

**Hyerim Cho**, University of Missouri iSchool

**Denice Adkins**, University of Missouri iSchool

**Jenny S. Bossaller**, University of Missouri iSchool

**Heather Moulaison-Sandy**, University of Missouri iSchool

## **What Is Genre? How Do We Talk about Fiction?: Assessing Complementary Schemas and Approaches**

### **Abstract**

Describing fiction is notoriously problematic. One aspect of this is the designation “genre” which can mean different things to different communities involved in the book trade, such as publishers/authors, readers, and information professionals. Using sources created by all three (i.e., ProQuest and *New York Times Book Review*, Library of Congress bibliographic records, GoodReads categories, New York Public Library catalog user-supplied content, and Amazon categories), aboutness and isness metadata are harvested for two novels. Ultimately, we find that although the vocabularies each have different purposes and intended audiences, the compilation of terms from different sources presents a coherent image of a title. Future work in knowledge organization (KO) should consider how best to harness information from multiple sources in providing library users with a more robust perspective on fiction titles.

### **Introduction and Problem Statement**

Fiction classification is problematic; the problems with knowledge organization (KO) and fiction titles has been acknowledged in the KO literature for at least 30 years (e.g., DeZelar-Tiedman 2011; Ranta 1991). The process for identifying facts about a non-fiction title is well established (if not nonetheless arbitrary at times). Indeed, capturing metadata about subjective elements relating to the reading/listening/viewing experience of fiction is notoriously fraught. KO researchers are not alone in their frustration, or their interest in fiction. Fiction classification has likewise been studied for many years by a number of communities outside of KO, including the Readers Advisory (RA) community (e.g., Dali 2014; Ross 2000; Saricks 2005; Scott 1995), as well as those looking specifically at audiovisual (A/V) formats (e.g., video games: Lee, Clarke, Cho, and Windleharth 2017).

One particularly troublesome aspect of classifying fiction is the notion of genre. In KO, genre is typically defined as “categories of works that are characterized by similar plots, themes, settings, situations, and characters” (Library of Congress 2011, 1) with the focus on fiction (c.f., Young and Mandelstam 2013). Not all communities view genre in this same way.<sup>1</sup> Further, although there are viable proposals for automating genre classification (c.f., Gupta, Agarwal, and Jain 2019), these are also limited by the changing nature of popular fiction, specifically changing genre standards and modern tendencies toward genre-blending. Examples include the merging of historical fiction novels and romances, or the merging of science fiction and supernatural. Scot Smith (2007) even claimed that “genre might be dead, that many of recently published novels

---

<sup>1</sup> For example, RA literature consistently considers genre to be a kind of novel-length fiction product, with genre fiction being divided into categories such as romance, Westerns, and historical fiction.

no longer fit into the predictable categories we typically designate for books.” (para. 5) Genre classification, therefore, emerges as being thorny and convoluted, and potentially reflective of the classifier’s perspective more than other approaches to organizing fiction.

Multiple schema have been proposed that provide either intentional or de facto ways to describe genre in fiction, including Library of Congress Genre/Form Terms for Library and Archival Materials (LCGFT), the Book Industry Study Group’s BISAC Subject Headings List (Martínez-Ávila 2016) that are used by online and brick and mortar bookstores, and various individual schemes such as genre color wheels (Scott 1995), faceted classification (Pejtersen 1991), and algorithm development (WhichBook.net 2021). Less formally, Wikipedia divides genre fiction (i.e., novels) into 1) crime, 2) fantasy, 3) romance, 4) science fiction, 5) Western, 6) inspirational, and 7) horror, creating an explicit taxonomy of fiction sub-genres. Social tagging systems, another user-generated system, allows users to create and apply their own tags, which when considered in the aggregate, form a folksonomy (Rafferty 2018; Spiteri and Pecoskie 2016), yet folksonomies are not meant to serve in the organization and retrieval of titles specifically by genre, and none is used across industries. In short, multiple genre schema and terms have been proposed across book industry domains (booksellers, librarians, database repackagers), and even social media users, but all are in effect using different language to describe subjective elements beyond the aboutness of the title.

### **Rationale for this Work and Approach**

In this short paper, we discuss as a work-in-progress our analysis of descriptive terms applied to fiction across the book trade; specifically, we begin by focusing on genre and expand out to the larger, more complex system of KO for fiction. To begin, the *New York Times Book Review* as indexed in ProQuest contains information in a field entitled “Genre” that, upon investigation, is unrelated to the isness of the book review (it is a *book review*, but no term reflecting that is evident in the Genre field). Further, the terms supplied to describe the book reviews actually describe aspects of the *titles* being reviewed, especially the title’s format (e.g., books, poetry) and the title’s aboutness (e.g., Surgeons, Heart). Our curiosity was piqued – in what way do these nontraditional terms to describe genre support a reader’s use of the system, or more specifically, the understanding of the title? How do other approaches throughout the book trade work to support access to genre? To the novel as a whole?

