IJDC | Peer-Reviewed Paper

Putting the Trust into Trusted Data Repositories: A Federated Solution for the Australian National Imaging Facility

Andrew James Mehnert
CMCA, University of Western Australia

Andrew Janke
CAI, University of Queensland

Marco Gruwel
MWAC, University of New South Wales

Wojtek James Goscinski
MeRC, Monash University

Thomas Close
MBI, Monash University

Dean Taylor
BITS, University of Western Australia

Aswin Narayanan
CAI, University of Queensland

George Vidalis
MeRC, Monash University

Graham Galloway
National Imaging Facility

Andrew Treloar
Australian Research Data Commons

Abstract

The National Imaging Facility (NIF) provides Australian researchers with state-of-the-art instrumentation—including magnetic resonance imaging (MRI), positron emission tomography (PET), X-ray computed tomography (CT) and multispectral imaging – and expertise for the characterisation of animals, plants and materials.

To maximise research outcomes, as well as to facilitate collaboration and sharing, it is essential not only that the data acquired using these instruments be managed, curated and archived in a trusted data repository service, but also that the data itself be of verifiable quality. In 2017, several NIF nodes collaborated on a national project to define the requirements and best practices necessary to achieve this, and to establish exemplar services for both preclinical MRI data and clinical ataxia MRI data.

In this paper we describe the project, its key outcomes, challenges and lessons learned, and future developments, including extension to other characterisation facilities and instruments/modalities.

Submitted 06 December 2018 ~ *Accepted* 22 May 2019

Correspondence should be addressed to Andrew Mehnert, CMCA, The University of Western Australia, (M010) 35 Stirling Hwy PERTH WA 6009, Australia. Email: andrew.mehnert@uwa.edu.au

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

Introduction

Characterisation facilities provide researchers with access to instrumentation and expertise, to probe and measure the structures and properties of matter at the micro, nano and atomic scales. Characterisation is essential across natural, agricultural, physical, life and biomedical sciences and engineering, and encompasses a diverse range of techniques including optical, electron, X-ray and ion-beam techniques, magnetic resonance imaging (MRI), positron emission tomography (PET), mass spectrometry, ultrasound and cytometry (image, mass, flow).

In many characterisation facilities, the de facto policy with respect to data management and archiving is that it is the user's responsibility; i.e. the expectation is that the user will take their data away with them. This, of course, has several drawbacks. For the facility they include the inherent security and virus infection risks, the inability to monitor the quality of the data acquired and difficulty tracking outcomes such as publications and data reuse. For the researcher and their host institution, they include a failure to follow best practices with regard to data management and/or to meet legal and funding obligations, difficulty collaborating with other researchers and institutions, and the impracticality of moving and analysing the data as advances in characterisation technologies, e.g. cryo-electron microscopy and lattice light-sheet microscopy, generate ever-larger volumes of data. More generally this policy does not support the reuse of data.

The provisioning of a trusted data repository (TDR) service for instrument data offers a solution to these problems. A TDR service is essential for sharing data and ensures that data created and used by researchers is:

> 'managed, curated, and archived in such a way to preserve the initial investment in collecting them. Researchers must be certain that data held in archives remain useful and meaningful into the future. Funding authorities increasingly require continued access to data produced by the projects they fund, and have made this an important element in Data Management Plans. Indeed, some funders now stipulate that the data they fund must be deposited in a trustworthy repository' (CoreTrustSeal, 2016a).

The National Imaging Facility (NIF)[1] is a 200 million AUD Australia-wide network of geographically separate characterisation facilities providing state-of-the-art capability for the characterisation of animals, plants and materials for the Australian research community. During 2017, NIF nodes across four states collaborated on a national project to define the requirements and best practices necessary for a federated network of trusted data repositories for NIF and to establish exemplar services for both preclinical MRI data and clinical ataxia MRI data (Mehnert, 2018). The project "Delivering durable, reliable, high-quality image data" was jointly funded by the Australian National Data Service (ANDS) and Australian Research Data Services (RDS)[2] under the Trusted Data Repositories Program. Here *high-quality* refers to a NIF user's expectation that an animal, plant or material can be scanned and from the resulting data reliable outcomes/characterisations can be obtained (e.g. signal, volume, morphology) over time and across NIF sites; *durability* refers to guaranteed long-term availability of the data; and *reliability* means that the data is useful for future researchers.

