PITSCHI: A CLOWDER-BASED END-TO-END DATA MANAGEMENT TOOL

WORK PACKAGE 4: BIG-DATA ELECTRON AND CORRELATIVE MICROSCOPY FROM INSTRUMENT TO PUBLICATION

Hoang Anh NGUYEN

AUSTRALIAN CHARACTERISATION COMMONS AT SCALE

OCTOBER 2022

DOI: 10.5281/zenodo.7183431

Dr Hoang Anh NGUYEN
Research Computing Centre, The University of Queensland

Contact:

Prof David Abramson
Research Computing Centre, The University of Queensland
david.abramson@uq.edu.au

Prof Roger Wepf
Centre for Microscopy and Microanalysis, The University of Queensland
r.wepf@uq.edu.au

Australian Research Data Commons

This project is supported by the Australian Research Data Commons (ARDC) and the following partners. The ARDC is enabled by NCRIS.

Contents

Executive summary

Data management at research facilities that operate a range of big-data-producing instruments has become ever more critical as the amounts of data generated by such instruments keep increasing thanks to technological advances. In this report we present the data management tool Pitschi. Based on the data management framework Clowder, Pitschi has been designed for the Centre of Microscopy and Microanalysis at The University of Queensland as an end-to-end solution that supports the entire research data lifecycle by storing, indexing, and annotating data generated at the centre. The workflow in Pitschi is described, particularly how it is fully integrated with the university's infrastructure. Importantly, Pitschi can also be adapted to other institutional environments. This work was undertaken under Work Package 4: "Big-data electron and correlative microscopy from instrument to publication" of the Australian Characterisation Commons at Scale (ACCS) project.

1 Introduction

Scientific imaging instruments with modern, fast CMOS detectors are generating increasingly large datasets and volumes of data, hence, data management has become more critical. This is particularly true in the context of large multi-user facilities such as the Centre for Microscopy and Microanalysis (CMM) at The University of Queensland as it operates a wide range of instruments and many of them are big-data producers. A central data repository to store, index, annotate data not only allows its researchers to search, browse and retrieve their data easily, but also has the potential to harvest metadata to enrich these datasets.

A previous review on data management at electron microscopy facilities at Australian universities noted that while many of the facilities interviewed had trialled a range of proprietary and non-proprietary tools to assist them in managing data, it was often difficult to find a solution that was suitable to deploy across all the instruments of a facility, in particular for facilities like CMM that owned a diverse range of instruments.[1] It was recommended to develop a data management tool that would be suitable to the electron microscopy community (and beyond), in particular in the case of big data. CMM has extensively evaluated various repository options including IMS,[1] OMERO [2] and Clowder [3] to name a few. Eventually, we have decided to use Clowder for our repository implementation, which is named Pitschi (Particle Imaging depoT using Storage CacHing Infrastructure). Pitschi adheres to the FAIR data principles.[4] Pitschi is fully integrated with the instrument-booking system and the storage infrastructure at The University of Queensland. It provides end-to-end process data management, that is from capturing raw data to transferring them to storage collection and finally ingesting/indexing the data into the repository. As part of the ingest process, metadata of supported file types are extracted automatically. These metadata are then used to facilitate search and discovery. Once the data are ingested in Pitschi, they are available in various platforms such as high-performance computing facilities, personal computers, and processing platforms such as the Characterisation Virtual Laboratory (CVL).[2] Data transport is arranged transparently using the university's Metropolitan Data Caching Infrastructure (MeDICI).[5]

This report first introduces the context of the project, including CMM and the university's storage and computing infrastructure. It then presents the evaluation process of data repositories, which explains why we chose the Clowder data management framework. Lastly, the report details the implementation and deployment of Pitschi at CMM. While the report is tailored towards CMM workflows and university infrastructure, it can be used as a reference for other institutions. Importantly, we do believe that Pitschi can be adapted to other institutions with different computing and storage infrastructure.

This work was undertaken under the Australian Characterisation Commons at Scale (ACCS) project, in particular under Work Package 4: "Big-data electron and correlative microscopy from instrument to publication". This report is intended as the conclusion of the work planned for deliverable WP4.9: "Prototype and illustrate current and potential data-management software for the big-data electron microscopy community".

---

1 https://imagic.ch/en/imagic-ims
2 https://www.imagingtools.org.au/characterisation-virtual-laboratory

2 Background

This section provides background information on the project, including the computing and storage infrastructure at The University of Queensland (UQ), the Research Infrastructure Management System (RIMS) and the Centre for Microscopy and Microanalysis (CMM).

2.1 UQ computing and storage infrastructure

Similar to other large universities, the computing and storage infrastructure at UQ is complex consisting of multiple computing and storage resources spreading across different locations. In addition, the archival data and some of those computing resources are housed in an external commercial data centre. The Metropolitan Data Caching Infrastructure (MeDiCI) was designed by the Research Computing Centre (RCC) at The University of Queensland to provide seamless access to data across multiple sites, regardless of where they are created, manipulated and archived. MeDiCI, a distributed file system, holds copies of data on campus until they have not been required for some time.[5] It allows data currently being accessed by researchers to be cached on campus, and moves data not accessed for a while back to the Polaris Data Centre.[3] In addition, researchers can access their data from various platforms, either via direct mount from workstations and clusters or via a number of standard protocols such as NFS, Samba, SMB and ISM's GPFS native clients.[5] The schematic diagram of MeDiCI is shown in Figure 1. At the time of writing this report, UQ operates two main High-Performance Computers (HPCs), namely Tinaroo and Wiener. The third system, Bunya,[4] is planned to enter production later in the year. While Tinaroo is a traditional HPC cluster,