As a point of departure, this paper uses the “Genre” terms assigned to ProQuest records for book reviews for a set of 10,080 reviews and examines the reviews in the aggregate. Within the set, there are 27,468 unique descriptors listed in the Genre field. Table 1 shows the most frequently used terms, which correspond mostly to a title’s form. Terms in this column include genre as described in literary studies (e.g., poetry, essays, novels), as well as genre divisions within mass-market publishing (e.g., romance, mystery, science fiction). This column also includes subject terms (e.g., World War II, refugees, Black culture), serving overall as an amalgamation of many of the elements of fiction that could be of interest to readers.

Table 1. Terms used over 200 times in the “Genre” metadata column of ProQuest to describe the contents of 10,080 NYTBR reviews.

Genre Metadata	Times Used
Books	8,782
Nonfiction	3,625
Novels	2,866
Autobiographies	716
Biographies	645
Short stories	356
Essays	256
Mysteries	232
Children’s literature	211
Children’s picture books	210
Poetry	209

How indicative can these be of the reviews, or of the fiction being described? After review, we identified a subset of two specific fiction titles for which ProQuest supplied genre information (*Dream with Little Angels* and *Disappearance at Devil’s Rock*). Genre and subject terms assigned to these titles were manually harvested across additional sources: MARC subject and genre terms in Library of Congress bibliographic records, user-assigned tags from GoodReads, Amazon categories, and tags available through the BiblioCommons interface of the New York Public Library (NYPL). The results demonstrate the challenges of identifying genre and the strengths of using multiple sources for genre identification in future KO initiatives in libraries.

### Results and Discussion

By drawing from the sources enumerated above, varied perspectives begin to be revealed. Figure 1 presents a word cloud of the individual words used to describe the novel *Dream with Little Angels* by Michael Hiebert. Aggregating the individual descriptors used in each source (ProQuest, LCSH/LCGFT, GoodReads, NYPL, Amazon) reveals some words used more commonly than others, such as Fiction (used 10 times across multiple sources), Mystery (used six times), and Thriller (used four times). Figure 2 shows a similar visualization for the novel *Disappearance at Devil’s Rock* by Paul Tremblay. Figure 2 demonstrates that Fiction is used 11 times, and Horror, Psychological, Mystery, and Books are used four times each.

Figure 1. Word cloud visualization of individual words used in descriptors associated with the novel *Dream with Little Angels*.

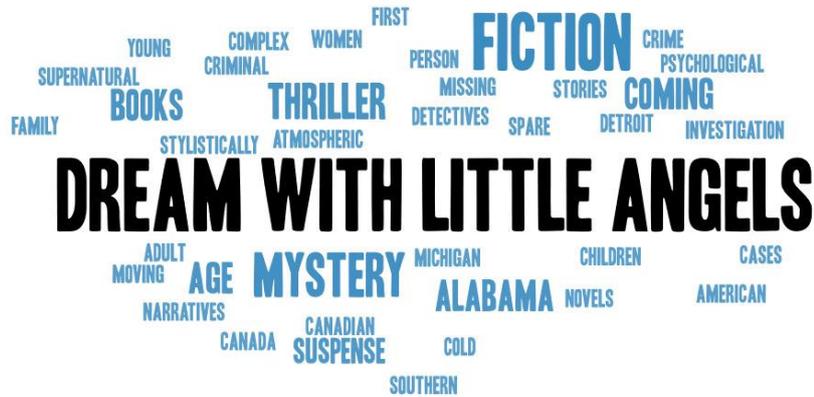


Figure 2. Word cloud visualization of individual words used in descriptors associated with the novel *Disappearance at Devil's Rock*.



Figures 3 and 4 present word clouds of the terms used to describe the novel *Dream with Little Angels* and *Disappearance at Devil's Rock* respectively. Descriptors are color-coded by source; descriptors that came from Library of Congress subject terms are in purple, Library of Congress “Other Subject” terms in red, Library of Congress Genre/Form terms in teal, GoodReads in green, the New York Public Library in blue, from Amazon categories in orange, and ProQuest in gray. For *Dream with Little Angels*, the only repeated terms come from GoodReads, with two instances each of the categories “Mystery,” “Thriller,” and “Mystery Thriller.” For *Disappearance at Devil's Rock*, the only repeated terms were Mystery (3 times) and Thriller (2 times), as assigned in GoodReads.

Figure 3. Word cloud visualization of descriptors associated with the novel *Dream with Little Angels*, color coded by source.



Figure 4. Word cloud visualization of descriptors associated with the novel *Disappearance at Devil's Rock*, color coded by source.



The sources reviewed (LC Subjects, LCGFT, NYTBR, GoodReads, NYPL, and Amazon) show varied approaches to classification which were created and used for their intended applications. For instance, the MARC 650 and 655 fields are intended to be used by libraries, and Amazon categories are intended to support online consumers. However, the content normally used in those fields is insufficient for work with library users. Libraries sometimes incorporate other databases (e.g., the BiblioCommons discovery system which includes user reviews and tags; some link to Amazon or other

booksellers) into their catalogs in order to strengthen retrieval for fiction. The NYPL has adopted appeal factors and read-alikes into the catalog, as well as incorporating metadata from NoveList (a commercial RA database, <https://www.ebsco.com/novelist>), and including reviews and other paratextual elements by Baker & Taylor (<https://www.baker-taylor.com/>). Subject headings, which speak to aboutness, provide information as to topics and settings. Genre terms help to understand isness, or the experience of the book. User-generated content, as from GoodReads or BiblioCommons, also help determine what the experience of reading this book will be like, without the constraints of a controlled vocabulary.

