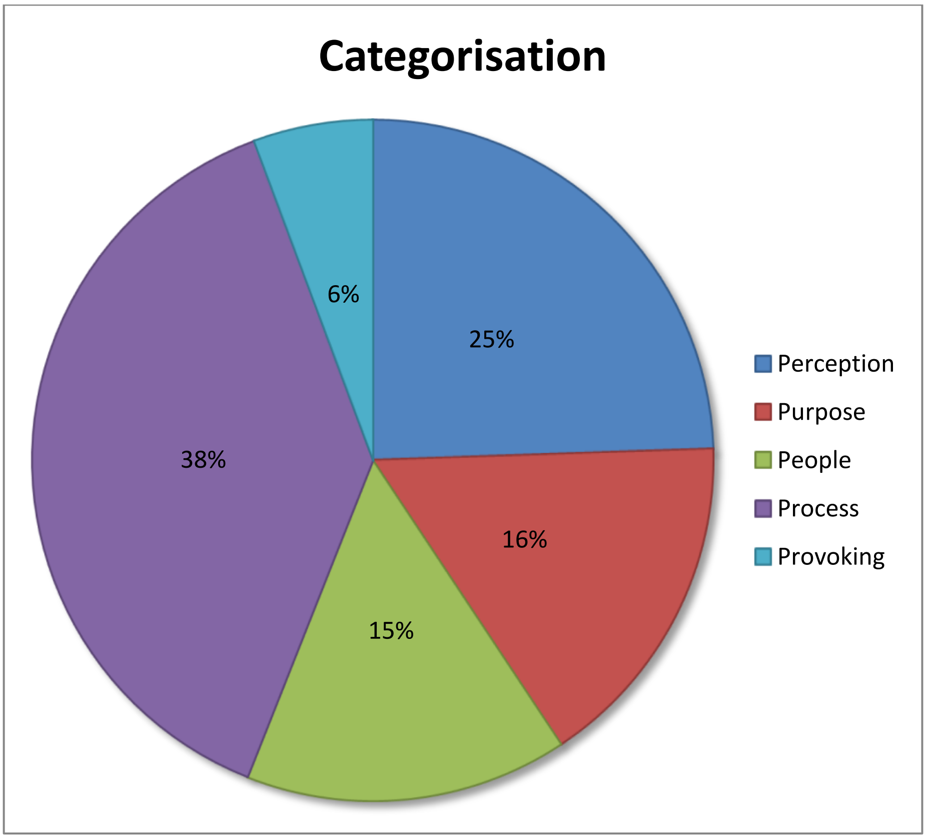
We also added an extra “catch all” theme, to grab anything interesting that didn’t quite fit.

- Provoking

As stated, the themes are deliberately vague. They were presented to the interviewers with minimal explanation, so that they could perform their own classification of their interview transcripts. At this stage, the interviewer is the key person in interpreting the interview data. The resulting classification was then collated to produce some initial statistics.

## Some Interim Results

The collation of the classification data provides an opportunity for some interim results:



## Initial Quantitative Observations

The aim of qualitative analysis is to allow the content of the data to inform the analysis, though at this stage some quantitative indicators can be observed. Simply looking at the proportion of the categorization quantitatively, the Process theme has generated about double the response of any of the other themes. This could indicate that this is the most important theme, at least in the view of the interviewees. However, until an examination of the actual content is made, this is only a hypothesis. On the other hand, the Provoking theme is small, at around 1/3 the size of the other themes. On the face of it, this is a good sign – our initial themes look to have covered a large proportion of the data, with only a few falling into this catch-all. Again however, the content will need to be examined.

## Inductive Thematic Analysis – The Next Step

The second inductive phase is directly informed by the data and involves unpicking, adding and rejecting themes. Each theme generated by the process requires stringent investigation to establish its full relationship with the data. A deep understanding of what each "new" theme actually means is required. Then each theme can be summarised as a series of recommendations to take forward into the later stages of the project.

# 2. Inductive Thematic Analysis

The Initial deductive thematic analysis phase took the interview transcripts and categorized them into deliberately vague themes; this second stage uses this output categorization as a starting point for a more inductive process, using the data to inform a more concrete classification.

## Starting Point

The process requires reading the deductive classification provided by the interviewers, and pulling out common threads and ideas that are apparent across the data, as well as exemplar quotes or ideas. The full notes for this process are provided in appendix A. Once this has been done for each of the initial categories, the results can be examined in order to provide a reclassification – one which provides a fuller picture of what the data means.

The starting point is the initial classification:

| Perception | Purpose | People | Process | Provoking |
|---|---|---|---|---|
| **…** | … | … | … | … |

Then we must examine each category in turn.

## Perception

The perception category was intended to capture users' current thinking about data management within the university. It produced a wide range of classifications from the interviewers, reflecting the wide range of experiences of the user base.

Several key themes emerged from the category:

### Usage

Data usage amongst interviewees was extremely diverse, but so too was their data management "policies". This ranged from full reliance on the "default" university services (to the point where some interviewees were not aware that anything but the standard services existed):

*"Almost nothing is actually saved on our hard-drives, whether internal or external, partly because ISS does a good job of limiting our access to these drives and machines"*

All the way to completely bypassing the university systems (where some users also showed that they were unaware of the full scope of services available):

*"The only reason that I'm able to do what I do is that we have our own computing officer so if I had to use the university system under Windows then it would be terrible. So if I was restricted to the university system, it would be unmanageable".*

There was a general sense that the services the university provided were too broad and a more tailored service would be much more usable.

## Misconceptions

Aside from the misconceptions about what services the university can offer, there were some general misconceptions about what data entails. In particular, there is a common conception that data is a final result and not the intermediate stages, or that it is numbers rather than notes or qualitative data.

*"Historians when they think of data tend to think in terms of numbers and graphs."*

## Space

Although a simplistic argument, one of the main recurring themes was the issue of space. Researchers like to keep everything, and some funders are starting to insist on this. However, despite there being "not enough space" on university systems, most interviewees were happy to maintain their long term data separately (albeit with greater risk of loss) – see Longevity:

*"So we are talking about a good amount of data, more than I get storage on my university computer!"*

However, a more pressing issue with space is when there isn't enough *working* space to complete tasks:

*"There is an arms race between the laboratory equipment which is generating data and the machines which are analysing it".*

## Backups

Users employ a large variety of backup policies, but almost all at least have one (even if it is just "let the university do it for me"):

*"...provided the university's backed everything up, so it's the university that'd be in trouble if it didn't work"*

While other users perform their own backup procedures:

*"I run a daily back up from my home computer and my data is stored in the work computer as well."*

All users regard their data as highly important, and are keen to ensure its safety.

*"Losing all my data at the moment would be catastrophe because the grant is running out and my funders are expecting a return"*

*"It would be career suicide if I lost all of my data"*

However, there is definitely scope for advice on how best to actually accomplish a safe and secure backup regimen.

## Security

Given the importance users place on their data, it is a little odd that there is much less importance placed on security of data. This may be partially down to the sharing mentality of researchers, who associate security with stealing or access rather than data being destroyed or corrupted.

*"Well they could steal my research; they'd certainly be able to do that. But there's nothing that could be damaging to the institution."*

An exception to this attitude occurs when the data is sensitive to a third party:

*"Some of my interviewees were quite keen for them to not be able to be identified"*

In general, most if not all interviewees were happy for the university to be responsible for the security, much more so than for their backups.

