UQ Library's Journey Towards Fair Research Outputs - Advancement of Research Data Management and Digitisation Services

Rebecca Deuble  
*The University of Queensland, Australia*, r.deuble@library.uq.edu.au

Mandy Swingle  
*The University of Queensland, Australia*

Dulcie Stewart  
*The University of Queensland, Australia*

Fei Yu  
*The University of Queensland, Australia*

https://docs.lib.purdue.edu/iatul/2019/fair/1

This document has been made available through Purdue e-Pubs, a service of the Purdue University Libraries. Please contact epubs@purdue.edu for additional information.
Abstract

The growth of data and technology has deeply impacted on the ways that scientific research are practiced, bringing new challenges to academic libraries. In Australia, the recent worldwide movement to FAIR – findable, accessible, interoperable, and reusable - research, has seen institutions shift towards developing relevant policies and standards, so as to remain competitive in the global economy.

This paper will discuss how The University of Queensland (UQ) Library has taken significant steps to achieving the FAIR principles for research outputs, within Digitisation and Research Data Management (RDM) services.

The institutional repository, UQ eSpace is critical in supporting these services. It is used by researchers to publish datasets underpinning research publications, with librarians providing support to curate and enrich metadata records. A collection of licences can be applied, including two UQ specific click-through licences. There has also been a focus on delivering educational programs to promote best-practice RDM within the UQ research community. More recently, an integrated research data management system was developed and rolled-out across the University, enabling researchers to manage project data across the whole research lifecycle – a significant step towards FAIR data.

Similarly, the FAIR principles are applied to digitised material, including heritage collections, to increase discoverability and use. Digitised objects are assigned persistent identifiers when uploaded into UQ eSpace, ensuring recognition over time. Licence and rights agreements are applied to enable reuse and correct acknowledgement of digital material. Metadata standards are also used to describe, manage and retrieve information about the digital objects and data, ensuring FAIR outputs.

The UQ Library’s continued efforts in adapting and evolving their services, is critical in ensuring that researchers can work towards producing FAIR research outputs that are of world standard.

Keywords: integrated data management system, research data management, digitisation, digital objects, datasets, metadata
Introduction

The University of Queensland (UQ) is a large, research-intensive university located in Brisbane, Australia, that is committed to pioneering and adopting the highest standards of excellence in digitisation and research data management (RDM). To ensure that these services are consistent with the FAIR principles and other major global drivers, UQ has invested in developing two crucial pieces of infrastructure. These include the institutional repository, UQ eSpace, and the more recently developed UQ Research Data Manager (UQ RDM) system. Both were developed in-house, are researcher-centric, and designed to meet the requirements of the Australian Higher Education sector and broader research landscape, which includes: Excellence in Research Australia (ERA)\(^1\), the National Science and Innovation Agenda\(^2\) and the Digital Access to Collections framework\(^3\) [Brown et al. 2018].

UQ eSpace is the authoritative source for research outputs of UQ staff and students, and the archival home of UQ Higher Degree by Research theses in only digital format [UQ Library, 2019]. It captures metadata for more than 10,000 research outputs annually involving UQ affiliated researchers. Since 2013, UQ researchers have been able to deposit datasets underpinning their publications into UQ eSpace, enabling compliance with various journal data sharing policies, and national funding requirements such as the Australian Research Council's (ARC) Open Access Policy.\(^4\) Digitised material has also been deposited since 2007 - including theses, books, journals and unique cultural and heritage material.

While digitised materials have been captured since 2007, the Digitisation Service was not officially established until 2014. Digital objects are captured, curated, evaluated and deposited by a unique team of staff, including a photographer, metadata creator and digital curator. The service is well supported by a policy that defines digitisation as being the conversion of analogue items into digital format, and its related processing to make the material usable [UQ Library, 2013]. Some examples include scanning items such as maps, architectural plans, manuscripts, photographs, 3D objects, theses, charts and audio tapes. These services are provided for UQ research, teaching and learning purposes, and the application of the FAIR principles helps to extend and enhance the re-use of the collection across all discipline areas and formats. For example, many of the special collections attract researchers both nationally and internationally.

