Data and Services Discovery projects - Transformative Data Collections

Title

Enabling access to the David Scott Mitchell digital collection for digital humanities research.

Approach

1. Establishment of multi-disciplinary project team.

A lean project team was established to oversee the project and communicate deliverables to stakeholders. The team comprised project sponsors, collection subject-matter experts (librarians) and technical staff. External collaboration and input were provided from UTS eResearch team and Macquarie University Associate Professor, Steve Cassidy (Dept. of Computing).

2. Internal scoping workshop and project updates

With SMEs and stakeholders, the team ran an internal workshop to share project information and discuss options for scoping of the DSM collection.

Feedback from the session included:

1. Importance of providing context for the DSM collection at the welcome page of each platform. Some of the collection may appropriate indigenous IP and reference sensitive issues.
2. Updating of the Library's Open Data pages to include new platforms.
3. Confirmation of what is in scope for the DSM collection, namely bibliographic records, digitised images and OCR text. A small pilot sample should suffice for this project.
4. Request to run a further workshop at the conclusion of the project. Topics to discuss include considerations for any existing policies, collection management, digitisation, operations and services as a result of this project.

3. UTS hack session

A data hack session was held with the UTS eResearch team to extract DSM data (catalogue records, digital assets and OCR text), transform to Oxford Common File Layout [1](OCFL) and create JSON-LD representations using the Portland Common Data Model [2](PCDM).

---

[1] https://ocfl.io/

[2] https://pcdm.org/

Publishing DSM data to an OCFL specification improves interoperability between digital repositories, including UTS RO-Crate[3]. By using PCDM to describe the relationships between objects, the DSM collection is generalised for consumption by machines/scripts whilst still providing archival context.

The hack session delivered a successful prototype workflow for publishing "Book as data" using an OCFL repository with RO-Crate metadata. The specific collection item used in the session is viewable as an RO-Crate artefact[4]. This result provides the dataset in its most base data form, abstracted from systems, software, applications. Further work is required to automate the extract, transform and publishing processes.

4. Macquarie University information session with Steve Cassidy

An information session was held with Associate Professor Steve Cassidy to discuss opportunities in using Jupyter Notebooks[5] and Voyant tools[6] as publishing platforms for researchers.

Notebooks examples were demonstrated during the session, including the work undertaken by Tim Sherratt (University of Canberra). It was recommended that the project document Library APIs via Notebooks as a means to engage researchers with development capabilities.

Voyant tools is used within Alveo[7] as a web-based reading and analysis environment for digital texts. Using Voyant as a prototype to upload a small corpus of DSM OCR text would demonstrate an interesting 'tools' test-case, enabling insights into the collection using data visualisation and text analysis.

5. Recruitment of contract developer

A contract developer was engaged as part of the project, to work with the project teams in delivering the three platforms. The developer has expertise in in python and data science, which is a good fit for the project deliverables.

6. Development of extract, transform and load (ETL) scripts for platform integration

The following two outputs were produced before the final report.

Output One: Jupyter Notebooks

Jupyter Notebooks as a medium for creating API documentation, which researchers can reference to see how to interact with the ALMA (catalogue) and Rosetta (digital

3 https://researchobject.github.io/ro-crate/

4 https://data.research.uts.edu.au/examples/ro-crate/examples/src/samples/IE4783007/ro-crate-preview.html

5 https://jupyter.org/

6 https://voyant-tools.org/

7 http://alveo.edu.au/

repository/preservation system) APIs to retrieve bibliographical information for DSM, as well as OCR data.

A collection of four notebooks have been published onto the State Library's GitHub, with documented steps on accessing the APIs.

The first two notebooks explore the ALMA API, with a step by step guide to retrieve bibliographical information for specific books. These notebooks have been used to demonstrate how bibliographical information can be loaded into dataframes, to be used for research projects.

The remaining two notebooks document the usage of the Rosetta API for retrieving book ALTO files and gleaning the ALTO files to obtain the OCR text data.

