BIBFRAME Instance mining:
toward authoritative publisher entities using association rules

Jim Hahn

Head of Metadata Research

University of Pennsylvania

USA

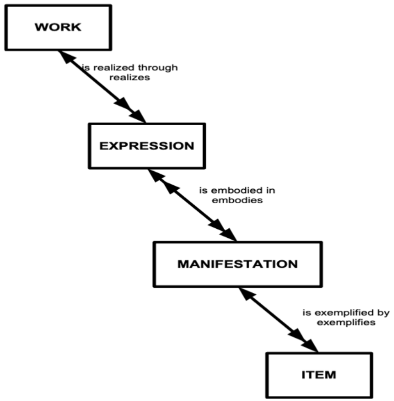
Penn Libraries

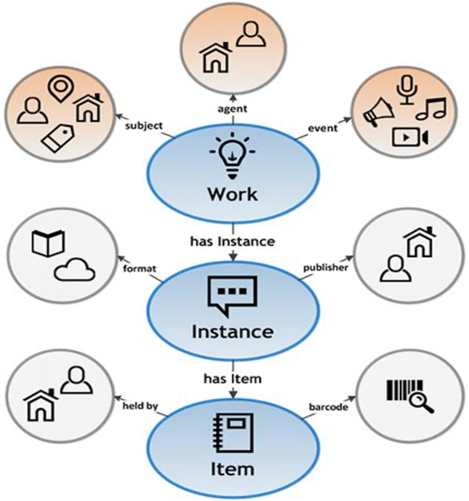
BIBFRAME and Share-VDE

- Much of the source metadata in Share-VDE currently comes from converted and transformed MARC. The converted records use BIBFRAME and are clustered around the BIBFRAME Work descriptions.

- The Share-VDE staff have a process for development of Instance clusters. This presentation introduces a complementary approach to discovering the BIBFRAME Instance clusters within the dataset of Share-VDE.

Figure 5.1 Relationships between Work, Expression, Manifestation, and Item





https://www.ifla.org/files/assets/cataloguing/frbr-lrm/ifla-lrm-august-2017_rev201712.pdf

BIBFRAME Ontology

- BIBFRAME: upper level bibliographic ontology implements LRM
  - **LRM conceptual model**
  - **BIBFRAME data model**

https://www.loc.gov/bibframe/docs/bibframe2-model.html

BIBFRAME

- What is an instance entity description and why is it important? Instance descriptions are linked to other Works, and Instance descriptions include format description data and publisher description data.

Unambiguously identifying an Instance relies in part on publisher data in a bibliographic description. The publisher data in these descriptions varies in terms of quality and has mostly not been converted from strings to controlled identifiers such as VIAF, among others.

Previous Work with Publisher Entities

- with the transition of a shared catalog to BIBFRAME linked data, there is a need for identifying the canonical Instance description for clustering in BIBFRAME.
- A fundamental component of Instance identification is by way of authoritative publisher entities.

Previous work in this area by OCLC research (Connaway & Dickey, 2011) proposed a data mining approach for developing an experimental Publisher Name Authority File (PNAF).

The Publisher Name Authority File (PNAF)

- The OCLC research was able to create profiles for "high-incidence" publishers after data mining and clustering of publishers.
- As a component of PNAF, Connaway & Dickney were able to provide detailed subject analysis of publishers.

PNAF Algorithm

Select records on Language code

Filter sets by ISBN Prefix

Contents of subfield b of MARC 260 extracted and deemed publisher name

Normalize the publisher names (clustered using Levenshtein Distance Value)

- Each automatic methodology worked to generate clusters of items based on an assigned publisher, first via ISBN prefix and then via further matches of 260 $b data, leading to a robust database of high-incidence publishers. Though the process could not be fully automated on a global scale, some 1,854 high-impact publishing entities were profiled by their publishing output, with detailed differences emerging between the profiles (Connaway & Dickey, 2011).