Herein we describe the project, its key outcomes, challenges and lessons learned, and future developments including extension to other characterisation facilities and instruments/modalities.

[1] The National Imaging Facility (http://anif.org.au) is supported by the Australian Government's National Collaborative Research Infrastructure Strategy (NCRIS).

[2] The Australian National Data Service (ANDS), Research Data Services (RDS) and the National eResearch Collaboration Tools and Resources (Nectar) NCRIS projects combined on 1 July 2018 to become the Australian Research Data Commons (ARDC).

NIF Trusted Data Repositories Project

The project involved collaboration between NIF nodes at the University of Western Australia (UWA), University of Queensland (UQ), University of New South Wales (UNSW) and Monash University. The broad aim of the project – to enhance the quality, durability and reliability of data that is generated by NIF – was motivated both by NIF's desire to enhance the quality of the data associated with the use of its facilities, and the desire of the funders to facilitate the establishment of TDRs for institutions and disciplines and to learn what is needed to move beyond simple data storage services. Here, two types of trust need to be distinguished: Trust in the data repository services (the container) and trust in the quality of the data contained therein (the contents). In the former case, given that NIF comprises a number of nodes across Australia, this required the development of federated TDR services. In the latter, this required the development of a NIF-wide agreed process for acquiring quality or trusted data from instruments and uploading it to a TDR service.

Scope

From the outset it was recognised that a variety of technical approaches in relation to the acquisition and management of data existed at the different NIF nodes, and that the project needed to develop a solution that recognised these differences, ensured interoperability and was technically feasible. Moreover the solution would need to have the ability to make data available to a variety of compute platforms, including high-performance computing, cloud computing and GPU clusters. For these reasons, and given the 12-month funding time frame, it was decided:

1. Not to mandate a particular software platform for implementing the exemplar TDR services;
2. To leverage, or more specifically extend, existing open source instrument data management software platforms to implement the exemplar TDR services;
3. To focus on magnetic resonance imaging (MRI) instrumentation and data, but with the understanding that project outcomes could be easily adapted to, or serve as a basis for, other instruments/modalities;
4. Not to seek certification for the repository services but rather be guided by the requirements needed to attain certification.

Goals/Objectives

The following five objectives were defined to address the aim and scope of the project:

1. To develop a list of requirements necessary and sufficient for a basic NIF TDR service;
2. To develop a NIF-wide agreed process (NAP) necessary and sufficient for acquiring quality or trusted data, hereinafter called *NIF-certified data*, from instruments and uploading it to a NIF TDR service;
3. To develop instantiations of the NAP for repository services supporting two exemplar data collections, chosen to reflect both preclinical and clinical imaging within NIF, and

also the difference between data acquisition for a specialist repository and a more general repository;

4. To implement the exemplar repository services across the participating nodes using a mix of platforms, specifically MyTardis[3], ImageTrove[4] and XNAT[5];

5. To assess the resulting repository services against a relevant international agreed metric, the CoreTrustSeal[6] "Core Trustworthy Data Repositories Requirements" (CoreTrustSeal, 2016b).

Outcomes

Figure 1 provides a summary of the NIF TDR Project. A detailed presentation of the project outcomes is provided below. All project documents are hosted in NIF's Customer Relationship Management (CRM) system. The software developed during this project is hosted on GitHub[7].

Trust in the Data: "NIF agreed process for acquiring trusted data" (NAP)

The NAP (National Imaging Facility, 2017a) is a list of requirements that must be satisfied, and the process that must be followed, to obtain high-quality or NIF-certified data, suitable for ingestion in a NIF TDR service. The term *certified* is preferred over *trusted* to distinguish the trust in the data (contents) from the trust in the repository service (container).