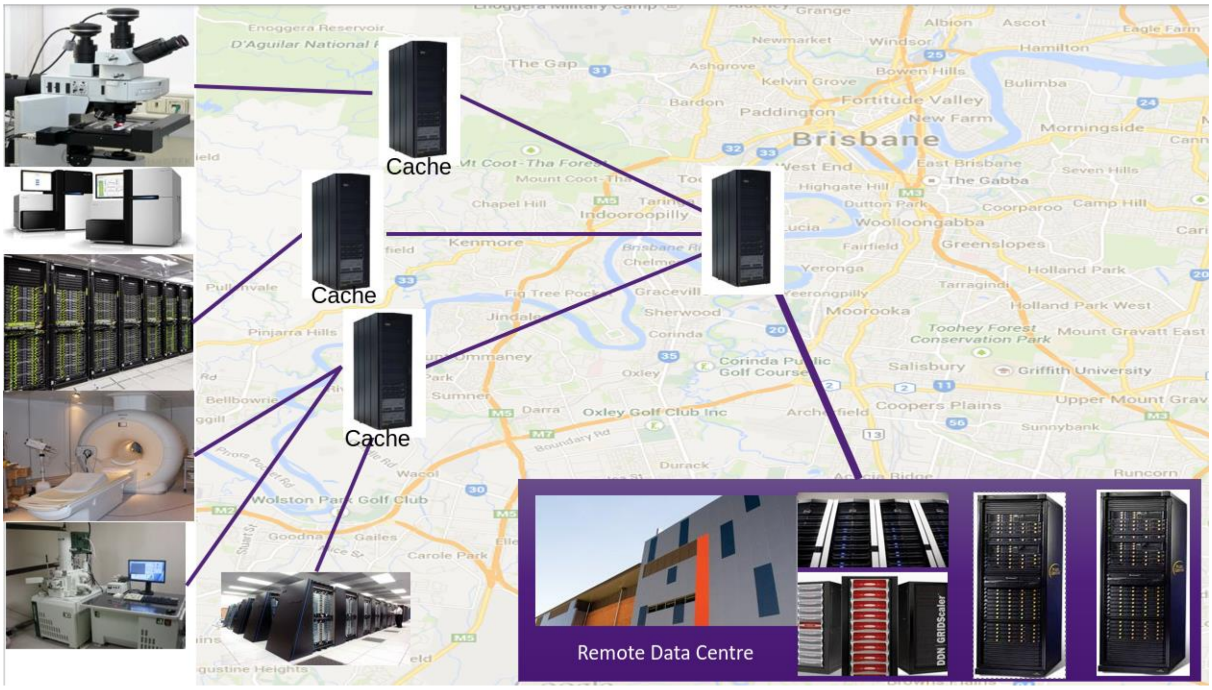

Figure 1 The data fabric (MeDiCI) at The University of Queensland.[5]

[3] https://rcc.uq.edu.au/data-storage
[4] https://rcc.uq.edu.au/article/2022/06/uq-acquires-new-supercomputer

focusing on solving tightly coupled parallel jobs requiring a very large number of computer cores to be applied to a given problem at the same time, Wiener is a HPC with a lot of graphical processing units (GPUs) focusing on accelerating image processing and artificial intelligence. More detailed information about the HPCs can be found in the RCC website.[5]

Research storage provision is handled via a system called Research Data Manager (UQRDM).[6] The researchers request storage via the UQRDM website with detailed project metadata such as grants, ethics, collaborators, and the amounts of storage required and projected during the lifetime of the project. Each collection is provisioned with 1TiB of capacity by default, but it can be increased later by request (TiB, tebibyte; 1 TiB = 1024^4 bytes ≈ 1.1 TB, terabyte). Once the storage collection is provisioned, it can be mounted to various platforms such as workstations and HPCs thanks to MeDiCI. This workflow is illustrated in

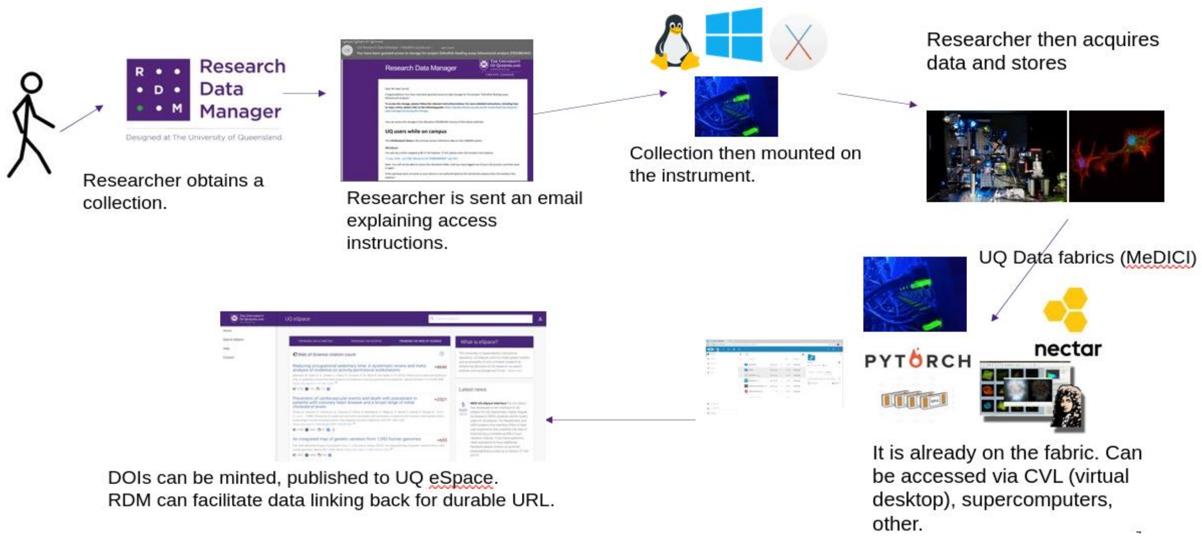

Figure 2 A typical workflow at The University of Queensland.

Figure 2.

2.2 Research Infrastructure Management System (RIMS)

RIMS is UQ's centralised system for instrument management, booking and accounting. This is a customised version of the booking system Stratocore (formerly referred to as PPMS).[7] For CMM, RIMS helps manage instrument booking and usage, users and billings, as well as standardise various processes, including user training and incident management. From the users' point of view, RIMS make it easy to reserve a slot for an instrument.

2.3 Centre for Microscopy and Microanalysis (CMM)

The Centre for Microscopy and Microanalysis (CMM) is The University of Queensland's

[5] https://rcc.uq.edu.au/high-performance-computing
[6] https://rdm.uq.edu.au
[7] https://www.stratocore.com/our-solution/technical-features

microscopy, characterisation and analysis core facility and part of the national network of Microscopy Australia facilities. CMM has more than 37 academic, professional and administrative staff and operates on 5 main sites on the UQ St. Lucia campus, including:

- the faciilty at the Australian Institute for Bioengineering and Nanotechnology (AIBN);
- the facility at the Queensland Bioscience Precinct (QBP);
- the faciltiy in the Faculty of Science Chemistry building;
- UQ Materials Performance ad advanced engineering; and
- the facility at Hawken Engineerig building

CMM supports a diverse range of instrument platforms, including light, xray and electron microscopy, and microanalysis techniques, electron and X-ray diffraction and imaging mass spectrometry methods. These instruments vary greatly in terms of configurations and detectors types, the amounts of data they generate and the capture procedures.

