

## Musings on Faceted Search, Metadata, and Library Discovery Interfaces

Kelley McGrath

To cite this article: Kelley McGrath (2023) Musings on Faceted Search, Metadata, and Library Discovery Interfaces, *Cataloging & Classification Quarterly*, 61:5-6, 439-490, DOI: [10.1080/01639374.2023.2222120](https://doi.org/10.1080/01639374.2023.2222120)

To link to this article: <https://doi.org/10.1080/01639374.2023.2222120>



© 2023 The Author(s). Published with license by Taylor & Francis Group, LLC.



Published online: 21 Jun 2023.



Submit your article to this journal [↗](#)



Article views: 1777



View related articles [↗](#)



View Crossmark data [↗](#)

# Musings on Faceted Search, Metadata, and Library Discovery Interfaces

Kelley McGrath 

University of Oregon Libraries, Eugene, OR, USA

## ABSTRACT

Faceted search is a powerful tool that enables searchers to easily and intuitively take advantage of controlled vocabularies and structured metadata. Faceted search has been widely implemented in library discovery interfaces and has provided many benefits to library users. The effectiveness of facets in library catalogs depends on a complex interaction between facet vocabularies, metadata quality and structure, and the library discovery interface's capabilities. This article provides a holistic overview of challenges for optimally implementing facets in library catalogs. This supports a systematic approach to refining and enhancing the capacity of faceted search to improve searching and exploring bibliographic metadata.

## ARTICLE HISTORY

Received March 2023

Revised May 2023

Accepted May 2023

## KEYWORDS

Faceted search; faceted vocabularies; metadata quality; library discovery interfaces; library catalogs; usability issues

## Overview

Faceted search is a powerful tool that enables searchers to easily and intuitively take advantage of controlled vocabularies and structured metadata. It can make searches more efficient, help users better understand options, and guide users who are browsing or exploring. It has been widely implemented in library discovery interfaces and has provided many benefits to library users. Tunkelang says that the essence of faceted search is the combination of text search of unstructured content and faceted navigation of structured content organized into component facets.<sup>1</sup> Facets are familiar to most users as they frequently encounter them on commercial websites. The Nielsen Norman Group notes the ubiquity of facets on contemporary retail websites and states that many users are frustrated when they are absent. They point out that only the largest ecommerce websites with the most diverse products, such as Amazon and Walmart, retain formerly common alternatives like scoped search and advanced search in addition

**CONTACT** Kelley McGrath  [kelleym@uoregon.edu](mailto:kelleym@uoregon.edu)  University of Oregon Libraries, Eugene, OR 97403, USA.  
© 2023 The Author(s). Published with license by Taylor & Francis Group, LLC.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

to facets.<sup>2</sup> As will be discussed below, libraries also have very large and diverse inventories.

When implemented well, faceted search enables users to more quickly and accurately perform successful searches. It increases both precision and recall. Faceted search is ubiquitous in contemporary library discovery interfaces, but there remain challenges for optimizing the implementation of faceted search in library catalogs. These include scale, computational constraints, user interface design, metadata quality and complexity, the diversity of library resources, and the theoretical and practical impossibility of completely populating consistent, accurate facets. The effectiveness of facets in library catalogs depends on a complex interaction between facet vocabularies, metadata quality and structure, and the library discovery interface's capabilities.

There are two main areas of difficulty. The first relates to systems design. This can further be subdivided into problems stemming from technical challenges and computational limits, and the challenges of designing user interfaces to present complex data. The second relates to the data that populates the facets. This includes characteristics of bibliographic metadata, such as the complexity and heterogeneity of information that libraries attempt to record about bibliographic resources. It also includes the design of controlled vocabularies and the definitions and data structures used for recording bibliographic information. Limits on the time, funding, and expertise available to populate bibliographic metadata, as well as theoretical and practical limits on the data it is possible to record, also present challenges. System design issues and data issues are deeply intertwined and interact in complex ways. They cannot be considered in isolation. For example, some of the shortcomings of using the Library of Congress Subject Headings (LCSH) for faceted navigation are due to the fact that it was designed to work in the card catalog environment where users access the subject headings only through a left-anchored, alphabetical list. It is not possible to evaluate the effectiveness of a controlled vocabulary or metadata structure independently of the technology that is used to interact with it.

This article provides a holistic overview of challenges for optimally implementing facets in library catalogs. It also includes illustrative examples and case studies, covering topics such as the complexities of dates and musical medium of performance, as well as challenges for using LCSH to populate topical subject facets. This overview is intended to support a systematic approach to refining and enhancing the capacity of faceted search to improve the process of searching and exploring bibliographic metadata. It is hoped that this exploration will generate ideas and encourage experimentation that will enable the development of more powerful and easier to use library discovery interfaces.

## **The many benefits of facets and faceted navigation**

Faceted search is a flexible, powerful, and user-friendly way to explore result sets. It enables users to choose limiters that are relevant to their initial search. Facets support an interactive experience where users have the opportunity to progressively refine their search while avoiding zero result sets. The ability to independently combine facets at whim lets users intuitively and effectively explore result sets to find the most pertinent resources. Facets also provide an overview of the search space and the resources that may potentially be relevant.

