The importance of tools in the data lifecycle

Malcolm Wolski, Louise Howard and Joanna Richardson
Griffith University

Abstract

Purpose – This paper outlines principal implications for institutions, particularly universities, in supporting the increasingly complex tools which are used in the data lifecycle.

Design / methodology / approach – The discussion paper draws upon the experience of the authors in this domain at the institutional, national and international level.

Findings –Support for research tools by universities has high-level implications, ranging from financial, strategic, and compliance through to capacity, capability, and connectivity. The large number of existing tools highlights the need to evaluate them against standardised checklists to determine suitability and levels of resources required for support. Librarians and other information professionals need to expand their current support for research tools beyond the discovery phase to the entire data lifecycle.

Practical implications – Universities can use this paper to assess their maturity in supporting tools in the data lifecycle. Librarians, in particular, can broaden their knowledge of the various categories of tools which support specific aspects of that lifecycle.

Originality / value – While much attention is currently being focused on supporting researchers with their data management requirements, there is a general lack of literature on how to support tools as a critical element in enhancing research outcomes.

Keywords: Research support, Research lifecycle, Research infrastructure

Article classification: General review

1. Introduction

In writing about economic growth, economists have long recognised technological advances as the “key driving force”, with innovation an important pillar (Kim and Nelson, 2000, p. 1). In 2014 the World Economic Forum outlined strategies for fostering innovative-driven entrepreneurship in Europe. The Obama Administration (White House, 2015, p. 2) updated its strategy document on innovation to “to drive economic growth and shared prosperity”. In the same year, the UK updated its policy on research and development, which has innovation as one of its cornerstones (United Kingdom, 2015), and the Australian government announced its national agenda for science and innovation (Australia, Department of Industry, Innovation and Science, 2015). The vision in the Asia Pacific region is to achieve significant innovative economic growth by 2025 (APEC, 2015).

If innovation helps to drive higher productivity growth (Jamrisko, 2016), then the research which underpins that innovation is critical to its success. As a result of increased focus, the research environment in higher education has changed significantly over the last decade. The first investment wave saw substantial resources channelled into developing research infrastructure, e.g. servers, storage, and high performance computing (HPC). However, there has been growing recognition that

“Research outputs, whether data, software, methods or publications, are critical inputs to future research and underpin innovation” (O’Brien, 2016).

This paper examines the role of tools as a critical element in enhancing research outcomes. While much attention is currently being focused on research data management, less attention is being paid to the implications for institutions, particularly universities, in supporting the increasingly complex tools which are used in the data lifecycle.

2. The importance of tools

Ten years ago, as eResearch was gaining prominence in Australia, tools were being mentioned as an integral part of the support panorama. Denison et al. (2007, p. 2) identified “a need to access diverse data sources, specialist instruments, software and other analytical tools, sample populations for surveys and trials, and specialist skills that require high-quality network access to undertake data- and simulation-intensive research”. Jarotka et al. (2006, p. 253) discussed the importance of “technologies and tools for supporting small and large scale research collaborations across time and distance”. Applebe and Bannon (2007) outlined the role of tools in establishing—and the subsequent use of—a national computing grid. For Lawrence (2006, p. 394), “the end user has to have something that is appropriate to the level of that end user and is deployable”.

In more recent years, attention has been focused on managing and leveraging the vast amounts of data now being generated for research. This has resulted in new methods, e.g. tools and compute, being developed to manipulate, analyse, process and preserve data.

In Australia, for example, the government’s *Public Data Policy Statement* (Australia, 2015) not only reinforces the importance of data but also the importance of tools: “… where possible, make data available with free, easy to use, high quality and reliable Application Programming Interfaces (APIs)”. An even more specific example is the government’s *Soil and Water – Capability Statement* (Australia, Department of Industry, Innovation and Science, 2015b, p. 1): “… developing tools for primary producers to integrate and understand data and information on soil and water from a variety of sources.”

The Australian Government has also invested in developing online environments, especially in virtual laboratories, which “draw together research data, models, analysis tools and workflows to support collaborative research across institutional and disciplinary boundaries” (NeCTAR, 2016, p. 2). This is echoed by the European Commission (Andreozzi *et al.*, 2016) in its discussion of the importance of Virtual Research Environments (VRE).

In its submission to the European Commission, the High Level Expert Group on Scientific Data (2010, p. 1) contextualised the importance of tools by emphasising that “Scientific research is supported by its infrastructures: technical *tools* and instruments and socio-economic systems for organising and sharing knowledge”. Within a proposed knowledge creation cycle for resolving society’s major challenges, Dozier (Hey *et al.*, 2009, p. 16) discusses the “... development of *new knowledge types* and *new tools for acquiring that knowledge*.” Furthermore, Ahmed (2016) notes that “Modern developmental challenges require powerful research tools, skills and orientation to ensure the production of excellent research”. Nielsen (2011) has foreshadowed “new tools for collaboration that will enable discoveries to happen at the speed of Twitter”.

Fiona Tweedie (2016, p. 3) reinforces this new thinking about tools in her statement: “Effective use of digital tools enables new avenues of research…” In the same vein, a recent editorial in the *eLife* journal (Schekman *et al.*, 2015, p. 1) announced the introduction of a new article type—called Tools and Resources—to highlight new experimental techniques, datasets, software tools and other resources. Crouzier (2016, p. 4) has found that Open Science will require "Innovative digital tools that facilitate communication, collaboration, and data analysis".

In discussing the data-intensive nature of scientific discovery, contributors to *The Fourth Paradigm* (Hey et al., 2009) noted that "The discipline and scale of individual experiments and especially their data rates make the issue of tools a formidable problem" (Bell, p. xiii) and "We have to do better at producing tools to support the whole research cycle—from data capture and data curation to data analysis and data visualization" (Gray, p. xvii). More recently, there is the issue of data provenance, particularly with regard to reproducibility and context, and hence the need for software tools to address this (Munafo, 2016).