## Longevity

Related to space, backups and security is the issue of longevity. Users talked about their data usage much more in this context than in the previous specific contexts. It allowed users to express the full life-cycle of their data.

The general consensus is that researchers prefer to never throw anything away:

*"all of my previous research jobs as well, I have all of my PhD data, which is all still active data it all still feeds in to everything that I do."*

*"I've still got data hanging around from my PhD which were in old-fashioned tape record form".*

## Sharing

Sharing of data with external partners or the public is becoming much more common, and some funders are starting to insist data is stored in common repositories. Newcastle researchers are using shared data from other institutions:

*"We are about to pull the data from the European centres"*

However, sharing with partners or internally sometimes leads to lowest common denominator sharing policies.

*"...then forward the file to my research assistant via email, although I think we've transferred some from memory stick to memory stick directly"*

Sharing facilities are often inadequate, both in scale and security, and this prevents some users from trying:

*"If someone else in the school that had data or had a project that would link quite well with my project and we had a way of sharing then we would be happy to do that and we would share"*

## Training

There was a common thread throughout the data that training was either not available, or users were not aware of it. A majority of users would like more training opportunities on data management.

*"I don't think it was ever mentioned".*

*"I think I'd benefit from a bit of guidance as to where things can be archived, how and what is safe and legal, and ethical"*

Generally, those who have found the training on offer have found it useful:

*"for new staff, and or, new researchers is to get an idea of things like encryptions, the policies, programmes, just a bit of induction to all the university offers because I wasn't aware of encryption until I attended a training session"*

## Summary

**Perception**

- Usage
- Misconceptions
- Space
- Backups
- Security
- Longevity
- Sharing
- Training

# Purpose

The purpose category captured the users' motivation for using data, and the value they attach to their data. Already there are overlaps with the previous category, *perception*, and more general themes are beginning to emerge from the analysis.

## Data Generation

Although users talked more about their data collection requirements in the *Process* category, several users emphasised the connection between data collection and the value of the data. If the data is hard to reproduce, or takes significant time, then the value of the data is perceived as greater.

*"You could quantify it in terms of what it cost to get it"*

*"if the experiments take longer time and money then the data from it is very valuable, if the experiments can be easily and quickly repeated then we they are less valuable"*

Interviewees also noted that often data storage can be strongly tied in with how the data is acquired, for example if a specialist piece of hardware has its own integral storage.

## Value

Users were keen to distinguish between academic value and commercial value, with most classifying their work in the former category rather than the latter. Most users valued outcomes from data higher than the data itself – publishing papers or books is seen as the main outcome of research and often the data is a stepping stone.

*"I think data's just the means to an end"*

*"The end product is papers; in the current project I promised my funders at least three plus a book."*

Having said this, many researchers were quick to point out that the intrinsic value of data is changing, especially as many funders are starting to ask for data to be put into repositories or similar.

*"Putting your data in a national archive or repository?" – "funders are pushing for it and I can understand why"*

*"data is feeding in to regional records, national records that are leading towards hopefully national, policy settings"*

Many users pointed out that different stakeholders in the research might have different value judgements of the data:

*"My funders are expecting a return - think if I lost all my data now I couldn't deliver on my results for my funders, I probably wouldn't get funding from them in the future"*

Many users echoed their sentiments on backups and keeping data safe, outlined in the *Perception* category – highlighting the cost of repeating work to regenerate data.

## Longevity

Users also spoke more about the longevity of their data, again indicating that they don't want to throw anything away. They added to this emerging theme by indicating the importance of data lasting longer than a projects lifetime, and being reused as a basis for future work:

*"You need the baseline there to begin with and that's your starting point. You then build on that to create long term data sets and those long term data sets are absolutely 100% more valuable than any individual data set because you've got that change over time and you can see"*

However there is a significant minority of researchers, who, although wishing to keep the data, do not see the value of using it in future work:

*"...by and large, each thing finishes and you move onto the next Thing"*

Indeed some users perceive reusing data as damaging:

*"I get nervous about going back to data that's pretty old, because you can always say, oh that's dated, and that doesn't tell us anything new"*

## Summary

| Purpose |
|---|
| - Data Generation<br>- Value<br>- Longevity |

# People

The people category provided a straightforward categorisation of who was involved in the research. More interesting was how they collaborated with others with regard to data.

## Who Has Access to Data

Internally, users by and large had a list of people who were working on their project(s). When it comes to sharing the data with others, there were several schools of thought. Some thought publication and presentations were the main sharing medium:

*"My most recent paper is already being used by people"*

*"I mean journal access is pretty good and anything that we don't have free access to or a subscription to, you can just go to the library and request and within a few days, and vice versa for collaborators"*

Others preferred face to face communication and discussion, either with colleagues or by attending conferences:

*"I'd go and talk to that person about, much more than wanting to access to their own data"*

Many researchers expressed an interest in sharing data more publicly, and were aware that funders are moving more to an open model for data access:

*"If someone asked you for it they could have it?" – "Yes"*

*"When someone comes in and asks if you've got anything and you've got it you give it to them"*

## Responsibility

The second major topic that arose in the category was responsibility. Interviewees were fairly unanimous when it came to who owned the Intellectual Property of their research:

*"When I signed my IP form it was with the university, so I'm not aware that EPSRC has any intellectual property rights over anything that I Produce"*

*"As far as I'm aware the data belongs to the university"*

Interviewees were less unanimous when it came to who was actually responsible for the data. Many choose to use the university services provided:

*"I've got them on my H drive"*

However a significant number rely on their own facilities, either due to a perceived lack of adequate facilities provided by the university:

*"I'm a bit biased and spoilt because the computing people we have ...I don't ever have any interactions with ISS....they're very good. The reason why they're getting this 32 GB stuff is because we asked them".*

*"In terms of size, complexity, managing, etc. It's too straight-jacketed. Whereas we have two computing officers who are very helpful and who can help you solve problems"*

Or simply for convenience:

*"A lot of people have their own personal computers that they would prefer to use."*

There was a general consensus that training was desired, but that it was not (apparently) available.

## Summary

| People |
|---|
| - Who Has Access to Data?<br>- Responsibility |

# Process

The process category produced the largest amount of feedback, and interviewees catalogued their data process from creation to dissemination, and also long term and repeated usage. This category contained a lot of overlaps with previous categories, and some strong themes are emerging.

## Data Generation

The overriding theme emerging from the data generation process for interviewees is the diversity. The range of data generation methods in use is large, with some requiring nothing more than a pen and paper or recording equipment, whilst some requires highly specialised equipment, techniques and data management facilities:

*"So I will use this programme, first of all I will write a little script to run inside this programme and so I set this up on a special server machine in my department. It will then instantaneously create these massive files stored in the virtual memory. And then once it's done that it, I can use the programme to search through these files and the output is actually quite small, and then I see the results"*

Also, data is not always generated within the university; some is generated out in the field and as such has its own data management challenges:

*"...because I did a lot of research abroad where computer technology wasn't as advanced"*

The rate of generation is also highly variable, some research generates large data sets quickly, for example while running a particular experiment, whilst others generate data more sporadically, such as compiling interview data from a large user base. However, in the majority of cases the predictability of data use is seen to be difficult, with many interviewees having to change their data storage mid project.

