Data and Services Discovery projects - Institutional Role in a Data Commons

Title

Institutionally supported policy and data infrastructure to ensure management and reusability of sensitive imaging data through controlled access using XNAT

Approach

Activity 1: a security audit of our existing infrastructure deployment of XNAT on AWS, a platform recently certified by the Australian Cyber Security Center up to PROTECTED.

By default XNAT integrates with local LDAP and our deployment has only been accessible from within the university's intranet. In an effort to make XNAT more FAIR, the University of Sydney and Monash University engaged QCIF for the development of an AAF plugin using OpenID Connect last year.

Opening up our Imaging Data Service to the wider internet and adding AAF authentication for external users has the potential to drastically increase the attack surface if not done properly. We arranged for an external security firm to perform penetration testing and hosting security assessments on our Imaging Data Service. The first round was performed at the end of July 2019, allowing three months for remediation before a second round was conducted in October 2019 to assess said remediations.

The first part of testing was of the XNAT software itself, being directly applicable to all deployments. Results were 1 Informational, 12 Low, 12 Medium, 2 High, and 0 critical. A more immediate outcome of this work is a security patch in the open source codebase, resulting in the releases of 1.7.5.4-6. This directly benefits all users of XNAT, including the seven other Australian institutions that currently have deployments.

The second part of testing was a configuration review of our deployment, including both XNAT and Clinical Trials Processor (CTP). There are a couple different deployment options that have been used across Australia, but we explored using commercial cloud AWS, located in Sydney. In January 2019, the Australian Cyber Security Centre (ACSC) awarded AWS PROTECTED status, allowing it to handle certain types of sensitive data. While providing a good foundation, additional work is needed to ensure that a deployment on such infrastructure is also using best practice. The results were 2 Informational, 22 Low, 16 Medium, 0 High, 0 Critical.

The details of these results will be circulated privately to interested parties due to their nature and are not attached to this report.

Activity 2: an audit of current workflows and sensitive data risks: we documented workflows of researchers and facilities for handling data on XNAT, capturing risks and capability gaps.

Issue 1 Tools: The results of the audit were that researchers use the tools they have available in an ad hoc manner. In this case, they have our general research storage (RDS) and imaging storage (XNAT), as well as commercial cloud solutions. Researchers often implement their chosen approach in an ad-hoc and unsupported manner.

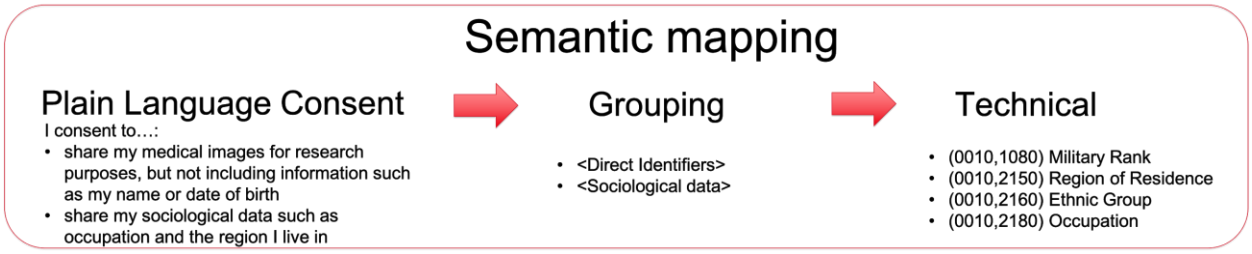
Issue 2 Culture: There is a disconnect between the Ethics and technical best practice, using relying on older approaches. Often a “Cloud is bad” mindset was seen. This mindset is understandable if set up improperly, but a challenge as our institution has a Cloud First strategy.

Issue 3 Reporting: After Ethics are awarded there is not significant oversight or reporting on compliance of the approach, usually due to a lack of tools to do so in a digestible manner.

Activity 3: Discussion on three use cases: Building on Activities 1 & 2, different levels of identification were required by three use cases. In the clinical context the DICOM standard is used for both imaging data and metadata, and allows us to build a standard framework. We sought to reduce risk providing baseline classification and agreement on minimum identification profiles.

De-identification of data is a difficult balance, particularly when mixing data from both university and clinical environments. On the one hand there is the desire for provenance and auxiliary data, which helps to expand research questions, ensure QA/QC, and help with reporting. One the other hand this same data can be potentially identifying. To add to the complexity is the requirement for Duty of Care, where incidental medical findings during the course of research need to be relayed to healthcare professionals for the patient’s sake.

