

# Semi-automated methods for BIBFRAME work entity description

Jim Hahn, Penn Libraries, [jimhahn@upenn.edu](mailto:jimhahn@upenn.edu)

**SWIB21 / Semantic Web in Libraries Conference / December 1 2021**

# Overview

This presentation will report an investigation of machine learning methods for the semi-automated creation of a BIBFRAME Work entity description within the RDF linked data editor Sinopia.

The automated subject indexing software Annif was configured with the Library of Congress Subject Headings (LCSH) vocabulary from the Linked Data Service.

A dataset comprising 9.3 million titles and LCSH linked data references from the IvyPlus POD project and Share-VDE was used as the training corpus.

Semi-automated processes were explored to support and extend, not replace, professional expertise.

# RDF Editors

Describing library resources with the BIBFRAME vocabulary and its core entities of Work, Instance, and Item is a resource intensive process.

Cataloging in linked data RDF editors with BIBFRAME involves careful selection of, and referencing to, external authority entities.

Creating external authoritative links is essential to produce an accurate context while describing the BIBFRAME Work entity in an RDF editor.

# Sinopia RDF Editor

Title label

ä

Civil Rights ×

Edit

Language: English

Subject of the Work 

ä

Enter lookup query

ä

Lookup

Civil rights--United States--Popular works@en



African Americans--Civil rights--History--20th century--Juvenile literature@en



Civil rights--United States--Periodicals@en



Civil rights--United States--Juvenile literature@en



State action (Civil rights)--United States--Cases@en



Civil rights--United States@en



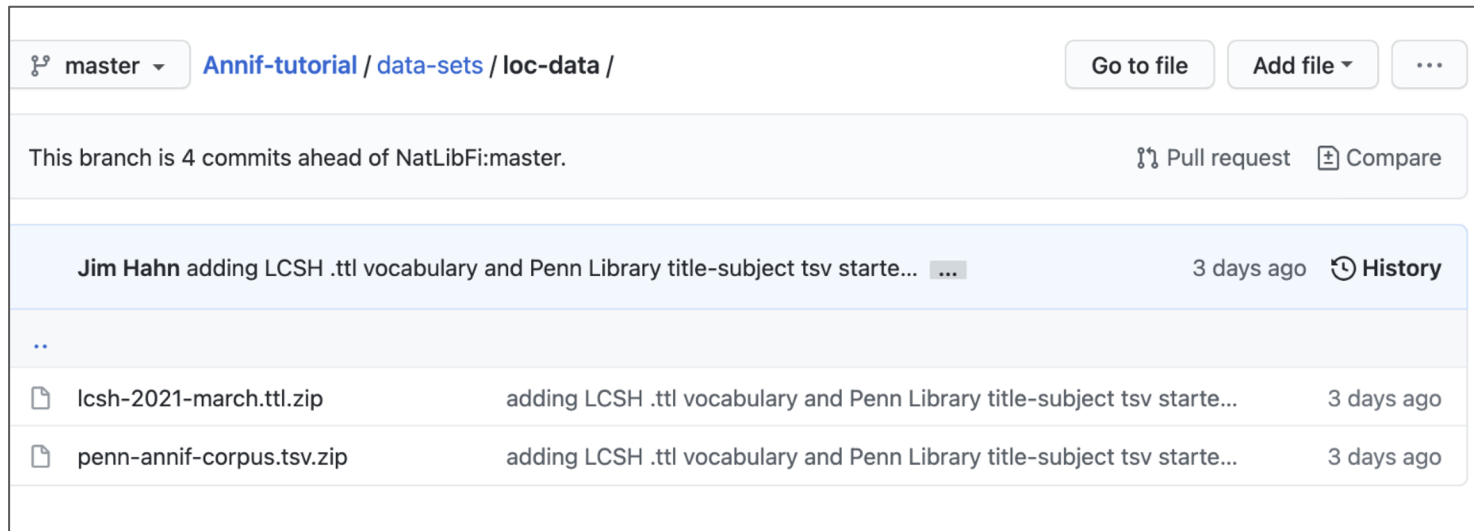
# Annif

Annif is an open-source machine learning software used to generate subject suggestions in linked data.