*"It's very rare to fully know the extent of data at the beginning."*

## Analysis

Again the main theme emerging is diversity, with a large range of techniques used to analyse data. The use of specialist software and hardware is even more apparent than in then data generation phase. Some of the analysis tools are bespoke (e.g. programs specifically written for the purpose), but many are off the shelf software packages, or a combination:

*"...when it comes to analysis I use some of the stats programmes that the university have, so things like  SPSS and things like minitab, but then Primer as well, I use Primer a lot and Primer's not available through the university. And then things like RJS and stuff like that but that's all available through the university so mostly it's university systems."*

*"we also use a very specific programme called Metamorph, for looking at the .tifs, just because it allows us to change contrast, brightness, things like that"*

*"Several of our instruments have their own custom software"*

## Collaboration

Interviewees definitely see collaboration and data sharing as key elements of their data process, and again have very diverse methods for achieving this. Internally, shared space provided by the university is a common sharing strategy, or even simply reverting to email. Externally however, the university systems are rarely used as to share with collaborators, with many interviewees using partners systems or a third party storage facility:

*"We use collaborators' systems (e.g. dropbox)"*

*"We tend to use their systems"*

Interviewees who do host shared data tend to roll their own solutions, such as svn style repositories:

*"I put data into repositories to enable others to access it rather than having to maintain it."*

There are an increasing number of funders and organisations that provide centralised storage for sharing data, which many interviewees see as the future of collaborative data sharing:

*"Funders are pushing for it and I can understand why. And, obviously I'm trying to pursue it, partly because I promised it to funders, partly because I think it's a good idea"*

*"(send data) To the funders to put it on the Economic and Social Data Service"*

*"It will be archived in ERIC which is the regional record centre, and it will be archived in DASHH which is the Database for Seabed Species and Habitats - it's a marine national database."*

There is still a (rare) issue when data is simply too big to share in a convenient way:

*"A friend of mine was recently sent his data from the place in Finland where his data was kept. So he received 5 1TB hard drives in the post."*

## Storage and Longevity

Storing data was a significant issue for many interviewees, once again a diverse range of storage policies were implemented. Local storage (i.e. users own computer) was widely used, especially when speed of access was an issue:

*"When you run a programme on a computer, you don't necessarily want it to send data across the network because it's slow, so you want to have a space on that computer that you can output to"*

Many users utilised university facilities for "day to day" storage, but not for archived data (mainly due to space restrictions. Researchers don't like to throw anything away. (Some interviewees used their email system as a data store.)

*"I'd love to store all of it on the university computer, but I don't have enough space for it, by far not enough space for it"*

A few interviewees were using external storage for their data, particularly if it was to be shared (see *collaboration* above). A significant number also stored a lot of data non-digitally, either physically or on long term backups like tapes.

In general, space was a major issue with interviewees, not just for archiving but also for ongoing research, and many interviewees are having to use their own storage solutions (with storage being fairly cheap, some thought it odd that the university didn't provide this).

*"...about 7GB, that's backed up on the university servers. And for my own work about 500GB. And that isn't including a lot of my CDs which are pictures, and that doesn't include the pictures because there are just too many of them, but if there's about 700MB on a CD, and there's probably a thousand CDs there."*

*"I have a 2nd year PhD student doing a lot of microscopy and she's got 26GB, so I think if you had a post-doc doing really high-end microscopy I'd probably add a zero to the end of that. Just because everything would have gone up in detail."*

*"There's a problem with the amount of space I require to actually produce this stuff in the first place. So I need a computer that has a lot of virtual memory, which is lacking."*

Keeping these large amounts of data backed up is also an issue, with many interviewees allowing the university system to backup their day to day data, whilst making multiple copies of their own data stores.

## Outcomes

The final phase of active data is dissemination, or converting the data into knowledge. Interviewees are fairly unanimous in that publication is the ultimate goal for research, but they differ on what constitutes the final form of their data. The process, unless it is itself a publishable outcome (i.e. it is novel and interesting), is seen as less important than the final results of the process.

*"So the final outcome from me is the result. By the process of finding them, is sort of set in stone so that's not really...it's the outcome that's important"*

*“I would say the final outcome of my research is theory and findings. I think data's just the means to an end”*

## Summary

| Process |
|---|
| • Data Generation<br>• Analysis<br>• Collaboration<br>• Storage and Longevity<br>• Outcomes |

# Provoking

The final category in the deductive analysis was a catchall to grab any interesting excerpts from the interviews that didn’t quite fit in the main four categories. There were relatively few entries in the provoking category, suggesting that the original categories were pretty good at covering all the topics that turned up. However, there was a significant theme that emerged in this category, which strongly overlaps with issues raised in several of the previous categories.

## One Size Fits All?

A major theme echoed throughout the data is that interviewees felt that developing a one size fits all policy could, if too restrictive, be counterproductive:

*“University tends to have large systems in which they try to fit everyone in. And we don't particularly fit that system.”*

*“When your job is to sit at a computer 8 or 9 hours a day, having to cope with something that you can't alter in any way is well, constraining.”*

*“Everybody gets 1 GB of storage and that's fine, it's all backed up, but then if you don't go over it, they don't give you anything to use that. It’s the same with the email system that they give you only so much storage and then just tell you to delete emails.”*

*“RDM shouldn’t be dictated by an institution as there is no one size fits all”*

*“Instead of looking one size for all, it is better to develop separate systems based on faculty.”*

*“The big thing, is university is not storing the data. The longer we go without storing, bigger the problem is.”*

## Summary

| Provoking |
|---|
| • One Size Fits All? |

## Developing Final Themes

The final themes from the analysis need to be well defined and informed by the full data set. The initial deductive analysis was used to loosely categorise the data in such a way that stronger themes could emerge. Having considered the deductive themes, the final task is to re-organise the key subthemes that came through in this initial analysis into the final well defined themes.

### Category Summaries

| Perception | Purpose | People | Process | Provoking |
|---|---|---|---|---|
| • Usage<br>• Misconceptions<br>• Space<br>• Backups<br>• Security<br>• Longevity<br>• Sharing<br>• Training | • Data Generation<br>• Value<br>• Longevity | • Who Has Access to Data?<br>• Responsibility | • Data Generation<br>• Analysis<br>• Collaboration<br>• Storage and Longevity<br>• Outcomes | • One Size Fits All? |

### Longevity / Life Cycle

From the summaries, there are already several apparent overlaps between categories. The first theme that runs across many categories is to do with the life cycle of data, in particular the longevity:

| Perception | Purpose | People | Process | Provoking |
|---|---|---|---|---|
| • Usage<br>• Misconceptions<br>• Space<br>• Backups<br>• Security<br>• Longevity<br>• Sharing<br>• Training | • Data Generation<br>• Value<br>• Longevity | • Who Has Access to Data?<br>• Responsibility | • Data Generation<br>• Analysis<br>• Collaboration<br>• Storage and Longevity<br>• Outcomes | • One Size Fits All? |

### Data Analysis

Similarly, how researchers generate and analyse data is a common theme. The difference in the use cases for researchers is particularly significant:

| Perception | Purpose | People | Process | Provoking |
|---|---|---|---|---|
| • Usage<br>• Misconceptions<br>• Space<br>• Backups<br>• Security | • Data Generation<br>• Value<br>• Longevity | • Who Has Access to Data?<br>• Responsibility | • Data Generation<br>• Analysis<br>• Collaboration<br>• Storage and Longevity | • One Size Fits All? |

| | | | | |
|---|---|---|---|---|
| • Longevity<br>• Sharing<br>• Training | | | • Outcomes | |

