Research Tools: Important Drivers for Innovation in Research Repository Architecture

Malcolm Wolski*, Joanna Richardson**

Abstract

Online tools are critical to undertake successful research. Researchers use software tools as an integral part of research to process, manage, and integrate data from multiple sources. However, while institutional repositories are tackling challenges around supporting research data, little attention has been paid to the implications for repositories in supporting the increasingly complex tools which are used in the data lifecycle. Tools and workflows can play an important role in building quality repositories. This increasing use of tools has implications not only for researchers but also the institutions who manage those repositories. This paper suggests strategies for institutional stakeholders, particularly libraries, on how to implement solutions which will ensure interoperability at all levels of research repository architecture.

Keywords: Institutional Repositories, Research Data Lifecycle, Metadata Standards, Data Workflows, Research Ecosystem, Scholarly Communication Tools

Introduction

In their origins, institutional repositories had a key role in managing and disseminating the scholarly outputs of their community members, principally publications and subsequently learning objects. However, as Johnston (2012) wrote in her report on the Open Repositories 2012 conference, "It has become clear that institutional repositories must not only manage scholarly publications, but [also] the data that was created through observation and experimentation or collected and published, in order to support the 're-' activities: review, reuse, replicability, and reproducibility".

In more recent years, attention has indeed 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, being developed to manipulate, analyse, process, and preserve data. Tools, for their part, are being regarded as an integral part of the research support environment. A number of key government documents internationally support this tenet, ranging from the High Level Expert Group on Scientific Data (2010), the Canadian Social Sciences and Humanities Council (Canada, 2014), and Australia's Public Data Policy Statement (Australia, 2015) to the European Commission (Andreozzi *et al*., 2016).

However, while institutional repositories are tackling challenges around supporting research data, little attention has been paid to the implications for repositories in supporting the increasingly complex tools which are used in the data lifecycle. This paper is intended to fill that gap.

Methodology

The authors undertook a brief, critical review of the literature. According to Webster and Watson (2002, p. xiv), a literature review is appropriate, for example, when investigating "an emerging issue that would benefit from exposure to potential theoretical foundations. The author's contribution would arise from the fresh theoretical foundations proposed in developing a conceptual model." In their typology of reviews, Grant and Booth (2009, p. 94) have defined a critical review in terms of several key attributes: typically narrative and typically resulting in a hypothesis or model. The purpose of this type of review is to compare and evaluate a number of perspectives.

* Director, eResearch Services, Griffith University, Brisbane, Australia. Email: m.wolski@griffith.edu.au
** Library Strategy Advisor, Griffith University, Brisbane, Australia. Email: j.richardson@griffith.edu.au

While the authors did not develop a model as the end product, they did use the critical review to identify potential strategies for addressing the topic of this paper.

Related Research

In examining the literature on this topic, the authors have focused on three main areas: institutional repositories, research data lifecycle, and research tools.

Institutional Repositories

The purpose for hosting an institutional repository has changed over the past fifteen years. Institutional repositories evolved from a need to archive and preserve scholarly materials (Johnson, 2002), specifically research publications (Jain, 2011; Crow, 2002). Crow (2002, p. 3) described it as an application scholarly in scope, in which the content is "… cumulative and perpetual; and open and interoperable". Lynch (2003, p. 238) expanded the scope of institutional repository beyond just an archive application, describing it as a set of services providing a stewardship role.

However, more recently Lynch (Poynder, 2016, p. 13) has expanded his thinking, "Technology has moved on quite a bit in the last fifteen years, and it may be that it makes more sense to think about how to do this in a way that involves more shared or collective platforms and services rather than highly distributed approaches". He goes on to foreshadow, "No question that there are going to be commercial software suppliers and also cloud hosted solutions for many, perhaps all, of the IR functions, and some of these are or will be excellent choices for many institutions" (Poynder, 2016, p. 15).

For the purpose of this paper, in theory any repository that holds institutional research assets, e.g. data, is an institutional repository, even though it may not be centrally supported.

Research data repositories are difficult to build because of a number of key factors as shown in Fig. 1.

