# Making Connections with Linked Data:
## Identities & Identifiers at University of Victoria Libraries

Prepared by
**Robbyn Gordon Lanning**
University of Washington MLIS candidate

**Information School**
UNIVERSITY *of* WASHINGTON

Project sponsor
**Dean Seeman**
UVic, Head, Resource Description & Discovery

**University of Victoria** | Libraries

## THE PROJECT

Linked data empowers libraries and archives to connect rich stores of local data to global audiences through the use of common identifiers and standards. The University of Victoria (UVic) Libraries possesses over 2.9 million records, each of which uses text strings as resource identifiers. In order to participate in the semantic web, UVic Libraries must prepare its metadata, transitioning its records from textual representations of people, places, and subjects to internationally accessible Uniform Resource Identifiers (URIs).

To assist UVic in meeting this challenge I implemented a study reconciling randomized datasets from each of UVic's four record repositories to linked data identifiers. This research culminated in a report providing detailed reconciliation results, suggestions to overcome obstacles encountered, and best practice recommendations for future reconciliation work. By assisting UVic Libraries in preparation for transitioning its "strings" to "things", this project has helped enrich the exposure, interoperability, and accessibility of UVic's information holdings.

## RECONCILIATION PROCESS

- Extraction of randomized datasets containing personal names from each of UVic's 4 record repositories.
- Exported data files imported to MarcEdit (.mrc files) or OpenRefine (tab delimited files).
- Data runs through assortment of selected retrieval parameters, and the programs work to match the data with identifiers from linked data vocabularies.
- Data is then exported to Excel for organization and manipulation.

In total 8 datasets of 50 personal names were processed. While this data represents just a small percentage of the records in UVic Libraries' holdings, it is substantial enough to provide us with insights into the challenges reconciliation of UVic's various records may present.
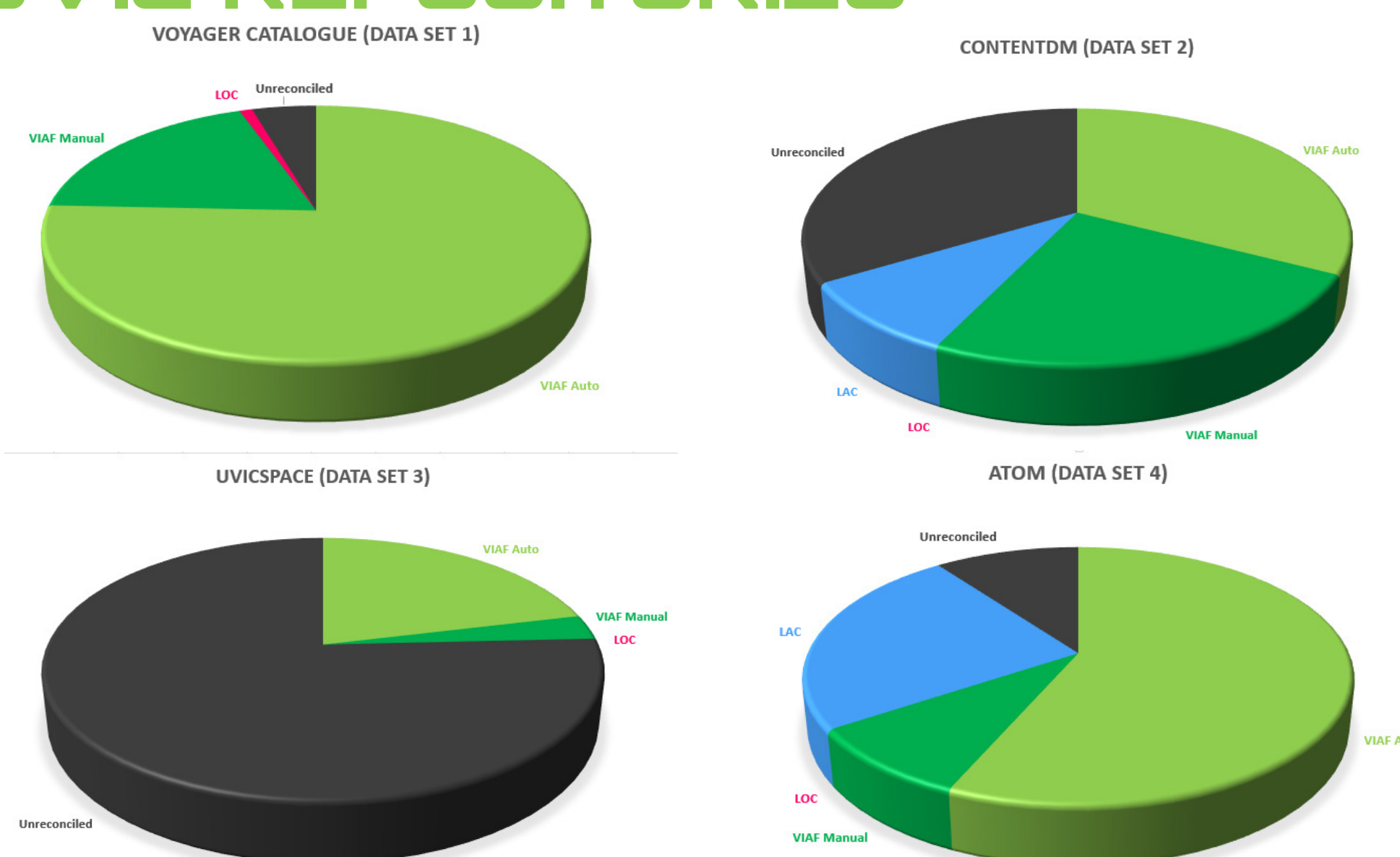
## CHALLENGES IDENTIFIED