## Responsibility

Another clear theme that emerges from the data is that of responsibility – which is largely captured by the existing people category, but is touched on by other categories as well.:

| Perception | Purpose | People | Process | Provoking |
|---|---|---|---|---|
| • Usage<br>• Misconceptions<br>• Space<br>• Backups<br>• Security<br>• Longevity<br>• Sharing<br>• Training | • Data Generation<br>• Value<br>• Longevity | • Who Has Access to Data?<br>• Responsibility | • Data Generation<br>• Analysis<br>• Collaboration<br>• Storage and Longevity<br>• Outcomes | • One Size Fits All? |

## Sharing and Collaboration

The theme of collaboration is apparent across most of the initial categories:

| Perception | Purpose | People | Process | Provoking |
|---|---|---|---|---|
| • Usage<br>• Misconceptions<br>• Space<br>• Backups<br>• Security<br>• Longevity<br>• Sharing<br>• Training | • Data Generation<br>• Value<br>• Longevity | • Who Has Access to Data?<br>• Responsibility | • Data Generation<br>• Analysis<br>• Collaboration<br>• Storage and Longevity<br>• Outcomes | • One Size Fits All? |

## Diversity

The most apparent theme from the data is actually more subtle to define from the summaries, but it is extremely significant. It touches on almost every aspect of the data, and is the overarching idea behind much of the responses from interviewees:

| Perception | Purpose | People | Process | Provoking |
|---|---|---|---|---|
| • Usage<br>• Misconceptions<br>• Space<br>• Backups<br>• Security<br>• Longevity<br>• Sharing<br>• Training | • Data Generation<br>• Value<br>• Longevity | • Who Has Access to Data?<br>• Responsibility | • Data Generation<br>• Analysis<br>• Collaboration<br>• Storage and Longevity<br>• Outcomes | • One Size Fits All? |

# 3. Inductive Thematic Analysis Findings

The Inductive Thematic analysis process essentially reclassifies the data into much more concrete categorisations; these new categories, outlined below, can be distilled into recommendations to be taken forward in the project. The purpose of the recommendations is not to implement policy directly, but rather to invoke a discussion. What are the implications of these recommendations, given that they are essentially the "zeitgeist" of the user base of any policy which may arise?

## Diversity

The first thing that comes across from reading the transcripts is the large diversity in data usage by the interviewees. Data is often generated using specialist techniques, equipment and software, which can vary from project to project, nevermind from user to user. This specialist data collection has repercussions throughout the data life cycle, e.g. file formats, storage linked directly to equipment etc.

Even users with "normal" data needs use and store their data in radically different ways, and many already have de-facto data management policies that are adhered to and work. The expertise already exists to a large extent within the research groups themselves.

Coupled with this, there is a strong sense that any push to one-size-fits-all style institution wide "enforcement" would be strongly resisted, even amongst those who are using ISS systems already. Users want the best solution for them, which they feel does not always fit in with the services provided by the university.

Certainly, some of this can be put down to user ignorance of what the university offers, but we should not dismiss the data management expertise of the users who have "rolled their own" solutions – they have a strong link with their data and experience of managing it in a way that works for them.

From this we can take forward the following recommendations:

- Any policy that is institution wide should not insist on strict use of university services, provided researchers can satisfy their funders that they are doing data management correctly, they should be allowed to. The policy should be a series of guidelines, and an opportunity to suggest solutions that the university can provide. Even a service that provides advice on which hardware / sw to use for those who wish to go it alone would be appropriate.
- Allow easy integration of central services with what researchers are using. Provide simple interfaces so that where applicable users can integrate their own policies with those of the university.
- This does not mean that the institution should "wash their hands" of users who do not wish to use the services. There is a strong case for a more integrated approach between support and research, rather than a separation. Many researchers have found great benefit in having support staff imbedded in their research groups, rather than in a separate building/office – researchers have the expertise to manage servers etc. so allow them training to do so, and make them aware of what the university provides, and vice versa have support staff take real interest in research groups activities – even become members of the groups. There needs to be a relationship

between individuals from both sides, the "single point of contact" policy of the help desk is not always suitable for this kind of support. Locating support and research in the same space would seem like the ultimate goal of this recommendation.

- Essentially, users who do not use university services want more control. The university should provide sufficient training to those who want it to allow them this control.

## Data Analysis

How users categorize data is also quite variable – some see data as the final results while others take the opposite view that data is the intermediate stages and the final result is no longer "raw" data. These are just differences in terminology, and most interviewers got across the message that as far as we are concerned, "it's all data".

What did persist however was the idea that data on a local machine was different and less final that data on a server or to be published. Indeed one commonly occurring issue, that of "space" seemed a much more pressing issue for data during processing (there is a spate category for longer term data storage below), where data is actively being worked on. This is mainly due to memory restrictions on a single machine, with not many users utilising cloud or clusters for processing.

- Users are largely unaware of hpc style processing solutions, or they are but do not know how to use them. The availability of such systems, and good training in using them should be provided.

## Longevity / Life Cycle

Another overarching theme is the longevity of data. People do not want to throw data away. This is more than the issue of space or capacity. Users are managing their own long term archives without any service prevision from the university, whether it be on their own hard drives, home computers, DVDs / CDs etc. Even email is saved off and preserved in external systems such as Gmail. Many users use their university space up to its limit and then archive their data off. The argument against this being something that the university should provide as it would cost too much is largely negated by users' awareness that they can buy the physical storage very cheaply and they are doing the transferring of files themselves.

Funders are becoming more insistent on users keeping data for longer periods. There are a few national/international data storage facilities being used, but they are not yet commonplace.

Another issue is accessing old data, and converting it to modern data formats, even for things like word or excel.

- Long term archiving is an issue that is currently overwhelmingly done by users themselves. This is a separate issue from day to day storage space, and should be dealt with separately by the institution (archives need not be on active servers but on other media such as tape, DVD. Users would be fully responsible for them rather than administrators etc.) – If possible the institute should provide a separate archive service, or at least have good advice available for users who need it.

- Users would be willing to pay for extra services such as archiving, but also for services such as mass file conversion. In particular, video and audio file compression.

## Responsibility

Largely coming from the "People" category but spread across several others is the theme of responsibility, or who should be doing what. It ties in somewhat with the diversity category, in that there is some dependency on the research and the user as to who is responsible for what.

Users who use university systems expect that they are backed up and maintained, i.e. that the university takes full responsibility for data integrity.

Most users assume that the university controls their IP, however only a couple remember signing anything to that end.

What is common is that this is an area that most users would like more information on what the university expects from them.

- The policy should include guidelines as to what the university services guarantee and what is the minimum expected of users who choose not to use the university services.
- Again, training is a big opportunity to clarify this situation.

## Sharing and Collaboration

Many projects are shared across research groups and even institutions. Often research that directly affects the public shares its data with the participant or the public at large, and of course after publication, data is more and more being shared with the research community. While national and international archives do exist, they are not yet commonly used.

To deal with this, many users share their data on an individual basis, a common way is through home pages, or project websites. Sharing with collaborators often involves emails, data sharing services, or users having access to external servers and vice versa or sometimes physical posting of hard-drives. Some projects are hosted externally, on servers such as Google Code or GitHub. These are essentially international data / code repositories.

Users who want to share data are very keen to link their data with their publications or profile, while those who do not want to share don't wish to be forced to do so.

The recommendations essentially cover two scenarios, sharing data during collaboration and sharing data post publication.