3. What is a tool?

Formal definitions of (research) tools tend to be contextualised within specific disciplines. For example, in biomedicine and agriculture, research tools are described as:

> … any tangible or informational input required in the process of discovering a drug, a medical therapy, a diagnostic method, or a new crop variety. In short, anything that a researcher needs to use or access in the course of research—such as an assay, a genomic database, an animal model, crop germplasm and so on—may be classified as a research tool (Clift, 2007, p. 79).

Based on the authors' investigation, there is no generic definition of "tool". However, despite its non-"big science" focus, the Canadian Social Sciences and Humanities Council (SSHC) has a very useful definition, which is more extensive than just the social sciences and humanities: "… vehicles that broadly facilitate research and related activities. Social science and humanities tools enable researchers to collect, organize, analyze, visualize, mobilize and store quantitative and qualitative data and creative outputs. Tools can be created as part of a research or related undertaking, or purchased off the shelf (2014)."

The SSHC goes on to outline examples of tools eligible for funding support:

- tools that facilitate research and knowledge mobilization, such as the development of—and research costs related to—bibliographies, geographic information systems, or video gaming;
- tools related to the creation and/or cleansing of a data set, database or administrative data, or to the creation of a website;
- tools that facilitate access to holdings or collections, such as the development of—and research costs related to—repository guides, inventories, documentary materials, or special indexes; and
- standard instruments and equipment—such as computers and mobile devices—linked directly to the development of the tool, and tools that have a large research or knowledge mobilization component.

One might logically ask whether tools can be considered part of an organisation's infrastructure. As seen above, the High Level Expert Group on Scientific Data (2010) has supported this concept. In discussing information infrastructures, Borgman (2007) asserts that traditionally initiatives have tended to be oriented toward the purely technical aspects of infrastructure, i.e., infrastructure *of* information, whereas the focus should be on infrastructure *for* information, which encompasses information practices within their specific social context and discipline. In the context of this paper, tools would therefore be included within the concept of research infrastructure.

A recently released report on Australian research training has shown satisfaction levels with the level of infrastructure provided (around 80%) but significantly less with the levels of support for that infrastructure. In addition, there is a wide range between disciplines in regard to that support, e.g. 63% in education but up to 86% for commerce and education (McGagh *et al.*, 2016, p. 70). It could be argued that many eResearch tools can be regarded as infrastructure (e.g. online image processing software, a Genomics Virtual Laboratory) and the skills to use these as not only a notable skill on a researcher's CV but also a skill transferrable to industry if a researcher moves across to an industry placement.

4. Categories of tools

When considering what types of tools might be available or potentially useful, it is helpful to reflect on trends which are driving the evolution and adoption of new tools. For example:

- As research activity scales up and collaboration increases, researchers have to move from the desktop to online tools
- There are now many common, free-with-subscription option-based solutions, e.g. SurveyMonkey, Dropbox, Figshare, and FreeMapTools
- Governments and other funding agencies are investing in larger capacity, community-based research infrastructure, which bundles data, methods, tools, and systems
- Data science courses and other training programs are upskilling researchers to self-develop tools, e.g. software carpentry
- There is an increasing importance of tools to handle the large volumes of data, of which a growing percentage is collected and processed in real time.

One way to categorise tools is to look at the source. These may be either purchased off the shelf, be free to the consumer, be developed on a community basis, or be developed by an individual researcher. An alternative perspective is to look at tools from the other end of the spectrum, i.e. the user community. This helps determine service priorities and support arrangements. The tools may:

- Be used for single use, i.e. developed for a single person use (e.g. self-developed Python script or complex tables built with spreadsheet tools)
- Be used for a specific project (specific for a purpose for a group)
- Be used for a community, including potential industry use (e.g. Virtual Research Environments)
- Have impact across the broader community (e.g. SPSS or a survey tool solution or clinical trials tool)

Tools can be categorised according to where and how they are used in the research lifecycle. Willison and O'Regan (2007, p. 400) have identified six facets of the research process:

> … embark on inquiry and so determine a need for knowledge / understanding; find / generate needed information / data using appropriate methodology; critically evaluate information / data and the process to find / generate them; organise information collected / generated; synthesise and analyse new knowledge; and communicate knowledge and understanding and the processes used to generate them.

It is also useful to differentiate between tools used by researchers and those used in a laboratory setting. Laboratories often have staff responsible for managing the facilities and these staff play a role in developing and maintaining workflows within those laboratories. It is the authors' experience that many of these staff tend to be located outside the institutional library and IT organisational units.

Individual researchers, like most information workers, use a variety of tools to work with data, whether they are documents, tables, images, etc. Some of these are available through their university, but many are obtained from other sources.

Kramer and Bosman (2016) have extensively researched the use of tools by researchers. Of the 20,663 responses in a 2016 online survey, researchers accounted for 14,896 and librarians for 1,517. The average number of tools reported per person was 22 (see Figure 1).