### **Conclusion and Future Work**

Given the wide range of interpretations for describing fiction and for labeling genre, our impression is that combing sources would only support better access to fiction going forward. This is despite the fact that each system is built for its own audience -- each can have a use or a meaning if able to be isolated. For example, ProQuest terms are useful for separating out content about books from the myriad of other content published in the New York Times, but less useful for helping a reader sift through the multitude of book reviews. Library of Congress Subject Headings are primarily used to indicate topics addressed.

The findings ultimately underscore the advantages to combining multiple systems to describe works of fiction. Already, genre terms are mixed and reused when sources cross from one medium to another, such as, from the original publishing source to an aggregator. As an example, in downloading NYTRB from ProQuest, we noted a column called "Genre." (See Table 1.) This column appeared to have been applied by ProQuest indexers, since it was more complete than the metadata supplied by the NYTBR, which was often limited to "Fiction, Nonfiction, Crime," and so forth.

Ultimately, we put forth as a topic of further study the question of how best genre should or could be considered. What conclusions can be drawn about the kinds of information labeled as "genre" in common information systems? Is genre more fluid in the minds of others than it is in library systems? Is the library's insistence on separating aboutness and isness in its controlled vocabularies a bona fide support for retrieval, or rather does it suggest a degree of unwarranted pedantry on the part of libraries? This short paper asks many more questions than it answers, as it begins to isolate some of these questions; further study should be carried out to address these and other questions to support access to fiction, especially in library systems.

### **Acknowledgement**

This research is supported by ProQuest TDM Studio Digital Research Support Program. The research team was given access to TDM Studio software for six months as an award for winning their research proposal competition.

### **References**

- Dali, Karen. 2014. "From Book Appeal to Reading Appeal: Redefining the Concept of Appeal in Readers' Advisory." *The Library Quarterly* 84 no. 1: 22-48.

- DeZelar-Tiedman, Christine. 2011. "Subject Access to Fiction: An Application of the Guidelines." *Library Resources & Technical Services* 40, no. 3: 203–10.  
<https://doi.org/10.5860/lrts.40n3.203>.
- Gupta, Shikha, Mohit Agarwal, and Satbir Jain. 2019, January. "Automated Genre Classification of Books Using Machine Learning and Natural Language Processing." In *2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*: 269-272. IEEE.
- Library of Congress. 2011, June 6 (rev.). "Frequently Asked Questions about Library of Congress Genre/Form Terms for Library and Archival Materials (LCGFT)." Retrieved from [https://www.loc.gov/catdir/cpsol/genre\\_form\\_faq.pdf](https://www.loc.gov/catdir/cpsol/genre_form_faq.pdf)
- Lee, Jin Ha, Rachel Ivy Clarke, Hyerim Cho, and Travis Windleharth. 2017. "Understanding Appeals of Video Games for Reader's Advisory and Recommendation." *Reference & User Services Quarterly* 57 no. 2: 127-139.
- Martínez-Ávila, Daniel. 2016. "BISAC: Book Industry Standards and Communications." *Knowledge Organization*, 43 no. 8: 655-662.  
<https://doi.org/10.5771/0943-7444-2016-8-655>
- Pejtersen, Annelise Mark. 1991. "Icons for Representation of Domain Knowledge Interfaces." *Advances in Knowledge Organization*, 2: 175-193.
- Rafferty, Pauline M. 2018. "Tagging." *Knowledge Organization* 45 no. 6: 500-516.  
<http://www.isko.org/cyclo/tagging>
- Ranta, Judith A. 1991. "The New Literary Scholarship and a Basis for Increased Subject Catalog Access to Imaginative Literature." *Cataloging and Classification Quarterly* 14, no. 1: 3–26.
- Ross, Catherine Sheldrick. 2000. "Making Choices: What Readers Say About Choosing Books to Read for Pleasure." *The Acquisitions Librarian* 13 no. 25: 5–21.
- Scott, Paul. 1995. "Wheels Within Wheels: Genre Fiction in Public Libraries." *Australasian Public Libraries and Information Services* 8 no. E: 191-194.
- Smith, Scot. 2007. "The Death of Genre: Why the Best YA Fiction Often Defies Classification." *ALAN Review* 35 no. 5: 43-50.  
<https://doi.org/10.21061/alan.v35i1.a.5>
- Spiteri, Louise F. and Jen Pecoskie. 2016. "Affective Taxonomies of the Reading Experience: Using User-Generated Reviews for Readers' Advisory." In *Proceedings of the Association for Information Science and Technology* 53, no. 1: 1-9.  
 doi:10.1002/pr2.2016.14505301032
- Young, Janis L. and Yael Mandelstam. 2013. "It Takes a Village: Developing Library of Congress Genre/Form Terms." *Cataloging & Classification Quarterly* 51, nos. 1-3: 6-24.