Linking FAST and Wikipedia

Rick Bennett
OCLC Research, OCLC, Dublin, Ohio, US.

Eric Childress
OCLC Research, OCLC, Dublin, Ohio, US.

Kerre Kammerer
OCLC Research, OCLC, Dublin, Ohio, US.

Diane Vizine-Goetz
OCLC Research, OCLC, Dublin, Ohio, US.

Abstract:

This paper describes a research project to develop automated techniques for linking FAST (Faceted Application of Subject Terminology) to Wikipedia articles. The research is motivated by libraries' interest in connecting library resources such as authority files to non-library linked data resources such as GeoNames and DBpedia (a dataset containing structured data extracted from Wikipedia). Of the approximately 183,000 non-subdivided topical headings in the FAST vocabulary, 76,000 terms were matched to Wikipedia article titles with 95% accuracy. Wikipedia links in the FAST authority file and FAST linked data enable people and software applications to take advantage of information in both of these resources.

Keywords: Faceted Application of Subject Terminology, FAST, Linked Data, Wikidata, Wikipedia.

Introduction:

This paper describes a research project to develop automated techniques for linking topical headings from FAST (Faceted Application of Subject Terminology) to Wikipedia articles. The research is motivated by libraries' interest in connecting library resources such as bibliographic and authority files to non-library linked data resources such as GeoNames and DBpedia (a dataset containing extracted data from Wikipedia). As the library profession has become more familiar with linked data principles, interest in accessing non-library resources has increased.[1] There are a number of demonstration projects in the library domain that connect authority files to other resources. For example, Entity Facts, a data service of the German National Library, includes information from the GND (Integrated Authority File) and links to other sources such as VIAF (Virtual International Authority File), IMDb, and Wikipedia.[2] Wikipedia links in the

FAST authority file and FAST linked data enable people and software applications to take advantage of information in both of these resources.

The FAST vocabulary is derived from the Library of Congress Subject Headings (LCSH), the largest and most widely used controlled subject vocabulary. In OCLC WorldCat, there are more than 250 million LCSH and FAST topical subject headings applied to bibliographic records. Like LCSH and FAST, Wikipedia is a general knowledge organization system that covers all topics. Since its start in 2001, Wikipedia has increased in both size and popularity. The English language Wikipedia has grown from about 20,000 articles at launch to more than 5 million articles today,[3] and the site consistently ranks as one of the most visited websites in the world.

The breadth of Wikipedia's subject coverage and the free-form nature of the articles make automated matching a challenging undertaking. The scope or coverage of a concept or entity in FAST can differ considerably from the scope of a Wikipedia article. Also, the nature of the information is quite different. Subject heading entries typically consist of an authorized heading and one or more variant forms of the heading (topical main headings average 2.8 words). In contrast, Wikipedia entries average about 560 words per article.[4]

To assess the potential for matching these diverse resources, a small random sample of non-subdivided FAST headings was selected and manually matched to Wikipedia articles. The pilot revealed that various templates used in Wikipedia could be employed to improve the accuracy of the matching. Automated techniques were developed to match non-subdivided FAST topical headings to Wikipedia article titles. This project did not attempt to extract meaning from the content of the articles.

FAST Subject Vocabulary:

FAST is an enumerative, faceted subject heading vocabulary derived from the Library of Congress Subject Headings. FAST was developed to meet the need for a general-use subject vocabulary that is easy to learn, apply, and control.

FAST consists of eight categories of terms, or facets, which cover key attributes of information resources (topics, persons, organizations, events, geographic places, titles of works, time, and form/genre). As a fully enumerative system, all authorized FAST headings, except chronological headings, are established in an authority file, eliminating the need to synthesize headings according to a set of rules.