Currently, further Jupyter Notebooks are being written, for conducting data science and natural language processing (NLP) projects. The NLP project entails processing the OCR text data retrieved from the ALTO files, and performing Named Entity Recognition, as a starting point for enhancing the usability and accessibility of the library's digital experience.

Links:

1- Jupyter Notebooks Collection: https://github.com/slnsw/ETL-projects/tree/master/Notebook%20Projects
2- Notebook for displaying bibliographical information along with the book cover: https://github.com/slnsw/ETL-projects/blob/master/Notebook%20Projects/2-Alma2.ipynb
3- Notebook for retrieving ALTO files and extracting the OCR text: https://github.com/slnsw/ETL-projects/blob/master/Notebook%20Projects/3-Alto_retrieve.ipynb
https://github.com/slnsw/ETL-projects/blob/master/Notebook%20Projects/4-Retrieving_Text.ipynb

Output 2 – Creating RO-Crates under the OCFL for UTS researchers

The code written to implement the ETL pipeline retrieves data from Rosetta, along with metadata from ALMA and creates RO-Crates for each book. Researchers can view book metadata and browse the ALTO files and scanned image contents in a user-friendly format. The code is scalable and has run successfully over a hundred books within the David Scott Mitchell collection.

Links:

1- GitHub repository for ETL pipeline code https://github.com/slnsw/ETL-projects/tree/master/RO-Crate

The following outputs are in progress, expected to be completed over the following weeks.

- Updates to State Library web pages - Open Data & Data sets
- Voyant tools - developing tools to integrate Voyant into pipeline for user-friendly computer-assisted text analysis. Additionally, exploring visualisation and dashboarding solutions within Jupyter Notebooks to enable an interactive coding experience for researchers.

FAIR

See FAIR assessment.

Collaboration and coverage

As a result of publishing to an OCFL repository, we've participated in an international effort to promote long-term object management best-practice approaches within digital repositories. Specifically, we've collaborated with the UTS eResearch team to publish select DSM books onto RO-Crate, which provides a platform for greater exposure with the national research community. Jupyter Notebooks will be published and marketed on the State Library's Open Data pages, which are accessible by local and national research communities.

Target consumers for these datasets are digital humanities researchers and historians. Indirectly, combining these datasets with other data will open the possibility for cross-disciplinary consumption from other subject domains. This is made possible by using standard data schemas such as PCDM.

Sustainability

At the conclusion of the project, an internal workshop will be run to discuss outcomes and next steps for the Library. Topics to discuss include considerations for any existing policies, collection management, digitisation, operations and services as a result of this project.

From a governance perspective, this initiative may continue within existing operational strategy groups. Operationally, technical teams are able to support and build on both platforms developed; OCFL and Jupyter Notebook publishing of curated Library collections.

Learnings

1. Timeframe and scope: Time constraints combined with lean operational resourcing to plan, consult and deliver outcomes has been a challenge. Scope and consultation were

deliberately limited to ensure sufficient time to deliver outcomes for the project and Library.
Recommendation: More time to undertake research to determine the information and data needs of researchers, specifically with cultural collections.

2. Organisational change management: The Library has limited capabilities and experience in the areas of Open Data within the eResearch community. Appreciating the benefits in undertaking and investing in similar initiatives is yet to be determined.
   Recommendation: Further workshops with research users demonstrating benefits and a change program to inform and collaborate with key staff to continue investing in eResearch initiatives.
3. Technology and capabilities: Developers with capabilities across data, information management and ETL processes are still emerging. State Library technical teams are upskilling in these areas.
   Recommendation: Explore training program for internal technical team to become skilled in data ETL processes, modeling and information management.

Impact

Due to the maturing nature of cultural datasets in eResearch, anticipated impact is currently unknown. Digital humanities and historians are target audiences for the State Library's cultural datasets, which in turn may contribute to social and cultural impacts. The publishing of these datasets on national and international platforms using data and archival standards will ensure persistence of information increasing discoverability and citation.

Report prepared by: Euwe Ermita, State Library of New South Wales
Date: 07/10/2019