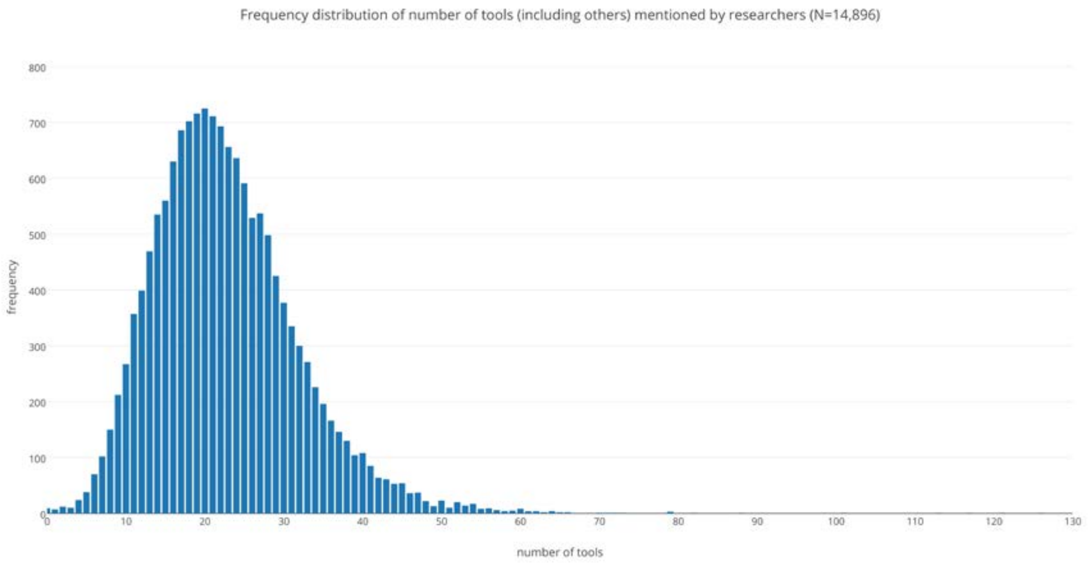

Figure 1: Frequency distribution of number of tools mentioned by researchers

Breaking down the research cycle into 30 phases, Kramer and Bosman then grouped these into seven higher level phases: preparation, discovery, analysis, writing, publication, outreach, and assessment. Their investigations have also highlighted the recent growth in tools as nominated by the researchers surveyed (see Figure 2). Their list of tools now exceeds 600. (https://101innovations.wordpress.com/about-1/).

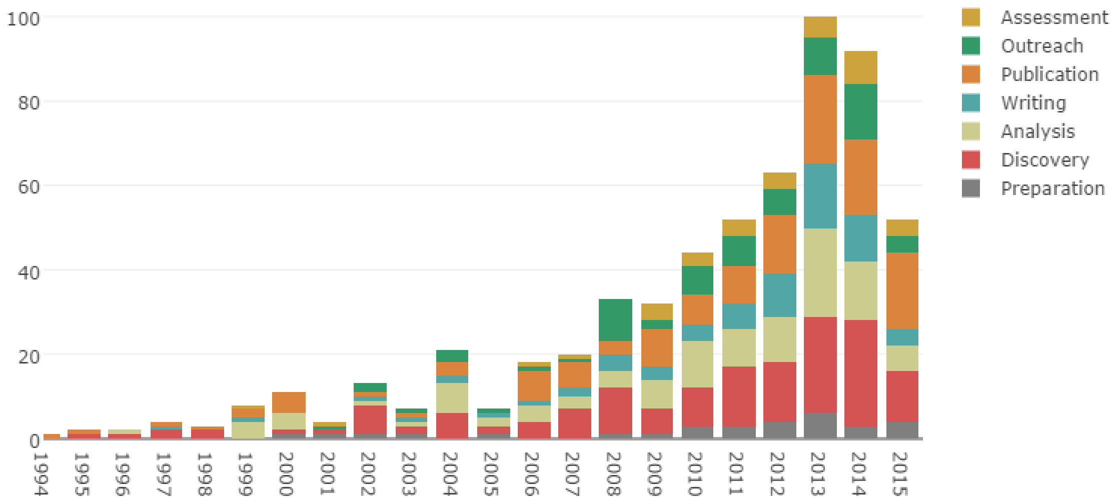

Figure 2: New tools by research phase, 1994-2015

A quick analysis of the 600+ tools shows that approximately 70% of tools are single phase tools, 20% cover two phases, and only a few cover more than one phase. The percentage of new, single use tools in the last four years is close to 80%. A possible explanation is that these were developed to solve a specific problem rather than to support a whole workflow. This high number of single phase tools is also contrary to Gray's earlier point about needing to produce tools to support the whole research cycle (Hey *et al.*, 2009, p. xvii).

This single phase tool development is not reflected in some of the larger, more feature-rich toolsets such as virtual research environments (VRE) (also referred to as virtual laboratories (VL) or Science Gateways). In Australia, for example, VLs were initially funded by national grants to address specific challenges, particularly in the areas of computation and infrastructure; however, they are now being expanded to solve a range of challenges in the data lifecycle (Lenzen *et al.*, 2014; Evans *et al.*, 2015). These may also service a different community than what is often referred to as the "long tail of research", which could make up a significant proportion of the respondents to the survey.

Although Kramer and Bosman refer to it as a *research lifecycle*, they have catered for data in their breakdown of the various phases / stages. Their breakdown of the research cycle aligns closely with what is commonly represented as the *data lifecycle* (see Figure 3).

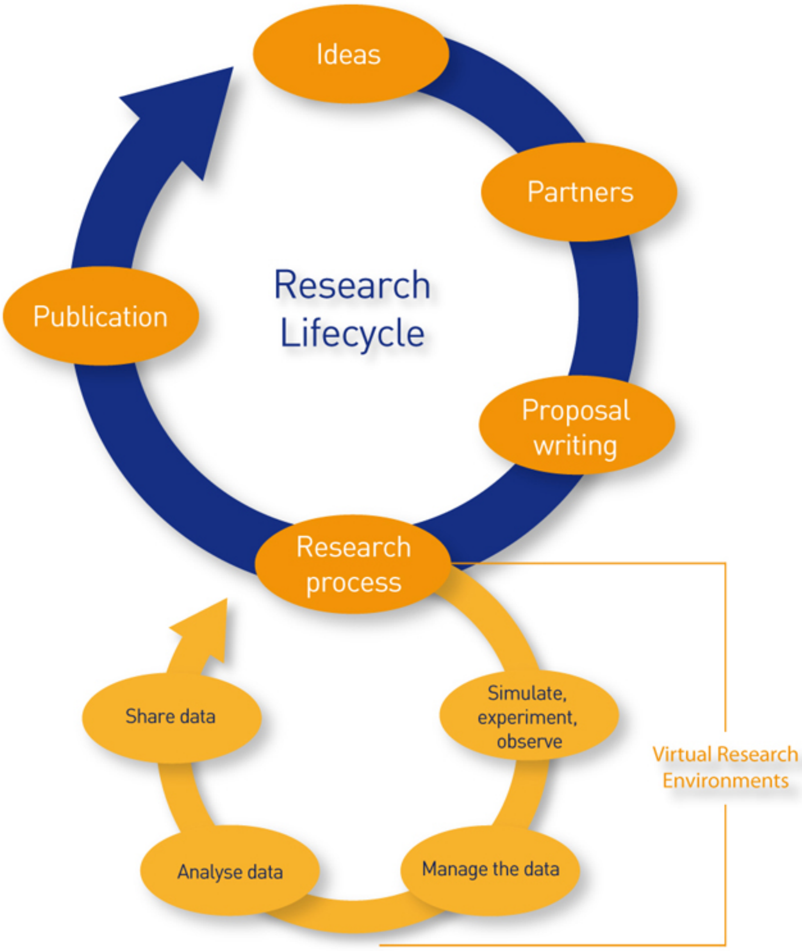

Figure 3: Research lifecycle diagram, incorporating data lifecycle

The JISC data lifecycle (Figure 3), as represented by the role of Virtual Research Environments, is reflected in the writing and analysis phases in Kramer and Bosman's survey. In these two phases Kramer and Bosman have encompassed the research activities of:

- experiment & collect/mine/extract data
- share notebooks / protocols / workflows
- analyse
- visualize
- write (+ code)
- cite
- translate

These activities are also key to the reproducibility and provenance of the research at a later date.

5. Applicability for universities

A Research Universities Futures Consortium (2012, p. 6) review of American research universities found that: “The fragility of research administration and leadership is not fully understood in the university community or by sponsors and stakeholders. As the number and complexity of research programs increase, the capacity of systems and operational support often lags, putting the institution's research enterprise at risk".

Universities have always provided and supported research tools. For example, many libraries provide research tools, technical services units have provided laboratory tools, and centres and schools have provided their own tools (e.g. purchased specific software). However, as more data is born digital and in large volumes, there is an increasing need for tools to manage, manipulate, and move data along the research lifecycle, both at the individual researcher level, the research group level, and in some cases the institutional or research community level. There are a number of reasons why university offices and faculties should take more interest in the tools being used within their institutions. These include:

- Financial implications – cost of acquisition and maintenance
- Strategic alignment – institutional support for tools that underpin strategic research priorities
- Compliance – meeting regulatory or funder requirements etc.
- Productivity – amount of researcher time spent on IT and data management related tasks rather than on actual research tasks (e.g. if data is readily accessible and re-usable, it saves time)
- Capacity – investment in tools for scaling up research, e.g. pipelines for processing larger quantities of data
- Capability – effective use of appropriate tools; development of broader eResearch skills among researchers and postgraduate candidates (McGagh et al., 2016, p. 39)
- Connectivity – ability to connect researchers with national and international research infrastructure / software activities
- Reproducibility – ability to reproduce the research produced by a tool at a later date
- Industry application - applicability of tools for industry use

In the following section, the authors will discuss their own university as an example not only of how complex tools are being used but also of some of the support implications.

6. Griffith University as a case study

Griffith University, one of the top 50 "universities under 50" in the world (QS University Rankings, 2016), is a multi-campus university of more than 43,000 students situated within Queensland, Australia. Research strengths range from the creative arts, social sciences, and humanities to science, engineering and health. Through its research and teaching, it is regarded as one of Australia's most innovative tertiary institutions and one of the most influential universities in the Asia-Pacific region.

The key features of the University's current structure were introduced in 1997, when the majority of support services were organised as centralised, multi-campus offices. As a result, the Division of Information Services (INS) is one of the most centralised and integrated information services models in Australian higher education. The Division integrates eResearch, library, and information and communication technology into a single organisation. A divisional restructure in 2016 saw the formation of seven portfolios. Of these, two have a predominantly library focus: Library and Learning Services and Information Management. eResearch Services provides access to specialist eResearch technologies as well as library and information professionals. Three portfolios cover various aspects of enterprise information technology: IT Architecture and Solutions; Enterprise Information Systems; and Information Technology Infrastructure. The seventh portfolio is responsible for planning and engagement.

While faculties and research centres have their own methods and processes for supporting research tools, INS currently supports research tools in several key ways:

1. Provides free or subsidised license to common software packages, e.g. SPSS, NVIVO, MATLAB, FLUENT, MAPINFO, ARCGIS
2. Manages/operates some national tools for all Australian researchers, e.g. Biodiversity and Climate Change Virtual Laboratory (BCCVL), QUADRANT (a participant management tool), TerraNova (a data sharing service for climate change adaptation data)
3. Facilitates access to other national tools, e.g. Genomics Virtual Laboratory, Terrestrial Ecosystem Research Network (TERN), and Australian Urban Research Infrastructure Network (AURIN)
4. Participates in developing and supporting internally funded tools for a special purpose, e.g. microscopy image management solution and a crowdsourcing application to transcribe handwritten documents.
5. Offers software carpentry workshops, which has a flow-on impact on self-developed tools

Anecdotally library and IT staff are aware that researchers are using other tools without any support from the central support services. This may stem from the fact that most of the discussion about data management, for example, within the University is between the Office for Research, the Library, the IT organisational units, and the University Executive level. As such, the technical services/facility management community tends not to be included in high-level conversations about strategy, policy and institutional infrastructure development.

In terms of the use of tools for research at Griffith, it has been observed that there are typically three categories of researchers. There are the high-end users, who are typically well resourced through funding via grants and facilities and who also have built up a solid skill base. For example, big data users are common in this group. They typically look after themselves or have the capability to acquire what they need. Then there is a middle group who have a requirement for tools but typically lack either the funding, facilities, or the skills (or any combination thereof) to obtain easily what they need. Finally, there is the third category which, because of the nature of their research, typically gets by with readily available, consumer-based products such as Excel or even common statistical packages such as SPSS. Researchers may move between these groups by virtue of winning grants or scaling up their research, or alternatively they may have low requirements locally. This latter group can be thought as the "long tail" of researchers (Bristow *et al.*, 2010, p. 18).

Researcher responses in an information-seeking exercise undertaken at Griffith University in early 2016 indicated that initial selection of tools is based on what they are familiar with or what their immediate colleagues use regularly. When the utilisation of additional research tools is considered, it is more often because researchers must meet legislative or grant requirements, rather than an ambition to leverage tools for greatest result.