## **Some challenges for implementing facets and faceted navigation in library discovery interfaces**

When libraries moved from card catalogs to online public access catalogs (OPACs), the functionality of the catalog was improved in many ways. In particular, keyword search provides access to many parts of the catalog record, such as free text notes, that could not be searched in a card catalog. However, many of the capabilities of the card catalog that supported browsing and provided context have never been successfully reproduced in online catalogs. From the early days of computerized library catalogs, there have been calls for modifying OPAC functionality and adjusting cataloging practice to improve the discovery process. In 2023, Coyle could still note that contemporary library discovery interfaces do not provide as much context as card catalogs did nor do they exhibit the “modicum of conversation” or back-and-forth that a card catalog could.<sup>3</sup> There are two main ways that library catalog functionality could be improved. The first is to change the system functionality, so that it works better with the existing data. Massicotte represents one of many calls to rethink the way that we present subject headings to users.<sup>4</sup> The second is to change our controlled vocabularies or cataloging practices to better utilize the affordances of computerized interfaces. For example, Nahotko advocates for new knowledge organization structures that take better advantage of features of digital interfaces, including faceted navigation.<sup>5</sup> Both approaches have potential for improving the usability of facets in library catalogs and thus the overall functionality of our discovery interfaces.

### **Scale**

Maintaining accurate, consistent, and completely-populated facet values and presenting them effectively at scale can require significant resources. Tunkelang lists three aspects to consider when scaling faceted search: 1) number of documents; 2) number of facet values per document; and 3) searchable text

per document.<sup>6</sup> Although the number of characters and amount of searchable metadata in most MARC bibliographic records is relatively small, most library catalogs contain a large number of records, each of which potentially has many facet values associated with it. A search interface with a large number of records combined with a large number of facets creates user interface design challenges. It also requires significant effort to create and maintain vocabularies and to populate records with accurate metadata to support facets. Finally, it comes with high computational costs.

It is challenging to present large numbers of complex facets in a user interface that is still easy to comprehend and utilize. The Norman Nielsen Group says that there is a tradeoff where the “extra power of faceted navigation also adds interaction cost by presenting users with more options to comprehend and manipulate.”<sup>7</sup> No arrangement of a long list of possible facets will be optimal for all users.

Vocabulary maintenance and metadata creation and quality control to support faceted navigation consume significant financial and human resources. Many libraries found that when they introduced faceted navigation in their OPACs and discovery interfaces, it exposed much inconsistent, incorrect, and missing metadata that required remediation. Over the past fifteen years, the Library of Congress has developed several new faceted vocabularies, such as the Library of Congress Genre/Form Terms for Library and Archival Materials (LCGFT),<sup>8</sup> the Library of Congress Medium of Performance Terms for Music (LCMPT),<sup>9</sup> and the Library of Congress Demographic Group Terms (LCDGT).<sup>10</sup> Multiple MARC fields have been added or expanded that can support structured data from these vocabularies or other sources and that can be used to populate facets. Although this expands the possibilities for using faceted search in library discovery interfaces, it also requires catalogers to take the time to add these values to new records. In addition, to improve recall, there must be some sufficiently accurate and scalable method for populating existing records with relevant facet values.

Even with quality metadata and the best possible user interface design, there are challenges for implementing facets on a large scale. As Tunkelang explains, facets make the presentation of search results much more computationally demanding. In a faceted search system, query processing consists of two steps. The first step of retrieving the set of records that matches the query is common to information retrieval systems and can be performed relatively efficiently. It takes significantly more computational resources to identify the associated facets. Many faceted search interfaces also compute the count associated with each facet value, which increases the load even further.<sup>11</sup> This constraint has led some major library discovery interfaces to provide incomplete recall both in terms of the number of facet values presented to users and in terms of the percentage of records that are used to generate the associated count.

## ***User interface design***

There are many challenges for populating and presenting facets in a way that is easy to understand and use. Some user interface design decisions are unrelated to the metadata used for faceting, such as whether the facets should be placed on the left, the right, or the top. However, in many cases, design decisions interact with the available metadata and users' anticipated needs. Such decisions include what facets to provide, how many facet values to display, and whether and how to implement Boolean search strategies.

### ***Number and choice of facets***

Tunkelang discusses this in terms of information overload and users' scarce attention. He states that both the number of different facets and the number of facet values in a given facet have the potential to overwhelm users.<sup>12</sup> There are many different facets that could potentially be useful to some subset of library users and many facets generated from library metadata have long lists of potential values.

Tunkelang suggests three considerations to help decide which facets to display.<sup>13</sup> The first is to favor displaying facets that have what he calls high coverage, i.e., there are values associated with the facet for most or all records in the collection. In a library context, the vast majority of records have some sort of date of publication coded in the 008 fixed field, so this is an example of a facet with high coverage. In contrast, only a small number of records have the date of creation of the work coded in the 046 field. His second recommendation is to favor facets that produce what he calls a "high-entropy distribution of values" in the result set. By this he means that the facet values are more evenly distributed rather than lopsidedly clustering around a single value or few values. In the context of many libraries in the U.S., language would be a low information facet because the vast majority of materials are in English. The values in a date of publication facet would likely have a more balanced distribution. Finally, he recommends consolidating facets with similar or overlapping values, particularly if the distinctions are sometimes made arbitrarily. He gives the example of merging authors, editors, and other contributors into a single facet, which is the approach commonly taken in library discovery interfaces.

Tunkelang more recently suggested criteria for selecting facets to display dynamically in response to a query.<sup>14</sup> The first is popularity, defined as the facets most often wanted by users doing the same or a similar search. On an existing site this can be measured, but is also influenced by the current presentation of the facets and cannot necessarily be obtained for less common searches. Coverage is mentioned again and also utility, which is defined as having a meaningful effect on the search results.