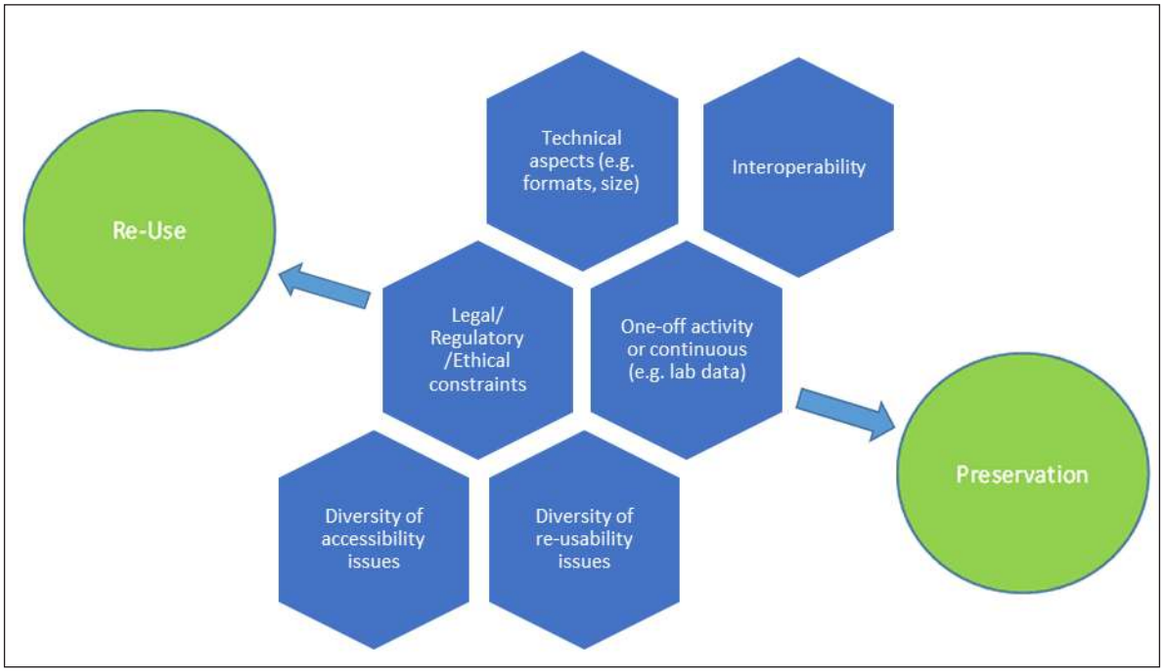

Fig. 1: Factors Affecting Research Data Repositories

The factors include, but are not limited to:

- Technical aspects: e.g. file size, file type, speed, diversity.
- Legal /regulatory aspects: intellectual property, licenses, other restrictions.
- A "collection" is all the artefacts collected and created as a result of a research activity (e.g. a research project or laboratory).
- A "collection", or parts of it, may be an input to another activity or part of a larger collection.

- Re-usability issues: provenance and quality of the data, software versions, trusted data services (data and tools), security, privacy, ethics.
- Accessibility issues: data must be in an accessible format and fit for purpose.
- Interoperability/ standards issues: sharing and machine-to-machine applications.

Whereas Faundeen (2017) discusses these factors in regard to building trusted digital repositories, Wolski, Howard, and Richardson (2017a) have explored in depth some of these factors in the context of online research data services.

Research Data Lifecycle

Much has been written in the literature not only about the research lifecycle but more especially about the need for institutions to support researchers as they progress through the entire cycle (Pryor, 2012; Rice & Haywood, 2011; Sergeant, 2006). In recent years, the proliferation of data, especially so-called big data, has given rise to the concept of the (research) data lifecycle (Stall, 2016; Wissik & Ďurčo, 2016; Gundersen, 2016).

Fig. 2 is an excellent representation by Jisc of this lifecycle.