Given the challenges for the University to support this range of differing needs, both eResearch Services and Library and Learning Services (within the Division of Information Services) partnered to investigate the tools currently used by Australian researchers. This was a timely exercise given the investment by the Australian Government in data, applications and infrastructure in recent years. The first step was to undertake a crowdsourcing exercise, distributed among a number of Australian research communities and library networks. The objective was to gather as much information as possible across a broad range of disciplines in relatively short timeframe. This was not intended to provide a definitive list of tools.

It was quickly discovered, however, that there were a number of lists already published on websites, many specific to a discipline or a purpose. The work by Kramer and Bosman (2016) at the University of Utrecht, for example, had already surveyed 20,000 researchers and people supporting researchers, as a result of which they had compiled a database of over 600 tools. It was therefore agreed to contribute to the work at Utrecht rather than creating another resource.

Notwithstanding, the sheer number of these tools raises a number of issues discussed below.

7. Discussion

7. 1 Governance Issues

There is a need for a holistic institutional look at the approach to supporting research in general and tools specifically. This is based on the fact that universities are responsible for enterprise-wide areas such as security, risk, compliance, financial management, and policy, which then cascade down to research centres, for example, and ultimately to individual researchers. Major funding and resourcing is also distributed at the enterprise level. There are flow-on implications for support. For example

- The evolution of cloud has removed the need to have software applications bound to the desktop reducing the need for software deployment and license management
- Services can now be acquired and purchased directly from a provider by a researcher / research group
- As a consequence, key issues which commonly need to be resolved include ownership, service availability, authentication / authorisation / access, and location of data

There is the importance of clearly articulated, and frequently updated, institutional policies to address these issues as part of a wider strategy to not only mitigate risk to the organisation but also to maximize the potential of any tools which are deployed.

Not only is governance important at the enterprise level, but it is also important at the faculty / research centre level. For reasons of sustainability, for example, a head of a research centre may dictate specific standards to be used. In addition, as members of their respective institutional community, scholars are frequently subject to regulatory and legal requirements affecting their organisation. Of course, legislative and regulatory requirements will vary depending on the relevant jurisdiction such as privacy legislation dictating where data can be stored.

There are also emerging issues such as the intellectual property rights of data surrounding data produced by equipment and software.

As daunting as this may all seem, there are opportunities to tap into national and international initiatives, as discussed later in this paper.

7. 2 Service and support issues

7.2.1 Culture and practice

Understandably researchers tend to be focused on key parts of the research lifecycle – successful grant applications and subsequent publications with significant impact, for example.   The fact that legislative requirements may influence the use of tools more than leveraging tools for greatest result may mean that researchers do not currently recognise the potential application of the broader range of research tools at other key stages of the research lifecycle. It may also explain why researchers are often more familiar with the various stages of the broader research lifecycle than with those of the more specific data lifecycle, hence the relatively new focus on data management.

Librarians have traditionally played a role in supporting researchers at specific points of the research lifecycle, particularly at the early stages of information discovery. With the growth of the use and importance of research tools throughout as well as an increasing emphasis on the data lifecycle, there is an opportunity for information professionals to extend this support to other aspects of the research lifecycle. Tenopir *et al.*, (2012) have identified the provision of data services, including management and curation, as another new role for the library. However, despite the inclusion of these responsibilities within the scope of academic librarians, there does not appear to be a corresponding increase in knowledge and awareness of the broader range of research tools.

For example, Griffith's close integration of IT and library services within the Information Services area provides a key opportunity for integrated research support activities to occur; however, this has only been extensively leveraged in a small number of cases. This has resulted in distinctly differing knowledge levels of research tools between librarians who partner closely with eResearch and IT colleagues and those who work more independently. Based on the authors' experience, it would appear that there are varying degrees of knowledge among research support librarians as to what (and when a) tool could be mapped to specific functions within the data lifecycle, let alone the more complex challenge of how tools can work together to meet a researcher's needs.

As librarians more actively engage in supporting researchers, especially in terms of their data, they may need to consider expanding their current definition of research (support) tools beyond the discovery phase to encompass the entire data lifecycle. Borchert and Young (2010) noted that researchers are now seeking training and support for collaboration tools, data management, and statistical and qualitative analysis tools.

A recent information-seeking exercise in January 2016 by Library and Learning Services with Griffith researchers across a cross-section of disciplines confirmed that the latter have similar needs, identifying learning about available tools and platforms and how to use them as one of their highest training needs. The same exercise explored the use of various research tools within Griffith and, as such, also highlighted the limited knowledge of research tools, and the limited awareness of the need for such knowledge, within some librarians supporting research. In response to these findings, and to the demand within the research community, cultural change around attitudes towards knowledge of tools supporting the data lifecycle is beginning to occur. This is evident in activities such as all Griffith research support librarians undertaking the Australian National Data Services' training program *'23 Research things'*, which includes a focus on data tools. Additionally, a number of Griffith's Library and Learning Services staff have undertaken data carpentry courses to extend their knowledge of the full lifecycle of data-driven research. Further engagement with research tools and how they map to both the data and research lifecycle is needed in order to better establish this aspect of research support.

A pilot undertaken in early 2016 with a large Griffith research centre to examine research data practices of staff found that researchers identified a preference for largely traditional tools that integrated well with each other and enabled them to move their data easily between systems. An additional program of work is underway, through the Griffith Graduate Research School, to examine the breadth of tools currently in use by Griffith Higher Degree Research candidates, with a particular focus on understanding the multi-disciplinary application of tools in use. These parallel activities demonstrate an institutional understanding of the importance of how research tools are implemented and managed.

In their investigation of how behaviour influences data management practices, Wolski and Richardson (2015) have examined some of the factors which have an impact on the adoption of technology by researchers. These include the time which has elapsed since graduate training as well as local group norms.

7.2.2 Support

There are also a number of practical issues that universities need to consider in terms of the type and level of services to support research tools.

Not every tool needs a high level of reliability /robustness and sometimes "good enough" is acceptable. There are reasons as to why some tools may require higher attention and support range across a number of areas. The first consideration is compliance to regulatory or legal requirements. This may dictate the need for services to ensure requirements are met and risks have been minimised.