Tunkelang also discusses several ways to mitigate information overload for facets that generate a high number of values.<sup>15</sup> His first suggestion is to exploit hierarchy when possible. Some types of facets lend themselves to this approach more than others. For example, a geographic facet might allow users to navigate between the continental, country, state, and city levels. However, introducing too many layers of hierarchy creates its own usability problems. It also makes maintenance of the facet vocabulary more complicated, as it is necessary to keep track not just of the terms, but also the hierarchy. Some potential vocabularies, such as LCSH, do not have an existing hierarchy that is fit for this purpose and it is not easy to retrofit a large vocabulary with a suitable hierarchy. Lacking an existing hierarchy, Tunkelang proposes that facets could either be grouped arbitrarily (e.g., alphabetically A-D, E-H) or statistically by clustering similar values.<sup>16</sup> An approach taken in many library catalogs is to limit the facet values shown to the ones that are most common in a given result set. For example, Ex Libris's Primo Back Office presents the top 20–50 values for each dynamic facet. The facet values displayed “are derived from the values stored in the Facets section of the 2,000 top-ranked PNX records in the search results. Once the system determines which values to display for each category, it will count the matching records from the first 50,000 results per slice and display the count next to the facet value.”<sup>17</sup> Alternatively, Tunkelang says it might be possible to only show the facet values that occur more frequently in the result set than in the overall collection.<sup>18</sup>

The labeling of facets may also be challenging. Some facets may contain similar values and be difficult to differentiate in a way that is easily understood by the casual user. For example, a language facet could combine all languages associated with the manifestation into a single language facet as is typically done. Alternatively, multiple types of languages could be distinguished, such as written languages, spoken or sung languages, subtitle languages, and caption languages. A separate facet could also be created for the original language of the work.

There may be a disconnect between labels used in library catalogs and user vocabulary. It is also not clear that users think of some categories, such as genres, in the same way as librarians tend to. Crowdsourced sites often to use tags that cover a multiplicity of uses without distinction or in overlapping ways. For example, Upton describes the tag “dark romance” as more of a “content warning.”<sup>19</sup>

### *Ordering of facet values*

In addition to decisions about how many facets and facet values to display, decisions must be made about how to order the values within each facet. There are three main approaches. One is to present the facet values in a fixed order, commonly alphabetical order.

**MESH subjects**

Hypertension (27)

Hypertension–therapy (19)

Hypertension–complications (9)

Hypertension, Renal (10)

Kidney Diseases (73)

Kidney Diseases–complications (13)

Kidney Diseases–diagnosis (11)

Kidney Diseases–etiology (8)

Kidney Diseases–pathology (8)

Kidney Diseases–physiopathology (9)

Kidney Diseases–therapy (24)

**Figure 1.** Facets for MeSH shown in alphabetical order.

**MESH subjects**

Kidney Diseases(73)

Hypertension(27)

Kidney Diseases–therapy(24)

Hypertension–therapy(19)

Kidney Diseases–complications(13)

Kidney Diseases–diagnosis(11)

Hypertension, Renal(10)

Hypertension–complications(9)

Nephrology(9)

Kidney Diseases–physiopathology(9)

Urologic Diseases(8)

Kidney Diseases–pathology(8)

Kidney Diseases–etiology(8)

**Figure 2.** Facets for MeSH shown in ranked order.




An alternative is to present the facets in a ranked order where the values that are most common in the result set are shown first. Finally, a hierarchical or nested approach may be used to group facet values. The optimal decision depends on the facet values and user needs and is subject to system constraints.

For example, a topical Medical Subject Heading (MeSH) facet might be easier to scan if it were organized alphabetically (Figure 1), but if only the top five facet values are displayed by default, it will potentially hide or deemphasize the most frequent values that would be highlighted in ranked results (Figure 2). Even if the default order is optimal for most use cases, it is desirable to allow users to toggle between these two options. In this particular case, a nested presentation with the top-level terms in ranked order might be better than either of the options shown, but this option is not supported by the discovery system used.

### **Boolean operators**


Some use cases for facets call for more sophisticated strategies that emulate the Boolean operators AND, OR, and NOT. These are more difficult to present to users in an easily understood way. Ex Libris's Primo supports all three in some situations. Selection from different facets always performs an AND search. Sequential selection of values from a single facet by clicking hyperlinked terms performs an AND search. If a user searches for "biography" and then clicks the topical facet value for "United States," the user gets just the biographies that have a subject term for United States. At this point, the user can then look at the topical facet again and select another subject, such as "presidents." They might then select an additional facet, such as "generals," which will produce a list of resources about American presidents who were also generals (Figure 3). On the other hand, simultaneous selection of values from a single facet using checkboxes performs an OR search (Figure 4). If the user has searched for biography and selected the topical facet "United States" as before and then selects both "presidents" and "generals" simultaneously using the checkboxes, they will get biographies about either presidents or generals or people who were both (Figure 5).

Primo follows good user design in that all of the facets exhibit the same behavior. Nevertheless, as Tunkelang points out, "Users are notoriously bad at inferring Boolean logic from subtle cues."<sup>20</sup> There are a couple of situations where library metadata exacerbates this challenge. One is the heterogeneity of terms that are included in some library facets. A notorious example is the common implementation of topical subject facets where different types of topical aspects are intermixed in a single facet list. Returning to the biography example, a user might naïvely select three of the first four facet values presented, "1900–1999," "United States," and "politicians."