Prior to Pitschi, data management was done in an *ad hoc* way and mostly relied on users to manage their data. Each collection was mounted permanently on each instrument for storing instrument data. A special credential for read-only was used by all the users so that they could copy the data from these instrument collections to the UQRDM collections of their projects. A simplified diagram of how instrument collections work is illustrated in Figure 3.

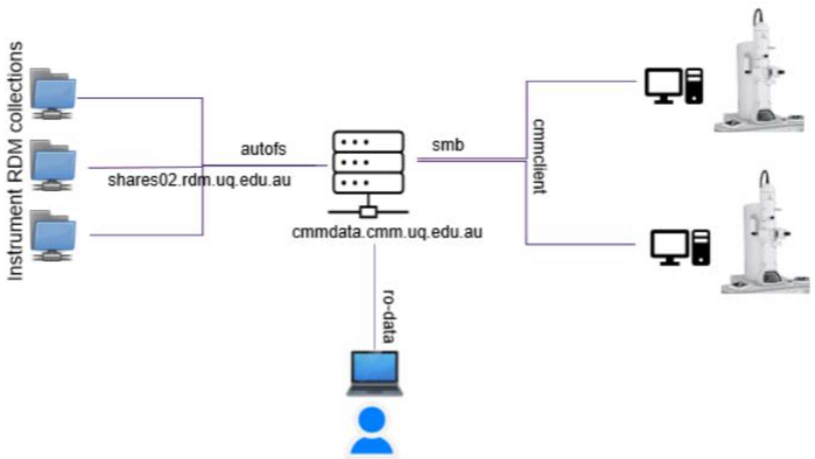

Figure 3 Instrument collections at the Centre for Microscopy and Microanalysis (CMM).

While the current configuration has met CMM's operational requirements so far, there are few limitations. The first limitation is security. Anyone with the read-only credentials can see all the data, and worse, data in the instrument collections can be modified from the instrument computers. Secondly, it is not possible to search for files or folders. This will make it unwieldy for users to manage their data, especially when there is lots of data. Third, there is no data management standard, as it is done *ad hoc* and by individual users.

As a result, there needs to be a central repository for data from CMM instruments. This central data repository to store, backup, index, annotate data not only allows researchers to search, browse and retrieve their data easily but also has the potential to harvest metadata to enrich these datasets and share datasets upon demand. To overcome the limitations of the current configuration, the central repository needs to address the shortcomings of the existing model. Thus the data repository needs to meet the following requirements:

- it should adhere to the FAIR principles;[4,6]
- it should automate the data movement from the instrument computer to the user's UQRDM collection. This is to ensure that only people in the project have access to the data. The solution should make use of UQ's storage infrastructure, such as MeDICI;
- it should automate metadata extraction. The metadata extracted from ingested files can be used later to enrich datasets. As CMM operates a variety of instruments, it should be relatively straightforward to add support for different data types;
- it should be user-friendly to use and to deposit data; and
- it should be reliable and secured.

3 Clowder

Clowder, an open-source data management framework from the National Center for Supercomputing Applications (NCSA) in the USA,[3] was selected to implement CMM's data repository. We selected Clowder over a wide range of data management systems such as iRODs,[7] MyTardis,[8] OMERO,[2] XNAT [8] and CKAN,[9] mainly because of the following reasons:

- Clowder offers support for a wide range of data formats across various disciplines;
- Clowder provides powerful search capabilities. The search can be done by item (directory, files), names and metadata;
- Clowder is extensible by adding support for new file types is relatively straightforward;
- Data ingest in Clowder is flexible: data can be ingested from the Web client or moved to the server side and then registered via the server-side APIs. The latter method is particularly important as it means MeDiCI can be used for data movement.

Figure 4 shows the data model in Clowder. [3] *Space* is the highest level in its data hierarchy. Spaces enforce access control and allow space administrators to share datasets and collections with specific users using role-based access controls. Each space defines three levels of default permission: viewer, editor and administrator. Custom permissions can also be defined by administrators. Under each space, there are *datasets* and *collections*. Datasets are the main resource types and contain files, folders and metadata. Each dataset has a name, description, owner and licence. Optionally, datasets and files can be tagged with keywords for discovery, receive comments by users and include generic metadata added manually by users or automatically by metadata extractors. Collection provides a way to organise datasets.

[8] http://www.mytardis.org

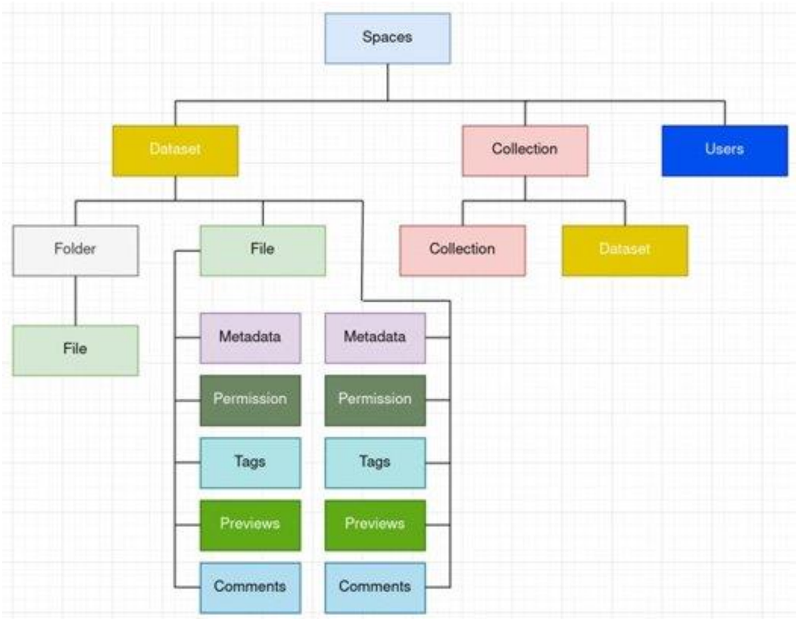

Figure 4 Schematic representation of the data model in Clowder.[3]

4 Pitschi—a central repository for CMM

4.1 Mapping Clowder data model to RIMS projects

In Pitschi, the UQRDM collections of all the projects are mounted as NFS in the nodes running Clowder. Each Clowder space is mapped to a project in RIMS, and each dataset in Clowder is mapped to a folder in UQRDM. Files and folders in the dataset correspond to the files and folders structure in the given dataset folder in UQRDM. The project details in RIMS are synchronised weekly automatically to the Pitschi database. These details include the project name, description, and project members. Users can then log into Pitschi via Australian Access Federation.[9] In order to specify the UQRDM collection for each RIMS project, we have added a special field in RIMS project profile called "UQRDM Collection #". Pitschi only synchronises projects (together with users of the projects) with a valid field setup to its database.

