Contents lists available at ScienceDirect

International Journal of Information Management

journal homepage: www.elsevier.com/locate/ijinfomgt

ERDMAS: An exemplar-driven institutional research data management and analysis strategy

Matthew I. Bellgard

eResearch Office, Queensland University of Technology, Brisbane, 4001, Australia

ARTICLE INFO

Keywords:
Research data management
Researcher-driven
Policy and guidelines
Exemplars
Hybrid Cloud
Analytics

ABSTRACT

Devising fit-for-purpose research data management strategies within a university is challenging. This is because the five 'Vs' for generated research data; its Volume, Variety, Velocity, Veracity and its Value must be constantly considered. Invariably, a combination of data V's for any given research endeavour determine how best to manage it appropriately addressing archiving, compliance, security, privacy, sharing, reuse and so forth. As such, institutions are faced with defining, shaping and refining strategies and practicies to ensure there are consistent and adequate research data management polices and guidelines in place for their researchers. FAIR data principles are very important for embracing open data opportunities, but more broadly, research data management practices need to be established in a comprehensive way. Additionally, new ICT options have rapidly become available where institutions can make considered choices on whether to continue to use 'on prem', private Cloud or public Cloud infrastructure. If a hybrid approach is adopted, then the potential impact on existing institutional research data management strategies must be continually assessed and revised accordingly. Getting the balance right between developing a relevant institutional policy on the one hand yet also dynamically catering for the eclectic research data management and analytics needs of researchers and their evolving interactions with external collaborators on the other, must be continually navigated. In this manuscript, an exemplar-driven research data management and analytics conceptual framework is introduced. A key feature of this framework is that it is couched in two dimensions. On one axis is the 'standard' linear approach of developing the research data management policy, guidelines, procedures, audit and risk assessment and an options matrix. Importantly, a second axis comprising a researcher-driven focus is introduced where exemplar research activities are used to define 'classes' of research data management and analysis requirements. This exemplar-driven dimension enables an ongoing system-wide comparative review to occur in parallel that can continually inform policy and guidelines refinement.

1. Introduction

Research Data management (RDM) has always been a challenge for research institutions (Jones, 2012; Maican & Lixandroiu, 2016; Ozmen-Ertekina & Ozbayb, 2012; Pinfield, Cox, & Smith, 2014; Vilminko-Heikkinen & Pekkola, 2019). However, with the ever pervasive role digital disruption plays in all research activities and the variety of storage, analytical and sharing technology choices available to researchers the challenge does not abate. The traditional approach within an institution is to devise and then refine every few years its RDM policy, guidelines, procedures, audit and risk and options matrix in a contemporary manner. In many cases, the "horse has bolted" where viable, but not necessarily institution-endorsed RDM solutions, have become available to researchers, that unless systematically monitored, might not be compliant to prevailing university policies. For example, for some institutions, national funding agencies and governments, research outputs must reside in the same country as where it was generated (data sovereignty), however, global Cloud-based data sharing platforms which easily facilitate cross-country collaborations, may not comply at the time they are used. In some instances, collaborating researchers may have little choice if they wish to engage in multi-institutional cross-jurisdictional initiatives. An ongoing understanding of the nuanced and evolving RDM issues for any given research endeavour is necessary to inform policy refinement, ensuring it is constantly relevant.

The four 'Vs' for generated research data; its Volume, Variety, Velocity and Veracity (Gandomi & Haider, 2015) (adapted from the three 'V's (Laney, 2001)) must be continually taken into consideration. Invariably, a combination of data V's for any given research endeavour, at any given time point, determine how best to manage it appropriately

E-mail address: matthew.bellgard@qut.edu.au.

https://doi.org/10.1016/j.ijinfomgt.2019.08.009
Received 28 May 2019; Received in revised form 28 August 2019; Accepted 29 August 2019
Available online 21 September 2019

addressing archiving, compliance, security, privacy, sharing, reuse and so forth. For instance, for many research projects, there are many phases of a research project, as in a large scale whole genome project, where a first phase might be experimental design that generates small quantities of research meta data compared to a second phase generating volumes and varieties of *-omics* data, through to a third phase that is focused on analysis and interpretation of data (Hunter et al., 2019). As this example highlights, consideration must be given to the data life cycle for each phase of the project, not just the data life cycle of entire project. Each phase will require different approaches to research data management. In addition, there a number of research projects that are 'spin-offs' of original projects where additional volumes, varieties may not have been initially envisaged, but these requirements may change depending on the nature of collaborative interactions, scientific discovery and innovation. As such RDM becomes a balancing act where a university must strive to maintain relevant RDM polices on the one hand, yet have a flexible and practical RDM environment for researchers on the other. Fortunately, data life cycle concepts can now be translated to tangible lifecycle polices and technical solutions available through Public Cloud offerings.

