Rethinking HathiTrust Metadata to Support Workset Creation for Scholarly Analysis

Katrina Fenlon, Timothy Cole, Myung-Ja Han, Craig Willis, Colleen Fallaw
University of Illinois at Urbana-Champaign (USA)

Research question

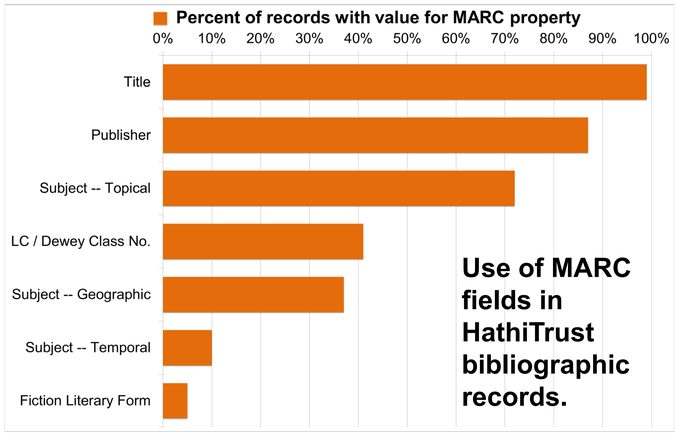
Existing bibliographic metadata records underlying the HathiTrust Digital Library are inadequate to support workset creation for scholarly analysis. How can we enrich metadata & item representations in the corpus to enable the creation of worksets supportive of scholarly analysis?

A MARC record *(select fields omitted for brevity)*:

```
<record>
 <leader>00788cam a2200229Ia 4500</leader>
 <controlfield tag="008">910204s1838    fr           000 1 fre d</controlfield>
 <datafield tag="100" ind1="1" ind2=" ">
   <subfield code="a">Sand, George,</subfield>
   <subfield code="d">1804-1876.</subfield>
 </datafield>
 <datafield tag="245" ind1="1" ind2="0">
   <subfield code="a">Indiana</subfield>
   <subfield code="c">par George Sand ...</subfield>
 </datafield>
 <datafield tag="260" ind1=" " ind2=" ">
   <subfield code="a">Paris :</subfield>
   <subfield code="b">Félix Bonnaire ...,</subfield>
   <subfield code="c">1838.</subfield>
 </datafield>
```

Limitations of MARC for workset creation

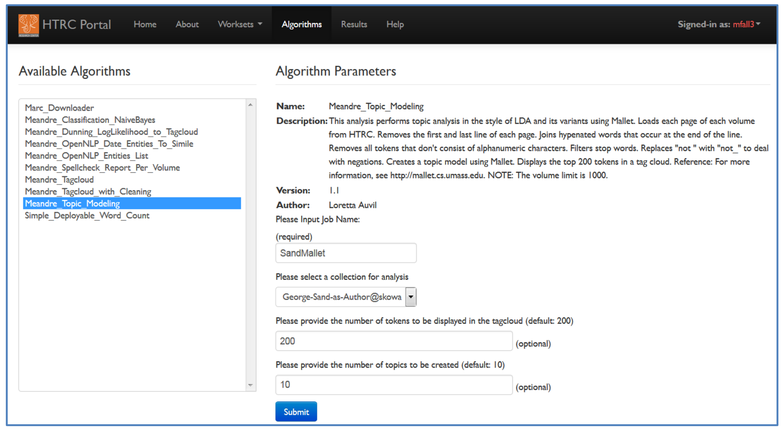
- Missing properties of interest – author gender, nationality, ...
- 'Subject' may be empty or describe form rather than scope.
- No means to link to thesauri, external authorities.
- Some fields require extensive parsing.
- Catalogers rarely use full expressiveness.