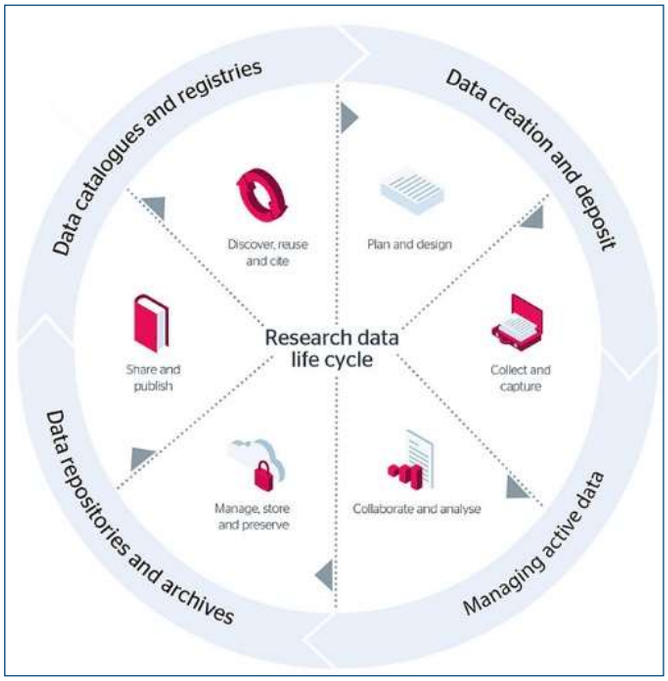

Fig. 2: Research Data Lifecycle (www.jisc.ac.uk)

The authors' underlying premise is that one needs to think of the research environment as a research "ecosystem", as epitomised in this classic illustration (Fig. 2) by Jisc of the research data lifecycle. Because, if not careful, institutions have a tendency to focus too much on objects and on end points rather on than the dynamics of the various interrelationships within the broader environment. It is within this context that research tools are quickly becoming an important focal point.

Tools

While it may be argued that there can be no data without supporting tools, there is, unsurprisingly, no single definition of what constitutes a research tool. Clift (2007, p. 79), for his part, has contextualised a definition based on the specific disciplines of biomedicine and agriculture: "… 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".

However, based on previous research, the authors have used the definition provided by the Canadian Social Sciences and Humanities Council (Canada, 2014), which can be applied more extensively than to just the social sciences and humanities: "… vehicles that broadly facilitate research and related activities. Social science and humanities tools enable researchers to collect, organise, analyse, visualise, mobilise 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".

The importance of tools as an integral part of the research environment is receiving increasing attention in the literature. For his part, Ahmed (2016) notes that modern developmental challenges require powerful research tools, skills and orientation to ensure the production of excellent research. Within a proposed knowledge creation cycle for resolving society's major challenges, Dozier and Gail (2009, p. 16) discuss the development of new knowledge types and new tools for acquiring that knowledge. Nielsen (2011) has foreshadowed new tools for collaboration that will enable discoveries to happen at the speed of Twitter. Crouzier (2016, p. 4) has found

that Open Science will require innovative digital tools that facilitate communication, collaboration, and data analysis.

A strong driver is the issue of data provenance, and specifically context and reproducibility. As Munafo (2016) notes, there is a need for software tools to address these issues.

Researchers use a variety of tools to work with data, some of which are available through their institution, while others are obtained from other sources. The seminal work done by Kramer and Bosman (2016) provides a useful starting point for appreciating the breadth of the challenge. Of the 20,663 responses to their online survey, researchers accounted for 14,896 and librarians for 1,517. The average number of tools reported per person was 22.

Kramer and Bosman have then divided the research cycle into 30 phases, from which they have created 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 Fig. 3). Their list of tools now exceeds 600. (https://101innovations.wordpress.com/about-1/).