The article contends that a main reason why RDM is such a significant challenge for research institutions is that 'analytics' is not explicitly acknowledged in RDM policy. The focus is on the storage, compliance, archiving, sharing and reuse of data. FAIR (Wilkinson et al., 2016) data principles confirm this focus to make data 'Findable, Accessible, Interoperable, and Reusable". As part of the proposed Exemplar-driven Research Data Management and Analysis Strategy (ERDMAS) maintains that the 'purposeful application of data' (Bellgard et al., 2017) is a key focus. For instance, that act of reusing data (secondary use of data Burton, Banner, Elliot, Knoppers, & Banks, 2017) for another research question will invariable change the *purpose* of that data and, as such, necessitate the need to generate another research data management plan that is likely different to the purpose of the original dataset. As such, there may well be other issues surrounding the sensitivity of the reuse of this data, that must be taken into consideration with appropriate data governance arrangements addressed including security, privacy, reconsent and access. This is especially the case in the secondary use of health data (Bellgard et al., 2018; Burton, Banner, Elliot, Knoppers, & Banks, 2017) and social media data (Bruns, 2019; Burgess & Bruns, 2015). It is argued that, unfortunately, a RDM-only perspective perpetuates an artificial divide between storage and compute which leads to a disconnect between RDM and analysis. Whereas, as soon as 'analysis' of data is considered at the outset, it becomes possible to determine the purpose and *value* of data (end-user driven). The *value* of data, via a data science lens, is essential to determine evolving management needs and reuse of research data (Lim et al., 2018). Hence it is important to always consider the five V's of data (Gandomi & Haider, 2015). As such it becomes feasible to devise ERDMAS.

2. Exemplar-driven Research Data Management and Analysis Strategy (ERDMAS) Conceptual Framework

An Exemplar-driven Research Data Management and Analysis Strategy (ERDMAS) Conceptual Framework is shown in Fig. 1. One of the key features to this framework is that it is represented in two dimensions, a bidirectional vertical and horizontal axis. Each axis is made up of a number of components. The vertical axis (Axis 1) is considered the Researcher-driven (end-user) perspective and the horizontal axis (Axis 2) is the institutional, policy-driven, perspective.

2.1. Axis 1: researcher-driven perspective of ERDMAS

From the top of the page to the bottom there are three distinct activities encompassed by dashed-lined boxes. Each *project* square represents the variety of research projects across an institution. Each square might represent a new or existing research project, comprising multiple data sources, software systems with its own legacy governance, data life cycle processes and how the data is to be used. The red dashed lines represent those research projects that are deemed *exemplar* research projects, shown in the middle layer. For each project a Requirements Scoping Brief (RSB) is created following an Agile and iterative refinement methodology. The RSB includes the subheadings: research client, problem statement, background, challenges, proposed solution, roles and responsibilities, budget, ethics, compliance, national benefit and so forth. The RSB is used to architect eResearch solutions available within the institutional operating context, working closely with researchers. Each RSB is also used to populate an ERDMAS Options Matrix (OM) that maps each completed RSB against contemporary data management and analysis issues. A populated OM provides an instant snapshot of overlapping/bespoke data management and analysis issues in common/specific to each exemplar project.

Training depicted on the vertical axis highlights the sharing of eResearch skills and resources to support and onboard researchers of latest technologies, utilise institution-approved research data management and analysis solutions, highlight compliance issues, develop timelines and deliver future-proofed, scalable solutions.

2.2. Axis 2: policy-driven ERDMAS

The populated RSB and OM as well as the delivered architected eResearch solutions can then be used to inform institutional policy, guidelines, procedures, audit and risk and refine the OM. It is multi-directional given the dynamic and evolving nature of policy development and refinement and the development of each RSB exemplar. In a systematic fashion each exemplar project can be assessed to comply with policy, identify gaps in policy, as well as identify gaps in proposed eResearch solutions. Training on the horizontal axis is important to unsure upskilling of key stakeholders (such as Ethics, IT services, Library) across the institution of shared ERDMAS challenges, compliance issues, archiving challenges, metadata requirements and so forth.

3. Application of ERDMAS at an institution

Within a two month period, over 38 RSBs have already been developed or are currently under development from research institutes, centres, groups, researchers, commercial entities, research infrastructure entities representing a diverse range of research data management and analysis problems across the institution. The institution's eResearch Office works closely with researchers to devise well architected eResearch solutions.