In January 2016, the FAST vocabulary contained more than 1.7 million headings. About 24% are topical subject headings. Topical headings in FAST are made up of topical main headings and main heading/subdivision combinations that have been used in WorldCat bibliographic records. Topical headings are the majority of subject headings used in cataloging.[5] The following are examples of FAST topical headings:
1. AIDS (Disease)
2. Beauty contests—Corrupt practices
3. Books in church work
4. Cities and towns
5. Commerce
6. Political science
7. Travel

8. Water
9. Water—Pollution
10. Water—Pollution—Environmental aspects
11. Water—Pollution—Environmental aspects—Measurement

In the list above, heading #2 and headings #9-11 are referred to as subdivided headings, meaning that they consist of a main heading plus one or more subdivisions. Approximately 183,000 (45%) of topical headings are non-subdivided headings. For this project, only non-subdivided headings were candidates for matching against Wikipedia article titles.

Wikipedia Article Titles and FAST Headings:

In Wikipedia, an article title indicates what the article is about and distinguishes it from other articles. Wikipedia provides its editors detailed guidance on choosing a title for an article. According to Wikipedia, good article titles will have the following characteristics: [6]

- **Recognizability** – The title is a name or description of the subject that someone familiar with, although not necessarily an expert in, the subject area will recognize.
- **Naturalness** – The title is one that readers are likely to look or search for and that editors would naturally use to link to the article from other articles. Such a title usually conveys what the subject is actually called in English.
- **Precision** – The title unambiguously identifies the article's subject and distinguishes it from other subjects.
- **Conciseness** – The title is no longer than necessary to identify the article's subject and distinguish it from other subjects.
- **Consistency** – The title is consistent with the pattern of similar articles' titles.

These guidelines echo many of the principles governing the formation of Library of Congress topical subject headings:[7]

- **Recognizability** / **Uniform heading** – Preference is given to terms in general use over technical terms or jargon where possible.
- **Naturalness** / **Uniform heading** – One heading is chosen to represent a topic; headings are based on standard, contemporary American English-language usage.
- **Precision** / **Specific entry** – Each subject is represented by the most precise term naming the subject, rather than a broader or generic term that encompasses it;
- **Conciseness / Unique heading** – Each heading represents only one topic. If a term could represent more than one concept, it is modified.
- **Consistency** / **Consistency** – Attempts are made to maintain consistency in form and structure among similar headings.

Despite this similarity in guiding principles, an article title and subject heading for a given concept can be quite different. For example, in Wikipedia scientific names are used for article titles about plants (e.g., Maize), but in LCSH and FAST the popular or common name is preferred (e.g., Corn). Wikipedia editors are advised to provide redirects (a page that has no content but sends the reader to another page) for alternative terms for concepts and names.[8] For example, when the term, 'Corn' is searched in Wikipedia, the article with the title, *Maize*, is displayed. Redirects are functionally similar to *see from* tracings in MARC 21 authority records, that is, they lead a user from a variant form to the used or authorized form. The FAST

authority record for Corn has the *see from* tracing, *Maize*. In matching FAST to Wikipedia, both authorized and variant forms are used to effect matches.

Wikipedia and the LCSH and FAST vocabularies differ in other ways. For example, they differ in the use of plural vs. singular forms of terms. Wikipedia articles generally use the singular form in article titles (Bird rather than Birds) in contrast to LCSH and FAST, which use the plural form for headings representing objects or classes of people (e.g., Coffeehouses; Trucks, Nurses).[9]

Matching Approach:

A copy of Wikipedia was downloaded on 2 November 2015. All automated matching was performed locally against this file. A pilot study conducted in the early phases of the research indicated that Wikipedia infoboxes,[10] templates used to foster consistency across articles and give a brief summary of an article's content, had the potential to streamline matching and improve accuracy. A small random sample of non-subdivided FAST headings was selected and manually matched to Wikipedia articles. For each attempted match, the information in Table 1 was recorded. Authorized headings and Wikipedia titles are shown in bold type; *see from* tracings are in italics.