We went through roughly 4,000 DICOM metadata fields which are the highest risk sources of patient identifying metadata, outside of the image itself. These were grouped first by commonality, e.g. (0008,0020-35) are all acquisition timestamps. The common groupings, where then grouped into conception groups such as <Direct Identifiers> or <Sociological data> to be mapped to plain language consent.



While this approach is what we want to use in a target state, we realized an additional non-orthogonal layer around international regulations was needed, as well as solutions around existing projects. To this end we created a AU/NSW minimum profile, on top of which additional groups could be applied and compared it to the two main standard profiles. We aim to expand this to GDPR and an updated HIPAA (the 2012 was included in CTP). The Region options act as a Blacklist, removing all subfields in their selections. The Patient Consents act as a Whitelist, allowing fields to pass through to the research repository.

Region (Required)	Patient Consent
- AU/State	- Sociological Data

- HIPPA	- Physical Description
- GDRP	

Activity 4: Creation of new policy: As part of tighter integration between the University and affiliated clinical sites, there is a large policy drive by the University to effectively and transparently manage bidirectional data flow. We sought to reduce systematic risk through creation of standardized workflows and patient consent templates. These are incorporated into the Operating Model of XNAT, increasing the integrity of research without being burdensome to researchers.

See description in activity 5

Activity 5: Enhancement of existing platforms: Based on Activities 3 & 4, we implemented tools to automate the desired workflows and force minimum rulesets for DICOM metadata along with updated reference architecture. These allow greater Interoperability and Reuse between deployments.

	Who drives the process?	***Configuration***	***Operation***
Previous	Researcher Led	Technical	Manual
Current	IT Led	Technical	Automatic
Target	Ethics Led	Grouping & Consent	Automatic

We implemented automatic de-identification at each clinical site using CTP scripts with profiles set up for each Project. We implemented the minimum AU/NSW list as well as additional fields needed for the project. In the future, this will develop standard scripts for each filter list which can be called from project filters for better maintainability.

To address Issue 3, we've enhanced the reporting in XNAT, particularly around identifiable metadata. Our future aim is to integrate with the Research Dashboard with our Ethics management system to allow feeding of these reports to easy digestion by Ethics committees along with control over automated filters based on the ethics forms.

Activities 3-5 ensured the data will be more Accessible and Reusable, as well as ensure efficiency and integrity of research through (semi)automated workflows and tools.

Monash Biomedical Imaging was a collaborator in this project. See the collaboration section for details of expanded collaboration during this project.

Outputs:

- The CTP and DICOMEdit scripts will be uploaded once they have been generalized to allow more automated deployment to other sites in the ARDC Platforms funded "Australian Imaging Service" (AIS) project (P19-B7). https://github.com/Australian-Imaging-Service
- The redacted security reports are available upon request, and will be disbursed through the AIS Project.
- As a result of the security audit, changes were made to the source code that are freely available to any institution seeking to deploy XNAT.
- Scripts are available to aid a standard deployment of XNAT that can be federated with the effort now taken up by the AIS Project.

FAIR

- See the corresponding FAIR assessment

Collaboration

There was considerable interest in the project from a number of other institutions. These institutions formed a consortium for a successful ARDC Platforms bid (AIS Project).

Sustainability

- The project continues in the form of the AIS Project, with an expanded consortium, and the possibility of others joining the project.
- The benefits of the security audit have been incorporated into the core open source project for all to benefit from.

Learnings

The big learnings from this project were the need for standard de-identification tools along with cultural change and education around these tools and approaches. Standardization of process and tools allows automation, which reduces human error or omission and de-risks the handling of sensitive data. Proving worthy of the trust that patients bestow on research institutions is vital to continue academic pursuits and improve health outcomes of our nation. Wider deployment of these tools and educating users at all steps along the process will help us as a community live up to patient's expectations.

Impact

What impact has this project had on research efficiency and integrity?

- The project has allowed us to automate movement of data from clinical sites to our central XNAT solution, moving over 30,000 patients' data in an automated manner directly from the instrument. This has improved research efficiency by allowing researchers quicker access to their data, as well as not needed to spend hours on manual tasks.
- This benefits the integrity of the researcher and institution, working to be better custodians of patients' personal data.
- As part of the related 2019 ARDC Project: Australian Imaging Service, 10 other Australian universities will deploy the same approach for automating their clinical de-identification, expanding on the work here.

Report prepared by: Ryan Sullivan, University of Sydney
Date: 05/06/2020