The OM has been populated and interestingly, not all issues are relevant to each exemplar project, however, upon populating this OM, common and bespoke ERDMAS issues have been rapidly identified. For instance, through this process it has been possible to define the common and different ERDMAS issues of a CT scanner installed in a pre-clinical setting versus one scanner installed in a hospital setting. One scanner's samples are at a four-times higher scanning rate, different end-users (embedded in a clinical setting versus co-located with a hospital for research purposes only), different proprietary software is installed on each instrument. However, transfer and storage of the data for further processing is common, the data generated format (DICOM) is the same, some of basic processing steps are similar as well as the nature of storage (store raw data once, analyse many times).

Obvious classes of exemplar projects might include commercialisation, citizen science, specialised instruments (e.g. NMR, DNA sequencers, Mass spectrometry), medical imaging, epidemiology, clinical trials, social media, photogrammetry, bioinformatics/genomics, sensor, robotics and automation, surveys and so forth. Using both the populated RSB and the OM the eResearch Office work closely with IT Services, Library and the Office of Research Ethics and Integrity to

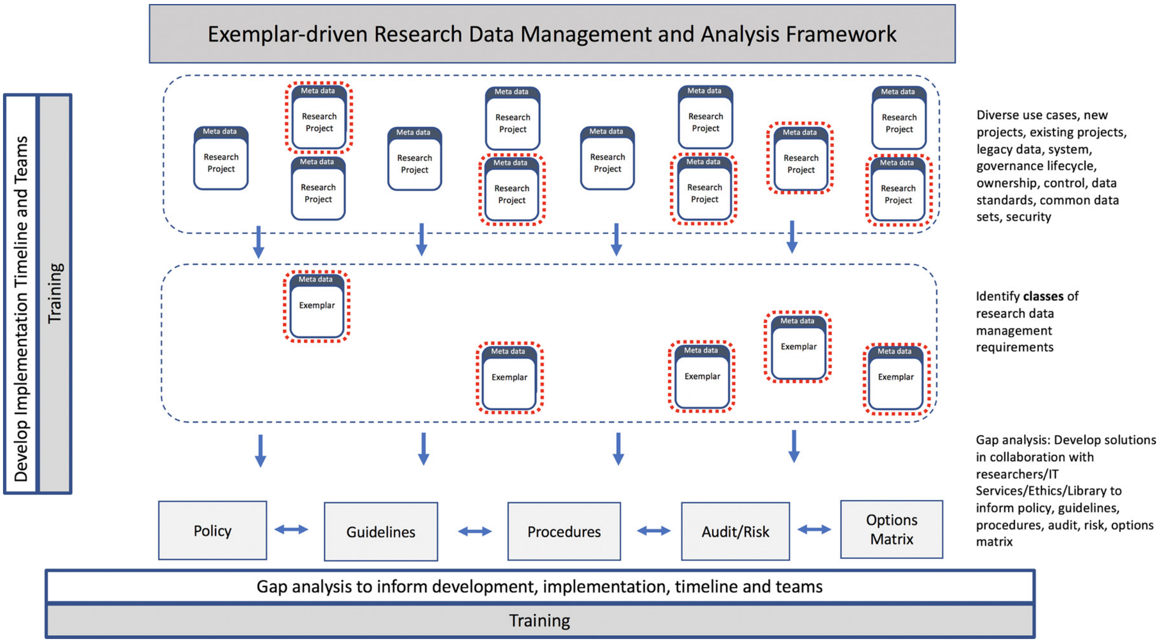
Exemplar-driven Research Data Management and Analysis Framework
Develop Implementation Timeline and Teams
Training
Meta data
Research Project
Diverse use cases, new projects, existing projects, legacy data, system, governance lifecycle, ownership, control, data standards, common data sets, security
Meta data
Exemplar
Identify classes of research data management requirements
Gap analysis: Develop solutions in collaboration with researchers/IT Services/Ethics/Library to inform policy, guidelines, procedures, audit, risk, options matrix
Policy
Guidelines
Procedures
Audit/Risk
Options Matrix
Gap analysis to inform development, implementation, timeline and teams
Training

Fig. 1. Exemplar-driven Research Data Management and Analysis Framework.

review/inform/revise/update the institution’s policy, guidelines, procedures, audit/risk documents.

4. Discussion

In this Research Note, a two-pronged strategy that enables a more agile approach to institutional research data management and analysis, called, ERDMAS: Exemplar-driven Research Data Management and Analysis Strategy is outlined. ERDMAS is inherently end-user (researcher) driven where the value of research data drives the RDM strategy, what the research data means and its purpose at a given time. Through this process as soon as research data is reused, its purpose invariably changes and so too must the strategy to manage and analyse the data. ERDMAS enables a systematic assessment and comparison of common ‘contemporary’ research data management issues as well as highlight nuanced differences. In summary, ERDMAS is a proactive approach and will help researchers and all other key institutional stakeholders in increasing research efficiency, improve research integrity, make research outputs more visible and enable collaboration.