Each data set acquired on an instrument is deemed to be associated with a Project that has a persistent identifier, hereinafter called the Project ID. For data to meet the definition of NIF-certified it must:

1. Have been acquired on a *NIF-compliant instrument*;

2. Possess *NIF-minimal metadata* including a cross-reference to the relevant instrument quality control (QC) data;

3. Include the native data generated by the instrument including the acquisition settings/parameters;

4. Include conversions to one or more open data formats where the native data is proprietary.

[3] MyTardis: http://www.mytardis.org
[4] ImageTrove: https://github.com/NIF-au/imagetrove
[5] XNAT: https://www.xnat.org
[6] CoreTrustSeal (https://www.coretrustseal.org) is an international community-based non-profit organisation offering core-level certification for a data repository.
[7] The project software is freely available for download from https://github.com/NIF-au/TDR

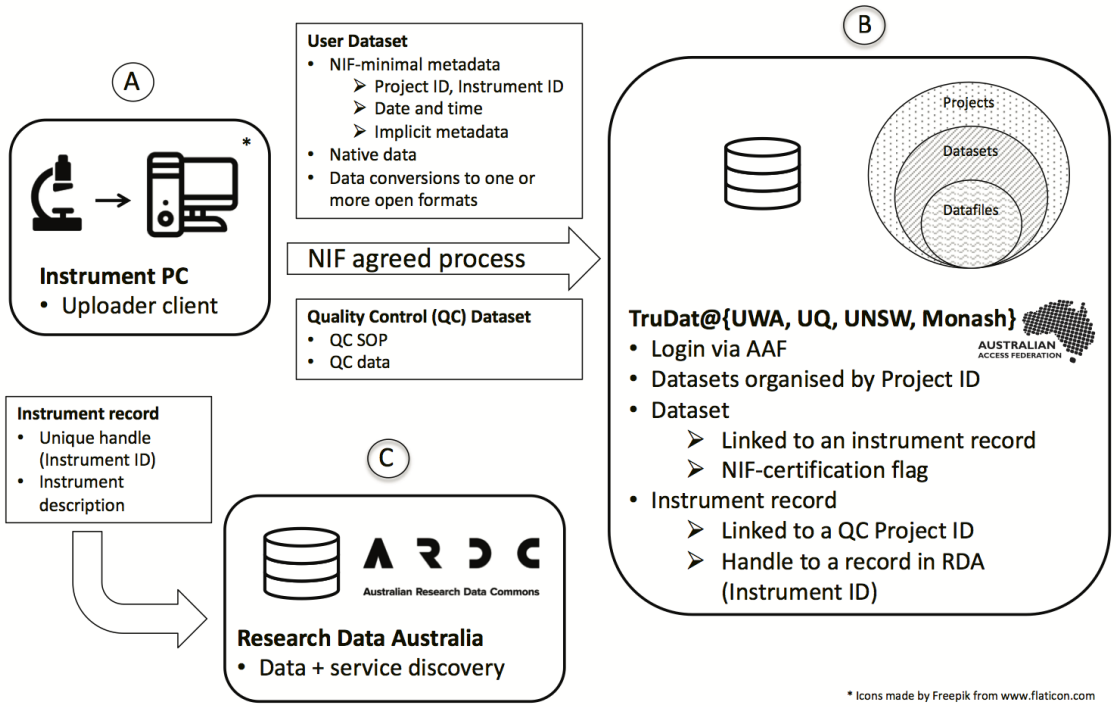

Figure 1. NIF TDR Project in a nutshell. (A) An uploader client is installed on the instrument PC or companion PC. The NIF agreed process (NAP) prescribes the necessary data and metadata needed to upload a NIF-certified user dataset or an instrument quality control (QC) dataset. (B) Each repository service organises data by Project ID and permits authentication using institutional credentials via the Australian Access Federation[8]. A NIF-certification flag indicates whether or not a dataset was acquired according to the NAP. (C) Each instrument must be registered in Research Data Australia[9] and possesses a unique handle that resolves to this record. The handle is also the Instrument ID.