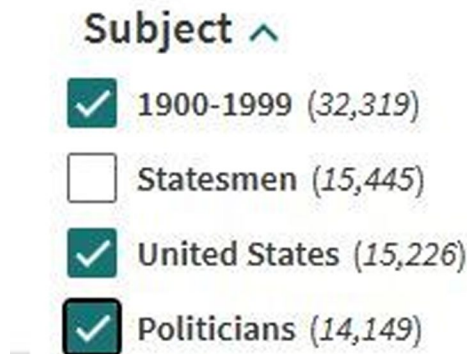


1  PRINT BOOK  
**Young Washington : a selection from George Washington, a biography**  
 Freeman, Douglas Southall, 1886-1953.  
 1966  
[Check for available services >](#)

2  PRINT BOOK  
**Biography of Andrew Jackson : president of the United States, formerly major general in the Army of the United States**  
 Goodwin, Philo A. (Philo Ashley), 1807-1873, author.  
 1832  
[Check for available services >](#)

3  PRINT BOOK  
**Lee and Grant, a dual biography**  
 Smith, Gene.  
 ©1984  
[Check for available services >](#)

**Figure 3.** Results of a search for biographies narrowed sequentially to the topic facet for United States, presidents and generals to retrieve biographies of Americans who were both presidents and generals.



**Subject** ^

1900-1999 (32,319)

Statesmen (15,445)

United States (15,226)

Politicians (14,149)

**Figure 4.** Multi-select option in Primo topical subject facet.

The user might expect to get biographies of twentieth century American politicians, but what they will actually get are all the biographies of twentieth century persons plus all the biographies of Americans plus all the biographies of politicians with only a subset exhibiting all three characteristics. This can potentially be remedied by splitting the topical facet into subcategories, such as topic, place, and time period, at the cost of increasing information load by presenting a larger number of facets. It is also not possible to split these categories cleanly in some vocabularies, such as LCSH. For example, LCSH marks many named chronological periods, such as “Middle Ages,” in a way that is indistinguishable from standard topics.

- 1



PRINT BOOK  
**Lyndon, an oral biography**  
Miller, Merle, 1919-1986.  
©1980  
[Check for available services >](#)



**Figure 6.** Boolean NOT option in Primo resource type facet.

Tunkelang proposes the heuristic that “facets that are typically singly assigned to documents (e.g., brand, document type) work well with disjunctive [OR] selection, whereas facets that are often multiply assigned to documents (e.g., consumer electronics features, topic) work well with conjunctive [AND] selection.” However, he simultaneously recommends consistent behavior within a given search interface, which is impossible to align with facets that individually benefit most from inconsistent behavior.<sup>21</sup> In an ideal interface, users would be able to intuitively choose to use AND or OR to combine facet values in the way that best supports their information need, but this would be complex to communicate.

Ex Libris’s Primo also supports the Boolean NOT operator in the form of a red checkbox with a slash through it that appears on mouseover (Figure 6). This functionality is not common on commercial websites and anecdotally public services librarians say that users have to be taught to use this function. It is particularly useful to remove unwanted formats, such as microforms or government documents. It works less well in situations where the facet value being removed applies to only part of the resource. For example, if a user is searching for fugues, but wants something new and decides to NOT out Bach from a creator facet, they will simultaneously remove any fugues by other composers that are part of compilations that include a piece by Bach.

### ***Exploratory search and search-free browsing***

In library catalogs and databases, a distinction is often made between known-item searches and other types of searches, such as subject searches. These other types of searches can be grouped in the overlapping categories of browsing and exploratory search. Facets can support users seeking

known items by helping them narrow large result sets to the item sought. This is especially useful when the query consists of short, generic search strings, such as “nature” for the journal *Nature*, or very commonly occurring terms, such as when a user is seeking a specific performance of Beethoven’s Ninth Symphony. However, faceted search has even greater potential to improve the browsing and exploratory search process.

Browsing and exploratory search encompass a variety of use cases ranging from users seeking a particular type of resource, such as recent fantasy novels or textbooks for learning Python, to users exploring an information space that is new to them. McKay, Buchanan, and Chang list a number of use cases for browsing, including “meeting loosely specified information needs, meeting well specified but hard to describe needs with recognition strategies, ... refining information needs in view of constructing a query, ... ‘social’ information seeking, and ... serendipitous discovery.”<sup>22</sup> Kules and Capra state that “uncertainty, ambiguity and discovery” are common characteristics of exploratory tasks.<sup>23</sup>

McKay, Buchanan, and Chang conducted a user study to try to empirically determine requirements for effective online browsing systems. They advocate for “purpose-built systems designed to facilitate serendipity and browsing” and emphasize the importance of search-free browsing.<sup>24</sup>

A faceted interface that works well for both searching and browsing could present the user with an initial screen that features facets as well as a search box. It should also enable a user who has performed a keyword search and then selected relevant facets to remove their search terms, so that they can see all the resources that fall into the category or categories that they selected.

Implementation of search-free browsing in library discovery interfaces faces several significant challenges, particularly if the discovery interface aggregates large amounts of metadata from many sources. These obstacles include computational constraints on scaling, heterogeneous facet values that are not consistently present or do not conform to a standardized vocabulary, either because the values come from nonstandardized sources like author keywords or because they come from multiple conflicting vocabularies that cover the same concepts, and the challenge of providing users with large numbers of options without overwhelming them.