The UQ RDM system was launched in January 2018, as part of UQ's initiative to establish a world-leading RDM service.\(^5\) Driven by various funding and compliance requirements, including the Australian Code for the Responsible of Research\(^6\), it was concluded that UQ needed a better way to allow researchers to practice good stewardship of research data. This innovative system was developed in collaboration between the Research Office, Library and central Information Technology Services. It enables researchers and Higher Degree by Research students to seamlessly manage their project's research data with their collaborators across the whole research lifecycle - from project conception to publication.

As of May 2019, there are over 4,300 unique users (UQ and external users) on over 3000 projects within UQ RDM. Development work is currently underway to integrate it with UQ eSpace, enabling researchers to create datasets for publications where the underlying data is managed via the UQ

---

\(^{1}\) [https://www.arc.gov.au/excellence-research-australia](https://www.arc.gov.au/excellence-research-australia)


\(^{5}\) [https://research.uq.edu.au/rmbt/uqrdm](https://research.uq.edu.au/rmbt/uqrdm)

RDM. Researchers who create datasets for publications through UQ RDM will be able to do so seamlessly by linking to existing documentation and project metadata. To reinforce these changes to the RDM service, there are plans to reflect this in University policy and procedures. This will ensure appropriate management of research data across all stages of the research lifecycle, with clear references to the supporting infrastructure (UQ RDM and eSpace). This approach is consistent with the national agenda for FAIR access, which aims to make all publicly funded research outputs, findable, accessible, interoperable and reusable [FAIR Access, 2017].

Aim

The aim of this paper is to demonstrate the steps that the UQ Library has taken towards achieving FAIR research outputs from a digitisation and RDM perspective. Examples of how the underlying infrastructure has been designed and adapted to meet the requirements of the FAIR principles will be shown throughout. References will also be made to the Digitisation policy and RDM-best-practice training programs that have been established.

Findable

The “Findable” principle, indicates that datasets, digital objects and their associated metadata should be easy to find by both humans and computers. It is important that the metadata underpinning these outputs are assigned with a globally unique and persistent identifier, and registered or indexed in a searchable resource such as a repository or online database [Wilkinson et al., 2016; Koster & Woutersen-Windhouwer, 2018].

In UQ eSpace, datasets and various digitised materials are easily findable through the application of relevant descriptive metadata within the Dublin Core schema. The elements within the schema are used to describe, catalogue and index resources. They also identify the dataset or digital object through elements such as “title”, “date created” or “publication date”, “author/s” or “creators” and “file format”. Additional elements have been created for complex digital objects; for example job numbers and scales for architectural plans.

Another key element of the metadata is the persistent identifiers (PIDs) that must be assigned to both digitised objects and datasets before they are made public. The PID (eSpace URL) uniquely identifies the dataset or digital object and ensures that it is managed and kept up to date over time. This means that the PIDs are findable, even if the data or digital object moves, the repository is upgraded, or the object is archived. This is also consistent with the UQ Library’s Digitisation policy objective to provide “easy and convenient access to the digitised content of its collections” [UQ Library, 2013]. In addition to the eSpace URL, the record underpinning a research dataset must be assigned a Digital Object Identifier (DOI) through the provider CrossRef (see figure 1). The DOI is assigned to the record, rather than the actual data file/s, as without this rich metadata, the data would be meaningless. It also ensures accessibility for those datasets that are “mediated”, that do not contain the underlying data - this will be explained further in the “Accessible” section.