Table 1. Matches between FAST and Wikipedia

FAST Identifier	FAST Term	Wikipedia Title	Match Type	Infobox Type	Match Status
fst00829260	**Bear dance**	**Tame bear**	redir_Bear dancing	None	No
fst00885485	*Canine teeth*	**Canine tooth**	redir_Canine teeth	anatomy	Yes
fst00885485	*Canines (Teeth)*	**Canine tooth**	redir_Canine teeth	anatomy	Yes
fst00885485	**Cuspids**	**Canine tooth**	redir_Cuspid	anatomy	Yes
fst01080963	**Psoriasis**	**Psoriasis**	article title	disease	Yes
fst01107947	**School-to-work transition**	**School-to-work transition**	article title	None	Yes
fst01195257	**Suites (Piano, horn, viola)**	NA	NA	NA	non-match
fst01200215	**Sancy diamond**	**Sancy**	redir_Sancy diamond	diamond	Yes
fst01429931	**Blogging**	**Joe Bloggs**	redir_Bloggs	None	No
fst01740955	**Children's writings, Latvian**	NA	NA	NA	non-match
fst01894827	**Ogon (The Russian word)**	NA	NA	NA	non-match

Using the sample results for guidance, two lists of infoboxes were developed, **Accept** and **Reject**. The Accept list permits matches to Wikipedia articles with particular types of infoboxes, for example, anatomy, disease, zoological and botanical taxa, language family, and ethnic group. The Accept infoboxes describe things that are likely to be in the FAST topical facet. The Reject list prevents matches of headings to article titles with a different set of infoboxes, for example, airline, airport, port, and river. The Reject infoboxes describe things that are unlikely to be in the FAST topical facet, such as airports which are in the corporate

facet, and rivers which are in the FAST geographic facet. A few infoboxes on the Reject list could be associated with the FAST topical facet, but the coverage differs greatly between FAST and Wikipedia in these areas so that matching would risk creating erroneous results. Matches to articles with no infoboxes were also accepted. The use of infoboxes to filter matches reduced the number of potential matches while increasing accuracy.

Other types of Wikipedia articles were excluded from matching. For example, lists are common in Wikipedia, and Wikipedia has an article "List of Italian-American women writers." This article could be associated with the FAST heading "Italian America women authors," however, it is a list of writers, not an article about writers; consequently, all "List of …" Wikipedia articles were ignored. Similarly, Wikipedia "Category: …" pages were excluded, since these are lists of category navigation links and not actual articles. Other Wikipedia pages that were ignored are administrative in nature, including those starting with Wikipedia:, Help:, Template:, and File:.

To account for differences in prefixes and suffixes (e.g., plurality, tense) all heading words were stemmed.[11] After elimination of Wikipedia headings associated with Reject infoboxes, stemmed FAST headings and *see from* tracings were matched to stemmed Wikipedia article titles and redirects, resulting in 76,000 matches.

Of the total matches, nearly 48% of headings matched an article title and 26% matched a redirect. Almost 19% of *see from* tracings matched an article title, and about 12% matched a redirect. The sum is greater than 100% because all possible match types were attempted (heading to article title, heading to redirect, *see from* to article title, *see from* to redirect). Examples of the matches by type are given in Table 2. A manual examination of a sample of matching headings indicates that about 95% of the matches were accurate.

Table 2. Example Matches by Type

Authorized Heading to Article Title			
FAST Heading	FAST *see from*	Wikipedia Title	Wikipedia Redirect
Librarians		Librarian	
Libraries		Library	
Library anxiety		Library anxiety	
Authorized Heading to Redirect			
FAST Heading	FAST *see from*	Wikipedia Title	Wikipedia Redirect
Library buildings		Library	Library building
Library historians		Library and information scientist	Library historian
Library research		Secondary research	Library research
Library schools		Education for librarianship	Library school
***See from* to Article Title**			
FAST Heading	FAST *see from*	Wikipedia Title	Wikipedia Redirect
Library administration	Libraries Management	Library management	
Library orientation	Library instruction	Library instruction	
Library shelving	Library stacks	Library stack	

See from to Redirect			
FAST Heading	FAST *see from*	Wikipedia Title	Wikipedia Redirect
Library information desks	Library help desks	Reference desk	Library help desk
Library storage centers	Deposit libraries	Legal deposit	Deposit libraries

About 66%, or 121,000, of non-subdivided headings failed to match a Wikipedia title or redirect. A sample of these headings was manually examined to: 1) find the rate of matches missed by the technique described above, and 2) characterize the FAST headings for which there are no corresponding Wikipedia articles. It was found that 15% of potential heading matches were missed. The dominant reason for failure to match was differences in terminology without *see from* tracings or redirects to compensate. Some examples are:

FAST Heading	**Wikipedia Title**
Cooking, Circassian	Circassian cuisine
Student Expulsion	Expulsion (education)
Vingt (Group of artists)	Les XX

Only a small proportion of the matches failed because the Wikipedia title was associated with a Reject infobox and was unavailable for matching.

