IFLA Satellite Pre-Conference of the Classification and Indexing Section
"Looking at the Past and Preparing for the Future"
Florence, Italy, 20-21 August 2009

FAST Headings as Tags for WorldCat

Diane Vizine-Goetz
OCLC Online Computer Library Center, Inc.

Abstract
This paper reports on an investigation to use Faceted Application of Subject Terminology (FAST) as a surrogate for tags in WorldCat, a global catalog of bibliographic records and location information for books, videos, music, and other types of materials found in libraries. FAST is a controlled vocabulary based on the Library of Congress Subject Headings (LCSH). FAST is applied to a copy of WorldCat to explore the potential of generating tag-like information for bibliographic records. The paper provides sample visualizations of FAST headings inspired by social tagging applications.

Introduction

The growing body of research on social tagging includes a few studies that compare user-entered tags and subject headings. Spiteri (2007) investigates the structure and content of tags evaluating them against guidelines for thesaurus construction and LCSH. She finds that most tags represent things (e.g., concepts, people, places, and organizations) and that nouns are the most common form of term. She reports a high degree of overlap between the concepts expressed in tags and LCSH, but notes that tags differ from LCSH in vocabulary and specificity of terms. Rolla (2009) compares the quantity and nature of user-entered tags in LibraryThing to LCSH in bibliographic records for a sample of 45 titles. He finds that there are many more tags for a title than subject headings. The tags express concepts not found in subject headings for each of the 45 titles; subject headings express unique concepts not expressed in tags for a little more than half of the titles. He also reports significant overlap between the concepts in tags and LCSH and that user tags are often either broader or narrower than LCSH. Both investigators suggest that subject headings can be improved by adopting some of the qualities of tags, e.g., use of more natural language and popular terminology.

This paper reports on an investigation to use Faceted Application of Subject Terminology (FAST) as a surrogate for tags in WorldCat. It does not compare user tags and subject headings, but instead seeks to generate tag-like data elements from a controlled vocabulary. In this study, bibliographic records containing FAST headings are clustered into work sets. The headings are then aggregated at the work level and treated like tags.

The paper begins with an overview of FAST followed by a discussion of the application of FAST to a copy of WorldCat that is clustered into work sets according to principles of the Functional Requirements of Bibliographic Records (FRBR) model. Examples of headings aggregated at the work level are provided, followed by sample visualizations of FAST headings. The paper concludes with observations on the potential application of FAST in cataloging and end user environments and suggests directions for further research.

Overview of FAST

FAST is a controlled vocabulary based on the terminology of LCSH. FAST retains the rich vocabulary of LCSH, the synonym and homograph control of its parent vocabulary, and the linkages among terms provided by LCSH's cross reference structure. FAST consists of headings which are categorized into seven subject facets and one form/genre facet:

Topic
Place
Time
Event
Person
Corporate body
Title of work
Form/Genre

Each heading belongs to only one facet, and facets may be used independently. All headings are enumerated in FAST except time[1] making FAST headings easier to apply and validate than headings from schemes that involve synthesis. FAST headings are established by faceting the Library of Congress Subject Authority File and LCSH in WorldCat records. A single LCSH may be broken up into multiple FAST headings:

LCSH

Psychiatric hospital patients—Massachusetts—Biography

FAST

Psychiatric hospital patients [Topic]
Massachusetts [Place]
Biography [Form/Genre]

LCSH

Women college students—Suicidal behavior—Fiction

FAST

Women college students—Suicidal behavior [Topic]
Fiction [Form/Genre]

Faceting does not always result in more headings. Chan et al. (2001) give an example in which 47 different geographic LCSH entries are faceted into 19 FAST headings. FAST headings typically have a simpler syntax than LCSH even though FAST retains the use of

subdivisions and the hierarchical structure of LCSH. A subdivision in FAST must belong to the same facet as the main heading. Headings are added to FAST according to their usage in WorldCat. Before a heading is added to the list, it is validated to eliminate errors and inconsistencies. In June 2009, the FAST file consisted of 1.6 million headings. For a detailed discussion of the history, structure and development of FAST, see Chan and O'Neill (forthcoming).