The intent is that these requirements are necessary and sufficient for a user, or re-user, to establish the provenance and quality of the data and to determine whether it is fit-for-purpose (e.g., for inclusion in a meta-analysis) and to make the data reliable, i.e. useful for future researchers.

A bespoke NAP is defined for each instrument or group of instruments; e.g. (National Imaging Facility, 2017c; National Imaging Facility, 2017d). A future aspiration is to implement project-specific data quality checks; e.g. computing quality measures post-ingest – such as the signal-to-noise ratio (SNR), contrast-to-noise ratio (CNR), shape, geometric distortion and coverage – using standardised and reproducible workflows.

[8] The Australian Access Federation (https://aaf.edu.au) is Australia's leading identity broker, enabling access to online resources for the Education and Research sector. It provides subscribers with a national single sign-on. Subscribers include all Australian Universities, as well as CSIRO and other government research agencies.

[9] Research Data Australia is a web service to help researchers and other interested parties find, access, and reuse data for research from over one hundred Australian research organisations, government agencies, and cultural institutions.

NIF-compliant instrument

An instrument is deemed NIF-compliant if:

1. It has been assigned a persistent identifier, hereinafter called the Instrument ID, that is identically a handle that resolves to the record in 2;
2. A record has been registered for the instrument in Research Data Australia (RDA)[9] that contains the Instrument ID and a detailed description of the instrument and links to related websites;
3. A documented quality control (QC) process exists for the instrument (standardised across NIF nodes that possess the same instrument), including definitions of quality assurance measures, e.g. SNR, and who has access to the QC data, e.g. access might just be restricted to the instrument/facility manager;
4. A QC Project ID has been defined for the instrument and the QC data collected, as well as the QC standard operating procedure (SOP) and revisions thereof, are routinely uploaded to this project in the associated NIF TDR service;
5. An SOP for moving the data to the receiving NIF trusted data repository service must exist. Where possible, this should be standardised across NIF nodes that possess the same, or similar, instrument. This should also document the client-side scripts/software for sending data and metadata to the NIF-trusted data repository service (platform-specific).

The RDA record[10] should be updated by the instrument/facility manager to reflect changes to software and hardware that do not constitute a new system or model from the standpoint of the vendor. These changes should be dated so that it is possible to know the software and hardware used to acquire a particular dataset. If upgrades to hardware and software constitute a change to the system or model-number, then the instrument is deemed to be a new instrument and must be given a new Instrument ID and be registered in RDA.

The intent is that NIF-compliance bestows a measure of quality on the data acquired on the instrument.

NIF-minimal metadata

This includes both the implicit metadata (typically embedded in the native data generated by the instrument) and the following explicit metadata:

- The date and time the data was acquired;
- A checksum for fixity checking;
- Project ID of the project to which the data belongs;
- Instrument ID of the instrument the data was acquired from;

[10] An example is the record for the for Bruker BioSpec 9.4T MRI at UQ: http://hdl.handle.net/102.100.100/50043

- A flag to indicate whether or not the data is NIF-certified;
- A short description of the conversion tools used to generate the open data formats.

The acquisition date and time can be used to identify, for any given user dataset, which QC data set within the QC Project is the most recent and relevant.

Open data formats

The reason for including conversions of the instrument data to one or more open data formats is to obviate the need for preserving/maintaining conversion tools over time. The selected open formats should be those that are well documented and widely used in the research community. In the case of MRI data, such formats include NIfTI[11], MINC[12] and DICOM (DICOM Standards Committee, 2018).