Library resources are heterogeneous, so it may be difficult to pick facets to display on the initial page of a search-free browsing interface. McGrath points out that LCSH has far too many top-level terms to be useful for topical browsing in this fashion, but classification schemes are more promising.<sup>25</sup> Search-free browsing interfaces may work better for smaller, clearly defined subsets of library resources. In 2018 McKay, Buchanan, and Chang published a review of two interfaces that offer search-free browsing: WhichBook, which covers fiction and poetry, and BookFish for young

readers.<sup>26</sup> In 2011 McGrath, Kules, and Fitzpatrick developed a prototype for moving images.<sup>27</sup>

### ***Complexity and multiple entity types***

One of the most significant difficulties for successfully implementing faceted search in library discovery interfaces is the scale and complexity of library metadata. Several structural features of library data make it challenging to implement easily understood facets that produce precise, accurate results. These include what Tunkelang describes as “multiple entity types,”<sup>28</sup> such as aggregates or compilations and the multiple levels described by the Functional Requirements for Bibliographic Records (FRBR)<sup>29</sup> and the IFLA Library Reference Model (LRM),<sup>30</sup> which are popularly known as the WEMI (work, expression, manifestation, item) stack. WEMI leads to superficially similar sets of facet values with different meanings, which are either conflated in a single facet list or produce a longer list of similar facets that must have easily distinguished and interpretable labels. Multiple entity types may also cause unexpected and undesirable results when certain combinations of facets are selected.

### ***Multiple entity types***

Tunkelang says that faceted search interfaces are normally based on sets of records for a single type of entity, such as books, which are associated with facets. If a second type of entity, such as authors, is introduced and multiple instances of the new entity (authors) can be associated with a single instance of the original entity (books), it can become problematic. The difficulty arises if one wants to provide access to facets associated with the second entity, such as the nationality or institutional affiliation of the authors.<sup>31</sup> Unless one has a way to make sure that the facets related to authors are forced to apply to the same author, faceting for Canadian authors associated with the University of Oregon may also bring back results where the facet values apply to different authors. For example, the result set might include a paper coauthored by an American working at the University of Oregon and a Canadian working at the University of Washington. It is much harder and more complex to design a search interface that provides effective faceting for multiple entity types. This is further complicated by the fact that some attributes of creators of bibliographic resources are not static and are associated only with a subset of that creator's works. For example, if a user is looking for symphonies by children, they do not want all the Mozart symphonies, only those that he wrote during his childhood.

### **Aggregates**

Aggregates, such as compilations or anthologies, suffer from a version of the multiple entity type problem. The resource as a whole has a certain set of attributes, such as an editor and date of publication, which can be used as facets. The individual works included in the anthology have a different set of attributes, such as author and date of creation, which can also potentially be used as facets. Because each anthology contains multiple individual works and those works have multiple characteristics, users may get misleading results. The Music Library Association's Music Discovery Requirements gives the example of a user searching for Beethoven symphonies and getting a CD that contains a Beethoven overture and a Mozart symphony, but no Beethoven symphony.<sup>32</sup>

Aggregates are not an easy problem to solve in our current flat MARC record environment and the current push for indefinite roundtripping between MARC and BIBFRAME prevents any near-term solution. A subfield for "materials specified" (\$3) has been introduced for many MARC fields. This allows catalogers to include a free text label for metadata that applies to only part of the resource. It may be helpful for humans looking at a catalog record, but it is not suitable for machine manipulation. The values are not standardized and are prone to typographical errors. The MARC 21 format does include a subfield that is intended to support this type of linking, the "field link and sequence number" (\$8). Over two decades ago, McBride advocated for the use of linking subfields to improve access to music,<sup>33</sup> but no systems that make input easy or that use these linking fields for discovery have been developed since then. Even if tools and models are developed that allow catalogers to accurately associate metadata with the appropriate works within a compilation and also enable discovery systems to present this data usefully, the immense corpus of existing records remains an obstacle to accurate retrieval as any automated remediation would be extremely challenging and error prone. McGrath has referred to this as the "Humpty Dumpty problem."<sup>34</sup> All the pieces may be present in the record, but it is hard to imagine how they can all be linked up again without manual human intervention.

### ***WEMI (work, expression, manifestation, item)***

There is yet another multiple entity lurking within bibliographic metadata. The IFLA Library Reference Model,<sup>35</sup> building on the Functional Requirements for Bibliographic Records or FRBR<sup>36</sup> describes four entities related to bibliographic resources: work, expression, manifestation, and item (WEMI). These are defined as follows.

- Work: “The intellectual or artistic content of a distinct creation.”
- Expression: “A distinct combination of signs conveying intellectual or artistic content.”
- Manifestation: “A set of all carriers that are assumed to share the same characteristics as to intellectual or artistic content and aspects of physical form. That set is defined by both the overall content and the production plan for its carrier or carriers.”
- Item: “An object or objects carrying signs intended to convey intellectual or artistic content.”<sup>37</sup>

All of these entities have multiple attributes which end users may be interested in and which could be presented as facets in a library discovery interface. For example, each entity has a date of creation, as well as one or more creators. As mentioned above, those creators also have relevant attributes. While troubleshooting some facets, the author once encountered a book that came up under a search for poetry with the creator demographic facet limited to African Americans and the original language facet limited to old French. This seems an unlikely combination, but the book contained a translation authored by an African American (an expression) of a text originally in old French (the work).