There are many FAST headings that do not have a corresponding Wikipedia article. These include headings with parenthetical qualifiers, language, nationality or ethnic group qualifiers, and various types of compound headings. For example, the following headings failed to match:
- Librarians in literature
- Librarians with disabilities
- Librarians' writings
- Librarians' writings, Swiss (French)
- Libraries (Computer program subroutines)
- Libraries and abused women

This does not necessarily mean that these subjects are not addressed in the body of a Wikipedia article, just that the concepts are not represented as article titles. Mining the content of articles for meaning was beyond the scope of this study.

FAST and Linked Data:

In late 2011, OCLC released FAST as linked data. The file is available in multiple representations including MARCXML and RDF. Individual FAST authority records can be accessed through a user interface and as linked data. The release of FAST as linked data enhances machine access to FAST.

> Linked Data is a publishing paradigm for making data…fully accessible and inter-linkable anywhere on the Internet. Linked Data uses the same common Web communications protocols as ordinary browser software to connect machine-readable data across distributed computers.[12]

A fundamental assumption about linked data is that the value and usefulness of data increases the more it is linked with other data on the web. Links enable people and machines to navigate between data sources and to discover and use additional data resources. In a summary of results from the 2014 OCLC Research Linked Data Survey,[13] Smith-Yoshimura reports libraries are using linked data in the following ways:

- Enrich bibliographic metadata or descriptions
- Interlinking
- As a reference source
- Harmonize data from multiple sources
- Automate authority control
- Enrich an application
- Dataset discovery
- Auto-suggest

A follow-up survey conducted in 2015[14] showed that the following linked data sources are used most often:

1. VIAF (Virtual International Authority File)
2. DBpedia (structured content extracted from Wikipedia)
3. GeoNames
4. id.loc.gov (Library of Congress Linked Data Service)

FAST authority records already include links to related Library of Congress Name and Subject authority records. Many FAST geographic headings also have links to the GeoNames geographic database, and records for personal names, corporate names, and titles of works link to the VIAF data file hosted by OCLC. In FAST linked data, the links to GeoNames and VIAF are represented as URIs to linked data services. The addition of thousands of links to Wikipedia further connects FAST to the top resources used by libraries in linked data applications. Table 3 provides the number of links in FAST to other linked data resources by facet.

Table 3. Links to External Resources in the FAST File

Facet Type						
	Persons	**Organizations**	**Events**	**Titles of Works**	**Topic**	**Geographic places**
FAST	698,103	362,382	12,461	63,071	407,350*	177,959
Links to External Files						
LC Subject Headings	22,945	8,414	5,422	85	217,569	46,288
LC Name Authority	675,157	353,953	6,916	62,986	0	121,923
VIAF	669,902	352,111	6,901	62,923	1	121,239
Wikipedia**	160,675	5	89	1	75,935	64
GeoNames	0	2	0	0	0	85,411
Total Links	1,528,679	714,485	19,328	125,995	293,502	374,925

*181,821 headings are non-subdivided;
**Wikipedia links to non-topical facets were extracted from VIAF

FAST as linked data is available via a number of methods.[15] The searchFAST interface[16] designed for catalogers and other metadata specialists includes the FAST URI as a permanent link in the form http://id.worldcat.org/fast/{fastid}. Although the primary use of this data will likely be software applications, human readable representations help people to better understand the content and nature of a resource. Linked data access via an SRU/SRW Application Programming Interface (API) is also provided. The API allows for machine access and Content Negotiation for FAST in MARC 21, HTML, or RDF/XML.[17] FAST authority records available in RDF format are based primarily on schema.org mappings. The FAST data set is also available for download from OCLC[18]