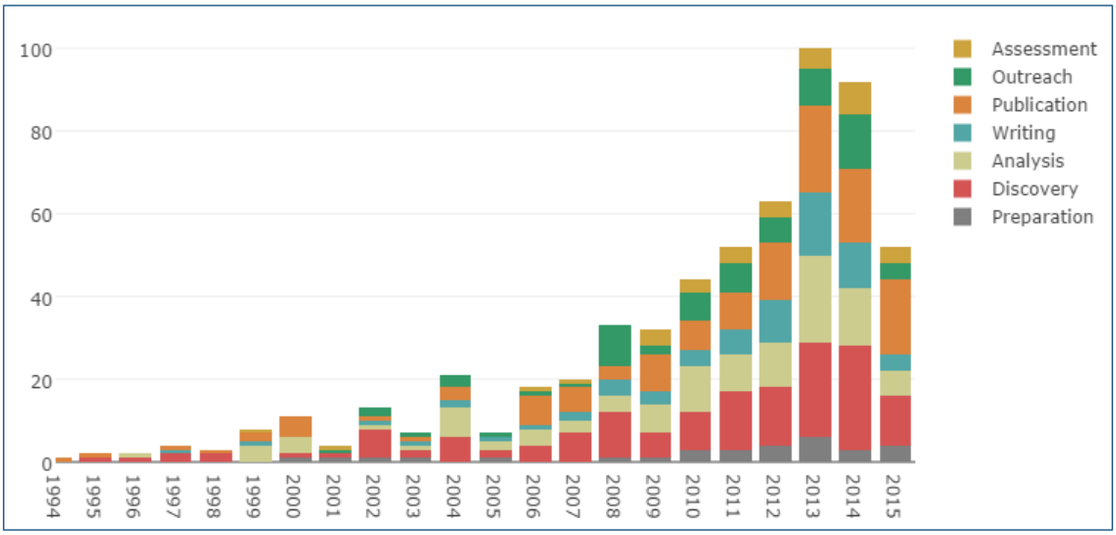

Fig. 3: New Tools by Research Phase, 1994-2015 (https://101innovations.wordpress.com)

The Challenge

In addition to the drivers mentioned above, the dramatic rise in the number and use of tools can be attributed to several other major factors:

- Tools are required as research activity scales up and collaboration increases.
- Researchers are moving 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.

The challenge, however, is that currently there is no single standard infrastructure architecture that would support an end-to-end workflow within the research data lifecycle, let alone the entire research lifecycle. In Fig. 4, Kramer and Bosman (2017) have provided examples of hypothetical workflows, which incorporate a variety of tools, thereby highlighting the changing nature of this environment.

Fig. 4: Changing Research Workflows (https://www.slideshare.net)

Fig. 5: Changing Research Workflows: Commercial "Suites" (https://s3-eu-west-1.amazonaws.com)

Fig. 5 provides two examples of the type of commercial activity which Lynch (Poynder, 2016) has foreshadowed.

The tools above are not necessarily single purpose. For example, some of the tools listed under "Outreach" could just as easily appear under "Discovery". The point made here is threefold: (1) as a part of the research (data) lifecycle, researchers tend to use more than just one tool; (2) tools are used at all stages of the lifecycle; and (3) these tools are generally not interoperable.

Therefore, interoperability becomes the key to creating an end-to-end workflow. This aspect has been reinforced by the recent work of the Coalition of Open Access Repositories (COAR) in developing a roadmap for the adoption of new functionality and technologies by repositories. Commenting on support for researchers' workflows, the COAR Working Group (2017, p. [9]) has noted that "the repository needs to be interoperable or integrated with other tools that authors are using offline and collaborative online tools …, as well as with journal submission systems such as the OJS platform. Additionally the repository should automatically recognise and autofill the metadata from [a] paper, dataset or other objects including author, title, date, and so on into the submission form". The COAR draft roadmap highlights the need

for the integration of additional functionality within the repository, such as applying - and discovering - licences; supporting assessment and peer review; and data mining.

The various components of the challenge can be summarised as follows:

- Tools are part of a data workflow.
- Repositories are a part of that workflow.
- Researchers use more than one tool in their workflow.
- Institutions do not provide all components of the end-to-end workflow.
- Data needs to move seamlessly between tools and repositories in that workflow.
- Interoperability becomes the key.
- How does one build workflows and infrastructure where interoperability is key?

Implications for Institutions

Institutions need to develop a whole-of-institution approach across all research repositories, major data workflows, and storage services. Within this approach, tools play an important role in supporting interoperability across research infrastructure architecture.

Given the active role that data repositories play in the research ecosystem, they need to be (re)designed to accommodate workflows, data re-use, and interoperability. They will need to deal with multiple classes of data and with data as part of a research project rather than merely as discrete objects. In terms of re-use and reproducibility, repositories will need to provide the context in which the data was generated. It is important to consider developing a number of solutions for their respective research communities, based on common standards and workflows, rather than attempting to design a single repository solution. In the current age of digital transformation, these new solutions need to accommodate the fact that, in many cases, the workflows and tools will extend beyond the institution. 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."

Given their current important role in supporting researchers, academic libraries also have a role to play in assisting in the development of these workflows from two important perspectives. First, they will want to ensure that the development and application of metadata standards meets best practice (or industry standards) and that the design of institutional repositories has the necessary standard, interconnecting services to workflows and tools so as to enable the movement of data. Second, they will want to ensure that any library-based repositories are included within the institutional research repository architecture. The authors would suggest that the interlinking of publications with associated data and research funding grants (where applicable) would be a logical starting point in establishing the traditional institutional repository as a core component within that architecture.