Another reason is the increasing drive towards the whole data lifecycle approach to information management. Given that a number of systems are usually involved across the data lifecycle, interoperability becomes important. Adopting data and technical standards are typical methods for ensuring interoperability. While data librarian services can help with the former, more services will be needed to assist with developing tools to meet IT standards both in design and development. This could range from services and support targeted at individual researchers (e.g. software carpentry, code repositories) to providing group solutions (e.g. data workflows within a laboratory to building a data bank). Of note is that in this last example, technical service staff/lab managers are a key stakeholder who should play an important role in data management in future as they will not only manage technical infrastructure, but also potential data infrastructure.

A more general question is whether the tools developed will be required to reproduce the results at some later date. If so, this raises other issues about how to cite/reference the tool and what services/support is available to archive the tool with the research data at some point so that the citation/reference is still valid. From a research quality point of view, it is also useful to know if there is documented evidence the tool in question has been through a quality assurance process in relation to the methods and processes used, e.g. peer review.

Another consideration is the number of users; this becomes complex if it is a community-based tool used by researchers across institutions. Clinical trials tools are examples of community-based tools which can be provided in-house or by utilising a service provided by another institution. Other tools, such as survey tools, can be developed in-house or by purchasing commercial services. The numbers of users within the institution is usually one determinant of the level and type of support; however, identifying those users within the institution utilising external services/tools is an ongoing problem, as most tool owners or the researchers' institutions either do not capture that information or will not share it.

The database of 600+ tools highlighted the scale of the problems discussed above. Two topics of interest are to determine a) which of these 600+ could be recommended to researchers and b) which of these could be used in combination for specific research domains to provide a workflow for data.

The fact that a library recommends external services and tools implies that its university has a certain level of trust in those tools or services. Libraries and IT service units within the institution strive to build "trusted" services through aspects such as confidence in their reliability and availability, commitment to sustaining them over the long term, providing defined services levels and dealing with support or conflict issues, and communicating about their use and changes to the services. This raises the question of how institutions apply similar criteria against external tools before they recommend them.

In a research setting, there are additional considerations. Confidence in a tool or service should result in information demonstrating that (1) there is evidence of some quality assurance on the outputs and, in particular, the underpinning research methods used (e.g. a peer review or equivalent community review); (2) it is citable; and (3) hopefully it has at least minimal version control for provenance and reproducibility purposes.

A key finding from Kramer and Bosman's work is that researchers not only use many tools but they also use them in combination, presumably to develop ad hoc workflows to suit their specific research needs. One consideration for institutions providing or supporting such tools is to look at common workflows or common elements of workflows rather than simply providing and supporting a range of tools as standalone solutions.

Support services can range from raising awareness and facilitating training, supporting applications, and providing advice to being an advocate for institutional researchers and representing the institution and research communities in nationally funded development programs.

7. 3 Resource issues

Similar to research data management, resourcing for tools has the same requirements in determining the roles and responsibilities of different units within the organisation. The growth in tools has, not surprisingly, coincided with the growth in the volume of data.

While institutions have responded with resources to the data management problem with infrastructure (e.g. storage services, repositories) and other services (e.g. data librarians), the same approach to support for tools has not been as evident. National funding programs such as the National Collaborative Research Infrastructure Strategy (NCRIS) in Australia have addressed this at the national level; however, within institutions the response has not been as coordinated or planned. Community-based responses are becoming more evident, ranging from fully developed tools (e.g. Omeka) to self-help (e.g. software carpentry courses).

Institutional resourcing to provide sustainable solutions is an ongoing issue as is identifying the responsible group within the institution to provide such support. In many cases, the tools in use are utilised by researchers from multiple institutions because of research collaboration. So, in effect institutions hosting tools may be providing a level of support for the collaborating researchers as well as their own researchers.

Along with its very useful definition of tools outlined above, the SSHC (Canada, 2014) provides a checklist which specifically targets the acquisition of any tools to support a project funding bid. This same checklist clearly has wider applicability when considering the level of resourcing and support services within the institution. Adopting their checklist provides useful criteria to determine the level of resources and support. The checklist asks:

- How will the tool meet the goals of the proposed project?
- Will the tool be integrated with and support the strategic plans of the centre/institution?
- In what way will the tool be a unique resource for the research community and what are the benefits?
- How could the tool have an impact across the wider research community?
- To what extent is the tool standards-based and interoperable?
- Does the tool have a clear purpose and audience?
- Is there a longer-term plan for sustaining the tool beyond its creation and initial use to meet the goal(s) of the proposed project?
- Does the tool need to be built or will it be purchased off the shelf?

If institutions respond by including tool development as part of their investment in research infrastructure, then these criteria would be a very useful start.

In terms of resourcing, another important driver is open science / open data. The European Commission, in particular, is driving a program of work in this domain. Many challenges need to be addressed, such as infrastructure, intellectual property rights, and content-mining, as well as inter-institutional, inter-disciplinary and international collaboration. The Facilitate Open Science Training for European Research (FOSTER) (https://www.fosteropenscience.eu/foster-taxonomy/open-science-tools) project has a portal that lists tools which can assist in the process of delivering and building on open science. It is, therefore, important for institutions to consider any additional criteria which may be required to ensure that relevant tools are compliant with open science principles, if this is a priority for the institution.

7. 4 National and international drivers and opportunities

Universities form part of a larger knowledge ecosystem. As already mentioned, research is undertaken not only at the local and national level but also at the international level. For example, as Wolski and Richardson (2014, p. 90) have observed, “Central planners need to regard their infrastructure as a node in a global IT ecosystem rather than as just local, physical infrastructure built only for use within their own institution”. At the time of writing, the Australian government has released its *National Research Infrastructure Capability Issues Paper* (Australia, 2016), in which it is acknowledged that research outcomes are becoming increasingly driven by access to complex research infrastructure (including analytical tools and services). As elsewhere in the world, it is important for the government to determine the key national research infrastructure needs “in a rapidly evolving global environment” (p. 4). Indeed, “maximising Australia’s investments in research infrastructure often requires linking into international projects and consortia” (p.8).