Title (245 \$a)	Library of Congress Subject Heading (650 \$0)
Machine learning for data streams	< <a href="http://id.loc.gov/authorities/subjects/sh97002073">http://id.loc.gov/authorities/subjects/sh97002073</a> >

<https://github.com/NatLibFi/Annif-tutorial>

# Library of Congress Linked Data Vocabulary



The screenshot shows a GitHub repository interface for the path `Annif-tutorial / data-sets / loc-data /`. The current branch is `master`, which is 4 commits ahead of `NatLibFi:master`. A commit by **Jim Hahn** is highlighted, titled "adding LCSH .ttl vocabulary and Penn Library title-subject tsv starte...". This commit includes two files: `lcsh-2021-march.ttl.zip` and `penn-annif-corpus.tsv.zip`, both added 3 days ago. The commit message for both files is "adding LCSH .ttl vocabulary and Penn Library title-subject tsv starte...".

master ▾ Annif-tutorial / data-sets / loc-data /

Go to file Add file ▾ ...

This branch is 4 commits ahead of NatLibFi:master. Pull request Compare

**Jim Hahn** adding LCSH .ttl vocabulary and Penn Library title-subject tsv starte... 3 days ago History

..

lcsh-2021-march.ttl.zip	adding LCSH .ttl vocabulary and Penn Library title-subject tsv starte...	3 days ago
penn-annif-corpus.tsv.zip	adding LCSH .ttl vocabulary and Penn Library title-subject tsv starte...	3 days ago

<https://github.com/jimfhahn/Annif-tutorial/tree/master/data-sets/loc-data>

# Baseline Training Data Composition

MARC Source	Number of Records in Set	LCSH id.loc.gov references in 650
Stanford	8258948	379605
Penn (SVDE Enriched)	5109592	1302499
Chicago	7648280	1682538
Duke	6704722	3929417
PCC Data Pool (SVDE Enriched)	4263628	2043020

<b>MARC Source</b>	<b>Sum of top 5 genres represented in 655 \$a</b>
Stanford University	music (38817); audio (34821); streaming (34683); scores (18394); fiction (16907)
University of Pennsylvania (SVDE Enriched)	films (58807); books (50583); electronic (48993); fiction (42754); streaming (23707)
University of Chicago	electronic (229266); books (228388); criticism (109294); history (107678); biography (66653)
PCC Data Pool (SVDE Enriched)	fiction (367074); periodicals (242067); history (174543); books (147182); works (136320)
Duke University	electronic (663452); books (608304); films (117252); videos (105641); internet (104231)



# How genre observations can improve prediction...

Genre distribution may influence the nature of subject data. Annif may be implemented as a set of APIs: 1) either all schools combined or 2) separate APIs based on genre type.

If an RDF editor receives a genre target, the Sinopia API selection could, as closely as possible attempt to use an end-point where that genre is prominently represented.

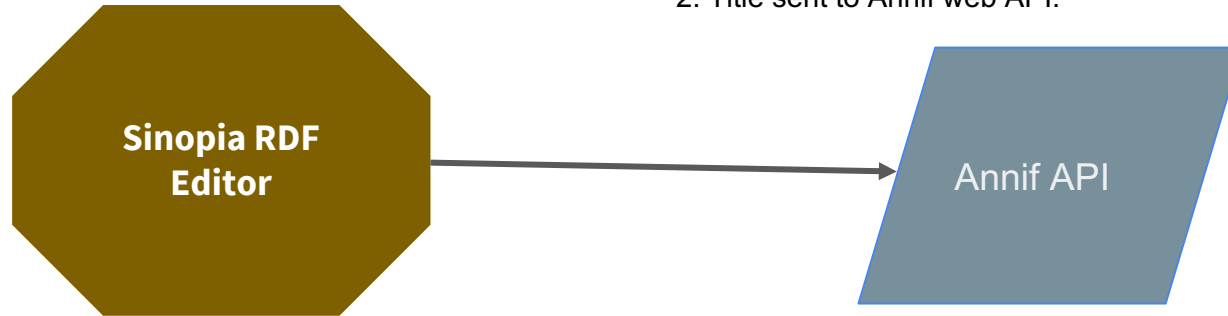
For example, if cataloging music, the editor may select the Stanford specific API as music is represented in the genre declaration more so than other collections. Similarly more film genres are declared in Penn collections and the editor may select that API for subject suggestions.