In their supporting role with researchers, librarians can encourage the latter—as part of good research data management practice—to move their data into institutionally-supported data repositories at the beginning of their project rather than as an afterthought at the end. As part of their outreach work in relation to data management, librarians need to educate their research communities that data needs to be "workflow ready". That is, the data is actionable (i.e. in the right format, standards compliant and accurate), the data is enabled and accessible for compute services, and the data is citeable.

Wolski, Howard, and Richardson (2017b) 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. An example of this would be participation in the development of national or research community metadata standards for use within institutional workflows. There are opportunities for research institutes and universities worldwide to sponsor keen staff to participate in current initiatives being undertaken, for example, by the Research Data Alliance (RDA), OpenAIRE, and the Coalition for Open Access Repositories (COAR).

From a resourcing perspective, at the local level, interested IT and library professionals could collaborate across institutions to identify the main workflows and tools used within their respective repositories, and then share common solutions. Finally, developmental work in regard to science gateways within one's own country may

generate ideas which can be incorporated into research repository architecture at the local level.

Conclusion

This paper has emphasized the role that tools play in supporting workflows which move data into repositories and has noted how this needs to be addressed earlier in the data lifecycle. It has demonstrated that tools are an important impetus for innovation in research repository architecture because of the driving imperative to include them in associated workflows. The authors have concluded with suggested strategies for institutional stakeholders, particularly libraries, on how to implement solutions which will ensure interoperability at all levels of research repository architecture, be that local, national, or international.

References

Ahmed, A. (2016). Supporting research excellence, *University World News*, issue 402. Retrieved from: http://www.universityworldnews.com/article.php?story=20160223220411616

Andreozzi, S., Arjona, A. B., Campos, I., Coelho, S., Dappert, A., Garavelli, S. … & Scott, M. (2016). *E-Infrastructures: Making Europe the Best Place for Research and Innovation*. Luxembourg: European Union. Retrieved from: https://ec.europa.eu/digital-single-market/en/news/e-infrastructures-making-europe-best-place-research-and-innovation

Australia (2015). Australian Government Public Data Policy Statement. Retrieved from: https://www.pmc.gov.au/sites/default/files/publications/aust_govt_public_data_policy_statement_1.pdf

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

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*. Oxford, U.K.: MIHR (pp. 79-88).

COAR Next Generation Repositories Working Group (2017). Next Generation Repositories – Introduction, Rationale and User Stories - Draft. Göttingen, Germany: COAR. Retrieved from: http://comment.coar-repositories.org/

Crouzier, T. (2016). *Science Ecosystem 2.0: how will change occur?* Luxembourg: European Union. Retrieved from: https://ec.europa.eu/research/innovation-union/pdf/expert-groups/rise/science_ecosystem_2.0-how_will_change_occur_crouzier_072015.pdf

Crow, R. (2002). The Case for Institutional Repositories: A SPARC Position Paper. Washington, DC: SPARC. Retrieved from: http://www.sparc.arl.org/resources/papers-guides/the-case-for-institutional-repositories

Dozier, J., & Gail, W. B. (2009). The emerging science of environmental applications. In Hey, T., Tansley, S., and Tolle, K. (Eds.), *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Redmond, WA: Microsoft Research (pp. 13-19).

Faundeen, J. (2017). Developing criteria to establish trusted digital repositories. *Data Science Journal*, 16, p. 22. doi: 10.5334/dsj-2017-022

Grant, M. J., & Booth, A. (2009). A typology of reviews: an analysis of 14 review types and associated methodologies. *Health Information & Libraries Journal*, *26*(2), 91-108. doi: 10.1111/j.1471-1842.2009.00848.x

Gundersen, L. C. (2016, February). Embedding Scientific Integrity and Ethics into the Scientific Process and Research Data Lifecycle. *American Geophysical Union Fall Meeting Abstracts*, abstract #PA12B-05.

