Survey Report: ICT in the Research Workflow

Survey Conducted: April 2012

Report Author: Salim Twalib

Report Co-Authors:

Alison Lynton
Andrew Buttsworth
Caroline Boyes
Florian Goessmann
Michael Ricketts
Shane Lynton

Date: August 3, 2012

Contents

Executive summary

An understanding of researcher requirements is vital for any institution that conducts and supports research. As part of Curtin University's ongoing efforts to improve our understanding of how researchers use Information and Communication Technologies (ICT) in research, the CITS eResearch Support team recently surveyed Curtin's research community. The survey was run in conjunction with Curtin University's Office of Research and Development to maximise its reach. Over a quarter of active researchers completed the survey.

The first section of the survey focused on research data during the life of a research project. The responses from the survey clearly outline the status quo of research data storage at Curtin University. Some findings included:

- The overwhelming majority of respondents collect some or all of their data manually (80%), followed by 35% of researchers harvesting data from the internet and 22% using sensors or instruments.
- Most of the respondents (87%) store some or all of their data as common file formats such as spreadsheets, PDF, text documents, etc.
- Nearly half of the respondents (46%) store some or all of their data as multimedia files which includes video and audio. While only a third of the respondents use databases to store some or all of their data.
- Nearly half of the respondents use less than 25 gigabytes of space per project.
- Curtin's network drives are poorly utilised, with only half of the respondents storing their data on I or J drives.
- Most of the researchers use flash/portable drives (76%) and desktop computers (75%).
- Dropbox, a non-institutional, commercial storage solution, is used by a significant (15%) proportion of respondents.
- Respondents identified enabling collaboration (61%) and remote access to data (56%) as the top two features they need in a storage system.
- Respondents collaborate mainly with other Curtin researchers (87%), researchers at Australian institutions (76%), industry partners (61%) and international collaborators (59%).
- Nearly all respondents require access to their data from home.

The second section of the survey focused on computational needs. Some findings included:

- Only a third (34%) of the respondents consider themselves to be undertaking computationally intensive processing on their research data.
- Respondents indicated that they use more than one device for processing. The majority (89%) of which use their personal computers, amongst others, for processing. This is followed by more than one third (40%) of the respondents who use dedicated servers or clusters on campus.
- More than two-thirds (70%) of the respondents believe that their current processing solution satisfies their requirements. Respondents who did not find their current solutions meeting requirements attributed it to insufficient working space (81%), long processing times (67%) and access challenges (25%).

The third section of the survey focused on data retention and data retention storage needs. Some findings included:

- Almost all (91%) respondents want to retain the data for themselves and others. Many (68%) also wanted to make their data citable (e.g., in publications) and (23%) wanted it publicly discoverable through portals such as Research Data Australia.
- The majority (42%) of respondents want to retain data for five to ten years while nearly a quarter want to retain it between two and five years and another quarter for more than 10 years.
- Most of the respondents (80%) indicated that their storage needs after a research project was similar to their storage needs during the research project.
- A data retention storage system has to provide for authorised access (78%), has to provide high security (49%) and has to support a specified retention time (37%). Only 16% of the respondents indicated that they would like the storage system to also provide open access.

The responses from the survey clearly show that Curtin's researchers do not have access to data storage that meets their unique needs. Many are storing their data in a way that places valuable research data in jeopardy and exposes Curtin to potential risks. The survey indicates that researchers require systems that allow for collaboration, remote access and secure storage for data spanning the life of their project and beyond. It is a starting point for further discussions and engagement with the research community to understand their ICT needs and develop or adopt solutions that meet those needs.

1. Introduction

The survey was developed to capture the ICT needs of Curtin University's researchers. It was targeted at the entire research community of Curtin University from all fields of research. The broad spectrum of the intended audience meant that the survey had to be structured in a generic and non-technical form. It was also important to make sure that the survey was both quick and easy to complete. To that end the survey was formulated around three key concepts:

1. Focus on capability, not technology.
2. Focus on gathering existing requirements, not future desires.
3. Utilize a generic research workflow that researchers could easily relate to.

These concepts enabled us to produce a fairly short and straightforward survey that captured the sort of capabilities that our research community requires access to right now. Figure 1 below highlights the structure of the workflow

Figure 1: Survey Structure

There were only 23 questions in the survey. The survey abandonment rate was low at less than 7%. The average time taken to complete the survey was approximately eight minutes and 171 researchers answered the survey. The majority of the respondents were from the Humanities (57%), Health Sciences (46%) and Science and Engineering (26%) faculties. 21% of the respondents were Higher Degree by Research (HDR) students.

The rest of the report details the survey results. The sections of the report will follow the structure of the survey.

2. Gathering Research Data

2.1 Source of research data

Question: What is the source of your research data?

Note: Respondents answered in more than one category.