- Sharing "internally" within the university is not a major problem, but sharing with collaborators is more difficult, particularly with very large data sets. The university should provide guidelines as to how to do this best, bearing in mind that other institutions may insist on certain technologies being used that this university does not recommend, the guidelines should allow some flexibility.
- The university should provide a method of sharing data post publication, which should be linked to publications and to researcher's profiles. This is very important for the external

view of the university. Ideally this would also fit in with researchers or projects existing web presence, perhaps by supplying an api?

# 4. Appendix A: Deductive to Inductive Analysis Notes

## Perception

1. Usage:
    a. Mathematical Results / intermediate, “diaries and letters and things like that”, Emails, Photographs, Scribbles / Notes, Audio Recordings, Word documents, PowerPoint presentations, occasionally an Excel spreadsheet, nvivo files, “effectively anything that is associated with the research”.
    b. Cross referenced databases – e.g. hover-tags showing when and where data was used/created etc.
    c. Use Version Control software:
        i. “it's all stored under SVN”
        ii. problems when the repository is unavailable
    d. Bypassing University Systems:
        i. “The only reason that I'm able to do what I do is that we have our own computing officer so if I had to use the university system under Windows then it would be terrible. So if I was restricted to the university system, it would be unmanageable”
        ii. “in terms of size, complexity, managing. It's too straight-jacketed. Whereas we have two computing officers who are very helpful and who can help you solve problems”.
        iii. “The university tends to have large systems in which they try to fit everyone in. And we don't particularly fit that system, then you find ways around that system, which happens a lot”
        iv. RDM Shouldn't be dictated by an institution as there is no one size fits all. Subjects are communities, not institutions.
        v. “I guess I can access my H-Drive from outside the university, but I usually just put them on my own USB, then, and take them home with me, or even email them to myself, as a way of having an extra copy”.
    e. Reliance / Expectation of University Systems.
        i. “the service should be provided by the university and system should standard and central”
        ii. “almost nothing is actually saved on our hard-drives, whether internal or external, partly because ISS does a good job of limiting our access to these drives and machines”
    f. Remote Access:

        i. "with smartphones and things like that, people are actually watching their experiments in the evenings".
    g. Wishes:
        i. "A system for storing, sharing, archiving, securing it. A system with appropriate access by people with and outside the university"
        ii. "A cloud based system would be fantastic"
        iii. "we are looking for a system for data management"

2. Misconceptions:
    a. Data is results / is not intermediate (or vice versa) – distinction "Intermediate" data Including:
        i. "thinking" i.e. notes, "scribbling"
        ii. Meetings
        iii. Blackboard photos.
        iv. Things like survey results, (i.e. pre analysis)
        v. Survey/Questionnaire designs themselves – i.e. the tools for getting more data.
        vi. "Historians when they think of data tend to think in terms of numbers and graphs."
    b. Misconception: Data is Quantitative, Not Qualitative.
        i. Data is not real until its analysed – linked to above.
3. Space:
    a. Local Space (for immediate processing) vs Stored Space
        i. Servers are adequate: "we've never had a problem of needing more space".
        ii. an arms race between the laboratory equipment which is generating data and the machines which are analysing it.
        iii. "The first place we keep in the hard drive and at a point we feel that there too much data in it so we put them in a another medium. but we don't have a procedure for doing it"
    b. Inadequate capacity (in either / both of the above)
        i. "I'd love to store it on the university computer, but I don't have enough space for it, by far not enough space for it."
        ii. "So we are talking about a good amount of data, more than I get storage on my university computer!"
        iii. "I try to use Gmail as much as possible as the university email is pathetic"
    c. Compression not required – "used to be software and devices that could compress data so you get more of it on to a relatively small drive but now that are unnecessary."
4. Longevity
    a. never throw anything away (but don't keep good track of where it's stored)
    b. Old data still in use:
        i. all of my previous research jobs as well, I have all of my PhD data, which is all still active data it all still feeds in to everything that I do.
    c. accessing old data
        i. emails from old versions of word

   ii. “I've still got data hanging around from my PhD which were in old-fashioned tape record form”.
   iii. uses open standards wherever possible

d. Relevance:
   i. “if you're doing something new, then it's quite hard to reuse material”
   ii. “cos it's through the reproducibility that, personally, I find confidence in the results”

e. Responsibility:
   i. when a project completes and there is no-one to maintain it any longer.
   ii. not sure what the actual mandate is at Newcastle.
   iii. rule of thumb I've always been taught is bascically you need to save your data for 5-7 years in order for if somebody does ask for it, and actually after that there's a bit of a grey area as to whether it's reasonable to ask

f. Value:
   i. “data has increasing value as we know more from it”

5. Sharing:
   a. Data outlives projects?
      i. Publications are based on data which are then cited, so indirectly certainly.
      ii. Data is only per project, i.e. not shared after
   b. Use other researchers data if it’s online, publicly available
      i. “we are about to pull the data from the European centres”
   c. Transferring between sharers:
      i. “...then forward the file to my research assistant via email, although I think we've transferred some from memory stick to memory stick directly”.
      ii. “someone  else in the school that had data or had a project that would link quite well with my project and we had a way of sharing then we would be happy to do that and we would share”
      iii. “they can get access through the repository. You can set permissions so that's fine”
   d. Publishing (i.e. along with papers and articles):
      i. “within the publication, if it's done well, you'll actually have all of the information you should need to repeat any experiments”
      ii. Destinations often have guidelines: “We're about to turn the data into a dataset, and potentially publish it in the Economics and Social Data Service Website and I think that there are more guidelines as to how the data should be stored”
      iii. “we need to be pretty sure that it, you know, we've got the data in an acceptable format before we actually submit it to them”
      iv. “all of that data is feeding in to regional records, national records that are leading towards hopefully national, policy settings for the agencies and for the government, as well as back in to the local community”
      v. “the opportunity to publish is actually something that most researchers would I think want to do. It's not as if they want to keep it to themselves. If

anything they'd like to make it available and justify the time and expense that somebody's spent working on it”

   e. Funders
      i. “I have a growing feeling that funders are going to ask for the raw data in the future.”
      ii. “data-sharing policy that you need to fill out for example with the BBSRC, saying that you'll make any large databases that you generate available and that you describe where you would store those”
   f. Not Sharing:
      i. Separate “Shared Drive” just for specific project
      ii. Don’t want colleagues to see / change data

6. Security
   a. Unimportant:
      i. “no one else would understand it”
      ii. “Well they could steal my research; they'd certainly be able to do that. But there's nothing that could be damaging to the institution.”
      iii. “we mostly rely on an honourable system”
   b. Selective, only “Sensitive” data “securely” stored.
      i. “most sensitive data on the university computer, so anything that would enable somebody else to identify who the individual was”
   c. Security vs Usability:
      i. “it's on my research memory stick which obviously is carried around and is the least secure, but it's just that it's very handy and I also have parts.”
   d. Anonymity:
      i. “some of my interviewees were quite keen for them to not be able to be identified”
   e. Responsibility:
      i. University’s Problem: “I'm not aware I'm using any security systems except whatever the university offers”
7. Training
   a. Little or none given.
      i. “I don’t think it was ever mentioned”.
   b. What’s needed:
      i. “think I'd benefit from a bit of guidance as to where things can be archived, how and what is safe and legal, and ethical”
      ii. Not aware of legislation (other than FOI and DPA).
      iii. “for new staff, and or, new researchers is to get an idea of things like encryptions, the policies, programmes, just a bit of induction to all the university offers because I wasn't aware of encryption until I attended a training session”;
      iv. “it is the lead recognised by the institutional. People need more education on the area of data management .”
   c. Quality:

      i. “And I still haven't quite figured out what I need to do, so if somebody says, 'this is encryption, we advise you do it, this is how you do it', would be really, really helpful”
   d. What’s available:
      i. know, SDU does data management and workshops on excel spread sheets but I am not aware of the best practices of the data management and data storage