This is further complicated by the fact that not all bibliographic models use all of the entities separately as described above and some introduce additional entities. *Resource Description & Access* (RDA) uses all of the LRM WEMI entities.<sup>38</sup> BIBFRAME, the proposed successor to MARC, posits somewhat different entities: Hub (which has no exact WEMI equivalent), Work (which includes both the LRM work and expression entities), Instance (or LRM manifestation), and Item.<sup>39</sup> Share-VDE’s BIBFRAME implementation replaces Hub with a different entity called Opus.<sup>40</sup> Other variations are conceivable, such as a combination of entities for Work plus Representative expression, Expression plus Manifestation and, where needed, a separate Expression entity for bundled expression attributes, as used in the prototype moving image search interface described in a 2011 Joint Conference on Digital Libraries presentation.<sup>41</sup> In the MARC bibliographic records used by most current systems, all of these types of information are stored in flat records. Reconciling all these different perspectives is difficult. Coyle notes that the disjoint nature of the LRM WEMI entities “makes it difficult to combine metadata using different entity definitions, such as the difference between BIBFRAME’s work-instance-item and RDA’s work-expression-manifestation-item.”<sup>42</sup> She advocates for “a minimally constrained set of classes and relationships that could form the basis for a useful model of created works” to help mitigate this clash in worldviews.<sup>43</sup>



### **Dates: a case study in complexity**

Dates are a good example of the challenges of presenting facets for complex information in a useful, comprehensible way. Dates may be related to all three of the categories of multiple entity types described above. In addition to dates associated with the whole resource described by the record, there are multiple dates potentially associated with the parts of the resource, with the WEMI entities related to the resource and its parts and with the agents related to the resource and its parts. It can be challenging to identify accurate, consistent date information. Some date information is impossible for catalogers to know. Some dates are impossible or impractical to ascertain with a level of specificity that aligns with the majority of library metadata. It is difficult to encode the complexities of dates in machine interpretable form and it is also difficult to unpack that data again to interpret it for discovery interfaces. The gradual, undirected evolution of the MARC 21 format in response to the needs of the moment has created additional difficulties for interpreting date data in MARC records. Finally, developing a user interface design that presents all of the date information associated with bibliographic resources in a way that is easy to understand and use poses its own challenges.

First of all, there are many types of dates that can potentially be associated with bibliographic resources, many of which involve multiple entity types. For example, different dates may be associated with the different levels of the WEMI stack. The most commonly used date facet in library catalogs is the date of publication or creation of the manifestation. Manifestation date is also the date that is most reliably encoded in machine-processable form in existing bibliographic records. A common use case for faceting on manifestation date would be to find the most recently published materials on some topic. However, for many date-related use cases, the date of the work more accurately embodies what users are seeking. A reprint of an older book may have a recent publication date, but will not meet a user's need for up-to-date information. Most users are interested in the date of release of a movie not of the DVD. Users may also be interested in the date of an expression. For example, a user might want a recent annotated edition of Shakespeare or be interested in early recordings of performances of Verdi operas. Expression dates are further complicated by the potential for multiple layers. An annotated edition of an English translation of Homer's *Iliad* is potentially associated with the date of the translation and the date of the annotated edition, both of which may be different from each other and earlier than the date of publication of the manifestation owned by the library. A recording of a performance of an arrangement for orchestra of a piano piece is associated with both a date of arrangement and a date of performance. There

may also be cases where users are interested in the date of production of a particular item that belongs to a particular manifestation.

In both MARC bibliographic and authority records, the date of the work or expression may be encoded in machine-actionable form in the 046 (special coded dates) field. These dates may also be encoded in textual form in the 388 (time period of creation) field. The 046 field is an excellent example of a jerry-rigged, Rube Goldbergesque MARC field. It has been modified and expanded over time to deal with various use cases and to accommodate an increasing variety of dates in increasingly nuanced ways. The history of the 046 field and attempts to use it to record dates for works and expressions is discussed in more detail in the appendix.

The net result of this agglomerative approach to developing metadata schemas is that the 046 field does not always clearly answer the question: what are the creation date or dates of the work or works contained in this book or other resource? It is not structured to answer this question and this situation is further complicated because this data has been recorded according to different standards at different times. Therefore, it is difficult to create a facet with values that are clearly defined and mean the same thing. If users are most interested in the dates of creation of the works within resources rather than the dates of aggregation or compilation, existing MARC metadata does not easily meet that user need due to a combination of inconsistent metadata and the need for complex logic to isolate the correct date or dates.

The challenges of using dates from the 046 field are compounded by the use of the Extended Date/Time Format (EDTF), a complex standard capable of conveying detailed information about approximate dates, date ranges, and degrees of certainty.<sup>44</sup> This data can be difficult to translate into a clear form for display and even more difficult to integrate into a list of facet values for dates.

Beyond these structural and technical problems that are theoretically resolvable, there are more intractable problems surrounding metadata quality and type. An ideal date facet would have complete coverage with accurate exact dates. These dates should be determined in a consistent manner, such that different catalogers will independently arrive at the same value. The dates should also be at the same level of specificity. In library discovery interfaces, dates in facets are most commonly specified at the level of the year.

Facets work best when precise values can be known and supplied for all the resources being described. This is much easier to achieve in most retail applications than in library catalogs. For example, there are many cases when a precise date is not known or is impractical to determine. Particularly for older titles, this may be because the information has been lost. For certain types of works, such as works that began as oral literature,

it is not clear that an exact date of creation even makes sense. For some works, it is only possible to determine an approximate date. For short time spans, such as a play written in 1667 or 1668, it is reasonable just to facet on both dates. Longer date spans could potentially be handled by creating a value in the date facet for each year of the time span. However, this creates usability and technical challenges. It is more difficult to see how to effectively incorporate such assertions as “written in or before 1984,” which are also permitted in the 046 field. It is also not clear whether and under what circumstances a value of “unknown” might be usefully included in a date facet.