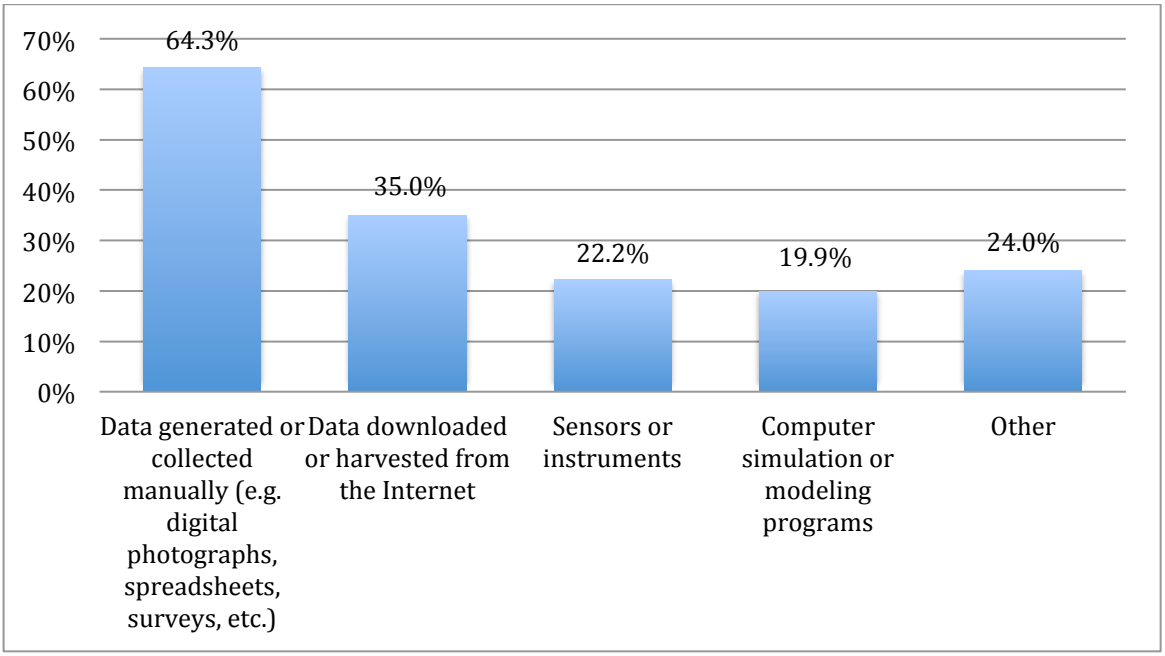

Figure 2: source of research data

- 171 respondents answered this question.
- 33.3% of respondents only use data that is manually generated or collected.
- 45.6% of respondents gathered research data from more than one source.
- 17.5% of respondents gathered data from three sources or more.
- There is a significant proportion of research data that is either collected by sensors (22.2%) or computer simulated (19.9%). Such sources usually produce sufficient data to warrant the need for automated data capture and ingest.
- Most of the comments provided for 'others' can assist in classifying them under the four main categories. Interviews, surveys and data from governmental resources such as the ABS were the most mentioned sources of data. If these were included in the appropriate categories then "Data generated or collected manually" would amount to around 80% and downloaded data will be at around 43% of respondents.
- Other comments included:

  Observation in virtual worlds
  Fieldwork to collect samples
  Health data through WA Data Linkage System
  Data from scientific experiments and clinical research
  Books, journals, personal experience and fields observation

Archival data (govt and private papers, pho to graphs etc) that has not been digitised.

Data supplied by the Health Departments in WA and NSW

Derived variables and processed data sets from raw data sets recorded by instruments.

Legal source documents (hard copy and soft copy) such as legislation, reported cases and journal articles

Anecdotal evidence, archives, texts, historical records, letters, diaries

Archival documents; physical remains of buildings

2.2 Data format

Question: Which of the following categories apply to your research data?

Note: Respondents answered in more than one category.

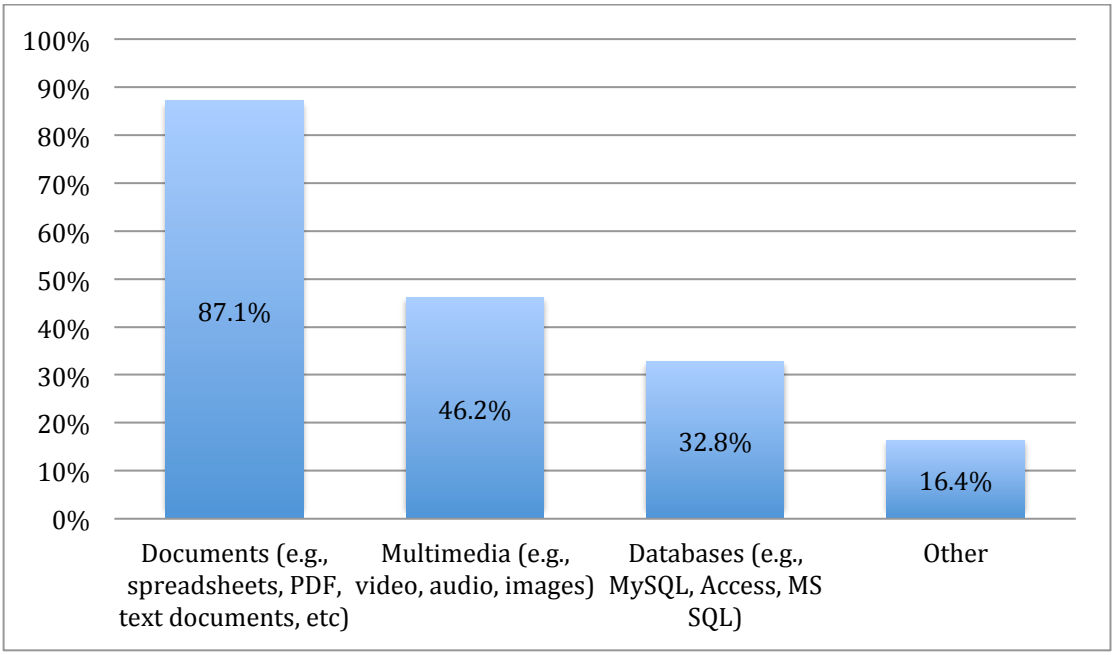

Figure 3: Types of research data

- 171 respondents answered this question.
- 28.7% of respondents have data only in the form of documents.
- 3.5% of respondents have data only as multimedia.
- 2.9% of respondents have data only in databases.
- Some of the comments provided by respondents that selected 'Others' include "data collected by acoustic and sonar systems", "digital recordings", "student assignment submissions", etc. which can fit in the three standard categories.
- Other comments included:

3D virtual world viewers

Rock samples

Art objects

Any data from whatever source

Physical remains of buildings

Motion analysis lab

All of the above plus filed interviews and observations

Not surprisingly, the majority of researchers use data in the format of documents such as spreadsheets, PDF, text documents, etc. Nearly half the researchers have data in multimedia format such as video, audio and images. Only third of the respondents use databases (47 respondents). They were then asked what type of databases they used. Their responses are depicted in Figure 4 (Note: respondents answered in more than one category.)