- Quality control and reliability of vocabularies.
- Ensuring long-term access to vocabularies.
- Disambiguating popular or common names and duplicate records.
- Safeguarding confidence of matches.
- Identifying inaccurate or absent metadata.
- Increasing collections knowledge in employees to decrease risk of human error.
- Time intensive manual reconciliation and verification processes.
- Encouraging scholarly buy-in/investment.

## THE DATA
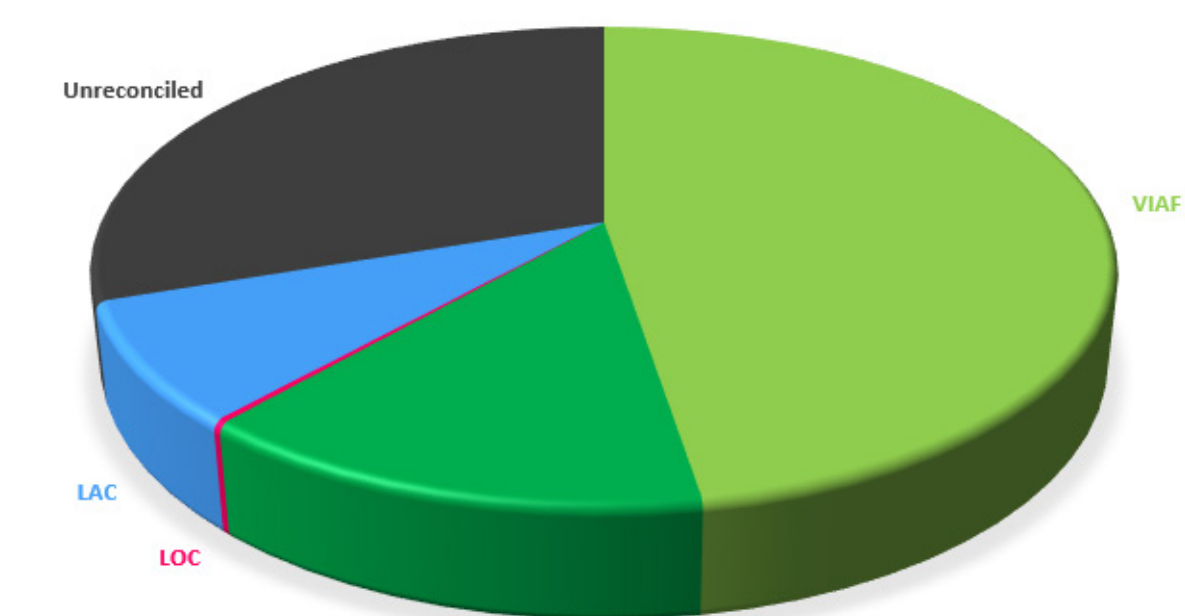
Overall nearly 70% of names were able to be matched with confidence to a VIAF identifier. Of the vocabularies accessed, VIAF and LAC proved to be the most useful to reconcile with UVic's holdings. These initial findings hold promise, but were also labour intensive to achieve, with each set taking anywhere between 3 and 16 hours to reconcile and manually verify.

## KEY QUESTIONS

- How difficult is reconciliation?
- To what extent can reconciliation be automated?
- How many identities don't have identifiers?
- How do the datasets differ (Voyager/UVicSpace/CONTENTdm/AtoM)?
- What challenges does the reconciliation process present?

## UVIC REPOSITORIES

VOYAGER CATALOGUE (DATA SET 1)

CONTENTDM (DATA SET 2)

UVICSPACE (DATA SET 3)

ATOM (DATA SET 4)

A total of 400 records were extracted from UVic's four repository systems: Voyager Catalogue, CONTENTdm digital library, UVicSpace institutional repository, and AtoM archival management system. The charts above highlight the success rates each dataset achieved in obtaining confident matches to VIAF and LOC identifiers, and LAC authority records through automatic and manual reconciliation processes using MarcEdit and OpenRefine.

## BEST PRACTICES

- Combination of VIAF and LAC identifies/authority records provides the highest occurrence of confident matches for UVic Libraries' records.
- Local content represented in AtoM and CONTENTdm repositories require additional access to or creation of identifiers local, provincial, and national significance.
- Integration of identifier creation into currently existing systems such as graduate thesis/dissertation submission and faculty publication processes.