Even when research could determine specific dates, the time and effort involved may be prohibitive. This is particularly true for aggregates, such as *The Norton Anthology of Poetry*, which may contain a very large number of aggregated works. In order to provide accurate information, the cataloger would have to determine and record the dates of creation of all the individual works within the anthology. Unfortunately, this information may be difficult to ascertain and time-consuming to record. Alternatives to recording precise dates include recording a date range or a named time period. For some resources, particularly literary anthologies, the creation dates of the aggregated works may be referred to only by a named time span, such as the Renaissance or the Middle Ages. The 2014 MARC proposal for the creation of field 388 (Time Period of Creation) listed five situations where it could be helpful to record a named time span.

- Historical or cultural periods are often difficult to date exactly, and specific dates may differ from location to location (e.g., the Renaissance has no exact beginning and ending dates and began earlier in Italy than in other parts of Europe).
- There is overlap in time periods (e.g., the dates of the late Middle Ages, the Renaissance, and the early modern period overlap and more information is needed for context).
- The significance of date spans differs from place to place or culture to culture (e.g., Middle Ages in Europe, Song dynasty in China).
- It is not possible to precisely date the creation of some works (e.g., *Beowulf*, the *Iliad*).
- Editors and publishers of aggregate works are often intentionally vague (e.g., it might be difficult and time-consuming to identify exact dates of creation for an anthology of fiction written during the period of World War I broadly defined to include the buildup to the war and postwar reconstruction).<sup>45</sup>

The difficulty with recording only a named time span is that it is not easy to see how to integrate those periods into the same facet as numerical

dates. Alternatively, creating two facets is likely to be confusing to users and both facets will suffer from a lack of recall.

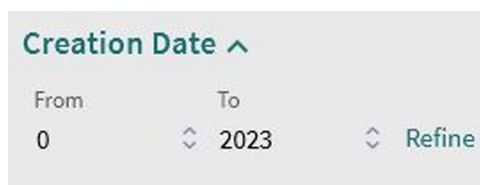
Catalogers could also choose to record a numerical date range that is approximately associated with the named time span or the dates covered by an aggregate. This approach comes with its own drawbacks and potential to produce misleading results. The fundamental problem is that for a date facet that is populated by years, there is no optimal way to map date ranges where not all dates in the range correspond to work creation dates. An anthology of 1950s science fiction stories that includes stories from every year in the decade, could accurately be mapped to each year of the decade (e.g., 1950, 1951, etc.). However, if an anthology that says it is a collection of twentieth century science fiction short stories does not include any stories from the 1910s, providing facet values for every year or even every decade of the twentieth century would not produce accurate results. A book that calls itself an anthology of Elizabethan drama might be given the date range of 1500–1600 based on the Library of Congress subject heading “English drama \$ Early modern and Elizabethan, 1500–1600.” However, if the contents were only from the late sixteenth century, a user selecting 1400–1550 would not actually want this book.

Depending on the system-supplied tools, it may also be difficult or impossible to map all the dates in a date range to individual years. For example, Ex Libris’s Primo has the ability to perform simple transformations of metadata values, including the use of regular expressions. This has enabled the Orbis Cascade Alliance to create an original date facet populated by date ranges such as decades or centuries (Figure 7). However, Primo’s tools do not reproduce the capabilities of a true programming language with features such as mathematical functions and loops. This makes it difficult or impossible to map date ranges to separate values for all the individual years within a range.

In addition to lack of information or imprecise information, there may also be multiple ways of defining and calculating a particular date, such as the date of a work. For a film, the date of production, date of original release, and copyright date could all potentially be used as the date associated with a work. There may be variation in what date is recorded based on what information exists, what information is available to the cataloger, and what sources are preferred. In film reference sources, it is not uncommon to see a one-year discrepancy in the date associated with a title, which is presumably based on the method used to calculate the date. There are less common situations where the definition used can have a significant impact. For example, Eisenstein’s film *Ivan the Terrible, Part II* was completed in 1946, but not released until 1958.<sup>46</sup>