Institutions should support opportunities to participate in national and international initiatives, which are tackling the high-level challenges discussed in this paper. The Research Data Alliance (https://rd-alliance.org/), for example, has a number of working and interest groups, which span all aspects of research data support. The work being undertaken by the FORCE11 Scholarly Commons Working Group (https://www.force11.org/group/scholarly-commons-working-group) on a scholarly communications ecosystem has produced the crowd-sourced database of tools discussed earlier in this paper.

Work done by the Broad Institute of Harvard and MIT on the GenomeSpace environment (http://www.genomespace.org/) exemplifies the type of development of tools being undertaken, into which other researchers can tap, without having to develop this functionality in-house (Broad Institute, 2012).

Tools, for their part, continue to evolve because of drivers such as advances in technology and changes in research methods. Therefore, there is a requirement for open standards to move data between tools within and between discipline communities and to allow tools themselves to interoperate. The Horizon 2020 Open Research Data Pilot (https://www.fosteropenscience.eu/content/horizon-2020-open-research-data-pilot) is investigating standards as part of its work in making research data openly accessible.

As a corollary, while one refers to “trusted data”, trusted services are as yet unexplored territories. In collaborating with colleagues external to one’s organisation, what assurance is there that the tools will operate as expected/stated? What is their ongoing support/sustainability in a cross-institutional environment? And what assurance does one have that the underlying method or process has been through some assurance, e.g. peer review, process? There is considerable scope for additional collaborative, international initiatives in this space.

8. Conclusion

In examining the important role which tools play in the data lifecycle, it is apparent that their sheer number and complexity highlights the need to evaluate them against well-established criteria so as to determine not only suitability but also levels of resources for support (if adopted). As institutions, especially universities, transition from a focus on research data management to research tools (or research cycle or research data lifecycle) management, there are some areas worthy of additional investigation.

In establishing evaluation criteria, the authors have suggested that an important criterion is that of the trustworthiness of a tool. Either nationally or at the research community level, some work needs to be done to determine what constitutes a "trusted tool". The purpose would be to provide sufficient confidence in a tool's functionality so that its outputs may be trusted.

The results of a 2016 survey have shown the use of large numbers of single phase tools, but no evidence of workflows based on them. It is known that researchers use tools in combination, but there is no evidence of interoperability between those tools. Therefore librarians and other support staff need to have a better understanding of their researchers' activities in order to determine not only what tools they are using but also to identify any common workflows. This knowledge will help underpin support for discipline needs within a university's strategic research groups and will be a useful input into infrastructure development and resource allocation.

Finally, the authors have advocated for institutions participating in national and / or international initiatives which are working on high-level challenges relating to the data lifecycle. Active involvement lifts the level of knowledge within the institution's support services and research community.

References

Ahmed, A. (2016), "Supporting research excellence", *University World News*, issue 402, available at: http://www.universityworldnews.com/article.php?story=20160223220411616 (accessed 13 September 2016).

Andreozzi, S., Burgueño Arjona, A., Campos, I., Coelho, S., Dappert, A., Garavelli, S. et al. (2016**),** *E-Infrastructures: Making Europe the best place for research and Innovation*, European Union, Luxembourg.

Appelbe, B. and Bannon, D. (2007), "eResearch-paradigm shift or propaganda?", *Journal of Research and Practice in Information Technology*, Vol. 39 No. 2, pp. 83-90.

Asia-Pacific Economic Cooperation (2015), *Policy Partnership on Science, Technology and Innovation Strategic Plan (2016-2025)*, APEC, Singapore, available at: http://mddb.apec.org/Documents/2015/PPSTI/PPSTI2/15_ppsti2_004.pdf (accessed 13 September 2016).

Australia (2015), "Australian government public data policy statement", available at: https://www.dpmc.gov.au/resource-centre/data/australian-government-public-data-policy-statement (accessed 12 September 2016).

Australia. Department of Industry, Innovation and Science (2015a), *National Innovation & Science Agenda*, The Department, Canberra, available at: http://www.innovation.gov.au/system/files/case-study/National%20Innovation%20and%20Science%20Agenda%20-%20Report.pdf (accessed 12 September 2016).

Australia. Department of Industry, Innovation and Science (2015b), *Science and Research Priorities. Soil and Water – Capability Statemen*t, The Department, Canberra, available at: http://www.science.gov.au/scienceGov/ScienceAndResearchPriorities/Pages/Soil-and-water.aspx (accessed 12 September 2016).

Australia. National Research Infrastructure Roadmap Expert Working Group (2016), "National Research Infrastructure Capability Issues Paper", available at: https://www.education.gov.au/2016-national-research-infrastructure-roadmap (accessed 12 September).

Borgman, C.L. (2007), *Scholarship in the Digital Age*; MIT Press, Cambridge, MA.

Bristow, R., Dodds, T., Northam, R. and Plugge, L. (2010), "Cloud computing and the power to choose", *Educause Review*, Vol. 45 No. 3, pp. 14-31.v

Broad Institute of Harvard and MIT (2012), "Researchers announce GenomeSpace environment to connect genomic tools", available at: https://www.broadinstitute.org/news/4129 (accessed 12 September 2016).

Canada. Social Sciences and Humanities Council (2014), *Guidelines for Support of Tools for Research and Related Activities*, SSHC, Ottawa, available at: http://www.sshrc-crsh.gc.ca/funding-financement/policies-politiques/support_tools_soutien_outils-eng.aspx (accessed 12 September 2016).

Clift, C. (2007), "Patenting and licensing research tools", in Krattiger, A., Mahoney, R. T., Nelsen, L., Thomson, J. A., Bennett, A. B., Satyanarayana, K., ... and Kowalski, S. P. (Eds.), *Intellectual Property Management in Health and Agricultural Innovation: A Handbook of best practices*, MIHR, Oxford, U.K, pp. 79-88.

Crouzier, T. (2016), *Science Ecosystem 2.0: how will change occur?*, European Union, Luxembourg.

Denison, T., Kethers, S. and McPhee, N. (2007), "Managing the soft issues in e-research: a role for libraries?", *Australian Academic & Research Libraries*, Vol. 38 No. 1, pp. 1-14. doi: 10.1080/00048623.2007.10721263.