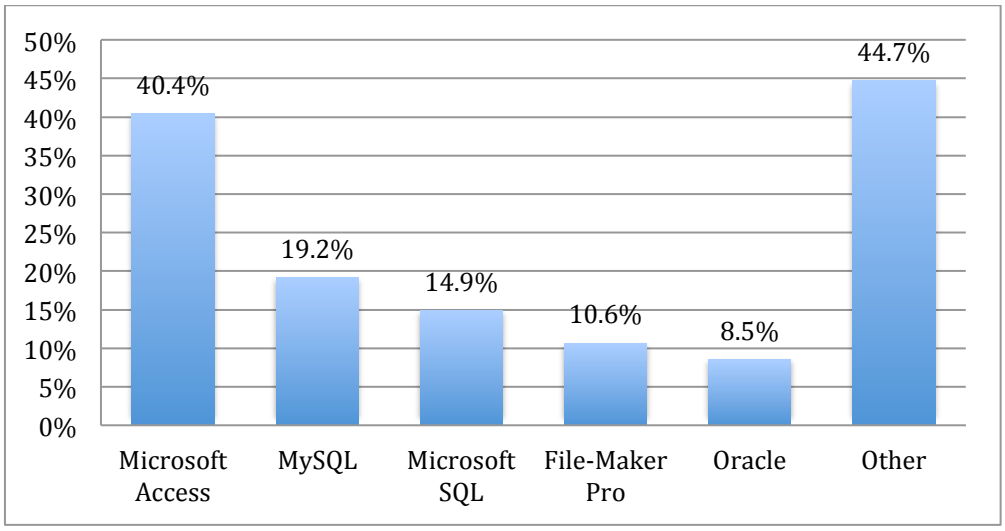

Figure 4: Databases

Interestingly, the majority of researchers who selected 'other' identified SPSS as a database. A count of the number of times SPSS was mentioned puts it at second place at 21% ahead of MySQL. Other databases include:

Ingres
NVivo
Postgres
Spreadsheets
MS Word and Excel
Flat files in SAS and Stata
Internet servers (e.g. SLIP)
specialised vegetation data-capture software

3. Working with Research Data

3.1 Data storage space

Question: How much storage space do you need to work on your research project?

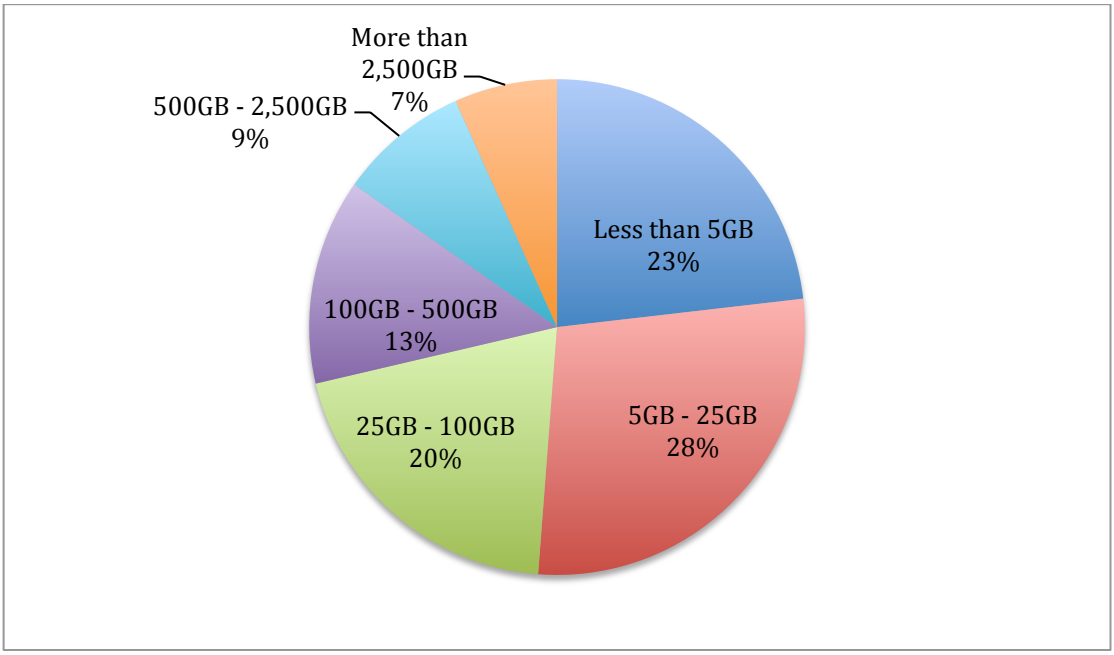

Figure 5: Storage space for data during project

- 164 respondents answered this question.
- 51% of respondents need less than 25 gigabytes of storage space for their research data per research project
- 71% of respondents need less than 100 gigabytes of storage space for their research data per research project.
- With an average of four research projects per year for each researcher, the storage requirements grow quickly.

3.2 Data storage locations

Question: While working, where do you store your research data?

Note: Respondents answered in more than one category.