4.2 Project registration

We modified the project registration process at CMM to include more information regarding the storage collection used by the projects. If there is no UQRDM collection for a project, we then request a storage collection to be created on behalf of the users. Otherwise, the users need to add a Pitschi service account as a collaborator to the collection storage. This service account will be used to ingest data into the UQRDM collection on behalf of users. The whole registration process in Pitschi is illustrated in Figure 5.

[9] https://aaf.edu.au

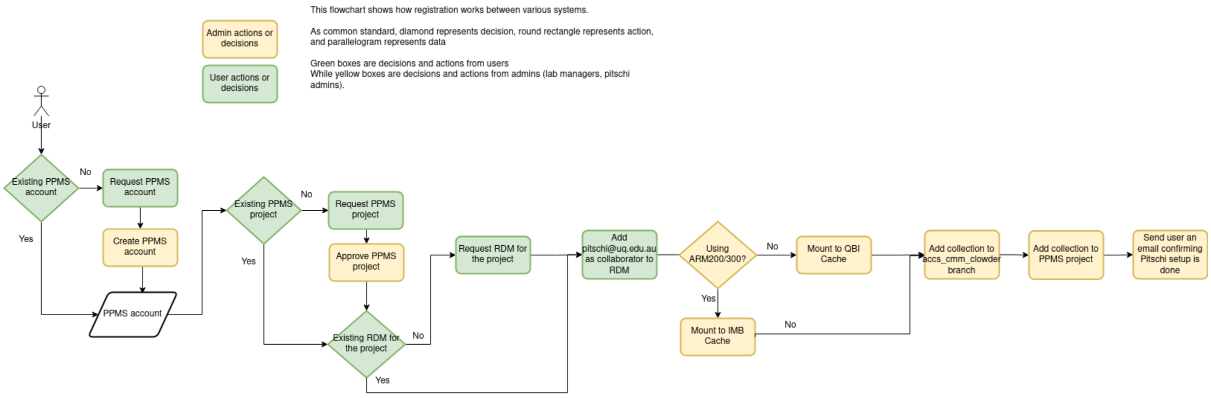

Figure 5 Pitschi registration process.

4.3 Data ingest

Pitshi supports two methods of data ingestion: from the Web or from the instrument computers. Either way, the data will end up in the UQRDM collection storage corresponding to the project and metadata of the uploaded files (of supported file formats) will be extracted automatically by Clowder.

Figure 6 shows how Web ingest works. The data is uploaded directly to Pitschi via the Clowder Web interface. The uploaded files are first landed in a temporary storage, then moved to the appropriate UQRDM collections. This method is recommended for small datasets. For big datasets, Pitschi makes use of MeDICI: data are ingested into campus caches from instrument computers using SMB protocols and then registered with Clowder via its API. With this method, the collection is mounted when a session starts, the data are written to the local hard drive first, then synchronised periodically to the mounted UQRDM collection. Pitschi keeps track of every single file synchronised to UQRDM for logging purposes. The next section discusses in more detail how data ingest from instrument computers works.

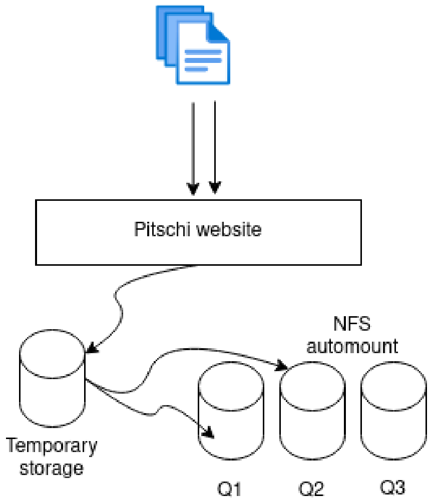

Figure 6 Web ingest of data with Pistchi.

4.4 Pitschi components

Figure 7 shows the overall architecture of Pitschi. On the server side, there are two main components, namely Clowder (and its dependents) and XAPI. Note, there are also multiple clients created to meet the requirements of the instrument configuration and dataflow. Table 1 lists the different components that have been developed for Pitschi. All of them have been made available in the Github repository of RCC.

Components on the server side include:

- XAPI: this API was created to 1) make querying booking information from RIMS easier and 2) hide the API key of RIMS away from users. This API pulls information from RIMS API every 10 minutes for updates regarding bookings, projects, users; and
- Clowder: this is a fork of the Clowder data management framework. It pulls data from RIMS API weekly to create newly created projects with valid "UQRDM Collection #" field and their users. Clowder uses a number of underlying components:
    - MongoDB: for database management;
    - ElasticSearch: for searching items and metadata;
    - various extractors to extract metadata out of ingested files; and
    - RabbitMQ: for communicating between Clowder and its extractors.

As stated earlier, CMM operates a wide range of instruments in the areas of electron microscopy, microanalysis, X-ray diffraction and imaging mass spectrometry. These instruments are varied in terms of computer configurations, acquisition software and acquisition procedures. As an example, while many instruments are attached to a single