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*. Luxembourg: European Commission. Retrieved from: http://ec.europa.eu/information_society/newsroom/cf/document.cfm?action=display&doc_id=707

Jain, P. (2011). New trends and future applications/directions of institutional repositories in academic institutions. *Library Review, 60*(2), 125-141. doi: 10.1108/00242531111113078

Johnson, R. K. (2002). Institutional repositories: partnering with faculty to enhance scholarly communication. *D-Lib Magazine, 8*(11), 1-7. doi: 10.1045/november2002-johnson

Johnston, L. (2012). Repositories: Not just about publications any more. *The Signal*, 20 July. Retrieved from: http://blogs.loc.gov/thesignal/2012/07/repositories-not-just-about-publications-any-more/

Kramer, B. & Bosman, J. (2016). Innovations in scholarly communication - global survey on research tool usage

[version 1; referees: 2 approved]. *F1000Research*, 5, 692. doi: 10.12688/f1000research.8414.1

Kramer, B. & Bosman, J. (2017). Changing research workflows - opportunities for researchers, librarians and publishers. Figshare. Retrieved from https://doi.org/10.6084/m9.figshare.4609423.v1

Lynch, C. A. (2003). Institutional repositories: essential infrastructure for scholarship in the digital age. *Portal: Libraries and the Academy, 3*(2), 327-336. doi: 10.1353/pla.2003.0039

Munafo, M. (2016). Scientific ecosystems and research reproducibility. Talk presented at Research Libraries UK Conference, London, 9-11 March. Retrieved from https://www.youtube.com/watch?v=TD2cUYVci28&feature=youtu.be

Nielsen, Michael (2011). Open science now! Talk presented at TEDxWaterloo, Waterloo, Canada, 3 March. Retrieved from http://www.ted.com/talks/michael_nielsen_open_science_now

Poynder, R. (2016). Q&A with CNI's Clifford Lynch: Time to Rethink the Institutional Repository? *Open and Shut*. Retrieved from: http://poynder.blogspot.com.au/2016/09/q-with-cnis-clifford-lynch-time-to-re_22.html

Pryor, G. (Ed.) (2012). *Managing Research Data*. London: Facet Publishing.

Rice, R., & Haywood, J. (2011). Research data management initiatives at University of Edinburgh. *International Journal of Digital Curation*, *6*(2), 232-244. doi: 10.2218/ijdc.v6i2.199

Sergeant, D. M. (2006). Using a Virtual Research Environment to present CRIS grouped to support the real researchers' research lifecycle. In Asserson, A., & Simons, E. J. (Eds.), *Enabling Interaction and Quality: Beyond the Hanseatic League (8th International Conference on Current Research Information Systems)* (p. 189). Leuven, Belgium: Leuven University Press.

Stall, S. (2016, February). Implementing and Sustaining Data Lifecycle Best Practices: A Framework for Researchers and Repositories. *American Geophysical Union Fall Meeting Abstracts*, abstract #OD23A-01.

Webster, J., & Watson, R. T. (2002). Analyzing the past to prepare for the future: Writing a literature review. *MIS Quarterly*, xiii-xxiii. Retrieved from https://web.njit.edu/~egan/Writing_A_Literature_Review.pdf

Wissik, T., & Ďurčo, M. (2016, April). Research Data Workflows: From Research Data Lifecycle Models to Institutional Solutions. In De Smedt, K. (Ed.), *Selected Papers from the CLARIN Annual Conference 2015, October 14–16, 2015, Wroclaw, Poland* (No. 123, pp. 94-107). Linköping, Sweden: Linköping University Electronic Press. Retrieved from: http://www.ep.liu.se/ecp/123/008/ecp15123008.pdf

Wolski, M., & Richardson, J. (2014). A Model for Institutional Infrastructure to Support Digital Scholarship. *Publications*, *2*(4), pp. 83-99. doi: 10.3390/publications2040083

Wolski, M., Howard, L., & Richardson, J. (2017a). A Trust Framework for Online Research Data Services. *Publications*, *5*(2), 14. doi: 10.3390/publications5020014.

Wolski, M., Howard, L., & Richardson, J. (2017b). The importance of tools in the data lifecycle. *Digital Library Perspectives*, *33*(3), 235-252. doi: 10.1108/DLP-11-2016-0042

Web References

https://101innovations.wordpress.com/2015/12/25/timeline-of-tools/

https://www.jisc.ac.uk/sites/default/files/research_data_life_diagram_0.jpg

https://www.slideshare.net/BaltimoreNISO/bosmankramer-changing-research-workflows