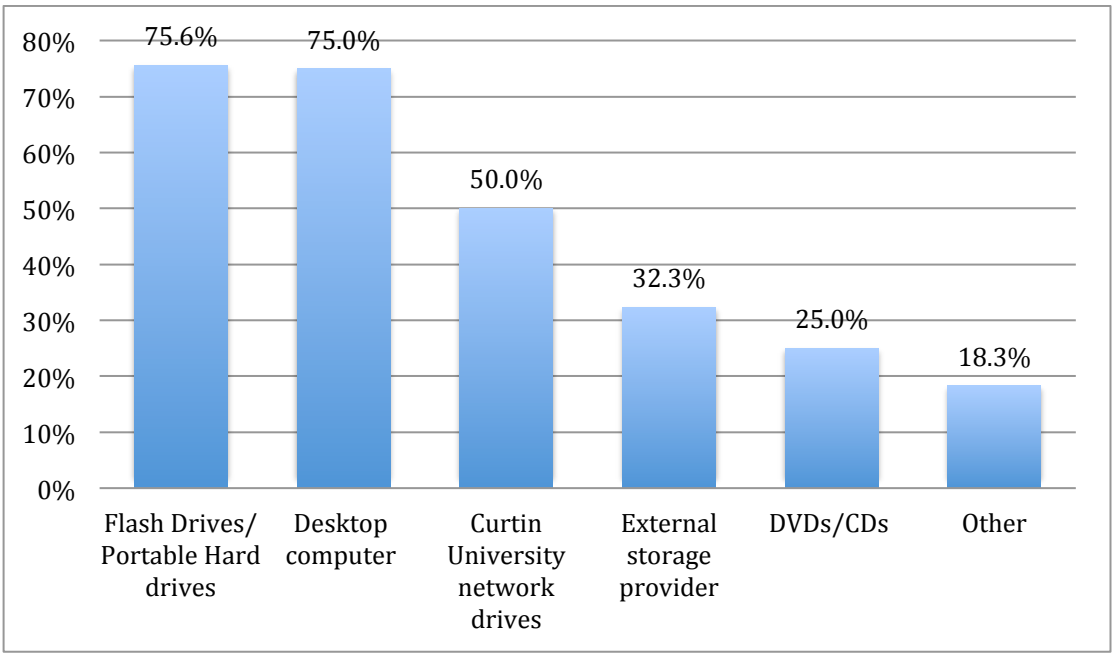

Figure 6: Current data storage locations

- 164 respondents answered this question.
- 85.4% of the respondents store their data in more than one location.
- 61.4% of the respondents store their data in more than two locations.
- 15.2% of the respondents store their data in Dropbox (External, commercial storage provider).
- 6.1% of the respondents store their data in the ARCS Data Fabric or the iVEC Petabyte Store (External storage provider).
- Other comments for external storage providers included SkyDrive, SugarSync, Files Anywhere, Flickr, YouTube, Mendeley, Pages and Numbers on iPad.
- For respondents who selected 'Other', comments included:

  External Sata RAID

  Inside the virtual world

  Research centre NAS box

  External servers, own drives

  Australian Super Computer ANU

  Dedicated website for image storage

  The ARC - used to be called WAGER at UWA

  Time machine drives on my desktop computer

The majority of researchers store their data on portable drives and desktop machines while working on their research data. Many researchers highlighted

security and confidentiality requirements for their data, yet seem to be unaware of the security risks involved with storing their data in this manner. In addition, there is a real risk of data loss if their portable hard drive or desktop machine experiences a failure. Only around half of the respondents indicated that they store their data on Curtin's network drives.

3.3 Requirements for working with research data

Question: What special requirements do you have for working with this research data?

Note: Respondents answered in more than one category.

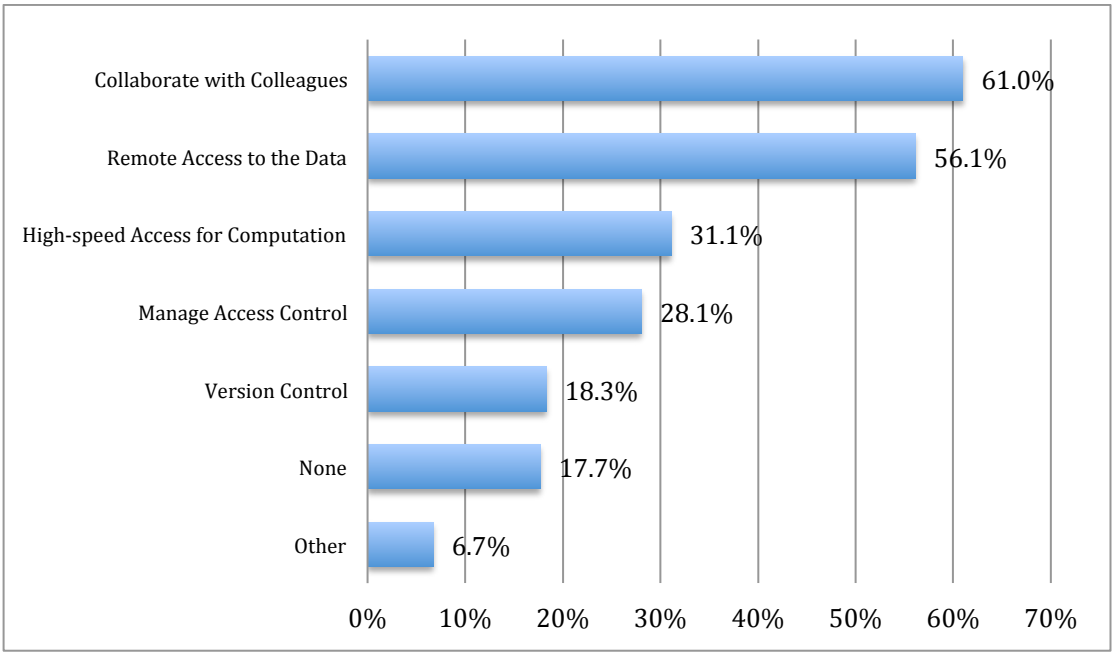

Figure 7: Requirements for working with research data

- 164 respondents answered this question.
- Some of the comments in 'other 'included:

Archive of raw data

Collaborate with clients outside Curtin

Large storage space

Wireless access

Specialised software

Lots of RAM

Normal IT services

Interface with CAD CAM machinery

Streaming video from web

For researchers, collaboration and accessibility are clearly the most important aspects of a data storage system. There is a clear trend towards an *anytime-anywhere-though any device-with anyone* approach to working with research

data. There are also a number of researchers that need to connect to their data for computational process, or with specific specialist equipment.

3.4 Access to research data

Question: Where do you need to be able to access your working research data from? (if respondent selected 'Remote Access to the Data' in section 3.3)

Note: Respondents answered in more than one category.

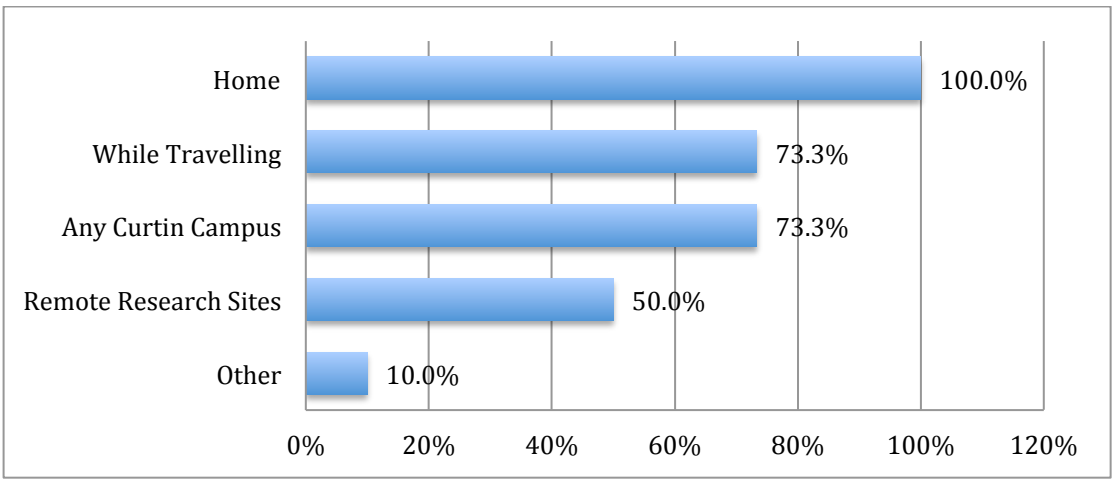

Figure 8: Remote access to data

- 30 respondents answered this question.
- Comments for 'Other' included:

  Other university campuses
  Industry partners
  24/7 worldwide

The responses in this category are not particularly surprising. All respondents want to be able to access their data from home and a large percentage need access while traveling and from remote research sites/areas. This indicates that a data store that is easily and conveniently accessible from anywhere in the world is of importance to researchers.

3.5 Collaboration

If respondent selected Collaborate with colleagues

Question: Who do you collaborate with? (If respondent selected 'Collaborate with colleagues' in section 3.3)

Note: Respondents answered in more than one category.

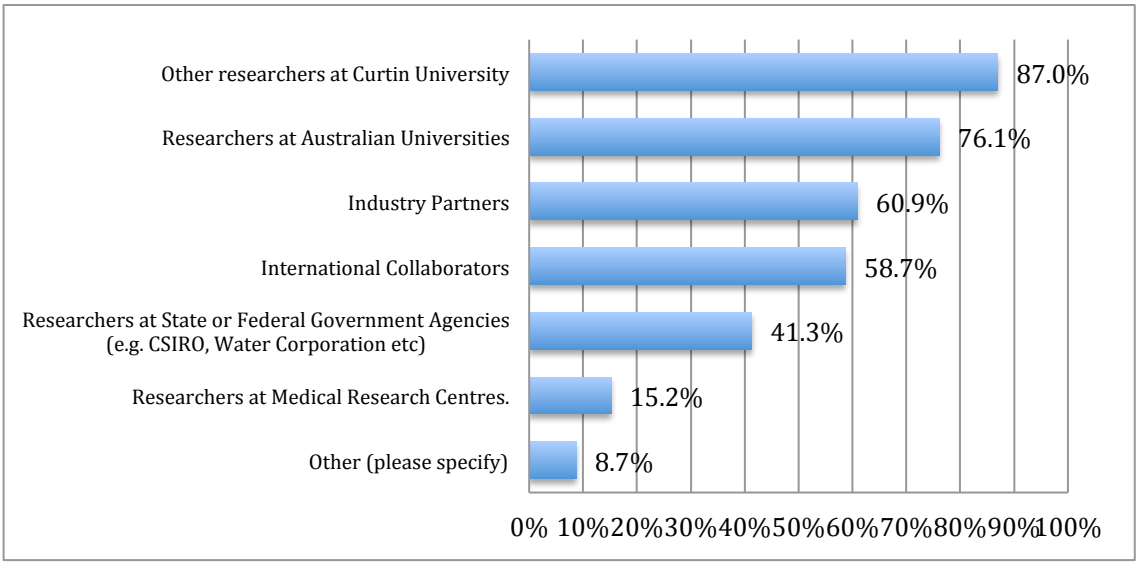

Figure 9: Collaboration

- 46 respondents answered this question.
- 86.9% of the respondents collaborate with industry partner, international collaborator, medical researchers, state or federal governmental agencies or others.
- Comments provided for 'Others' included:

  International Organizations. e.g. UNESCO (Paris)

  Micro businesses and remote communities

  Publishers - and other co -authors from international universities

4. Processing Research Data

Question: Do you use computationally intensive processes on your research data?

34.2% of the respondents (56) perform what they consider to be computationally intensive processing on their research data. The subsections below examine their computational infrastructure and needs.

4.1 Computational infrastructure

Question: What infrastructure do you use to process your research data?

Note: Respondents answered in more than one category.

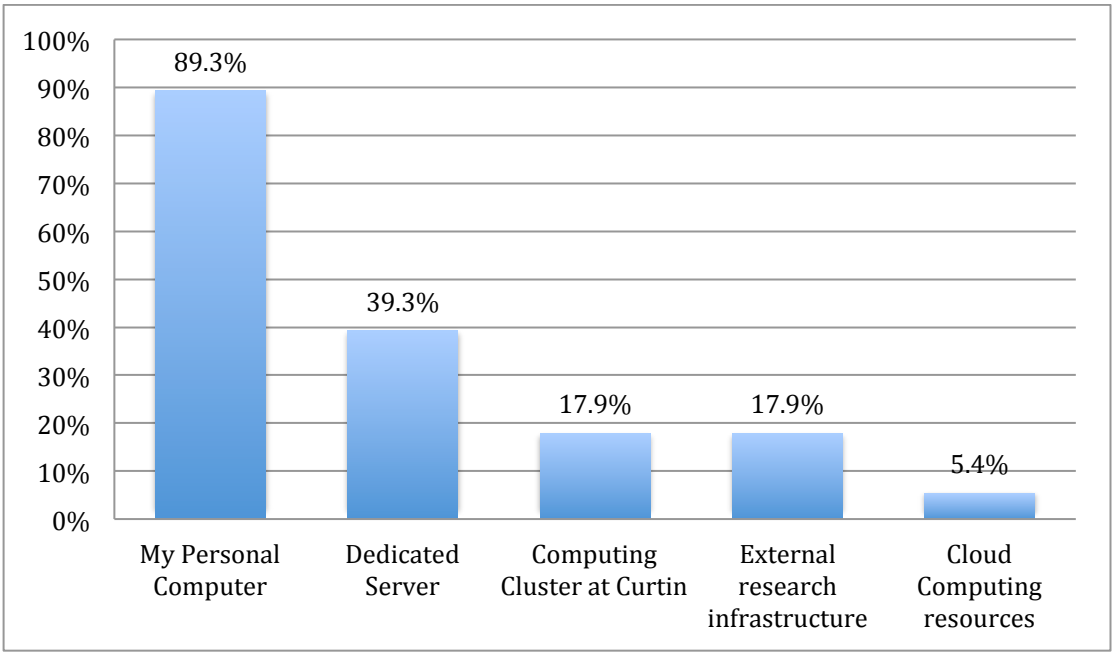

Figure 10: Processing infrastructure

- 56 respondents answered this question.
- 46.4% of respondents use only their personal computers to perform computationally intensive processing.
- Only 10 respondents (17.9%) used external research infrastructure mainly provided by iVEC and the National Computational Infrastructure (NCI).
- Only three respondents (5.4%) selected Cloud Computing resources. Accompanying comments were:

  auscope, ivec

  Nectar, Amazon EC2

  Possible in the future

Almost 90% of the respondents use their personal computer to perform computationally intensive processing. Nearly half of them use **only** their personal computers for such processing. How do they define computationally intensive processing? Nearly 20% used a mix of PC and a dedicated (on campus) research server. This indicates that there is significant potential to improve

research efficiency and reduce risk if some of this work can be moved from the desktop to a more robust and scalable solution.

4.2 Adequacy of current computational infrastructure

Question: Does your current solution meet your processing requirements?

40 respondents (71.4% of 56 respondents) are satisfied with their current computational infrastructure. Only 16 respondents (28.6% of 56 respondents) did not find their current computational infrastructure sufficient. Of those, a little more than half of them (9 respondents) used only personal computers to process their data. They were then asked: Specify why it doesn't meet your processing requirements?

Note: Respondents answered in more than one category.

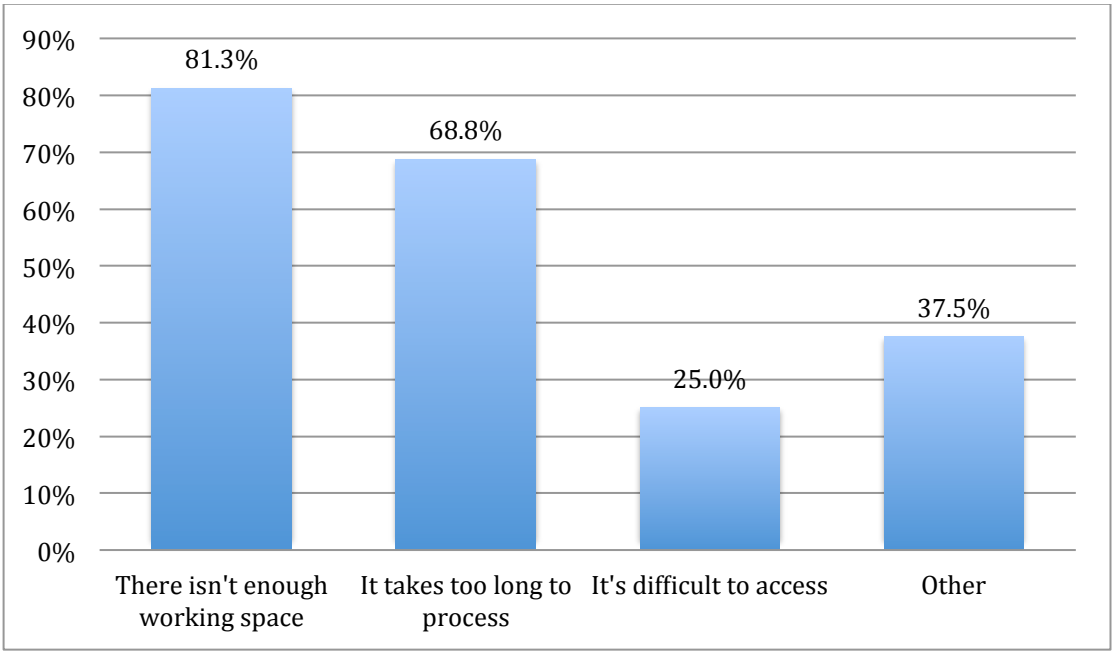

Figure 11: Current processing challenges

- 16 respondents answered this question.
- Some of the comments provided for 'others' included:

  Need collaborative workspace with high levels of security

  Not enough resources

  For my personal computer that I bought with my money, I am unable to get the latest versions of the software

  Not enough RAM available

  Is not fully functional

  My technical requirements are in excess of my knowledge.

5. Retaining Research Data

5.1 Reasons for data retention

Question: At the end of your project, what do you need to do with the data that supports your research result?

Note: Respondents answered in more than one category.