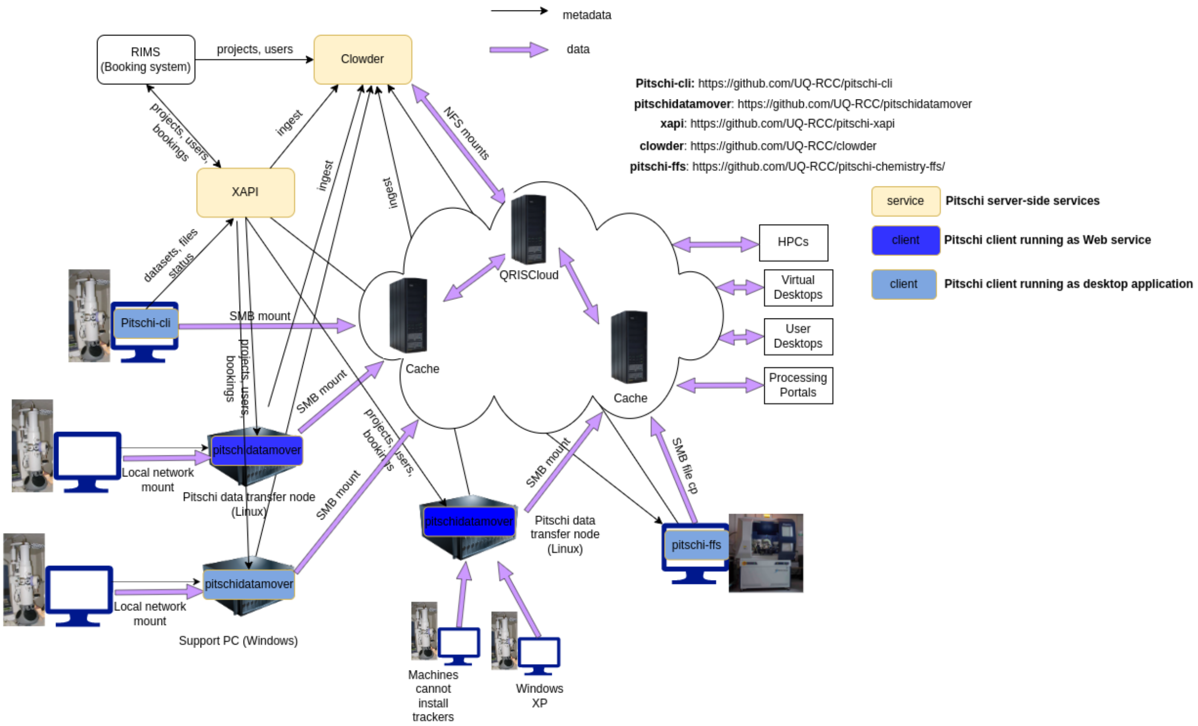

Figure 7 Pitschi architecture.

Table 1 List of the Pitschi components developed in this work.

Component	GitHub URL	Publicity
Clowder	https://github.com/UQ-RCC/clowder	public
XAPI	https://github.com/UQ-RCC/pitschi-xapi	public
Deployment script	https://github.com/UQ-RCC/ansible-swarm-clowder	public
pitschi-cli	https://github.com/UQ-RCC/pitschi-cli	private
pitschi-dashboard	https://github.com/UQ-RCC/pitschi-dashboard	public
Documentation	https://uq-rcc.github.io/pitschi-docs	public
Pitschi-datamover Web app	https://github.com/UQ-RCC/pitschidatamover-desktop	private
Pitschi fee for service client	https://github.com/UQ-RCC/pitschi-chemistry-ffs	private
wdf-extractor	https://github.com/UQ-RCC/wdf-extractor	public
dm3-extractor	https://github.com/UQ-RCC/dm3-extractor	public
BioFormats-extractor	https://github.com/UQ-RCC/bioformat-extractor	public

computer with relatively modern operating systems (Windows 7 and Windows 10), some instruments are hooked with multiple computers (separate microscope and camera computer, and even computers for various sensors). Some instruments are still connected to Windows XP, which cannot be exposed to university-wide network due to security concerns. Since we cannot create each client for each instrument, we categorise these instruments into four groups, based on the possibility that one client can be used for instruments in those groups.

- Group 1: Instruments with one or two computers. This group covers the majority of instruments in CMM. With this group, UQRDM can be mounted directly to the instrument computers, and data can be synchronised from local hard drive to the RDM. Pitschi-cli, a Windows client, is created to serve this group.
- Group 2: This group includes instruments with non-modifiable camera computers (as part of warranty requirements) or camera computer that cannot connected to the instrument network. With this, a third-party computer connected to the instrument network is used as a data mover.
- Group 3: Instruments with Windows XP. There are a small number of instruments with Windows XP.
- Group 4: Instruments, such as some using X-rays, can hold multiple samples from different projects. This requires a different way to distribute raw data to project collections.

At the time of writing this report, Pitschi only supports instruments in Groups 1 and 2.

4.1 Pitschi clients

Pitschi-cli

This consists of several Windows executables written in PowerShell. This client is designed to serve instruments in Group 1. These scripts are invoked by PPMS Tracker at user login or logout. Figure 8 describes all the steps in the Pitschi data-flow. In order to use the instrument computer, users need to log in with institution credentials. In case a user logs in without any booking, Pitschi will display a message indicating that the Pitschi flow is not enabled. If there is a valid booking and the user project has a valid "UQRDM Collection #" tag, the project UQRDM is then mounted, a project folder is created in the local disk, and a shortcut call "Today" is then created on the desktop pointing to this project folder. During acquisition, users need to save data to this "Today" shortcut folder and Pitschi will automatically synchronise data to the UQRDM project. Once the user logs out, Pitschi will finalise the synchronisation, then dismount the UQRDM collection from the computer. As shown in Figure 9, Pitschi will send the user two emails: the first email is sent shortly after logout confirming that the data are safely stored in UQRDM together with how to access the data; and the second email is sent once the data are indexed by Clowder.

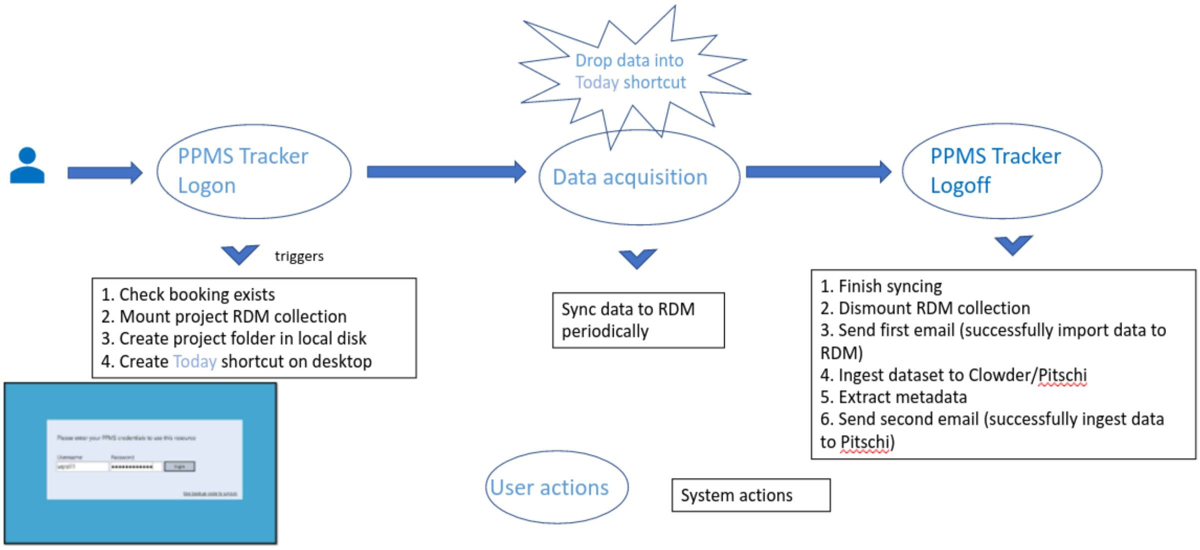

Figure 8 Pitschi data-flow with Pitschi-cli.