Trust in the Repository Service: "NIF requirements for a trusted data repository service"

This document (National Imaging Facility, 2017b) does not mandate a particular software platform, but rather is a checklist of requirements that a NIF trusted data repository should satisfy. This includes a self-assessment against the "Core Trustworthy Data Repositories Requirements" (CoreTrustSeal, 2016b) with an aspiration to achieve certification, as well as the following additional NIF-specific requirements:

- **Project ID**: All data in the repository must belong to one or more projects. Each project must have a Project ID. Only those users who are a member of a given project can access its data. The reason for organising data by Project ID, rather than by user identifier, is that the latter can change over time, cease to exist or be recycled as users come and go, graduate, change employment and so on.
- **Instrument ID**: All NIF-certified data in the repository must originate from a NIF-compliant instrument and thus be associated with an Instrument ID.
- **Quality control Project ID**: A quality control (QC) project must exist for each Instrument ID.
- **Authentication and interoperability**: The service must implement sign-on via the Australian Access Federation.[8] This means that it is possible to log into each service using home institution credentials; e.g. a UWA user can log into the UNSW, UQ or Monash services using UWA credentials. This permits users to download data from any of the repository services in the federation. It also permits the transfer of data between services.
- **Deployability**: The service must be easy to deploy at different NIF nodes. This, in part, means that all of the code, documentation, etc., must be accessible to and freely available across all nodes.

[11] NIfTI: https://nifti.nimh.nih.gov/
[12] MINC: http://www.bic.mni.mcgill.ca/ServicesSoftware/MINC

- **Service**: Each repository service must include a help desk service. At minimum this should be a local service provided by the host node. An aspirational solution is a NIF-wide centralised help desk service.

Exemplar Repository Services

Trusted data repository services were implemented for two exemplar collections:

1. Preclinical MRI data (with mouse brain data as an example) acquired across three different NIF nodes – UNSW, UQ, UWA – using a common instrument make and model: Bruker[13] BioSpec 9.4T MRI;
2. Clinical ataxia MRI data acquired using a Siemens[14] Skyra 3T MRI scanner in support of a Monash-proposed International Ataxia Imaging Repository (IAIR).

Table 1 lists the repository service implemented at each collaborating NIF node.

Table 1. Implemented TDR services.

NIF node	URL	Base software platform*
UNSW	https://mytardis.unsw.edu.au	MyTardis[3]
UQ	https://imagetrove.cai.uq.edu.au	ImageTrove[4]
UWA	https://trudat.cmca.uwa.edu.au	MyTardis[3]
Monash	https://ataxia.erc.monash.edu.au	XNAT[5]

*Supported by additional software and scripts developed in this project.

Key features of the services:

- **Easy to deploy:** UWA and Monash developed docker[15] compose scripts for the MyTardis and XNAT services respectively. The MyTardis script was used to deploy services at both UWA and UNSW;
- **Data upload/ingest:** UQ implemented client-side Python scripts, using the MyTardis RESTful API, for uploading NIF-certified data to ImageTrove. UWA/UNSW implemented a workflow using the MyTardis MyData client. Monash implemented an XNAT plugin for uploading non-DICOM files.

Self-assessment Against the "Core Trustworthy Data Repositories Requirements"

Each NIF node completed a self-assessment against the "Core Trustworthy Data Repositories Requirements". Requirement R0 is a statement of context. Requirements R1-R16 additionally require the specification of a level of compliance: 0 – Not applicable; 1 – The repository has not

[13] Bruker: https://www.bruker.com
[14] Siemens: https://www.siemens.com
[15] Docker: https://www.docker.com

considered this yet; 2 – The repository has a theoretical concept; 3 – The repository is in the implementation phase; and 4 – The guideline has been fully implemented in the repository. The responses were remarkably similar, with a variation across the nodes of no more than one for 11 of the 16 requirements. Table 2 lists the five requirements with the greatest variation. They reflect the local differences in the maturity of the service and level of commitment of the node.

Table 2. Comparison of self-assessments against the "Core Trustworthy Data Repositories Requirements": Responses with a variance greater than one across the field.