4.1. Recommendations and lessons learned

The challenge of implementing an institution-wide research data management strategy is not insignificant. The introduction of the ERDMAS has enabled discussions with key stakeholders across the institution in manageable levels of complexity. For instance, completing the RSBs enable careful delineation of ownership and custodianship of research data, especially when this data is subsequently used in other research projects or for teaching purposes. In the CT Scanner example outlined above, there is a further important difference articulated between the data from one instrument that is hospital-based versus the one based in the institution for research purposes. Given the clinical setting for the hospital-based instrument, the data needs to be carefully deidentified through a multistage process, some of which requires manual intervention. In addition, user controls and audit trails of data access must be carefully managed. The end-user focus of ERDMAS continually informs management policies, procedures and guidelines to devise sustainable data management and analysis solutions in an iterative manner of continuous improvement. It is recommended to use Cloud-based online collaboration tools to document the RSBs. The use of ‘tags’ provides keywords which allows search of previously defined RSBs that can be reused for other projects, particularly where solutions have been implemented. It is planned that the use of cloud formation templates could enable an automated approach to reuse the solutions, not just at the documentation stage.

Acknowledgements

The author wishes to acknowledge discussions with a number of university staff, particularly, K McMahon, C Brown, L Christian, R Steck, L Barner, A Walsh, G Barbour, M Boyle, D Nyholt, S Jayasinghe, P Wikstrom, A ter Hofstede, D Henderson, J Jacobs, S Bradbury.

References

Bellgard, M. I., Chartres, N., Watts, G. F., Wilton, S., Fletcher, S., Hunter, A., et al. (2017). Comprehending the health informatics Spectrum: Grappling with system entropy and advancing quality clinical research. *Frontiers in Public Health, 5*, 224. https://doi.org/10.3389/fpubh.2017.00224.

Bellgard, M. I., Napier, K. R., Bittles, A. H., Szer, J., Fletcher, S., Zeps, N., et al. (2018). Design of a framework for the deployment of collaborative independent rare disease-centric registries: Gaucher disease registry model. *Blood Cells, Molecules & Diseases, 68*, 232–238. https://doi.org/10.1016/j.bcmd.2017.01.013.

Bruns, A. (2019). After the “APIcalypse”: Social media platforms and their fight against critical scholarly research. *Information, Communication & Society*. https://doi.org/10.1080/1369118X.2019.1637447 In press.

Burgess, J., & Bruns, A. (2015). Easy data, hard data: The politics and pragmatics of twitter research after the computational turn. In G. Langlois, J. Redden, & G. Elmer (Eds.). *Compromised data: From social media to big data* (pp. 93–111). New York: Bloomsbury Academic.

Burton, P. R., Banner, N., Elliot, M. J., Knoppers, B. M., & Banks, J. (2017). Policies and strategies to facilitate secondary use of research data in the health sciences. *International Journal of Epidemiology, 46*(6), 1729–1733. https://doi.org/10.1093/ije/dyx195.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management, 35*, 137–144.

Hunter, A., Bowland, G., Chang, S., Szabo, T., Napier, K., Lum, M., et al. (2019). The bioplatforms australia data portal. *eResearch Australasia Conference 2018*(August), 2018 https://conference.eresearch.edu.au/2018/08/the-bioplatforms-australia-data-portal/.

Jones, S. (2012). eGovernment Document Management System: A case analysis of risk and reward. *International Journal of Information Management, 32*, 396–400.

Laney, D. (2001). *3D data management: Controlling data volume, velocity and variety.* Stamford, CT: META Group, Inc.

Lim, C., Kim, K., Kim, M., Heo, J., Kim, K., & Maglio, P. P. (2018). From data to value: A nine-factor framework for data-based value creation in information-intensive services. *International Journal of Information Management, 39*, 121–135. https://doi.org/10.1016/j.ijinfomgt.2017.12.007.

Maican, C., & Lixandroiu, R. (2016). A system architecture based on open source enterprise content management systems for supporting educational institutions. *International Journal of Information Management, 36*, 207–214.

Ozmen-Ertekina, D., & Ozbayb, K. (2012). Dynamic data maintenance for quality data, quality research. *International Journal of Information Management, 32*, 282–293.

Pinfield, S., Cox, A. M., & Smith, J. (2014). Research data management and libraries: Relationships, activities, drivers and influences. *PloS One, 9*(12).

Vilminko-Heikkinen, R., & Pekkola, S. (2019). Changes in roles, responsibilities and ownership in organizing master data T management. *International Journal of Information Management, 47*, 76–87.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data, 3*, 160018. https://doi.org/10.1038/sdata.2016.18.