8. Backups
   a. On users personal computers:
      i. “back-ups on my home laptop”
   b. Importance:
      i. “Losing all my data at the moment would be catastrophe because the grant is running out and my funders are expecting a return”
      ii. “It would be career suicide if I lost all of my data”
   c. Behaviour:
      i. “Very sporadically”  “Too Time Consuming”
      ii. “CDs are brilliant for that kind of thing, it's a static”.
      iii. Automated systems (e.g. Gmail) “It's a secondary back-up, without me having to do anything”.
      iv. “ i run a daily back up from my computer and my data is stored in the computer as well”
      v. “ I know there are data which does not have any backup...”
   d. Restoring:
      i. “haven't needed to yet...”
      ii. Use Version Control: “with SVN you can go back to previous versions”.
   e. Responsibility:
      i. “provided the university's backed everything up, so it's the university that'd be in trouble if it didn't work”.
      ii. “ I think we should get guidance of minimum standard that has to be practised for data management and also some support for online tools and structure within the university to follow a high standard of RDM. If this is going to become as the other requirement then everyone is going to find time and do it”
      iii. “I think [University Backups] are kept indefinitely”

9. Digital vs Physical
   a. Organisation
      i. “the advantage with nvivo is that everything is in one place, whereas, I might lose bits of paper and scribbles and that's why I quite like electronic storage”

10. Policy & Legislation
   a. Data Protection Act
   b. Freedom of Information Act
   c. Behaviour:

i. People like MRc they have standard clause on the data sharing etc , it is almost you have to tick the box, you have the standard text. Nobody really practises it.

## To People:

- any requirements my funder
- Royal Society is who I'm funded by now. BBSRC. That's pretty much it. There are some private institutions as well like the Lister Institute, and then a lot of foundations for funding people. So, Marie Curie, Human Frontier Science Programme, even other governments, like the National Institute of Health in the States.

## Process

1. Data Generation:
    a. Specialist Software
        i. Requires full access to "own" server-like architecture:
            1. "So I will use this programme, first of all I will write a little script to run inside this programme and so I set this up on a special server machine in my department. It will then instaneously create these massive files stored in the virtual memory. And then once it's done that it can use the programme to search through these files and the output is  actually quite small, and then I see the results"
    b. Intermediate data vs Results
    c. Rate
        i. I'd say more like hourly, I'd say
        ii. Matches project cycle – "towards the beginning there's relatively little, you then the exponential phase, where you know, it really takes off, and then towards the end, as let's say they're trying to either write up their data for publications, or they're looking at jobs, or they're writing their thesis, you'll see the amount of data accumulating kind of taper off"
    d. Predictability
        i. you have a set idea of the data that you're looking to find? - Both actually.
        ii. Very rare to fully know the extent of data at the beginning.
        iii. [Size]... No, that would be very difficult to predict
        iv. No, I mean to be honest, the best days are when you find something you did not expect
    e. Recording in an environment where technology isn't available:
        i. "because I did a lot of research abroad where computer technology wasn't as advanced"
2. Analysis
    a. Specialist sw
        i. when it comes to analysis I  use some of the stats programmes that the university have, so things like  SPSS and things like minitab, but then Primer as well, I use Primer a lot and Primer's not available through the university. And then things like RJS and stuff like that but that's all available through the university so mostly it's university systems.
        ii. we also use a very specific programme called Metamorph, for looking at the .tifs, just because it allows us to change contrast, brightness, things like that
        iii. Several of our instruments have their own custom software
    b. Specialist hw
        i. This requires specialist computing hard-ware
    c. Mobile access
        i. now with smartphones and things like that, people are actually watching their experiments in the evenings
3. Collaboration

- a. Centralised / External processing (Funded):
    - i. a lot of the grants buy time on computers elsewhere, so people have data - I mean a friend of mine was recently sent his data from the place in Finland where his data was kept. So he received 5 1TB hard drives in the post.
    - ii. National Grid
    - iii. funders are pushing for it and I can understand why. And, obviously I'm trying to pursue it, partly because I promised it to funders, partly because I think it's a good idea
    - iv. to the funders to put it on the Economic and Social Data Service
    - v. it will be archived in ERIC which is the regional record centre, and it will be archived in DASHH which is the Database for Seabed Sepcies and Habitats - it's a marine national database.
- b. Sharing:
    - i. Use svn style repos linked to the university. And also sometimes GITHub as well, which isn't linked to the university.
    - ii. Collaborators' systems (e.g. dropbox)
    - iii. we tend to use their systems
    - iv. data into repositories to enable others to access it rather than him not having to maintain it.
    - v. there is organisation who collect data from papers and put them in the archive
    - vi. Sharing data would be a...preparing PowerPoint presentations - your own interpretation of your data
    - vii. No, most of the things where you do collaborate, by the nature of the projects, you're doing something quite specific that somebody's asked you to do - but I would usually generate what you would see within the publication.
- c. Using others (Public or collaborators) data

4. Transport
    - a. Email
        - i. moving things across by email
    - b. Thumbsticks
        - i. I use memory stick, I carry my laptop so I carry using it.I rarely email them
    - c. RAS
5. Storage
    - a. Local
        - i. Intermediate data
        - ii. "when you run a programme on a computer, you necessarily want it to send data across the network because it's slow, so you want to have a space on that computer that you can output to"
    - b. Own Server
        - i. "I do keep it in my home area. We have our own little set up in the Maths and Stats school and on each of the computers that run the Linux operating system there's something called \data"

c. ISS managed
   i. "I usually put them on my H drive or I use RAS on wifi wherever I am"
   ii. "I'd love to store it on the university computer, but I don't have enough space for it, by far not enough space for it"
   iii. ISS managed S drives (shared by group)
   iv. almost nothing is actually saved on our hard-drives,

d. Mixed
   i. Most seem to use some or all options above
   ii. I'm not sure all of the labs use the S-Drive as much as we do.
   iii. once you're done here with the actual capturing the data, you know, some people might like to go home, see their kids, and then work on the analysis bit later on, which is, we personally don't have a problem with or discourage in any way.

e. Email
   i. when someone sends me data, I don't want to delete it. I might never download it somewhere else, but I don't want to delete it (Gmail)
   ii. 'cos it's easier to search (Gmail)

f. Non-University
   i. software projects are hosted on Google Code or GitHub
   ii. tools are developed locally but then are run on a number of systems including servers owned by the research group, school or collaborating institutions

g. Space
   i. there's a problem with the amount of space I require to actually produce this stuff in the first place. So I need a computer that has a lot of virtual memory, which is lacking.
   ii. "has enormous memory requirements, you can fill up a thumb drive very quickly"
   iii. I'd love to store it on the university computer, but I don't have enough space for it, by far not enough space for it
   iv. Compression unnecessary
   v. about 7GB, that's backed up on the university servers. And for my own work about 500GB. And that isn't including a lot of my CDs which are pictures, and that doesn't include the pictures because there are just too many of them, but if there's about 700MB on a CD, and there's probably a thousand CDs there.
   vi. even if you go out in the field and take 50 pictures I try very hard not to delete any of those pictures because you never know what you mind find from them.
   vii. 2nd year PhD student doing a lot of microscopy and she's got 26GB, so I think if you had a post-doc doing really high-end microscopy I'd probably add a zero to the end of that. Just because everything would have gone up in detail.