Other examples of relevant PIDs that can be applied to research datasets to increase their findability are ORCID - Open Researcher or Contributor ID and RAID - Research Activity Identifier. These are both endorsed and recommended by the Australian Research Data Commons (ARDC) - a national and transformative initiative driving access to high-quality data collections [ARDC, 2019a]. ORCIDs will enable the researchers who created the datasets to be identified more easily, while RAIDs will enable UQ to discover where the research activities are located, which researchers are collaborating and what resources are being used. While this information is not currently

7 https://ppl.app.uq.edu.au/content/4.20.06-research-data-management
8 https://orcid.org/
collected for datasets stored within the repository, the UQ RDM system stores ORCIDs for all UQ researchers that have linked their account through UQ eSpace. All UQ RDM projects are also assigned a RAiD. This means that there is the potential to add these PIDs to datasets published through UQ RDM in the future, thus increasing discoverability, and making data provenance clear and easy (ARDC, 2019).

Another factor that contributes to findability is the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) that is used by UQ eSpace. This facilitates the harvesting of the metadata underpinning datasets and digital objects, which allows records to be exposed through relevant search engines such as Google, and Primo – UQ Library’s search engine, and discovery services such as Trove – created by the National Library of Australia. Research datasets in UQ eSpace are compliant with the Registry Interchange Format: Collections and Services (RIF-CS) schema enforced by Research Data Australia (RDA) - the national registry for storing and managing dataset records [ARDC, 2019b]. The RIF-CS schema is an international standard, that enables descriptions to be harvested automatically for display and discovery in RDA (see figure 2).

---

*Figure 1 - Screenshot of a dataset record in UQ eSpace, showing the eSpace PID and DOI that has been assigned.*
In addition to RDA, datasets within UQ eSpace are indexed by the Data Citation Index (DCI) - a database produced by Clarivate Analytics. UQ was the second university in Australia to have dataset records ingested into the DCI, with the first batch of records appearing in December 2014 [Yu and Morgan, 2016]. As of April 2019, 268 records have been harvested by the DCI. The DCI provides a single point of access to datasets indexed in eSpace and other repositories across the world. These datasets are searchable within the Web of Science database where they are linked to relevant literature articles (see figure 3 for example of UQ dataset indexed in Web of Science). For dataset records to be accepted into the DCI, there are minimal standards that have to be met including validating against the DCI schema and satisfying the Clarivate Analytics repository evaluation, selection and coverage policies [Clarivate Analytics, 2019]. This highlights the high quality of the datasets that are stored within UQ eSpace.
Accessible digital objects and datasets are defined by metadata that are accessible, even when the data is no longer available, and retrievable by their identifier using a standard protocol [Wilkinson et al. 2016]. This protocol should be open, free and universally implementable. UQ Library aims to enhance the discoverability and accessibility of unique material that is authored by UQ students and researchers or that is held in specific collections [UQ Library, 2013].

Different levels of access are applied to datasets and digital objects in UQ eSpace. Both outputs can be open or mediated access, while restricted access can also be applied to digital objects. Figure 4 shows the number of datasets and digital objects that are mediated, open or restricted access within UQ eSpace as of May 2019.
Figure 4 - Number of datasets and various digital objects in UQ eSpace as of May 2019.

Open access refers to datasets and digital objects that are made publicly available for others to re-use. This means that the item is either attached to the record itself (i.e. the underlying metadata) or in the case of some open datasets, a link may be provided to an external repository where the actual data is stored. When uploading digital objects to the repository, digitisation staff always use file formats that allow easy access for users, and which are consistent with relevant international digitisation standards, as will be discussed within the “Interoperable” section. Similarly, researchers uploading datasets to the repository are strongly encouraged to use open-source document types such as CSV files to improve accessibility. The benefits of open-source document types are explained within best-practice RDM training sessions that are regularly conducted throughout the year. As mentioned, researchers who prefer to upload their datasets to external subject-specific repositories such as Pangea, are free to do so; however the Library still actively promotes linking this dataset to the corresponding metadata (a record) in UQ eSpace.