#	Requirement description	Monash	UNSW	UQ	UWA
R3	Continuity of access: The repository has a continuity plan to ensure ongoing access to and preservation of, its holdings.	1	4	4	4
R4	Confidentiality/Ethics: The repository ensures, to the extent possible, that data are created, curated, accessed, and used in compliance with disciplinary and ethical norms.	4	2	3	2
R9	Documented storage procedures: The repository applies documented processes and procedures in managing archival storage of the data.	1	3	3	3
R12	Workflows: Archiving takes place according to defined workflows from ingest to dissemination.	4	2	3	2
R13	Data discovery and identification: The repository enables users to discover the data and refer to them in a persistent way through proper citation.	1	2	3	2

Challenges and Lessons Learned

The cross-institutional nature of the project – four NIF nodes at four universities – posed two major challenges. The first was in terms of project management and execution. The second was in terms of accommodating institutional differences with respect to the existing technical approaches in relation to the acquisition and management of data. These challenges were solved as follows. A project lead was appointed at each node and one of these (UWA) was also appointed the Project Manager (PM). A Steering Committee was established with membership from the funding bodies (ANDS and RDS), each institution (UNSW, UQ, UWA and Monash) and NIF to provide oversight. The Committee met with the PM via Zoom[16] twice during the course of the project. Given the geographical separation between the nodes—four states with time zone differences of as much as three hours depending on the time of year—face-to-face workshops were held at each node in turn during the planning and initial implementation phases of the project. A final face-to-face closure workshop was also held. Overall project management was achieved using regular fortnightly minuted Zoom meetings to coordinate the participating nodes, to report on progress and problems, to problem solve, to reach consensus on various design and implementation issues, and to plan future work. Project documents were shared and worked on collaboratively using the Google[17] Docs web-based application. This strategy facilitated both timely completion of the project and the resolution of differences. One

[16] Zoom: https://www.zoom.us
[17] Google: https://www.google.com

of these differences was the choice of software platform to implement the exemplar services. This was accommodated by refraining from mandating a particular platform in the "NIF requirements for a trusted data repository service". Another was that the UWA node trains users to operate instruments and to acquire their own data, whilst at other nodes an instrument operator takes responsibility for scanning a user's sample. This was resolved by making the NAP sufficiently flexible to permit either client-side NIF-certification of data by the expert and trusted instrument operator, or server side validation via post-ingest filters.

Lessons learned from the project:

- That regular meetings and collaborative platforms for documentation and meetings were essential for the development of the NAP, NIF TDR requirements and QC SOPs;
- These documents will evolve with time and need to be adapted for different instruments;
- CoreTrustSeal's "Core Trustworthy Data Repositories Requirements" is a useful guide to the establishment of services, as well as a metric against which to assess services. Even without certification, it is useful in highlighting differences in the maturity and commitment of services offered by different nodes across the federation;
- The repository service should not be named after the underlying software platform used to implement it. It is better to use platform-agnostic branding, e.g. TruDat@UWA, in case it becomes necessary (performance issues, costs, etc.) to migrate to another platform in the future.

Future Developments

The project partners are committed to maintaining the NIF TDR services for at least ten years. The project team will meet annually post-funding (the inaugural meeting took place in June 2018), via Zoom or similar, to report on the status of each of the services and to take remedial action as needed. For example, UQ and Monash have agreed to host the data from any of the other services should the need arise. The meetings will also serve to progress the post-funding plan that includes:

- Integration of additional NIF instruments – to date this includes several additional MRI, MRI-PET and PET-CT scanners at UQ and Monash;
- Integration of Microscopy Australia[18] instruments – to date several electron microscopy instruments at UWA (which is both a NIF node and Microscopy Australia node) have been integrated;
- New national and international service deployments;

[18] Microscopy Australia (https://micro.org.au) is a rebranding of the Australian Microscopy and Microanalysis Research Facility (AMMRF). It is an Australia-wide network of geographically separate characterisation facilities supported by NCRIS that provides researchers with access to optical, electron, X-ray, scanning and ion-beam techniques, along with world-leading platforms such as atom probe tomography and advanced ion probes.