Pitschi-datamover web application

This Web application is created to serve instruments in Group 2 with a Linux third-party computer (referred to as "datamover") attached to the camera computer. With this, the Web application is installed in this third-party computer and users access this Web application from the camera computer to start and stop their acquisitions (or experiments). Figure 10 (panel a) shows the home page of the Web application once users log in using AAF. As shown in Figure 10b, a list of projects is presented when the user starts a new experiment. As part of the process, the UQRDM collection in the given project is mounted to the datamover computer,

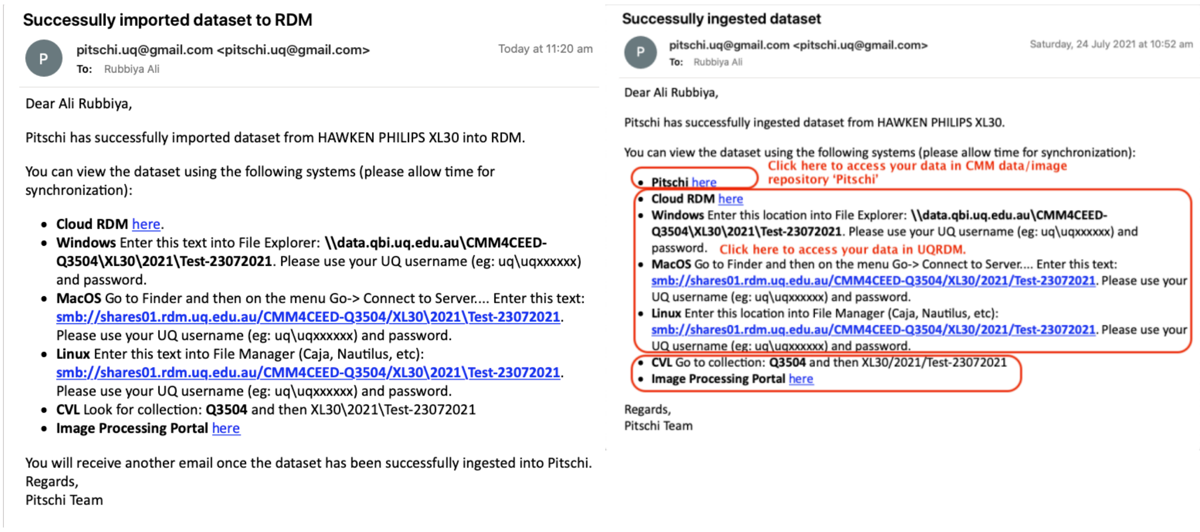

Figure 10 Emails sent automatically by Pitschi after user's logout.

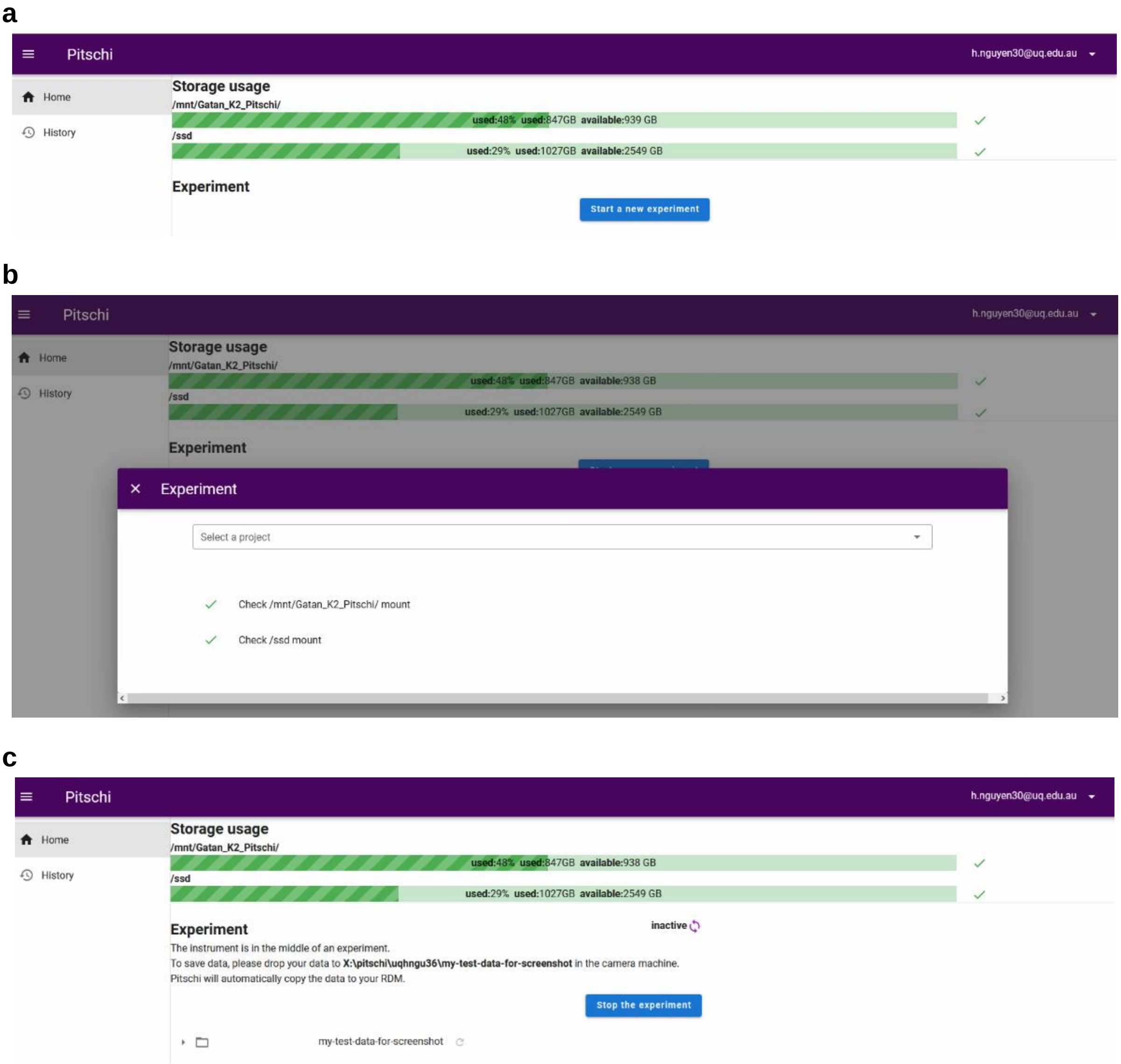

Figure 9 Snapshots of the Pitschi-datamover home page. **a**, upon user's login. **b**, when starting a new experiment. **c**, after the creation of a folder for a new experiment.