Mediated access datasets and digital objects are not made publicly available for re-use (i.e. are not attached to the metadata record), instead, an appropriate contact is listed, so that those who wish to request access to the item can do so. For mediated datasets, the contact is always an administrative email monitored by data librarians, to allow time to check whether the primary researcher’s are still at UQ. Similarly, for mediated digital objects, such as cultural and heritage material, the Library is listed as the contact as they are responsible for hosting the collection. Users can see details of the dataset or digital object record, but cannot access it directly. There are various reasons why access to a dataset or digital object may be mediated - these include sensitive health or ecological data, cases where the object has been digitised for preservation and is in copyright, and when researchers wish to monitor who is requesting access.

Restricted access applies to digitised material uploaded to UQ eSpace for preservation that may be culturally sensitive, such as audio recordings of taboo words from an Australian Indigenous language.
**Interoperable**

Digital objects and datasets that are interoperable must use an appropriate and accessible language for knowledge representation. The data and digital objects should also be integrated with other data and metadata, to allow for analysis, storage, archiving and processing workflows [Wilkinson et al. 2016; Koster & Woutersen-Windhouwer, 2018].

Datasets and digital objects stored within UQ eSpace must be accompanied by a set of minimal viable metadata that are standardised. Many of the metadata fields have been mentioned within the “Findable” and “Accessible” sections; however some specifically ensure interoperability - these include file format, subject classification schemas or keywords, and funding information for research datasets.

For example, when researchers are describing their datasets, they are required to specify at least one Field of Research (FoR) code from the Australian and New Zealand Standard Research Classification scheme. This is the schema that is used to allow research and development activity within Australia and New Zealand to be classified according to research activity. The same standard is applied by the digitisation team when digitising Higher Degree By Research theses and is endorsed by the ARDC as being “best-practice” [ARDC, 2019c].

In addition to FoR codes, researchers uploading their datasets to UQ eSpace can enter funding and ethics information, where applicable. While this is not currently mandated, integration with the UQ RDM will enable this information to be pre-populated through an Application Programming Interface (API) connection to the relevant institutional grant and ethics systems. Training for researchers currently using UQ RDM has focused on encouraging this information to be entered where appropriate within the research lifecycle, as they have the ability to update the project metadata at any point in time. Further, UQ RDM was recently integrated with the University grants databases via an API, allowing researchers to select relevant grants that they are listed on, to add to their project metadata. Once a grant is added, it auto-populates the relevant metadata such as the grant ID and funding body. Similar integration is planned for both the UQ animal and human ethics systems. Additionally, the quality of other metadata fields that are used to describe a dataset can be improved through UQ RDM’s ability to curate and describe the data at all stages throughout the lifecycle.

The Digitisation team similarly comply with relevant national and international standards to produce digital objects that are interoperable. This includes the recommended National Library of Australia image capture standards, the Federal Agencies Digital Guidelines Initiative (FADGI) and PREservation Metadata: Implementation Strategies (PREMIS). Technical metadata is embedded into the file, ensuring the material is accessible, shared through a broadly acceptable format, preserved, and has long term-usability [Wilcox, 2018].

Further to this, two file types are produced: a robust archival file for preservation and an access file. The archival or preservation file is often large due to the quality and high resolution of the scan. In contrast, the access file is much smaller, enabling easy access, thus not impacting on the user’s downloads. For example, when digitising a cassette tape, an archival Waveform (WAV) audio file is produced for preservation purposes, while access is given through an .mp3 file. This is consistent with the requirements specified in the digitisation policy, which states that digitisation must “comply with accepted standards for digital capture, storage, preservation and access” [UQ Library, 2013].

UQ eSpace records contain information about the dataset and digital file type. This information helps users understand the file size or type of file, which is important for downloading. For example,
an archival file of architecture plan may indicate that it is a TIFF file and is 1.29 gigabytes, whereas the access file of the same digital object may indicate that it is a JPEG file and 172 kilobytes.

