Title

A brain imaging database of rare and endangered Australian mammals

A conceptual demonstration of a national imaging data collection of Australia's unique fauna.

Approach

This project established a database and metadata framework to archive imaging data from preserved brain specimens in various bio-sample and museum collections. Aim of this development was to take the first step towards a comprehensive data description and national database for Australian native animal tissues.

The project was conducted in two milestones. Firstly, we have developed an XML based framework to describe imaging data from preserved bio-samples in collections. In this project we mainly focus on descriptors for MRI and CT data from brain and skull studies. However, long-term goal of this development is to eventually establish a generally accepted standard that meets the needs of national museums and collections for a metadata framework to describe general images taken from any tissues in their bio-sample archives. As such the definition provides quite general descriptors for preserved tissue datasets and can be extended to incorporate future needs.

Secondly, we have implemented a database for image volumes from preserved brain and skull specimens from (mostly endangered) endemic animals. The database contains MRI and CT volumes that were acquired in research collaborations with various collections over the past 5 years. The foremost aim was to transform our pre-existing digital data from a project based (only locally accessible) resource into a publicly available data collection for research and education. To improve findability and accessibility we have incorporated metadata from various sources (local repositories, museum database and imaging instruments) to provide searchable information about the specimen background and available imaging data. The repository will be continuously extended with newly acquired image volumes from collaborating Australian (NIF) imaging facilities to provide public educational institutions and researchers across the country with a searchable resource for teaching and comparative studies.

The details of our approach and achieved work-packages from both milestones are listed below:

1. **Establishment of an extensible metadata framework for volumetric imaging data from preserved animal specimens.**

- Stakeholder consultations and literature research.
  Several parties were consulted concerning needs and requirements for a national specimen image database in various individual meetings with representatives from the Museums (Australian Museum, Queensland Museum), related databases (Atlas of Living Australia), Imaging stakeholders (nodes of the National Imaging Facility, Australia), inter-state universities (Universities of Queensland, Western Australia and Melbourne) and Industry partners (Pedestal3d Pty.). Consultations revealed a clear need for platforms to consistently archive digital imaging data from preserved collection specimens. Data structure requirements and potential platform solutions were discussed within the "Animal, Plant and Material Imaging (APM) theme group" of the National Imaging Facility, Australia.
  A literature research showed that flexible and extensible metadata structures for the specific need to describe metadata of specimen collections are not well developed. Although some international initiatives have implemented platforms to deal with volumetric images from collections (with iDigBio, oVert and morphosource being the biggest projects in that domain) these generally rely on standards that are adapted from other domains (e.g. Darwin core made for biodiversity datasets) and therefore not well suited to describe important aspects of imaging data from collection sources.
- Review of other related standards
  Information and tags contained in other published standards have been reviewed. Main standards considered included DICOM (clinical image and transfer protocol standard), XCEDE (XML standard for brain imaging experiments) and the Darwin core (Biodiversity information). Datatypes for the specific needs of specimen data storage were extracted and adapted from these existing standards.
- Development of an extensible, XML based metadata description

Based on these learnings we have developed a metadata specification to describe image data from specimen digitization. To ensure flexibility, extensibility and compatibility with machine readers we implement the framework as a hierarchy and collection of XML datatypes. The definition represents metadata and data resources as a collection of one or more XML documents which are validated against a datatype within an XML schema. Datatypes and schema were implemented at UNSW by Dr Andre Bongers and Simone Zanoni (MSc) (Biological Resources Imaging Laboratory. UNSW) using the eclipse IDE. Datastructure development is still ongoing and will be finalized in Dec 2019. XML datatypes and exemplary XML instances will be made plublicly available once completed.

- Compilation of a documentation manual for the metadata framework
  Documentation to describe datatypes and metadata of the developed framework is currently being compiled. The work on the manual is ongoing and the document will be uploaded to public repositories in conjunction with the XML structures once finalized.