FAST in WorldCat

This study reports on the application of FAST to a copy of WorldCat.[2] Approximately 136 million LCSH in 62 million (46%) WorldCat records were faceted into 176 million FAST headings. The average number of LCSH per record is 2.19; the average number of FAST per bibliographic record is 2.84. The entire bibliographic file of 135 million records was then clustered into work sets using the OCLC FRBR Work-Set Algorithm.[3]

The algorithm collects bibliographic records into groups based on author and title information from bibliographic and authority records. Author names and titles are normalized according to the NACO Authority File Comparison Rules to construct a key for each bibliographic record (e.g., **plath, sylvia/bell jar** is the key for *The Bell Jar* by Sylvia Plath). All records with the same key are grouped together in a work set. The clustering process produced 98,960,368 work sets; 44% (43,105,432) had FAST headings. About 1.5 million different FAST headings are represented in the clustered bibliographic file.

The distribution of unique headings over the six major subject facets is shown in Table 1. Ninety percent of topical headings and 77% geographic headings are applied to multiple work sets. Nearly twice as many unique personal name headings than topical headings are applied; however, more than 50% of personal names are applied to a single work set. Of the top 50 headings, 26 are topics, 23 are place names, and one is a corporate body. Less than 2% of the headings account for 80% of the use.

Table 1.

Facet	**Unique FAST**	**% of FAST**	**Multiple Uses**	**%**	**Single Use**	**%**
Person	635,088	42%	310,766	49%	324,322	51%
Corporate body	346,535	23%	196,857	57%	149,678	43%
Event	9,964	1%	6,308	63%	3,656	37%
Title of Work	47,654	3%	30,138	63%	17,516	37%
Topic	343,748	23%	310,765	90%	32,983	10%
Place	136,936	9%	105,269	77%	31,667	23%
Total	1,519,925	100%	960,103		559,822	

The average number of FAST headings per work set is 2.55; work sets consisting of 3 or more bibliographic records average 3.2 headings per work. Work sets for fiction have slightly more headings than other work sets; they average 2.65 FAST per work, and 3.78

FAST for works with 3 or more bibliographic records. Records in work sets with 3 or more records are the most likely to be used by libraries; they average 25 holding codes, or holdings,[4] per record and 133 holdings per work set, compared to 11 and 14, respectively, for the entire file.

Aggregating FAST Headings

This section presents information on the FAST headings in six work sets. Table 2 provides the following information for each work set:

- Title of work and work key (column 1)
- Total holdings count for the work set (column 2).
- Total number of languages of publication in WorldCat (column 3). This number includes the language of the original plus languages of translations.
- Total number of bibliographic records in the work set, i.e., all records with the same author/title key (column 4).
- Total number of LCSH (column 5). This number is a count of all unique headings coded as LCSH in the bibliographic records. This count was not adjusted for variations in capitalization, punctuation, or subfield coding, or for spelling and typographical errors.
- Total number of unique FAST that resulted from faceting LCSH (column 6).

Table 2.

Title / Work key	Holdings	Languages	Records	Unique LCSH	Unique FAST
1. The Bell Jar *plath, sylvia/bell jar*	10,474	23	188	28	17
2. A Brief History of Time *hawking, s w\stephen w/brief history of time*	11,402	21	162	22	14
3. Fast Food Nation *schlosser, eric/fast food nation*	7,297	14	60	20	16
4. Housekeeping *robinson, marilynne/housekeeping*	3,757	6	53	20	14
5. Girl, Interrupted *kaysen, susanna\1948/girl interrupted*	4,363	10	30	16	10
6. A Savage War of Peace *horne, alistair/savage war of peace algeria 1954 1962*	2,450	3	28	2	2

Titles 2, 3, and 6 are titles studied by Rolla (2009, 184). They are included here so that the reader can compare results from the two investigations.

All of the example work sets include records for translations of the original title, and all are widely held by libraries; they average more than 6,600 holdings per work. The number of records per work set ranges from 188 records for *The Bell Jar*, a work of fiction, to 28 records for *A Savage War of Peace*, a non-fiction work. The Bell Jar has the greatest number of LCSH with 28; *A Savage War of Peace* has the fewest with 2. The number of unique LCSH tends to increase as the number of records in the work set increases. The average number of LCSH per work set is 18 and the average number of FAST is 12.17, much greater than the average of 2.55 FAST for the complete file. The FAST headings for each work set are given below. The top eight headings by holdings count are shown in bold type; individual headings are delimited by a semicolon.