and a folder for the new experiment is created and displayed (Figure 10c). Note that this folder is in a shared drive between the camera computer and the Linux box. During acquisition, the user needs to save data to this newly created folder. Pitschi-datamover synchronises the data periodically to the UQRDM until the experiment is stopped. With this, users also receive two emails from Pitschi (Figure 9), one to confirm the data is imported successfully and one to confirm the data is ingested to Clowder successfully. All the steps for this client are detailed in Figure 11.

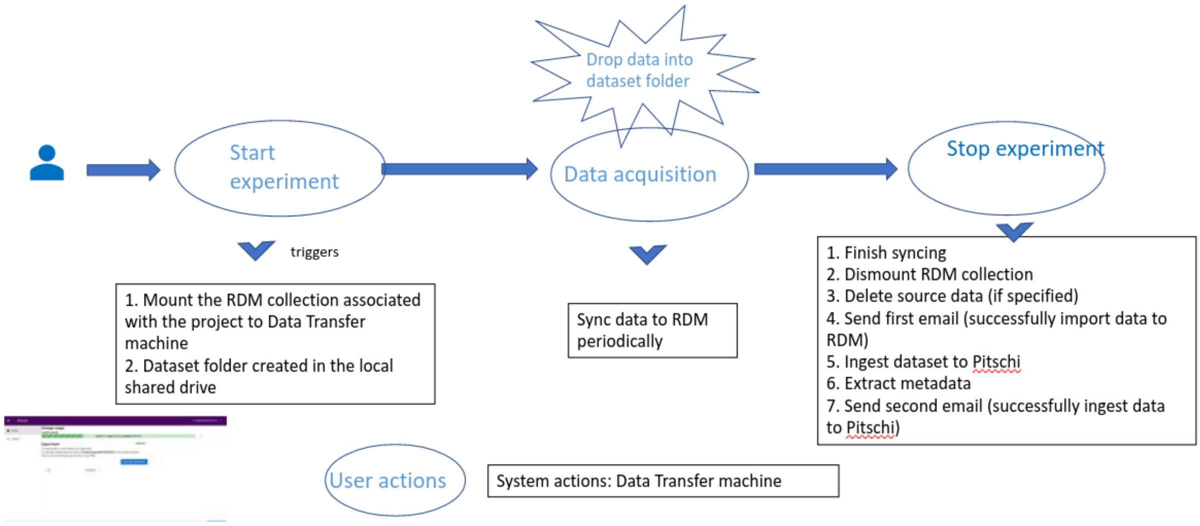

Figure 11 Pitschi data-flow with Pitschi-datamover web application.

4.2 Pitschi Deployment for server-side components

Figure 8 demonstrates how Pitschi server components are deployed on NecTAR. It uses Docker Swarm to manage services running as Docker containers inside the cluster. The main reason Docker Swarm was selected over Kubernetes is its simplicity to maintain. All the deployment steps are automated using ansible in this GitHub repository: https://github.com/UQ-RCC/ansible-swarm-clowder. All the steps to deploy the Docker Swarm cluster as well as deploying the serve- side components are documented in there.

As seen in Figure 12, the important services like Clowder and XAPI have both multiple instances running for redundancy. For security reasons, XAPI is only accessible inside the university's VPN, while Clowder is public-facing.