2. Establishment of a brain and skull image database of Australian endemic species

- Implementation of an XNAT based brain database platform
  We have developed an imaging database based on an instance of the open-source imaging informatics software platform XNAT. This choice was guided by the fact that XNAT is a highly extensible platform that was originally developed for brain imaging-based research and is now used for other large scale brain databases such as the human connectome project. The XNAT instance, including database, web-service and https-certificates was implemented from a docker distribution by Simone Zanoni (MSc), Image Analyst and network specialist at the Biological Resources Imaging Laboratory, UNSW in collaboration with Dr Tom Close, NIF Informatics Fellow, NIF node Monash University. The database currently runs on a virtual machine which hosted by IT services in UNSW's Mark Wainright Analytical Centre (MWAC) and is open for public access through a secure web-service.
- Extension of the XNAT platform for the project
  By default, XNAT provides datatypes that are very much centered on data handling for human neuroscience studies and related assessments of human behavior. To adapt the platform to the specific needs of imaging data from (museum) collection specimens, we have extended our XNAT instance to incorporate modified and new datatypes. These are partially based on the types defined in the general matadata structure from milestone 1 and incorporate additional background information about the specimen (including information about collection, location, preparation, etc.) as well as more detailed metadata about image acquisition and imaging preparation procedures.
- Implementation of workflows for data export from imaging instrumentation
  Existing imaging and data archiving protocols were refined and standardized. The majority of pre-existing data entering the project database was acquired on a Bruker Biospec 9.4T high field pre-clinical MRI in UNSW's Biological Imaging Laboratory. We have optimized and standardized the data acquisition and transfer protocols on the scanner, including conversion from proprietary to dicom formats. Python based scripts were developed to include consistent scanner information in exprted dicom files.
- Data collation and preparation for the database
  Pre-existing specimen image data to enter the database were collated from local instrument repositories and various local unstructured data archives (previously held in flat folder structures). MR data was reconstructed, exported and converted from proprietary data on the local imaging instrument into standard imaging formats (16bit dicom and nifti) using manufacturer and custom software on the pre-clinical MRI scanner at UNSW.
- Metadata collation and preparation for the database
  Metadata for each specimen was collated from multiple sources (including museum databases, lab books and in consultation with data curators). Stakeholders from multiple institutions were involved in this process, including Dr Sandy Ingleby and Dr Camron Slatyer (Australian Museum), Prof Ken Ashwell (School of Medicine, UNSW) and Dr Lydia Tong (Taronga Zoo Conservation Society). XML structures for metadata about scanner hardware and image acquisition and reconstruction parameters were derived from protocols and hardware setup at the local scanner (Dr A Bongers, UNSW)
- Acquisition of additional imaging data from newly established data sources:
  An additional collaboration was established with the Taronga Zoo Conservation Society. The society now serves as a valuable source for additional tissue samples for the brain database. From the collaboration

additional brain datasets were acquired on the UNSW MRI and curated for the database (currently fro brains of Koala, Pygmy sperm Whale and Quokka). (Dr A Bongers, UNSW, MWAC).

Outputs produced:

1. Metadata specification to describe 3D volumetric image datasets from specimen digitization processes. The metadata framework is largely finalized. An associated documentation manual is currently being compiled. The output will be published in Dec 2019, when the specification and documentation is completed.
2. The implemented brain database is publicly available in our XNAT repository through secure web-service of via the URL *aussiebrain.unsw.edu.au*. The DB currently contains ~50 curated datasets (from 11 species) and will be continuously extend with pre-existing and newly acquired image datasets in the future. Users and interested researchers can request access by registering for a database account on the website.

FAIR

The work in this project has strongly improved FAIRness of our existing specimen imaging data as well as optimized the protocols and metadata sets to process and archive imaging from future image acquisitions. The brain data entering this project were previously stored in an unstructured way on imaging instruments and in local data repositories in largely flat directory structures. The datasets were only findable by and accessible to researchers who were directly involved in the data acquisition the related research projects. Metadata was generally not included with the data but distributed in various repositories, including museum databases, imaging instruments, local lab notebooks, etc.