# Demo: test suggestions in terminal

```
$ echo "Machine learning algorithms build a mathematical model based on sample data" | annif suggest loc2-tfidf-new
<http://id.loc.gov/authorities/subjects/sh85079324> Machine learning 0.4599681794643402
<http://id.loc.gov/authorities/subjects/sh2002007921> Mathematical models 0.4102459251880646
<http://id.loc.gov/authorities/subjects/sh2008107143> Machine learning--Congresses 0.40729984641075134
<http://id.loc.gov/authorities/subjects/sh2009122874> Data structures (Computer science)--Congresses 0.399117648601532
<http://id.loc.gov/authorities/subjects/sh91000149> Computer algorithms 0.3975925147533417
<http://id.loc.gov/authorities/subjects/sh85014869> Blind--Travel 0.3957383334636688
<http://id.loc.gov/authorities/subjects/sh85003487> Algorithms 0.38545161485671997
<http://id.loc.gov/authorities/subjects/sh2008101223> Computer algorithms--Congresses 0.37680134177207947
<http://id.loc.gov/authorities/subjects/sh85117056> Sampling (Statistics) 0.37073814868927
<http://id.loc.gov/authorities/subjects/sh2009129432> Learning, Psychology of--Mathematical models 0.36924073100090027
$
```

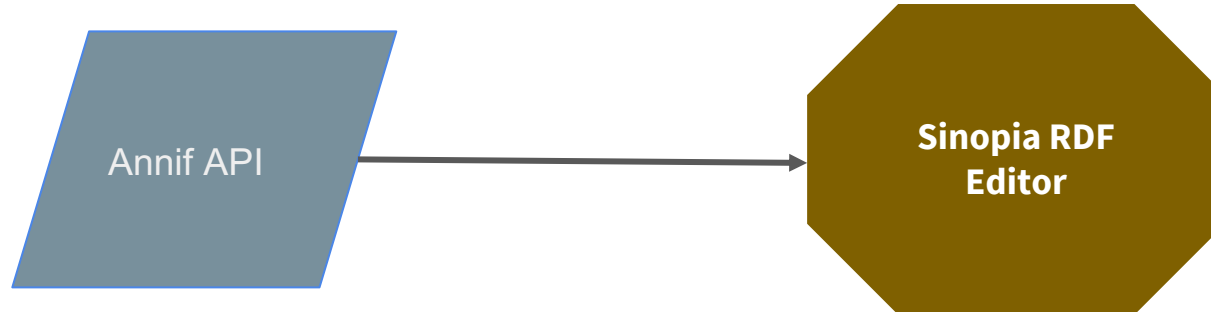
1. Cataloger types in a title in bf:Work description.

2. Title sent to Annif web API.



3. Annif web API returns suggestions with scores.

4. Business logic in Sinopia parses for suggestions over a pre-set threshold. Those that meet criteria are populated to subject field form. Cataloger selects relevant subjects.



# Outputs

The machine learning outputs, accessed by Annif web API, enable a feature for dynamically auto suggesting subject attributes based on a cataloger supplied title.

Semi-automation as a potential integration target is in contrast to completely automated cataloging and is a very specific use of machine learning.

# Next Steps

Future work will include a focus on user evaluation of Annif machine learning outputs to gather practitioner scoring as evaluation of the Annif API suggestions.

The ensembles of algorithms which Annif can support may also be a focus of future experimentation.

# Other Classification APIs

The Classify service from OCLC can **provide suggested authors (w/ VIAF authority links), FAST subject (w/ URIS), and LC number.**

Each of these may be used in the BIBFRAME Work Entity Description.

The Classify web service requires either an ISBN, OCLC number, UPC, or ISSN.

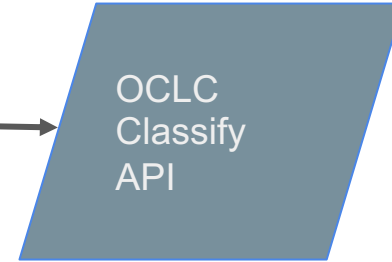
- The database provides access to more than 115 million classification numbers.
- The Classify database is current through September 2021

## OC LC Classify Web Service data flow

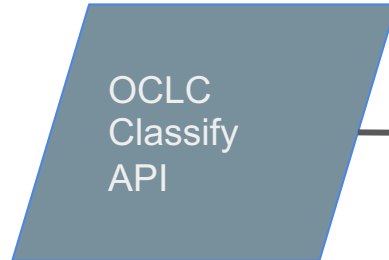
1. Cataloger types in an OCLC number or ISBN.



2. OCLC number or ISBN sent to Classify web API.



3. Classify web API returns properties.



4. Business logic in Sinopia evaluates and auto-suggests properties for each of the form elements matched. Cataloger selects relevant properties.





# Resources consulted

Annif Tutorial: <https://github.com/NatLibFi/Annif-tutorial>

Raptor RDF Syntax Library: <https://librdf.org/raptor/rapper.html>

OCLC Research, Experimental Classification Service:  
[http://classify.oclc.org/classify2/api\\_docs/index.html](http://classify.oclc.org/classify2/api_docs/index.html)