h. Not on a computer

i. that data is not part of a computer at all, it's me sitting with a blackboard

ii. That stuff is stored in a filing cabinet or piled up high on my desk

iii. Most of those blue folders, for instance, are my PhD notes, which are all in pencil shorthand so no one else can read them.

iv. I've got videos and cassettes that have been built up while I was at sixth form and university.

v. all of the volunteer data they record on sheet form

vi. Everything that I've got in hard copy should be on the PC, it doesn't get filed until it's on the computer, so everything hard copy-wise is on the PC, but it allows me to validate things if I need to go back to have a look at it. So there is definitely a place for the hard copy but it is all backed up electronically.

i. Concerting to digital:

i. "I take photographs of my blackboard and store them on my hard drive which I then back up to the university's my user area, but I do tend to use photographs for my blackboard and whiteboard."

ii. "I have started moving things, by scanning because we recently got a printer with a scanner on our floor, which is especially useful when you have your phone and things as you can send stuff to yourself and read your notes Online"

iii. then I have all of my field books as well and all of my notebooks which is all kept in the folders. It's then transcribed onto the PC, mostly on Excel spreadsheets.

j. Data Formats

i. MS office

ii. Pdfs

iii. Jpgs

iv. Audio

v. File conversion sometimes required

vi. Whatever is in use at the university

vii. we can use files coming from mechanical modelling tool in a electrical modelling

k. Structure

i. Per Project – decided by researchers

ii. Only important nothing is lost, i.e. can find anything quickly

iii. encourage fellows, and PhD students to organise data into folders monthly

l. Specialist sw

i. do use nvivo for data management / analysis but, we could do it manually if we fancied

m. Searching

i. Only important nothing is lost, i.e. can find anything quickly

ii. Relational database – tooltips / linked data

6. Longevity

a. Some data changing constantly, vs some not changing for years.

i. but source code is constantly changing

b. Intermediate data less important to keep:
    i. “I tend to just delete mine as I go, once I've got something I'll tend to write it somewhere appropriate and then just delete the existing files”
c. Accessibility / Future proof
    i. “when I started as a PhD studtent, the electronic devices then, you can't really use the material anymore, you know like 3.5 inch floppy disks.l
d. Project Length:
    i. “my current project which I started in 2007, although the data collection didn't really start till the Autumn of 2010, so it was data collection of about 2 months: October, November, mainly 2010, and then my research assistant started analysing the data when she started in May 2011. Data analysis is still continuing as we speak in March 2012 and will probably go on for quite a while. Because the work is quite inductive, it's necessary when writing a new paper, or looking at a new issue, to go back to the data and look for examples, and look for any relationships between different aspects or themes, what people are talking about. And the data collection for the second case study just finished earlier this month and the analysis is just beginning. And again, while the bulk of it will probably be done over the next 6 or so weeks”
e. Archiving
    i. there's data from my PhD still hanging about somewhere without me having a proper place where to archive them
    ii. no , I don't think I have deleted data
    iii. I was just talking to a colleague who said he was still going back to data collected 25 years ago so
    iv. I wanted to use the data which i used while i was researching but i often cannot find it

7. Backup
    a. Non Deletion
        i. at some point, we probably will run out of some space. I'll then ask the computer guy if we can make some back-ups, and then we'll probably do again, multiple formats, whether it's taope or DVD. I'll probably buy myself a big tetra-byte hard-drive
    b. Non ISS
        i. “the guys [local support] have a specific time of day in which data is backed up”
        ii. Very sporadically. I'm not very good at backing up
        iii. CDs are brilliant for that kind of thing
        iv. I guess I can access my H-Drive from outside the university, but I usually just put them on my own USB, then, and take them home with me, or even email them to myself, as a way of having an extra copy
    c. Keeping Track

        i. “I mean you copy stuff and I'm not quite sure how  much of it is duplicates or the originals, and this is going back like 10 years or a bit longer  of research so it's very hard to say with any confidence”
    d. University Responsibility
        i. ...provided the university's backed everything up, so it's the university that'd be in trouble if it didn't work
    e. Human Error vs Computer failure
        i. ad-hoc backup system which includes synchronising file systems across multiple computers so that his data is usable on any of his computers. This serves as a disaster recovery backup although not a backup for human error – deletions migrate across machines. For human error, Phil has an incremental backup to his desktop machine
    f. Problematic
        i. problem because many dont back up data and those who do dont remember where the data is saved.  so there are danger of losing the important data. It is stored in people's laptop,i use evernote, lab books and everywhere it is a mess

8. Outcomes
    a. Data or Process:
        i. “So the final outcome from me is the result. By the process  of finding them, is sort of set in stone so that's not really...it's the  outcome that's important”
        ii. , I would say the final outcome of my research is theory and findings. I think data's just the means to an end
        iii. Is all  about interpretation of data.
    b. Publication
        i. are two avenues when writing a paper, one can use these subversion programmes when anyone can access it and you sort of check in the paper.
        ii. But I choose not to do this because I like  greater control
        iii. The end product is papers; in the current project I promised my funders at least three plus a book.
        iv. you can't re-publish already published
9. Security
    a. Not a priority
    b. University Responsibility
    c. more interested in `safety' in that he can access the data at a later date

➔ People
    ○ Person who is leading is the project stores the data so data gets stored on a single location

Project Identifier: iridium
Version: v1 Date: 5.7.2012
Contact: Lindsay Wood (lindsay.wood@ncl.ac.uk)

# Purpose

1. Collection:
   a. Specialist equipment, i.e. hard to replicate:
      i. “we use some kind of an instrument like Hplc etc”
   b. Cost to obtain data
      i. “you could quantify it is in terms of what it cost to get it”
      ii. Technology has made getting data easier
      iii. “if the experiments take longer time and money then the data from it is very valuable, if the experiments can be easily and quickly repeated then we they are less valuable”
      iv. Process might cost money: “financial cost for the reagent we used to get collect and the staff time who spend on collecting it”
2. Value:
   a. Commercial vs. Academic
      i. Used only for publication
      ii. My data isn't commercially sensitive
   b. Destination / End result
      i. “you think about where the projects going to end up”
      ii. “Putting your data in a national archive or repository?” – “funders are pushing for it and I can understand why”
      iii. “The end product is papers; in the current project I promised my funders at least three plus a book.”
      iv. “I think data's just the means to an end”
      v. “data is feeding in to regional records, national records that are leading towards hopefully national, policy settings”
      vi. Making new processes: “There's an element of process there because you're finding that Methodology”
      vii. Data itself is final outcome.
   c. Loss / Repeatability
      i. “Losing all my data at the moment would be catastrophe”
      ii. “I'd be distraught if I lost any of them, but it is dependent on what you're working on at the time”
      iii. “It'd be painful and I would have to do the work again but would be possible”
      iv. Depends on repeatability:
         1. “if the experiments take longer time and money then the data from it is very valuable, if the experiments can be easily and quickly repeated then we they are less valuable”
         2. “most valuable is data which cannot be easily repeated”
         3. “within the publication, if it's done well, you'll actually have all of the information you should need to repeat any experiments”
   d. Stakeholders

i. “my funders are expecting a return - think if I lost all my data now I couldn't deliver on my results for my funders, I probably wouldn't get funding from them in the future”

ii. “collaborators, but they can get access through the repository. You can set permissions so that's fine”

e. Project dependant

f. Storage
   i. Keep together: “definitely more valuable when they're together”.
   ii. “data as a whole is more valuable”.
   iii. “usually talking in megabytes”;

g. Security:
   i. “Prior to being hacked...”
   ii. Some users do not encrypt, others do.

h. Judging Value not always easy (during collection):
   i. “we don’t know it until afterwards, sometimes we never find out results”
   ii. I think you have an idea, but one should perhaps have an idea but be prepared to change or just to do more things.