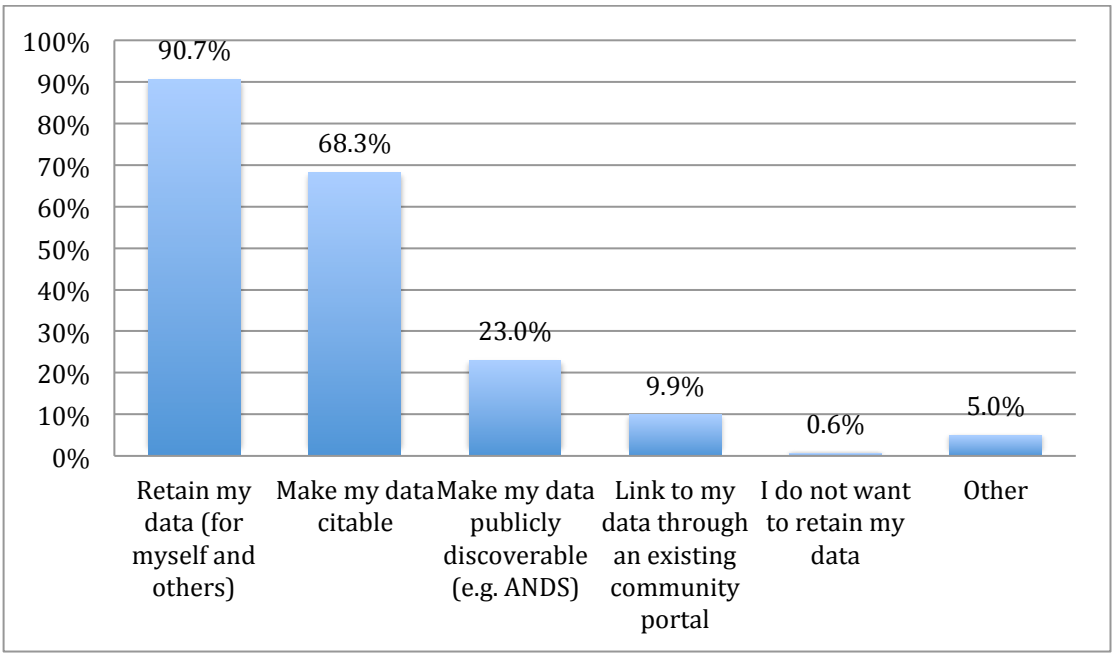

Figure 12: What to do with data after research project

- 161 respondents answered this question.
- 22.4% of the respondents want to only retain data for themselves and others.
- 42.9% of the respondents want to only retain the data for themselves (and others) and make it citable.
- 15.5% of the respondents want to make their data citable and publicly discoverable.
- Most of the comments for 'Other' were justifications on why data needs to be or might be retained:

  [Provide] CRC access

  [Data to be retained] As according to HREC conditions of approval

  No one else uses my data

  I do not wish to retain my data cancels out the other boxes, but they are not mutually exclusive. There should be a central and accessible data storage depot on campus

  Link my data to online web applications

  Retain my data as per ethics requirements specified by external sources for participant protection

5.2 Data retention period

Question: How long do you need the data that supports your research result to be retained for?

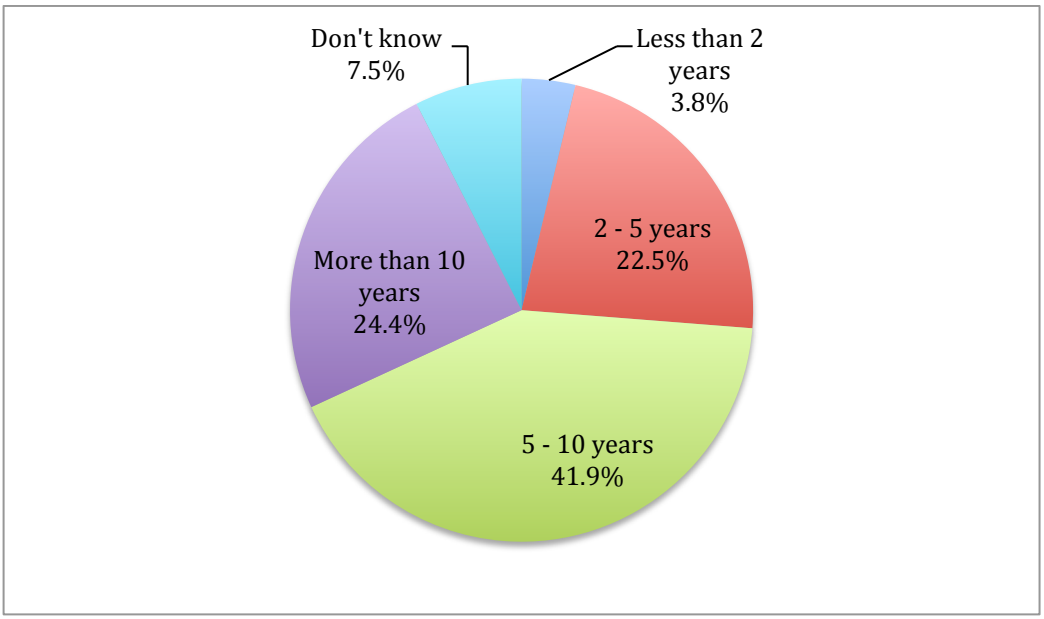

Figure 13: Retention period

- 160 respondents answered this question.

Given that the majority of respondents want to retain their data, this question tries to identify the length of time the data needs to be retained. The question captures the respondent's personal appreciation of data retention periods and in no way reflects any funding requirements or institutional data retention policies. A quarter of the respondents want to retain the data for five years or less but the majority of the respondents want to retain it for five to ten years. This question tackled the 'how long' without digging deeper to the 'why', which will highlight the true purpose of the retention period and therefore magnify its importance.

5.3 Long-term storage space

Question: How much storage space do you need for the data that supports your research result?

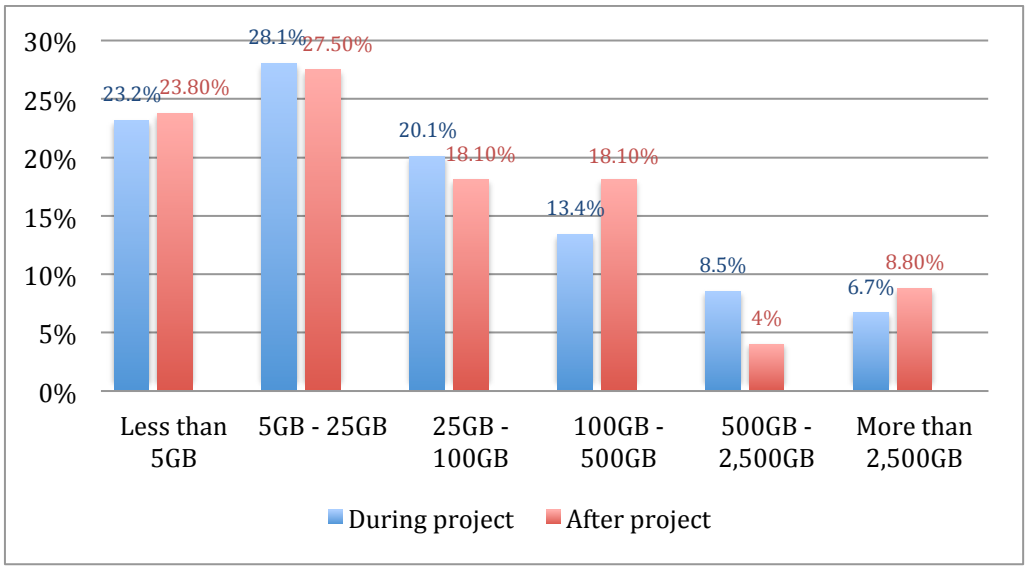

Figure 14: Storage space during and after project

- 160 respondents answered this question.