A large part of the data has now been moved to the publicly accessible XNAT based brain image repository developed in this project. Associated metadata is stored within the same database and linked to the specimen ID increasing findability of the datasets and enabling researchers across the country to access (and download) the data.

The extensible metadata framework developed in this project forms the basis of further improvements of 'FAIRness' of both, the specific brain data from this project and more generally any imaging data from preserved animal tissues. Our aim is to maintain and further extend this framework to eventually form a nationally accepted standard to archive imaging data of specimens from collections.

Collaboration and coverage

The project has established multiple new collaborations and fostered the extension existing collaborations-widening the coverage of the data collection and metadata definitions established in this project.

- The work in this project contributes to several initiatives on the national level. The Animal, Plant and Material Imaging (APM) of the National Imaging Facility (NIF), Australia has recently launched a national initiative to explore data curation concepts for digitized specimens from national collections. Major stakeholders of the APM theme group were consulted to give their input, and metadata structures developed in this may form a template to define national guidelines for more general descriptors within this wider scope. A national workshop is planned to define national metadata and data curation standards for museum specimens.
- New connections have been established to interstate institutions as potential contributors and beneficiaries of the data collection in this project. Discussions with the Australian Museum Sydney and the Queensland Museum are ongoing to establish common museum data curation platform. In project consultations with the Australian Academy of Technology and Engineering (applied.org) the Academy has expressed interest to participate in the project as a data consumer, namely to include data from the digital image collection in educational curriculums across the country.
- The project has taken pre-existing collaborations with the Taronga Zoo conservation society and the Australian Museum to a new level. The Taronga Zoo conservation society now contributes as a continuous source for animal brain tissues from their associated animal hospital and veterinary pathological services. The expertise of the facility will also provide additional input to the digital collection, including microscopic images from histological sections and animal tissues from other body regions. The collaboration with the Australian

Museum has been extended now entering discussions about the possibility to link the textural museum specimen database (Axiell EMU DB) to our project database.

- A new industry collaboration has been established with a start-up company Pedestal3D Pty. to add web-based visualization tools and a discovery layer to the data collection. This will increase findability and accessibility of the data retained in the repository.

Sustainability

Our long-term goal is to eventually establish a national data standard and repository to facilitate consistent storage of imaging data from tissues of animal samples of all major national collections. On this way we have already established a strong network with major relevant stakeholders who have all identified an urgent need for these capabilities. The width and common interest of the participating interest groups, including major national entities such as the National Museums and nodes of the National Imaging Facility ensures that the metadata framework will be refined to potentially form the basis for a national standard.

The strong and continuing collaborations with multiple collections and conservation societies will ensure a continuous stream of animal tissue specimens. We therefore anticipate that the brain database will grow quickly. However -as laid out in the 'Learnings' section below- the database platform might change in the future to improve data structures and allow for more flexible searches on various aspects of the datasets. We envisgae that the database will be linked to other resources that hold related information about the digital specimen in our database. Discussions to link the respective databases have already been initiated with the Australian Museum (to link the museum's EMU database) and the Atlas of Living Australia (to link to its biodiversity platform). Our project is also involved in related discussions with ozMammals initiated by the National Imaging facility to explore the possibilities to link imaging data to the genetic datasets available on Bioplatforms Australia.

Learnings

Learnings from the metadata framework implementation

From the project feedback from multiple consultation parties we have learned that there is a tremendous need in the stakeholder community to establish consistent curation and storage concepts for imaging data from specimen digitization processes in collections. This includes comprehensive metadata descriptions as well as database and visualization capabilities. Our background research shows that no flexible platforms and metadata frameworks exist to meet this need and enable 'big data' exploration from such specimen databases. Existing platforms and metadata frameworks are generally not focused or flexible enough to allow for searching of multiple aspects of digital representations of physical specimens in a standardized way. To date many museum repositories, hold metadata information in free text or very one-dimensional documents or tables and cannot properly store volumetric imaging data.