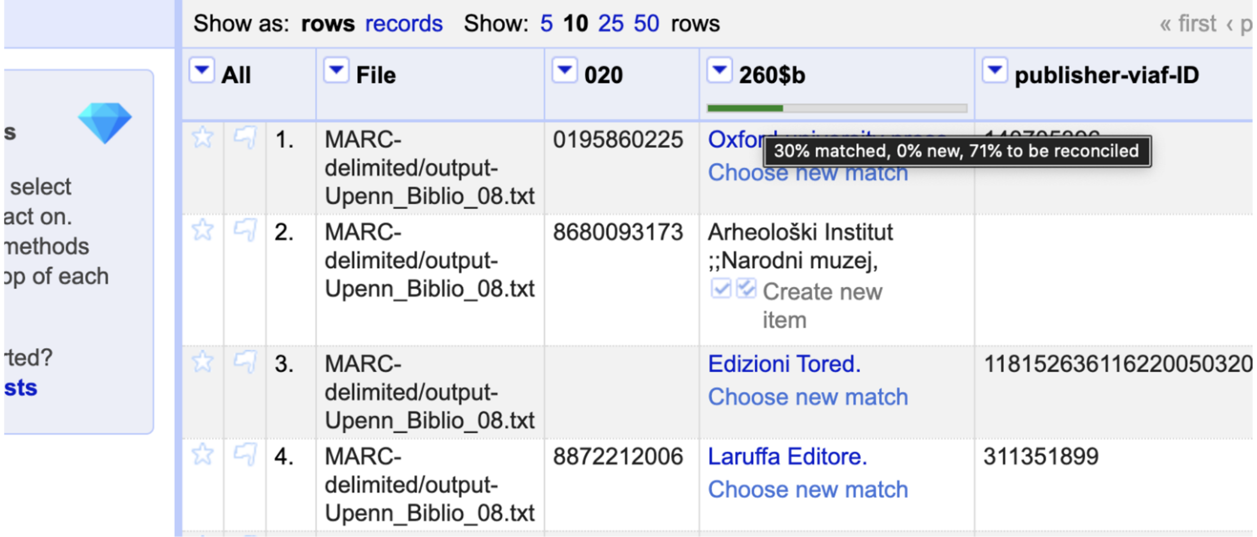
Replicating the PNAF prototype process for Penn Libraries metadata

Using the OCLC Paper (Connaway & Dickey, 2011) as inspiration, 5,109,592 MARC records that were sent to Share-VDE for enrichment and transformation were first clustered by strings in the 260$b field.

1. The algorithm “Fingerprinting - "Key Collision" method” from OpenRefine first clustered the publisher strings into a common string of near-matches (https://github.com/OpenRefine/OpenRefine/wiki/Clustering-In-Depth).

1. After this, reconciliation using VIAF and Conciliator was performed over the corpus of 260$b strings.

Replicating the PNAF prototype process for Penn Libraries



s

select
act on.
nethods
op of each

ted?
sts

Show as: **rows** records Show: 5 **10** 25 50 rows « first ‹ p

All	File	020	260$b	publisher-viaf-ID
☆ 1.	MARC-delimited/output-Upenn_Biblio_08.txt	0195860225	Oxford university press Choose new match	[illegible]
☆ 2.	MARC-delimited/output-Upenn_Biblio_08.txt	8680093173	Arheološki Institut ;;Narodni muzej, ☑ ☑ Create new item	
☆ 3.	MARC-delimited/output-Upenn_Biblio_08.txt		Edizioni Tored. Choose new match	118152636116220050320
☆ 4.	MARC-delimited/output-Upenn_Biblio_08.txt	8872212006	Laruffa Editore. Choose new match	311351899

30% matched, 0% new, 71% to be reconciled

Replicating the PNAF prototype process for Penn Libraries

- **Findings:** For semi-automated reconciliation, it was possible to reconcile VIAF entities to 30% of the 260$b publisher strings.

- **Completely automated reconciliation:**
  - 843,012 (16%) of the records were reconciled to a VIAF entity with .9-1.01 confidence.
  - 213,926 (4%) of the records were reconciled to a VIAF entity with .8-.89 confidence.
  - 174,454 (3.4%) of the records were reconciled to a VIAF entity with .7-.79 confidence.

- **ISBN Analysis of 5,109,592 MARC records :**
  - 2064994 (40.4%) have ISBN
  - 3044598 (59.6%) do not have ISBN
- **VIAF Reconciliation Relative to null ISBN:** 504836 (9.8%) Publisher entities with no ISBN were matched in a semi-automated process.

Can data mining and machine learning help with Publisher Entity-ification (Entification) when ISBN is not available?