The Bell Jar

American fiction; Autobiographical fiction; College students; **College students—Suicidal behavior**; **Depression, Mental**; Mental illness; **Plath, Sylvia**; **Psychological fiction**; **Suicidal behavior**; Suicidal behavior—Treatment; United States; Women authors; **Women college students**; **Women college students—Suicidal behavior**; Young adult fiction; Young women; Young women—Psychology

A Brief History of Time

Astronomy; **Big bang theory**; **Black holes (Astronomy)**; **Cosmography**; **Cosmology**; Einstein, Albert, 1879-1955; Galilei, Galileo, 1564-1642; **Hawking, S. W. (Stephen W.)**; Korean language; Newton, Isaac, Sir, 1642-1727; **Physics**; Planets—Origin; **Space and time**; **Time**

Fast Food Nation

Book clubs (Discussion groups); Consumer behavior; **Convenience foods**; Convenience foods—Social aspects; **Cookery, American**; Diet; **Diet—Health aspects**; **Fast food restaurants**; **Food habits**; **Food industry and trade**; Great Britain; Group reading; Packing-houses; Restaurateurs; Social history; **United States**

Housekeeping

1900 – 1999; **American fiction**; **Aunts**; Authors, American; **Domestic fiction**; **Eccentrics and eccentricities**; **Girls**; Idaho; Mothers; **Mothers—Death**; Pacific Northwest; **Psychological fiction**; Robinson, Marilynne; **United States—Northwestern States**;

A Savage War of Peace

Algeria; Revolution (Algeria : 1954-1962)

Girl, Interrupted

1900 – 1999; **Authors, American**; **Kaysen, Susanna, 1948-**; **Massachusetts**; **Mental health**; Mental illness; Mentally ill; **Mentally ill, Writings of the**; **Psychiatric hospital care**; **Psychiatric hospital patients**

Each set of headings is an aggregation of the unique FAST headings for a given work set. For example, for *Fast Food Nation*, 2 records with 26 holdings have the heading, **Book clubs (Discussion groups);** 9 records with 50 holdings have the heading **Cookery, American**, 36 records with 6,797 holdings have the heading **Fast food restaurants**, and so on. Aggregating the headings at the work level brings together all of the headings for a work, including correctly assigned headings and headings that may be inaccurately applied, e.g., **Cookery, American**.[5] Headings that pertain only to translations or particular editions are also part of the set, e.g., **Book clubs (Discussion groups)**, for a book club edition.

FAST as Tags

A tag cloud is a common way of presenting tags where more frequently used tags are emphasized using different font sizes or colors. The headings for *Fast Food Nation* are presented in Figure 1 as a tag cloud using holdings counts to weight the headings. The resulting cloud makes it easy to see that **Convenience foods** and **Fast food restaurants** are the primary subjects applied to this work and that **Packing-houses** and **Consumer behavior** are applied less frequently.

Lists and charts are other ways of presenting terms and their frequencies or weights. An alternate graphical visualization of the headings for *Fast Food Nation* is shown in Figure 2. It consists of a pie chart and a list of the top eight headings ranked by holdings count; only the top five headings are presented in the chart. The value eight was selected based on a study of user tagging by Shirky (2005). He finds that the top eight tags in social tagging applications represent a consensus view of a resource. Research is needed to determine if there is a similar threshold for subject headings and bibliographic records.

Conclusion and Future Directions

FAST headings, aggregated in the manner described in this study, can give librarians and other users access to a wider variety of headings than is available in individual bibliographic records. When FAST headings are presented as tag clouds, users can quickly see what subjects have been applied to a work. In a cataloging environment, access to aggregations of headings could lead to more efficient workflows, especially for the cataloging of new editions of existing works. Library staff members might also be more likely to notice and correct erroneous headings since they stand out among the correctly assigned ones. In end user environments, FAST could be used to improve browsing and navigation.

The next phase of this research will involve prototyping one or more of the scenarios described above and testing FAST with users. Future research may include exploring ways to increase the number and type of headings associated with work sets that have

only a few headings, e.g., *A Savage War of Peace*, and experimenting with techniques to supplement controlled headings with additional entry vocabulary, including words and phrases from user tags.

Figure 1.