Use of MARC Fields considered in combination

	% of all records	% of records w/ genre=fiction
LC Classification	43%	46%
Topic & geographic subject(s)	36%	9%
Topic subject(s) only	36%	14%
Geographic subject(s) only	1%	< 1%

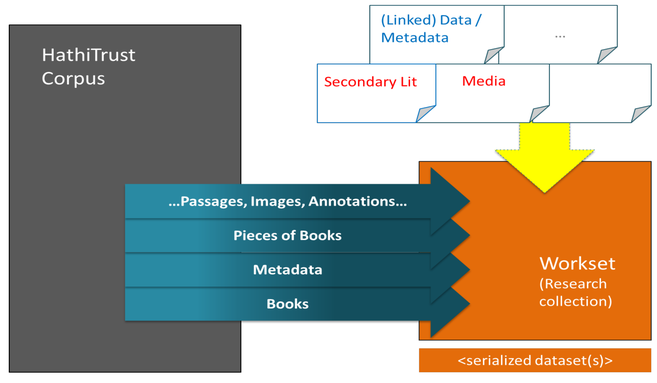
The HathiTrust Research Center



The HathiTrust (HT) is a repository of over 10 million volumes (3 billion pages) of text. The HathiTrust Research Center (HTRC) enables computational access to the HT corpus. http://www.hathitrust.org/htrc/

Why Worksets?

- Scholars regularly create worksets – selecting & gathering materials from disparate sources to answer specific research questions
- Scholars need sophisticated tools for the management and manipulation of "custom collections" of digital texts



Requirements (gathered from user studies):

- Allow scholars to gather not just primary items in the HathiTrust corpus (books), but also metadata & granular, intra-book content.
- Allow integration of external sources, such as linked datasets, secondary literature, and references.
- Allow identification and description of worksets so that they function as persistent, reusable scholarly resources.

Workset Creation for Scholarly Analysis objectives

- Allow HTRC users to formally gather selected subsets of the HathiTrust corpus together for computational analysis.
- Enable routine computational analysis across subsets of materials in the HathiTrust corpus
- Engage scholars in tool design
- Enrich metadata in the HathiTrust corpus
- Formalize the notion of worksets

Key questions being investigated by WCSA

- Given sparseness of HathiTrust records, can we enrich the corpus metadata by distilling analytics over full text?
- Can we deploy/modify off-the-shelf tools, e.g., to determine language(s) of the text, temporal / spatial coverage, etc.?
- Can we augment string-based metadata with URIs for entities – e.g., names, subjects, place of publication, etc.?
- Can we formalize the notion of worksets in HTRC, e.g., defining the necessary elements of a workset? How do we balance rigor with extensibility and flexibility?

Moving from MARC to an RDF-centric architecture

```
[ {
    "@id": "http://hdl.handle.net/2027/hvd.32044021104237",
    "@type": "http://schema.org/Book",
    "http://schema.org/datePublished": "1838",
    "http://schema.org/name": "Indiana",
    "http://schema.org/genre": "novel",
    "http://schema.org/inLanguage": "fr (French)",
    "http://purl.org/library/placeOfPublication": {
        "@id": "http://id.loc.gov/vocabulary/countries/fr" },
    "http://schema.org/author": {
        "@id": "http://viaf.org/viaf/46766944",
        "@type": "http://schema.org/Person",
        "http://schema.org/birthDate": "1804",
        "http://schema.org/deathDate": "1876",
        "http://schema.org/gender": "male",
        "http://schema.org/familyName": "Sand",
        "http://schema.org/givenName": "George" },
    "http://schema.org/about": [
        {"@id": "http://id.loc.gov/authorities/names/n79079314",
            "http://schema.org/name": "Réunion" },
        {"@id": "http://data.bnf.fr/ark:/12148/cb16549469x",
            "http://schema.org/name": "Roman d'amour français" } ]
 } ]
```

CIRSS
GRADUATE SCHOOL OF LIBRARY AND INFORMATION SCIENCE
The iSchool at Illinois
The Andrew W. Mellon Foundation