A potential obstacle on the way to establish a commonly accepted metadata concept and platform is the considerable diversity in the community. Our consultations revealed that the priorities for such a platform are different among different stakeholder groups. E.g. curators (from museums, conservation societies or herbariums), researchers (from different fields, neuroscientists, paleontolgists, etc.), imaging facilities, educational institutions etc. all seem to have different views about what they expect from a future curations system for digital specimen imaging data. This diversity can only be managed in a more extensive consultation with the user community to thoroughly define the scope and capabilities needed. To lay the foundations for a commonly accepted metadata framework (and potentially platform) we propose to bring all national stakeholders together with the aim to come up with a detailed requirement specification. Discussions are currently ongoing to organize such a workshop and potentially establish a related working group.

Learnings from the technology implementation

Our specimen image database is based on an instance of the open-source imaging informatics software platform XNAT. This choice was guided by the fact that the XNAT was originally developed for brain imaging-based research and is now used for large scale image repositories such as the human connectome project.

However, in the course of the project some design features of XNAT became obvious that limit its usability to store metadata of museum specimen images. Firstly, as XNAT is predominantly designed for neuroscience experiments

on human subject cohorts it exhibits a data model that is strongly (and rigidly) bound to projects, subjects and experiments (providing only two distinct hierarchy levels, 'projects' and 'subjects'). This structure is not well suited for the specific needs of museum specimen images where e.g. the notion of an 'experiment' (which is the only major datatype extension point in XNAT) does not exist. XNAT's predefined datatypes are quite focused on clinical standard modalities (MRI, CT and PET). Introducing descriptors for other instruments that are frequently used to image museum and tissue specimens but lie outside this realm (e.g. pre-clinical instruments, photographic or histologic images) is an involved process. These and similar limitations make it difficult to store specimen metadata in appropriate ways images to enable flexible data searches under various aspects. Therefore - although the current platform is a big step forward for us to improve findability, accessibility and reusability of our data - we are considering porting the database to a more suitable platform for the specific needs once the data amount increases.

Impact

This project works towards a common data standard to store digital image data from collection specimens. Previous and ongoing consultations show that there is a strong need from the community to establish such concepts and database implementations. Despite strongly increasing efforts to acquire image data from specimen collections from various stakeholders (Museums, Conservation Societies and government institutions) there is currently no platform available in Australia to store these datasets in a consistent way. The metadata framework developed in this project directly impacts on the development of such a platform. Specifically, our consultations within the framework of this project already helped to trigger the establishment of a national working group to develop common data curation concepts.

Commonly agreed metadata standards will have a tremendous impact on the value of digitized specimen data to society and research. They will help to link resources between data archives from multiple institutions and increase usefulness of the datasets for research and education. Eventually a well thought out and widely adopted metadata framework will allow for extensive data-mining far beyond of what is possible today.

E.g. comparative evolutionary or environmental studies can directly search, combine and compare multiple datasets of the same (or different) species from different temporal or environmental contexts. Curators may learn about preparation methods and deterioration of samples by comparing digitized specimen data from different sources and connecting them to the preparation history of the samples. Linking imaging datasets to other data resources (e.g. genetic or environmental databases) using standard metadata formats will allow for new research e.g. connecting imaging phenotype to genotype of animal tissues. Although it is just a small excerpt of potential data-mining possibilities this short list may indicate the tremendous impact of common imaging data definitions on the outcomes of research studies as pursued in this project. This project has also established direct collaborations which directly impact the educational sector. The Australian Academy of Technology and Engineering (applied.org) have expressed interest to introduce specimen imaging data into national school curriculums. The brain database developed in this project improves accessibility of the pre-existing brain data with immediate impact on evolutionary neuroscience research by allowing for international collaborations by merging multiple datasets.

Report prepared by:

Dr Andre Bongers,
Fellow of the National Imaging Facility, Australia
Biological Resources Imaging Laboratory, Mark Wainwright Analytical Centre, The University of New South Wales, Sydney
Date: 29/09/2019