A sample FAST authority record in MARC format with links to Wikipedia is shown in Figure 1. All Wikipedia links are coded with subfield 4 (Relationship code) equal to RM (Related mapping).[19] The weakest relationship code was chosen for coding the mappings, e.g., RM (Related mapping) vs. EQ (Equivalence), given that all matches were made programmatically. Users of the searchFAST interface will have the opportunity to vote on the correctness of the matches. The votes will be used to remove invalid matches. This effort dovetails with other OCLC Research projects designed to score and synthesize information from authority and bibliographic files and combine that information with non-library resources on the web.[20,21]

Figure 1. FAST Authority Record with Links to Wikipedia

001		fst00878993
003		OCoLC
005		20150624163456.0
008		041024nn anznnbabn || ana d
016	7_	fst00878993$2 OCoLC
040	__	OCoLC$b eng $c OCoLC $f fast
053	_0	QK495.G74 $c Botany
053	_0	SB191.M2 $c Culture
150	__	Corn
450	__	Corn plant
450	__	Indian corn
450	__	Maize
450	__	Zea mays
550	__	Zea $0 (OCoLC)fst01184138 $w g
688	__	LC (2014) Subject Usage: 313 (791)
688	__	WC (2014) Subject Usage: 8,739 (27,256)
750	_0	Corn $0 (DLC)sh 85032625
750	_4	Cornally $4 RM $0 (uri)http://en.wikipedia.org/wiki/Cornally $0 (uri)https://www.wikidata.org/wiki/Q5171075 □Remove incorrect match
750	_4	Flint corn$4 RM $0 (uri)http://en.wikipedia.org/wiki/Flint_corn $0 (uri)https://www.wikidata.org/wiki/Q3303407 □Remove incorrect match
750	_4	Maize$4 RM $0 (uri)http://en.wikipedia.org/wiki/Maize $0 (uri) □Remove incorrect match

An additional benefit to matching to Wikipedia is the ability to link to Wikidata, which acts as a central store for the structured data of Wikimedia projects including Wikipedia.[22] For example, each Wikipedia page with an entry in Wikidata uses the language links stored in Wikidata to populate the Languages section in the left column of a Wikipedia article. In this

example, Wikidata links the English Wikipedia article about Maize to corresponding articles in more than 100 additional languages.

Wikidata: https://www.wikidata.org/wiki/Q11575

These links may provide a mechanism for future linking of FAST and non-English controlled vocabularies.

Future Research:

FAST Linked data will be monitored for changes in usage after the addition of Wikipedia links, and statistics on match validity will be accumulated through the searchFAST interface. Positive outcomes, such as increased usage of FAST in linked data implementations, might lead to follow-up projects in which topical matching techniques are further refined or techniques for matching additional facets are developed.

References and Notes

[1] Smith-Yoshimura, Karen. "Linked Data Survey Results 3–Why and What Institutions Are Consuming (Updated)." Web log post. *Hangingtogether.org*. OCLC Research., 1 Sept. 2014. Web. 30 June 2016. <http://hangingtogether.org/?p=4155>

[2] *Entity Facts*. Deutsche Nationalbibliothek, 24 Mar. 2016. Web. 30 June 2016. <http://www.dnb.de/EN/Wir/Projekte/Abgeschlossen/entityFacts.html>.

[3] Wikipedia." *Wikipedia*. Wikimedia Foundation, 30 June 2016. Web. 30 June 2016. <https://en.wikipedia.org/wiki/Wikipedia>.

[4] Wikipedia:Size Comparisons." *Wikipedia*. Wikimedia Foundation, 10 June 2016. Web. 30 June 2016. <https://en.wikipedia.org/wiki/Wikipedia:Size_comparisons>.

[5] "MARC Usage in WorldCat: Zoomable Starburst of All MARC Tags - By Tags." *MARC Usage in WorldCat*. OCLC, n.d. Web. 30 June 2016. <http://experimental.worldcat.org/marcusage/viz/starburst-tags.html>.