If we can thoroughly reconcile publishers we can come very close to unambiguously finding the BIBFRAME instance to provide an authoritative canonical BIBFRAME instance in which to cluster instance entity descriptions….

Case Study: Publisher Association Rules

- **Why association rules**: a bibliographic description has a set of data points that may include a publisher name and also an agent such as a person or corporate body and it includes descriptions of subject areas

Association rules are produced using algorithms such as FP-Growth. FP Growth is an algorithm for discovering frequently co-occurrent items in a data set (Han et al, 2000; Li et al, 2008).

A rule can be defined as an implication, **X⟶Y** where X and Y are subsets of I(X,Y⊆I), and they have no element in common. X and Y are the antecedent and the consequent of the rule, respectively.

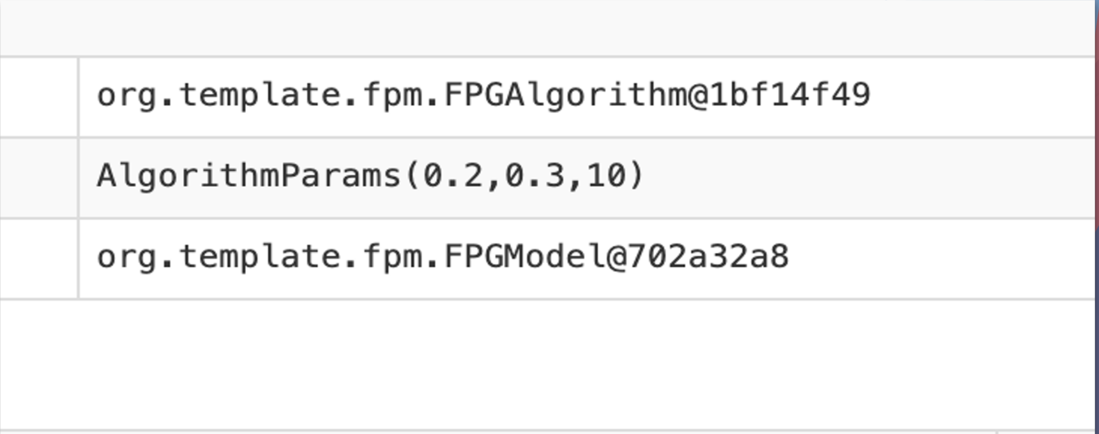
Eg: {Agent,Subject}=> {Publisher} ItemSet={Agent,Subject,Publisher}

https://www.kaggle.com/sajidcse/market-basket-analysis

Metrics common to Fp-growth

There are various metrics in place to help us understand the strength of association between antecedent and consequent. Here we have baselines set for **support** -- the probability of some item in the dataset; we also have a **confidence** baseline set, e.g. how sure are we that two things are happening consequently?

- Min Support 0.2
- Min Confidence 0.3



```
org.template.fpm.FPGAlgorithm@1bf14f49
AlgorithmParams(0.2,0.3,10)
org.template.fpm.FPGModel@702a32a8
```

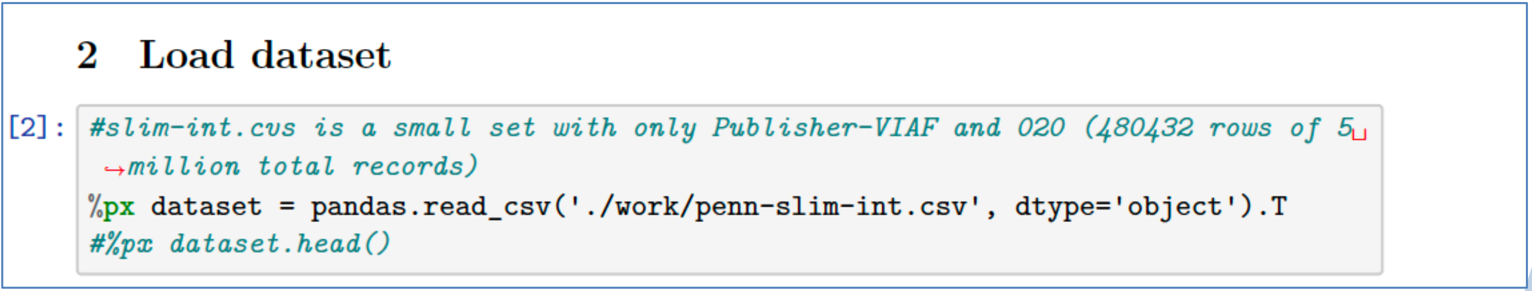
https://www.kaggle.com/sajidcse/market-basket-analysis