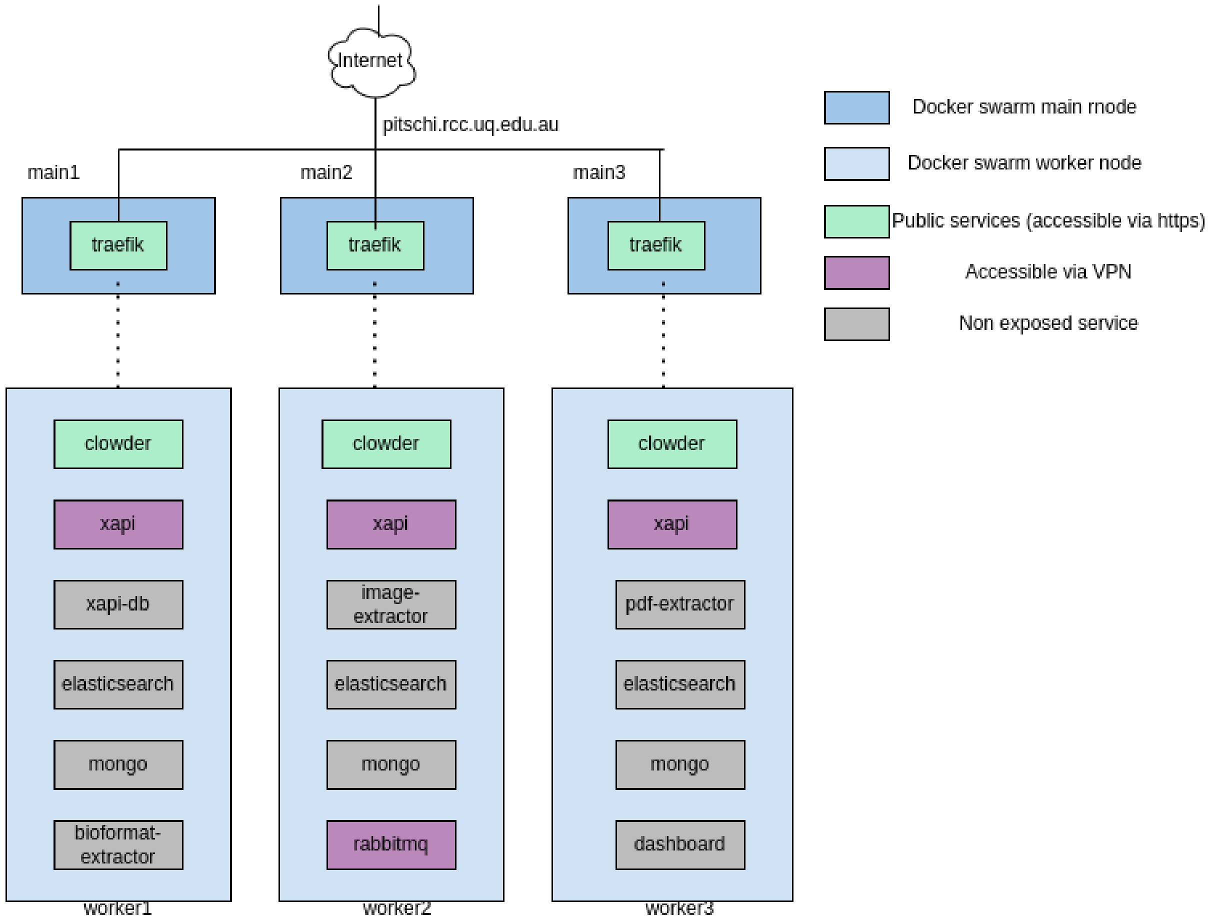

Figure 12 Pitschi server deployment.

5 How Pitschi can be adapted to other institutions

The current Pitschi implementation relies on two components specific to The University of Queensland. The first component is RIMS for booking management and the second component is MeDICI for data movement between the instrument computer to the storage collection. Pitschi can be adapted to other institutions by replacing these two components. Replacing RIMS with another booking system is quite straightforward as long as the new booking system provides an API. On the other hand, MeDICI can be replaced by point-to-point data-movement tools such as SFTP and Globus.

6 Conclusion

We have presented Pitschi, an end-to-end data management framework for the Centre of Microscopy and Microanalysis (CMM) at The University of Queensland. We use the Clowder data management framework to implement Pitschi. Importantly, we have integrated Clowder with the university's infrastructure such as the storage structure and booking systems. At the time of writing this report, Pitschi has been deployed on 18 instruments at the centre for 82 projects and 139 users.[10,11] Most notably, the two cryogenic electron microscopes—the biggest data generators at CMM—are supported by Pitschi. CMM users have deposited almost 75 TB of data with more than 620,000 files that belong to more than 270 datasets. Weekly training for users has been established by Hoang and Rubbiya Ali.

Currently, we are still rolling out Pitschi to more instruments across the Centre of Microscopy and Microanalysis and more clients are to be developed to serve instrument groups 3 and 4 (see section 4.4).

7 Contributorship

HN designed the work, developed Pitschi with the RCC team, and deployed it with RA at the Centre of Microscopy and Microanalysis, and wrote the final report.

8 References

1. Poger, D., van Schyndel, J., Nguyen, H., Silver, J., & Goscinski, W. J. (2021). Orchestration and management of data generated by big-data electron microscopy instruments: A Discovery report. *Zenodo*. https://doi.org/10.5281/zenodo.4744876
2. Allan, C., Burel, J.-M., Moore, J., Blackburn, C., Linkert, M., Loynton, S., MacDonald, D., Moore, W. J., Neves, C., Patterson, A., Porter, M., Tarkowska, A., Loranger, B., Avondo, J., Lagerstedt, I., Lianas, L., Leo, S., Hands, K., Hay, R. T., Patwardhan, A., Best, C., Kleywegt, G. J., Zanetti, G., & Swedlow, J. R. (2012). OMERO: flexible, model-driven data management for experimental biology. *Nature Methods*, *9*(3), 245–253. https://doi.org/10.1038/nmeth.1896
3. Marini, L., Gutierrez-Polo, I., Kooper, R., Satheesan, S. P., Burnette, M., Lee, J., Nicholson, T., Zhao, Y., & McHenry, K. (2018). Clowder: Open source data management for long tail data. In *Proceedings of the Practice and Experience on Advanced Research Computing* (pp. 1–8). https://doi.org/10.1145/3219104.3219159
4. Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J. G., Groth, P., Goble, C., Grethe, J. S., Heringa, J., 't Hoen, P. A. C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 160018. https://doi.org/10.1038/sdata.2016.18
5. Abramson, D., Carroll, J., Jin, C., & Mallon, M. (2017). A Metropolitan area infrastructure

[10] https://uq-rcc.github.io/pitschi-docs
[11] https://pitschi.rcc.uq.edu.au

for data intensive science. In *2017 IEEE 13th International Conference on e-Science (e-Science)* (pp. 238–247). https://doi.org/10.1109/eScience.2017.37

6. Jacobsen, A., de Miranda Azevedo, R., Juty, N., Batista, D., Coles, S., Cornet, R., Courtot, M., Crosas, M., Dumontier, M., Evelo, C. T., Goble, C., Guizzardi, G., Hansen, K. K., Hasnain, A., Hettne, K., Heringa, J., Hooft, R. W. W., Imming, M., Jeffery, K. G., Kaliyaperumal, R., Kersloot, M. G., Kirkpatrick, C. R., Kuhn, T., Labastida, I., Magagna, B., McQuilton, P., Meyers, N., Montesanti, A., van Reisen, M., Rocca-Serra, P., Pergl, R., Sansone, S.-A., da Silva Santos, L. O. B., Schneider, J., Strawn, G., Thompson, M., Waagmeester, A., Weigel, T., Wilkinson, M. D., Willighagen, E. L., Wittenburg, P., Roos, M., Mons, B., & Schultes, E. (2020). FAIR principles: interpretations and implementation considerations. *Data Intelligence*, *2*(1–2), 10–29. https://doi.org/10.1162/dint_r_00024

7. Rajasekar, A., Moore, R., & Vernon, F. (2007). iRODS: A Distributed Data Management Cyberinfrastructure for Observatories. In *American Geophysical Union, Fall Meeting 2007*.

8. Herrick, R., Horton, W., Olsen, T., McKay, M., Archie, K. A., & Marcus, D. S. (2016). XNAT Central: Open sourcing imaging research data. *NeuroImage*, *124*(part B), 1093–1096. https://doi.org/10.1016/j.neuroimage.2015.06.076

9. Wang, Z., Lin, G., Tan, H., Qinghong, C., & Liu, X. (2020). CKAN: Collaborative Knowledge-Aware Attentive Network for Recommender Systems. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 219–228). https://doi.org/10.1145/3397271.3401141