Thickshake: A Smarter Library Catalogue

Author: Mark Shelton | **Supervisors:** Joshua Hollick (Curtin HIVE), Dr Andrew Woods (Curtin HIVE), Debra Jones (SLWA) | **Program:** Pawsey Summer Internship (2017-18)

Introduction

The State Library of Western Australia (SLWA) holds more than one million items in its pictorial collection. The public expects to be able to easily search and browse these items. To make this possible, SLWA researches and describes each item's contents and context. This task is expensive and time-consuming. Recent advances in technology have opened the possibility of automating this task. We have developed **Thickshake**, a system that automatically improves library catalogue metadata (see **Figure 1**).

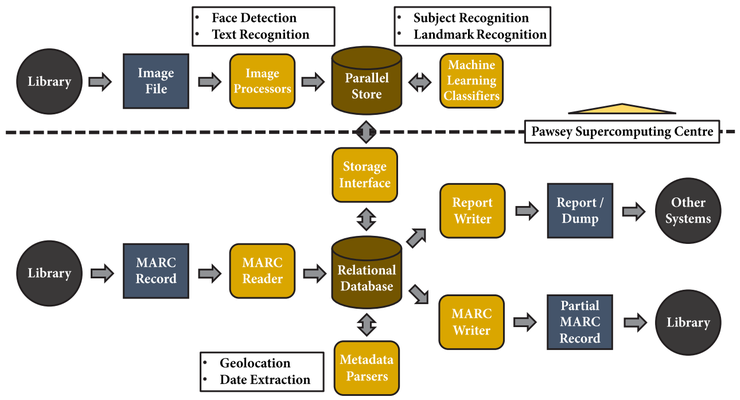

Figure 1. Thickshake system architecture.

Library Interface

MARC is the international digital format standard for library materials. MARC records contain items' titles, subjects, topics, and other cataloging details. However, MARC is complex and difficult to use. To address this shortcoming, we have developed a sytem that maps MARC records (pymarc) onto a relational database (SQLAlchemy, PostgreSQL). The interface works on a user-defined map. The interface creates partial MARC records that can be loaded back into the catalogue. It can also produce reports in formats (CSV, JSON, XML) for use in other systems.

Metadata Parsing

As MARC was designed for use by humans, MARC fields are relatively unstructured. We have developed a metadata parsing framework to extract useful, structured information from semi-structured MARC fields and load the new data back into the database. We have developed parsers to extract information like locations and dates. We use the Mappify.io API to geocode structured locations into coordinates, which we use in our OldPerth project to place SLWA photos on an interactive map (see **Figure 2**).

Figure 2. OldPerth map screenshot (geolocated items).

Image Processing

Images are a rich source of metadata, but processing images is computationally intensive. We have developed an image processing framework that runs on Pawsey's Athena GPU cluster. Image processing results are stored in a parallel-access HDF5 store (PyTables, pandas) which syncs with the relational database. We have developed two image processing functions: face detection (see **Figure 3**) and text recognition. Face detection uses Dlib, text recognition uses Tesseract and Hyperopt (optimised pre-processing), and both use OpenCV for image manipulation.

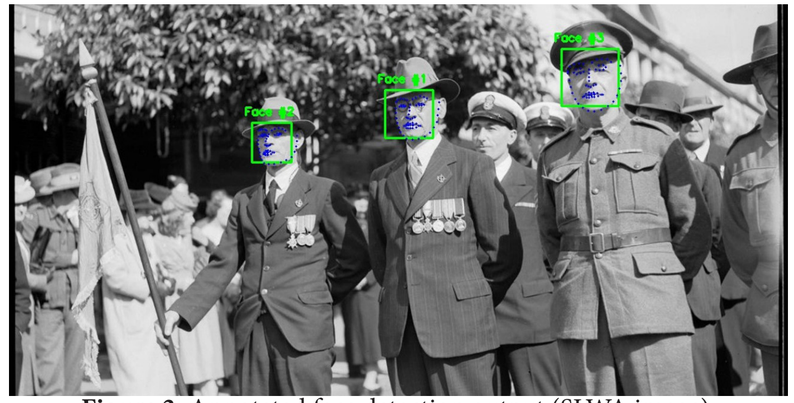

Figure 3. Annotated face detection output (SLWA image).

Machine Learning

Metadata parsing and image processing are performed on a per-item basis, but we can gain more insights by identifying trends across the catalogue. We started developing a subject recognition classifier that synthesises face detection results with existing metadata from the catalogue. The classifier runs alongside our image processing framework on Pawsey's Athena GPU Cluster, using GPU-enabled Tensorflow. We plan to continue developing this and other classifiers (e.g. landmark detection).

Conclusions

This project makes three primary contributions:

- An extensible and flexible platform for manipulating and augmenting library catalogue metadata.
- A suite of processing functions incl. geolocation, face detection, text recognition, and others.
- A system that leverages cutting-edge image processing and machine learning libraries on Pawsey Supercomputing Centre's Athena GPU cluster.

For more information, visit our project at:
github.com/markshelton/thickshake