Initial Findings

Rule generation indicated support for the following sets:

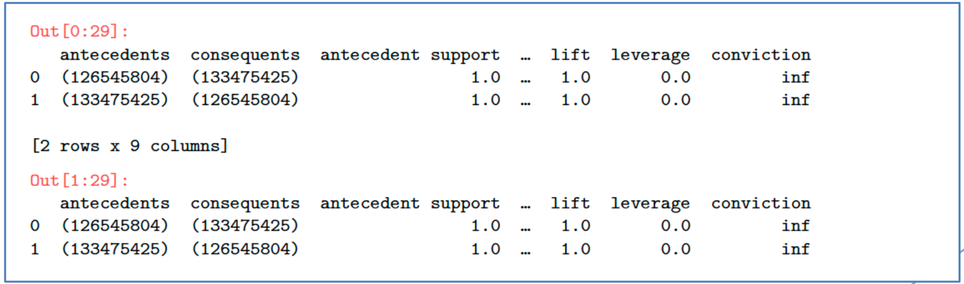
{publisher-VIAF-id, ISBN} – expected.

{main-entry-VIAF-id(100), publisher-VIAF-id} – a little more interesting.



2 Load dataset

```
[2]: #slim-int.cvs is a small set with only Publisher-VIAF and 020 (480432 rows of 5␣
      ↪million total records)
     %px dataset = pandas.read_csv('./work/penn-slim-int.csv', dtype='object').T
     #%px dataset.head()
```



```
Out[0:29]:
   antecedents  consequents  antecedent support  ...  lift  leverage  conviction
0  (126545804)  (133475425)                 1.0  ...   1.0       0.0         inf
1  (133475425)  (126545804)                 1.0  ...   1.0       0.0         inf

[2 rows x 9 columns]
Out[1:29]:
   antecedents  consequents  antecedent support  ...  lift  leverage  conviction
0  (126545804)  (133475425)                 1.0  ...   1.0       0.0         inf
1  (133475425)  (126545804)                 1.0  ...   1.0       0.0         inf
```

Other sets considered

- {020/ISBN, publisher-VIAF-id,650,main-entry-100-VIAF-id}
- {020/ISBN, main-entry-100-VIAF-id, publisher-VIAF-id}

Summary

Share-VDE has a process to for BIBFRAME Instance identification and clustering. The process does include some curation by hand if the algorithm does not fully match all publishers in records during the clustering algorithm process.

By using an association rule approach of pattern finding the present Fp-growth research presented here offers a supplement that may alleviate the need for extensive hand curation of publishers.

References

- Connaway, L., and Dickey, T. (2011). “Publisher Names in Bibliographic Data: An Experimental Authority File and a Prototype Application,” *Library Resources and Technical Services*, 55,4. Publisher Names in Bibliographic Data: An Experimental Authority File and a Prototype Application
- Li, H., Wang, Y., Zhang, D.,Zhang, M., Chang, E. (2008).”Pfp: parallel fp-growth for query recommendation,” *RecSys’08: Proceedings of the 2008 ACM conference on Recommender systems*, 107-114, Pfp | Proceedings of the 2008 ACM conference on Recommender systems
- Library of Congress. (2016). Overview of the BIBFRAME 2.0 Model. https://www.loc.gov/bibframe/docs/bibframe2-model.html
- Han, J., Pei, J., & Yin, Y. (2000). “Mining frequent patterns without candidate generation,” *SIGMOD Record (ACM Special Interest Group on Management of Data)*, 29(2), 1-12. Mining frequent patterns without candidate generation | ACM SIGMOD Record
- Riva, Pat, Le Boeuf, Patrick, & Žumer, Maja. (2017). *IFLA Library Reference Model: A Conceptual Model for Bibliographic Information*, IFLA Library Reference Model

Kaggle Notebook

Market Basket Analysis: https://www.kaggle.com/sajidcse/market-basket-analysis

Questions?

Jim Hahn
Head of Metadata Research
University of Pennsylvania Library
jimhahn@upenn.edu

Penn
Libraries