[6] "Wikipedia:Article Titles." *Wikipedia*. Wikimedia Foundation, 28 June 2016. Web. 30 June 2016. <https://en.wikipedia.org/wiki/Wikipedia:Article_titles>.

[7] "Session 2 History and Principles of LCSH." *Basic Subject Cataloging Using LCSH: Instructor's Manual (Revised 2011)*. Ed. Lori Robare. Washington: Library of Congress., 2007. 51-61. Web. 30 June 2016. <http://www.loc.gov/catworkshop/courses/basicsubject/pdf/LCSH_Instructor_2011.pdf>.

[8] *Wikipedia*. Wikimedia Foundation, n.d. Web. 30 June 2016. <https://en.wikipedia.org/wiki/Wikipedia:Redirect#Purposes_of_redirects>.

[9] "Session 3 History and Principles of LCSH." *Basic Subject Cataloging Using LCSH: Instructor's Manual (Revised 2011)*. Ed. Lori Robare. Washington: Library of Congress., 2007. 83. Web. 30 June 2016. <http://www.loc.gov/catworkshop/courses/basicsubject/pdf/LCSH_Instructor_2011.pdf>.

[10] "Wikipedia:List of Infoboxes." *Wikipedia*. Wikimedia Foundation, 4 Dec. 2015. Web. 30 June 2016. <https://en.wikipedia.org/wiki/Wikipedia:List_of_infoboxes>.

[11] An implementation of the Krovetz stemmer (Kstem) was used.

[12] "Why Linked Data?" *LD4L Linked Data for Libraries*. LD4L, 2014-2016. Web. 30 June 2016. <https://www.ld4l.org/node/11>.

[13] Smith-Yoshimura, Karen. "Linked Data Survey Results 3–Why and What Institutions Are Consuming (Updated)." Web log post. *Hangingtogether.org*. OCLC Research, 1 Sept. 2014. Web. 30 June 2016. <http://hangingtogether.org/?p=4155>

[14] Smith-Yoshimura, Karen. *Linked Data Implementations. OCLC Research Update ALA Annual 2016*. OCLC, 27 June 2016. Slide 89. Web. 30 June 2016. <http://www.slideshare.net/oclcr/oclc-research-update-ala-annual-2016>.

[15] "FAST Linked Data." *OCLC Research*. OCLC, 2016. Web. 30 June 2016. <http://experimental.worldcat.org/fast/>.

[16] "searchFAST." *OCLC Research*. OCLC, 2016. Web. 30 June 2016. <http://fast.oclc.org/searchfast/>.

[17] "FAST Linked Data API." *OCLC Developer Network*. OCLC, 2015. Web. 30 June 2016. <https://www.oclc.org/developer/develop/web-services/fast-api/linked-data.en.html>.

[18] "FAST Dataset Download." *OCLC Research*. OCLC, 29 Feb. 2016. Web. 3001 June 2016. <http://www.oclc.org/research/themes/data-science/fast/download.html>.

[19] "7XX - Heading Linking Entries-General Information." *MARC 21 Format for Authority Data.* Network Development and MARC Standards Office. Library of Congress, 20 Oct. 2014. Web. 30 June 2016. <https://www.loc.gov/marc/authority/ad7xx.html>.

[20] Mixter, Jeff. "Turning Bibliographic Metadata into Actionable Knowledge." Web log post. *OCLC Next*. OCLC, 29 Feb. 2016. Web. 30 June 2016. <http://www.oclc.org/blog/main/turning-bibliographic-metadata-into-actionable-knowledge/>.

[21] Washburn, Bruce, and Jeff Mixter. "Looking Inside the Library Knowledge Vault." *OCLC Research Library Partners, Works in Progress Series: Looking Inside the Library Knowledge Vault*. OCLC Research, 12 Aug. 2015. Web. 30 June 2016. <http://www.slideshare.net/oclcr/looking-inside-the-library-knowledge-vault>.

[22] "Wikidata." *Wikipedia*. Wikimedia Foundation, 7 June 2016. Web. 30 June 2016. <https://en.wikipedia.org/wiki/Wikidata>.