Evans, B., Wyborn, L., Pugh, T., Allen, C., Antony, J., Gohar, K. et al. (2015), "The NCI high performance computing and high performance data platform to support the analysis of petascale environmental data collections", in Denzer, R., Argent, R., Schimak, G. and Hřebíček, J. (Eds.), *Environmental Software Systems. Infrastructures, Services and Applications*, Springer International Publishing, Switzerland, pp. 569-577. doi: 10.1007/978-3-319-15994-2_58.

Hey, T., Tansley, S., and Tolle, K. (2009), *The Fourth Paradigm: Data-Intensive Scientific Discovery*, Microsoft Research, Redmond, WA.

High Level Expert Group on Scientific Data (2010), *Riding The Wave - How Europe Can Gain From The Rising Tide of Scientific Data. A Submission to the European Commission*, European Commission, Luxembourg, available at: http://ec.europa.eu/information_society/newsroom/cf/document.cfm?action=display&doc_id=707 (accessed 12 September 2016).

Jamrisko, M. (2016), “These are the world's most innovative economies”, *Bloomberg Markets* (magazine), 19 January, available at: http://www.bloomberg.com/news/articles/2016-01-19/these-are-the-world-s-most-innovative-economies (accessed 12 September 2016).

Jirotka, M., Procter, R., Rodden, T. and Bowker, G. C. (2006), “Special issue: collaboration in e-research”, *Computer Supported Cooperative Work (CSCW)*, Vol. 15 No. 4, pp. 251-255. doi: doi:10.1007/s10606-006-9028-x.

Kim, L. and Nelson, R. R. (2000), *Technology, Learning, and Innovation: Experiences of Newly Industrializing Economies*, Cambridge University Press, Cambridge, UK.

Kramer, B and Bosman, J. (2016), “Innovations in scholarly communication - global survey on research tool usage [version 1; referees: 2 approved], “*F1000Research*, Vol. 5:692. doi: 10.12688/f1000research.8414.1.

Lawrence, K. A. (2006), “Walking the tightrope: The balancing acts of a large e-research project”, *Computer Supported Cooperative Work (CSCW)*, Vol. 15 No. 4, pp. 385-411. doi: 10.1007/s10606-006-9025-0.

Lenzen, M., Geschke, A., Wiedmann, T., Lane, J., Anderson, N., Baynes, T., ... & Hadjikakou, M. (2014), “Compiling and using input–output frameworks through collaborative virtual laboratories”, *Science of the Total Environment*, Vol. 485, pp. 241-251. doi: 10.1016/j.scitotenv.2014.03.062.

McGagh, J, Marsh, H, Western, M, Thomas, P, Hastings, A, Mihailova, M. and Wenham, M (2016), *Review of Australia’s Research Training System. Report for the Australian Council of Learned Academies*, ACOLA, Melbourne, available at: www.acola.org.au (accessed 12 September 2016).

McKen, K., Pink, C., Lyon, L. and Davidson, M. (2012), “Research360: data in the research lifecycle”, available at: http://opus.bath.ac.uk/32292/ (accessed 12 September 2016).

Munafo, M. (2016). “Scientific ecosystems and research reproducibility”, talk presented at Research Libraries UK Conference, 9-11 March, London, available at: https://www.youtube.com/watch?v=TD2cUYVci28&feature=youtu.be (accessed 12 September 2016).

NeCTAR (2016), “Impact: collaborate connect contribute”, available at: https://nectar.org.au/wp-content/uploads/2016/04/NEC_004-Codex_230x260_FINAL-WEB.pdf (accessed 12 September 2016).

Nielsen, Michael (2011), “Open science now!” TEDx Waterloo, available at: http://www.ted.com/talks/michael_nielsen_open_science_now (accessed 12 September 2016).

O'Brien, L. (2016), "Remove the barriers to innovation by increasing access to research", *The Machinery of Government*, 11 March, available at: https://medium.com/the-machinery-of-government/remove-the-barriers-to-innovation-by-increasing-access-to-research-5329e2f4c6a2? (accessed 12 September 2016).

QS [Quacquarelli Symonds Limited] (2015), "QS University Rankings: Top 50 Under 50 2016-2017", *Top Universities*, available at: http://www.topuniversities.com/university-rankings/top-50-under-50/2016 (accessed 26 September 2016).

Research Universities Futures Consortium (2012), *The Current Health And Future Well-Being of The American Research University*, Elsevier, Philadelphia.

Schekman, R., Weigel, D. and Watt, F. M. (2015), "Recognizing the importance of new tools and resources for research", *eLife*, Vol. 4:e07083. doi: 10.7554/eLife.07083.

Tweedie, F. (2016), The data scientist, *share*, No. 24, p. 3, available at: http://www.ands.org.au/news-and-events/share-newsletter/share-24/the-data-scientist (accessed 12 September 2016).

United Kingdom. Department for Business, Innovation & Skills (2015), *2010 to 2015 Government Policy: Research and Development*, The Department, London, available at: https://www.gov.uk/government/publications/2010-to-2015-government-policy-research-and-development (accessed 12 September 2106)

United States. National Economic Council and Office of Science and Technology Policy (2015), A *Strategy for American Innovation*, The White House, Washington, DC.

Willison, J. and O'Regan, K. (2007), "Commonly known, commonly not known, totally unknown: a framework for students becoming researchers", *Higher Education Research & Development*, Vol. 26 No. 4, pp. 393-409. doi: 10.1080/07294360701658609.

Wolski, M. and Richardson, J. (2014), "A model for institutional infrastructure to support digital scholarship", *Publications*, Vol. 2 No. 4, pp. 83-99. doi: 10.3390/publications2040083.

Wolski, M., & Richardson, J. (2015) "Improving data management practices of researchers by using a behavioural framework", THETA 2015, 11-13 May, Gold Coast, Australia, available at: http://hdl.handle.net/10072/69141

World Economic Forum (2014), *Enhancing Europe's Competitiveness: Fostering Innovation-Driven Entrepreneurship in Europe*, WEC, Geneva.