The Digitisation team have also implemented controlled vocabularies to describe certain digital objects. For example, relevant vocabularies have been applied to digitised Australian Indigenous language recordings. These include the Library of Congress Subject Headings vocabulary, the Australian Institute of Aboriginal and Torres Strait Islander Studies (AIATSIS) Pathways Thesaurus, and Austlang language codes. This is a core principle of linked data and supports the FAIR concept of interoperability [Wilcox, 2018].

Reusable

The re-useable principle states that reusable data should “maintain its initial richness” [ARDC, 2019d]. Datasets and digital objects that are reusable are characterised by metadata that consist of a plurality of accurate and relevant attributes. Metadata must meet domain-relevant community standards, and include both a clear and accessible usage licence, and associated provenance information. One of the key roles of the reusability principle is the ability of humans and machines to assess and select data on the basis of criteria relating to provenance information about the dataset or digital object [Wilkinson et al. 2016].

One of the major strengths of UQ’s dataset form, is that researchers have the ability to select a range of licences, tailored to research data. While there is the option to assign one of the standard Creative Commons licences to the dataset, two UQ terms and conditions licence agreements have also been developed specifically for research datasets. These UQ licence agreements ensure that downloads and re-use of the data are properly acknowledged. Researchers who wish to use open-access datasets that are assigned one of the UQ licences, must agree to these terms and conditions through a click-through form that is displayed before the dataset is downloaded. An example of what is shown to the user is displayed in figure 5. These same licences will also be presented to researchers through the dataset form in UQ RDM, once integration with eSpace is complete.

![UQ eSpace](image)

**Figure 5 - example of UQ terms and conditions click-through agreement presented to users that are accessing an open-access dataset through UQ eSpace.**

While digital objects can have different licensing conditions, these all comply with the “relevant legislation and licence provisions” [UQ Library, 2013]. Digitised objects have permission or rights statements included in the item record within UQ eSpace. This allows users to understand how they...
can use the item. These right statements include copyright expired, licensed through Creative Commons, or protected by copyright. For those items protected by copyright, the record in UQ eSpace will indicate that users will need to contact the Library to obtain permission for re-use.

In addition to applying appropriate licensing, one of the key roles of the reusability principle is the ability of humans and machines to assess and select data on the basis of criteria relating to provenance information about the dataset or digital object [European Commission, 2018]. Provence refers to information regarding the origins, custody, and ownership of an item.

The UQ eSpace record of a digitised object provides provenance information in the abstract as well as the source field. The source field indicates exactly where in a collection the physical item can be found. This allows users to find information about the collection and understand where the original item is located. A date stamp in the record also indicates what year the item was uploaded.

UQ eSpace dataset records also have the ability to capture provenance information. For example, researchers can choose to upload a newer version of an existing dataset where changes to the underlying data or analysis are made. Researchers are strongly encouraged to include as much information as possible on the data analysis process to enable others to re-use the data, and potentially replicate the results. UQ RDM will further facilitate the capture of provenance information by allowing researchers to link key information about data collection and data analysis methods to their project. This can be done through uploading a dataset, which not only includes the raw data files, but also the relevant methodology and data analysis documentation that underpins a publication. Additionally, UQ RDM now allows researchers to request a digital research notebook for their project, to enable this day-to-day documentation to be stored and linked to the minimal project information (e.g. project name and collaborators). One of the key features is that there is the ability to archive the entire notebook into a PDF document when the research project is finished. This document could also be included in the underlying dataset/s that underpin publications.

Conclusion

The UQ Library continues to demonstrate their strong commitment to ensuring that digital objects and research datasets are findable, accessible, interoperable and reusable. This is evident through the ongoing development and enhancements of the supporting infrastructure – UQ eSpace and UQ RDM, and the relevant underlying policies and procedures.

---

9 [https://research.uq.edu.au/rmbt/drn](https://research.uq.edu.au/rmbt/drn)
References