- Integration with the national Research Activity Identifier (RAiD)[19] project;
- CoreTrustSeal certification;
- Support for the ongoing ARDC-funded "Characterisation Data Enhanced Virtual Laboratory" Project[20] that is, in part, expanding the existing remote desktop environment for the analysis of characterisation data (called the CVL or Characterisation Virtual Laboratory) into a federated service.

This project was one of three funded under the Trusted Repositories Program to identify issues that might be relevant to a wider national rollout. The results are being used to inform ARDC funded repository activity during the 2018/19 Australian financial year.

Conclusions

The benefits of this project for NIF users and the broader characterisation community include: Reliable and durable access to data, improved reliability of research outputs and the provenance associated with it, making NIF data more F.A.I.R. (Wilkinson et al., 2016; Australian National Data Service, 2017), easier linkages between publications and data and stronger research partnerships. The benefits for NIF include improved data quality, improved international reputation and the ability to run multi-centre trials. The benefits for research institutions include enhanced reputation management, a means by which to comply with the Australian Code for the Responsible Conduct of Research (National Health and Medical Research Council, Australian Research Council & Universities Australia, 2018) and an enhanced ability to engage in multi-centre imaging research projects. The project outcomes can easily be adapted to other instruments/modalities and facilities. Indeed, as noted above, several additional instruments have been integrated across the participating NIF nodes, including several instruments from Microscopy Australia.

Acknowledgements

We acknowledge funding from the Australian National Data Service (ANDS) and Research Data Services (RDS). We also acknowledge in kind, support from the National Imaging Facility (NIF). ANDS, RDS and NIF are supported by the Australian Government, through the National Collaborative Research Infrastructure Strategy Program (NCRIS).

References

Australian National Data Service. (2017). *The FAIR data principles*. Retrieved from https://www.ands.org.au/working-with-data/fairdata

[19] RAiD: https://www.raid.org.au
[20] Characterisation Data Enhanced Virtual Laboratory Project: https://www.ands-nectar-rds.org.au/researchdomainprogram, https://www.cvl.org.au/cvl-project

CoreTrustSeal. (2016a). *An Introduction to the Core Trustworthy Data Repositories Requirements*. Retrieved from https://www.coretrustseal.org/wp-content/uploads/2017/01/Intro_To_Core_Trustworthy_Data_Repositories_Requirements_2016-11.pdf

CoreTrustSeal. (2016b). *Core Trustworthy Data Repositories Requirements v01.00*. Retrieved from https://www.coretrustseal.org/wp-content/uploads/2017/01/Core_Trustworthy_Data_Repositories_Requirements_01_00.pdf

DICOM Standards Committee. (2018). *DICOM PS3.1 2018d – Introduction and Overview.* Retrieved from http://dicom.nema.org/medical/dicom/2018d/output/html/part01.html

Mehnert, A. (2018). National network of trusted data repositories. *National Imaging Facility Quarterly Newsletter, Q1.* Retrieved from http://anif.org.au/wp-content/uploads/NIFNewsletter_Issue1_2018_digital.pdf

National Health and Medical Research Council, Australian Research Council & Universities Australia. (2018). *Australian Code for the Responsible Conduct of Research.* Retrieved from https://www.nhmrc.gov.au/about-us/publications/australian-code-responsible-conduct-research-2018

National Imaging Facility. (2017a). *NIF agreed process for acquiring trusted data.* Brisbane: National Imaging Facility.

National Imaging Facility. (2017b). *NIF requirements for a trusted data repository service.* Brisbane: National Imaging Facility.

National Imaging Facility. (2017c). *NIF agreed process for acquiring trusted data: UNSW Bruker BioSpec 9.4T MRI, UQ Bruker BioSpec 9.4T MRI, UWA Bruker BioSpec 9.4T MRI.* Brisbane: National Imaging Facility.

National Imaging Facility. (2017d). *NIF agreed process for acquiring trusted data: Monash Siemens Skyra 3T.* Brisbane: National Imaging Facility.

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3, 1-9. doi:10.1038/sdata.2016.18