Figure 1 shows a comparison of changes in storage space needs during and after a research project. It is evident that storage space needs remain relatively constant over the course of a project. Indeed, an analysis of that data shows that 78% of the respondents have storage needs that are the same during and after a research project as depicted in Figure 16 below.

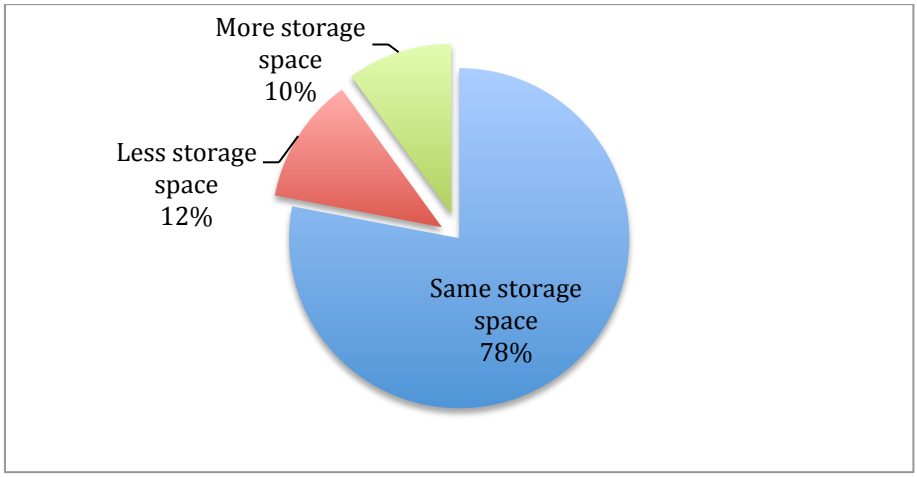

Figure 15: Change in storage space during and after project

5.4 Features of data retention storage

Question: What are the required features of the storage system that will hold this data?

Note: Respondents answered in more than one category.

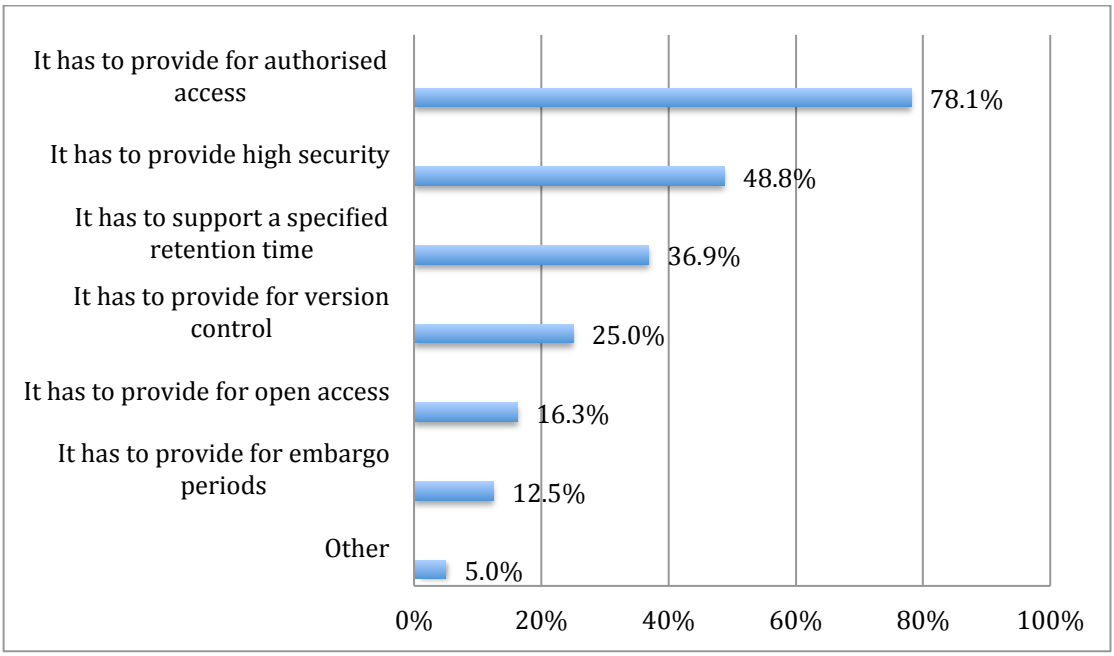

Figure 16: Features of data retention storage

- 160 respondents answered this question
- Comments given for 'Other' included:

  Rapid and reliable access
  None of the above - it just has to be available to me
  reliability and easy access
  It has to be able to be combined with data from sources on the internet
  and the current "destroy after 5 years" rule is nonsense
  Needs usage license included in metadata
  None of the above

It is evident that providing authorised access to the data and providing high security are the two most important features of such a data retention storage system. Interestingly, providing for open access to the data is not a feature highly demanded by respondents. There may be a trend towards open access in the future, as research funding bodies are showing a tendency towards requiring data to be publicly available as part of the funding conditions. There are also a growing number of publishers requiring open access to data as part of research submissions.

6. Next steps

Some of the respondents indicated interest in further discussions on their ICT needs. The CITS eResearch team is continuing that engagement and following up with the respondents.

7. Conclusion

The responses from the survey highlight a clear trend in the ICT needs of researchers. Research data storage is a primary concern for most researchers. The nature of the research work demands a number of features on storage solutions beyond local (within Curtin) access to the data. Enabling collaboration with Curtin and non-Curtin researchers, and remote access to data were the most important features. There is also a need to retain the data beyond the life of a research project. The majority of respondents highlighted that their storage needs remained the same even after the end of a research project. However the features of a data retention storage solution are different to the storage solutions used during the project. The top two features of a data retention storage solution were its ability to provide authorised access and high security for the data.

Capturing data processing needs are challenging. A researcher's perception of what comprises computationally intensive processing is relative; therefore, the survey provides a platform for further discussions which will reveal the true data processing needs.

The survey managed to capture significant information about fundamental ICT needs in the research workflow. It can form the basis of discussions with all stakeholders. It can provide invaluable insight into preliminary investigation of ICT solutions for researchers. However, it still remains a survey and will not replace a proper analysis and solution design with invested involvement from all stakeholders, especially the researchers.