2015-02-23 breakout: Entity resolution (strings to things)

facilitator: Steven Folsom

Themes Identified:

- Requires a mix of automated and manual methods
- Need tools to do this, e.g. present user with automated matches and allow them to make changes (this could then be used to tune the algorithm)
- There's a potential to open this up to communities beyond library professionals (crowd-sourcing/niche-sourcing)

UNEDITED NOTES

Table 1

DPLA: placename resolution

- matching against Geonames
- staff discomfort
- lack subject expertise in aggregated data

Entity recognition

- use entire record as context for resolution
- points vs. shapes in geo entity resolution
- crowdsourcing opportunity?
- OCLC - several passes through data, information from multiple sources (ISNI, VIAF, etc.)
- need public feedback for last 20%
- refine algorithms based on crowdsourcing feedback
- machine transformation and confidence rating – mark that is machine-generated, with date

Table 2

strings --> things

- need string info in perpetuity
- accuracy, testability of ambiguity
- places ... think maps ...
- people
- dates ... map interface
- subjects

libraries divide and conquer entity cataloging

post-processing tools

- human mediation
- less human mediation
- hybrid models – e.g., obit project
- akin to OCR post-processing

accuracy tools

- page rank algorithm
- BibFrame converter – work accuracy?

entity extraction

- from metadata – how structured is it?
- lots of text – algorithms better

how motivate users to take tools/data for a spin?

what if we had no metadata and started only with full text?

Table 3

challenges

- solutions – would be awesome

parsing MARC to find translaters and role

- roles as strings should be things

person reconciliation

- requires human review
- resolve ambiguity in identity, roles, contributions
- predicates restrict detail
    - e.g., performer vs. violinist

crowd sourcing

- simple problems or too complex, requires experts?

music parsing

image identity

Table 4

UCSD – mix of auto & manual review

CERL – name, spelling & disambiguation

HBS – URIs provided by authority vendor

Create local auth record/URI for strings with no auth?

Feed into LC or OCLC for needed authorities?

Improve cataloging tools with type-ahead entity resolution