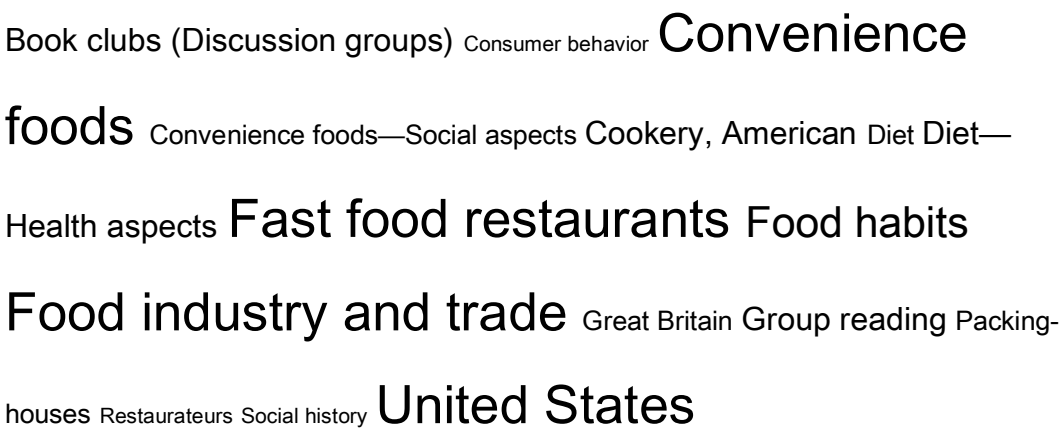
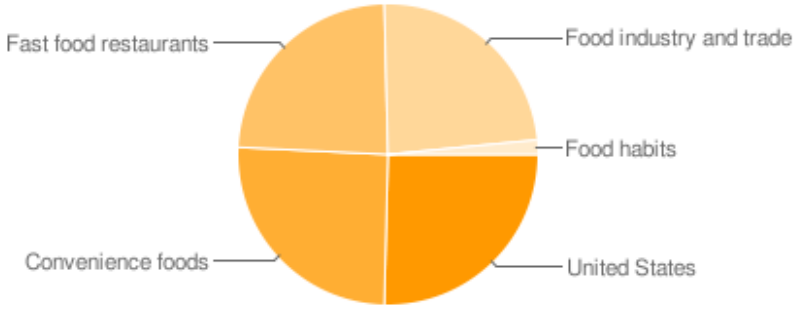


Figure 2.



1. United States (7235)
2. Convenience foods (7225)
3. Fast food restaurants (6795)
4. Food industry and trade (6789)
5. Food habits (444)
6. Cookery, American (50)
7. Diet--Health aspects (34)
8. Book clubs (Discussion groups) (26)

Acknowledgements
The author is grateful to Kerre Kammerer, Jenny Toves, and Harry Wagner for providing statistics on FRBR work sets and FAST headings, and to Roger Thompson for computer programming support. The author also thanks Terry Butterworth, Joan Mitchell, Larry Olszewski, Ed O'Neill, and Melissa Renspie for their assistance in preparing this paper.

References

Chan, Lois Mai, Eric Childress, Rebecca Dean, Edward T. O'Neill, and Diane Vizine-Goetz. 2001. A Faceted Approach to Subject Data in the Dublin Core Metadata Record. *Journal of Internet Cataloging* 4 (1/2): 35-47.

Chan, Lois Mai and Edward T. O'Neill. *n.d. FAST: Faceted Application of Subject Terminology, Principles and Application.* Westport, Connecticut: Libraries Unlimited, forthcoming.

O'Neill, Edward T. and Lois Mai Chan. 2003. "FAST (Faceted Application of Subject Terminology): A Simplified LCSH-Based Vocabulary FAST," paper presented at the World Library and Information Congress (69th IFLA General Conference and Council), Berlin, August 1-9, 2003. http://archive.ifla.org/IV/ifla69/papers/010e-ONeill_Mai-Chan.pdf (accessed July 27, 2009).

Rolla, Peter J. 2009. User Tags versus Subject Headings: Can User-Supplied Data Improve Subject Access to Library Collections? *Library Resources & Technical Services*. 53 (1): 174-184.

Shirky, Clay. "Dynamic Growth of Tag Clouds." Weblog entry. You're It! a blog on tagging. May 24, 2005. http://tagsonomy.com/index.php/dynamic-growth-of-tag-clouds/ (accessed July 27, 2009).

Spiteri, Louise. 2007. "The Use Of Collaborative Tagging In Public Library Catalogs." Final report : 2006 OCLC/ALISE Library and Information Science Research Grant Project, February 26, 2007. http://library.oclc.org/cdm4/item_viewer.php?CISOROOT=/p267701coll27&CISOPTR=271 (accessed July 27, 2009).

Notes

[1] Authority records for chronological headings are only made when they are necessary to create a reference. See O'Neill, Edward T. and Lois Mai Chan for more information on chronological headings.

[2] FAST headings were applied to a copy of OCLC WorldCat current through the end of March 2009.

[3] The algorithm is available for download at http://www.oclc.org/research/projects/frbr/algorithm.htm.

[4] A holding code identifies a library and indicates that the library owns an item. A holding code is associated with a bibliographic record when an item is cataloged.

[5] Rolla (2009, 182) cites **Cookery, American** is an example of a heading that does not accurately describe the subject of a work.