3. Planning – knowing about data requirements beforehand.
4. Longevity:
   a. “by and large, each thing finishes and you move onto the next Thing”
   b. “But again some people would have a programme of research which is consistent and it develops from previous data”.
   c. “nervous about going back to data that's pretty old, because you can always say, oh that's dated, that doesn't tell us anything new”
   d. “but it is something that I come back to sporadically”
   e. “if you spoke to me again in 10 years time I'd still have my PhD data hanging about”
   f. “the project I'm working on at the moment, the data set for the Big Sea Survey is absolutely essential. I would say in terms of me, my PhD data is essential, that really does underpin everything I do. I would say all of them”
   g. you still go back to it and you still use it...
   h. “you need the baseline there to begin with and that's your starting point. You then build on that to create long term data sets and those long term data sets are absolutely 100% more valuable than any individual data set because you've got that change over time and you can see”
   i. “Data value would decrease over time” vs “Increases over time”.
   j. “Data more valuable independently”
   k. “Never delete data, unless it is intermediate”
   l. “People've been doing them since the 60s and 70s”
   m. “it's a problem having too many of those rather than too few”
   n. Space requirements:
      i. “More or less a similar size”
      ii. “5, 6, 7 years. About 40GB...”
5. Qualitative vs. Quantitative?

   a. "but I'm dealing with people, and as you probably experience every day when you interview, people tell you very different things and you may not quite find out what you wish to find out..."

6. Storage
   a. Per Project
   b. Per date/time
   c. File conversion
      i. "my research assistant, who is transcribing them, couldn't read the files so she had to convert it"
   d. Data stored as publisher expects:
      i. "We know at the end of the data how we want to publish these"
7. Analysis
8. Teaching

To Perception:

- "So it isn't just numbers it's everything"

# People

1. Who?
    a. Researchers / internal collaborations
        i. Sharing data / programs
        ii. research assistant started analysing the data when she started in May 2011
        iii. prefer that it goes onto the S-Drive, so that it's immediately backed-up
        iv. “I'd go and talk to that person about, much more than wanting to access to their own data”
    b. Collaborators
        i. Sharing data / programs
        ii. “be experimental data, something that someone else has collected”
        iii. Restricted by what other collaborators use, e.g. dropbox.
        iv. “mean journal access is pretty good and anything that we don't have free access to or a subscription to, you can just go to the library and request and within a few days”
    c. Funders
        i. “guidelines as to how the data should be stored”
    d. Post Publication / Community
        i. “My most recent paper is  already being used by people”
        ii. “I've had my programmes asked for before, by a journal referee in fact.”
        iii. “data will be held within national  databases, but we will restrict access to that data until publication”
        iv. “The Marine Conservation Zones have recently done a Defra
        v. “data survey and they've tried to collate all of the data from around the UK on marine habits”
        vi. preparing PowerPoint presentations.
        vii. “I mean most of these [journals, papers] you do have to include the data and they would ask that you do that.
    e. University
        i. the publications that the institution care about
    f. Students
        i. “I just gave  a student my cuttings [to sort] for his dissertation”
        ii. “it’s an important point for an institution where we have research informed teaching”
    g. Public
        i. “If someone asked you for it they could have it?” – “Yes”
        ii. “when someone comes  in and asks if you've got anything and you've got it you give it to them”
        iii. National Archives
        iv. “about 150k reads”
        v. “He is contacted each year by people wishing to use the software but it no longer works or is maintained”

- vi. Cross referenced data base.
- h. Users / Beneficiaries / Participants
  - i. “educating local communities and volunteers”
  - ii. “puts his data into repositories to enable others to access it rather than him not having to maintain it”
- i. (Non ISS) Research Support
  - i. “I had to use the university system under Windows then it would be terrible”
  - ii. “I'll then ask the computer guy if we can make some back-ups”
- j. Commercial
  - i. Transcription services
  - ii. Have their own systems

2. Responsibility
   - a. IP
     - i. “when I signed my IP form it was with the university, so I'm not aware that EPSRC has any intellectual property rights over anything that I Produce”
     - ii. “As far as I'm aware the data belongs to the university”
     - iii. “Collaborators, if they generated data”
     - iv. “tries to maintain Uni copyright and writes grants which includes what license code will be released on”
   - b. Bypassing ISS
     - i. “I'm a bit biased and spoilt because the computing people we have ...I don't ever have any interactions with ISS....they're very good.  The reason why they're getting this 32 GB stuff is because we asked them”.
     - ii. “I had to use the university system under Windows then it would be terrible”
     - iii. “In terms of size, complexity, managing. It's too straight-jacketed. Whereas we have two computing officers who are very helpful and who can help you solve problems”
     - iv. a lot of people have their own personal computers that they would prefer to use
   - c. Security
     - i. University:
       - 1. “I've got  them on my H drive”
     - ii. Researcher:
       - 1. “If I lost all my data now I couldn't deliver on my results for my funders”
       - 2. “rely on an honourable system of people respecting each other”
   - d. Keeping Track
     - i. In reality each researcher / each phD student has precise more or less recollection where they put the data.
     - ii. most people, still use things that look more like this. And we, you know, kind of generate our own files and within each one of these you have...(shows notebook)

e. Training

   i. People need more education on the  area of data management

   ii. “I'm not sure what the actual mandate is at Newcastle but the rule of thumb I've always been taught is basically you need to save your data for 5-7 years in order for if somebody does ask for it”

## Provoking

1. One size fits all?
    a. university tends to have large systems in which they try to fit everyone in. And we don't particularly fit that system.
    b. when your job is to sit at a computer 8 or 9 hours a day, having to cope with something that you can't alter in any way is well, constraining.
    c. everybody gets 1 GB of storage and that's fine, it's all backed up, but then if you don't go over it, they don't give you anything to use that. Same with the email system that they give you only so much storage and then just tell you to delete emails.
    d. RDM Shouldn't be dictated by an institution as there is no one size fits all
    e. , i run a daily back up from my computer and my data is stored in the computer as well. I know there are data which does not have any backup
    f. instead of looking one size for all, it is better to develop separate systems based on faculty.
    g. the big thing, is university is not storing there data. longer we go without storing, bigger the problem is.
2. Clarity
    a. A generally, we can do it in a more organised way, but it is the time and commitment to do that.
3. Keeping central repositories upto date
    a. PhD thesis library – not upto date.
4. Sharing/Using LARGE data – current systems inadequate?
    a. So he received 5 1TB hard drives in the post
    b. just has this data sitting in these hard drives now, under his desk
    c. we tend to use their systems
5. Archiving
    a. Throw nothing away – but not necessarily on uni systems
    b. manual archiving
    c. Old file formats?
    d. “I wish I could Store all of my data on the university computer”.
6. Training
    a. Unaware of legislation
7. SVN style repositories
8. Security vs Access
    a. Researchers having to bypass security