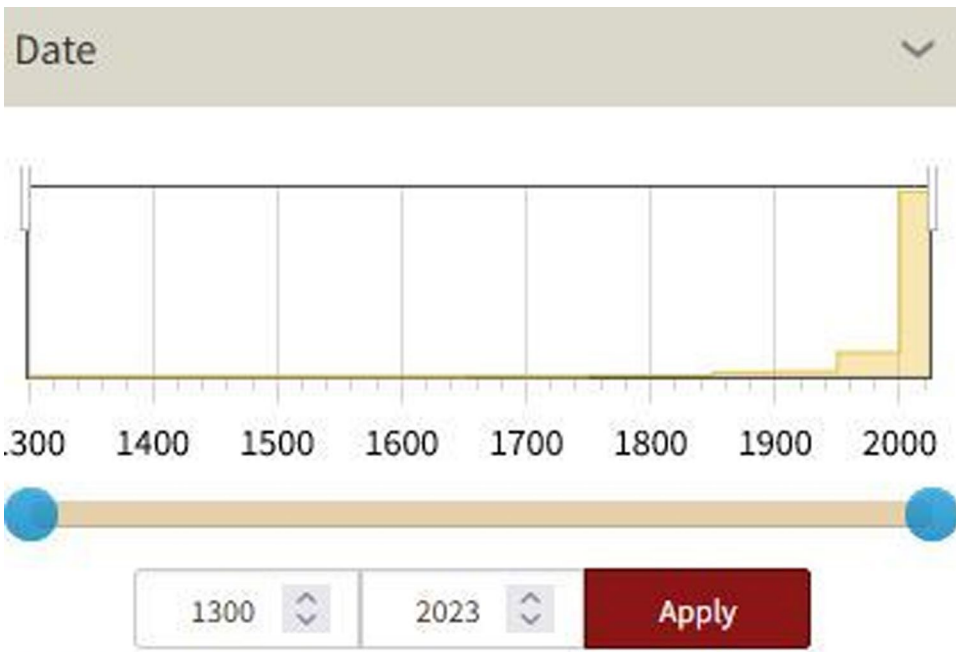


**Figure 7.** Orbis Cascade Alliance original date facet featuring predefined date ranges.



**Figure 8.** Example of Primo tool for manually entering date of publication ranges.

Finally, there are user interface considerations. There are two main ways that dates can be presented to users. Users may be asked to select the start and end date of a date range. This can be implemented either by having users manually type in their desired beginning and end dates (Figure 8) or by providing some sort of slider or other widget for selecting a customized date range (Figure 9). Alternatively, dates or date ranges may be presented as links for predefined ranges (Figures 7 and 10). The amount of granularity supported by a facet depends on the range of values in the data. Library metadata generally specifies dates at the level of years or sometimes in terms of a range of years. Because a long list of years would be unwieldy, facets are usually presented as ranges of years. Some interfaces combine these approaches. For example, WorldCat provides predefined date ranges for recent time spans, such as the last five years. For users for whom this is not sufficient, a customizable date limiter where years can be manually input is provided (Figure 10).



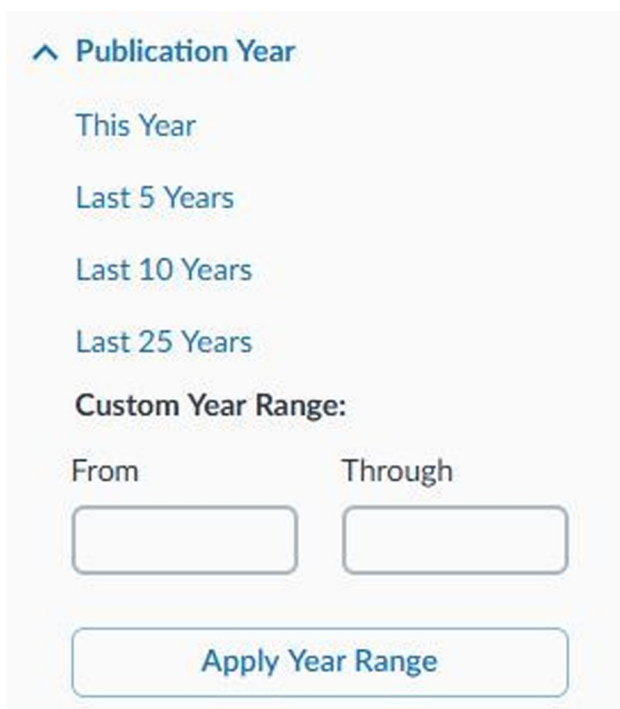
**Figure 9.** Example of date slider for date of publication ranges from Stanford University Libraries' Blacklight interface.

### ***Heterogeneity of library metadata***

Taylor's well-known definition states that facets are "clearly defined, mutually exclusive, and collectively exhaustive aspects, properties, or characteristics of a class or specific subject."<sup>47</sup> Facets are most effective when populated by values that share these characteristics. This requires quality metadata with the attributes defined by Park: completeness, accuracy, and consistency.<sup>48</sup>

Despite the best efforts of metadata professionals, several forces work against achieving the necessary level of data quality to support accurate facets. There are, of course, fiscal constraints that limit the amount of money that is available to hire catalogers or others to spend the time to add or perform quality control on metadata to populate facets. Due to these resource constraints, completeness will always be aspirational in any large database. Clearly defined facet definitions improve accuracy of assignment and increase ease of user understanding of the terms used.

Consistency, in particular, is also hobbled by the wide variety of approaches to recording library metadata and the sources of that data. This diversity is found both within standard MARC bibliographic records handcrafted according to cataloging rules and with metadata generated by other processes, such as vendor records and metadata from institutional repositories and digital asset management systems. Although



^ Publication Year

This Year

Last 5 Years

Last 10 Years

Last 25 Years

Custom Year Range:

From

Through

Apply Year Range

**Figure 10.** Example of combination of named date ranges and manual input for date of publication from WorldCat.

MARC bibliographic records created according to cataloging standards are more consistent with each other than records from other sources, there are nevertheless many variations. Cataloging rules and practices have changed over time and cataloger judgment varies. Although most MARC bibliographic records are created according to RDA/AACR, there are also records created following numerous other standards. For many fields in MARC records, there are also numerous potential controlled vocabulary sources. This leads to duplicative variants and the presence of near synonyms in lists of facet values presented to users. Figure 11 shows some of the subject facets retrieved with a search for Tolstoy. Tolstoy's name appears in two forms, Tolstoy and Tolstoj. The list includes both "Literature" and "Literatur," as well as both "Fiction" and "Nouvelles." It also includes some headings that lack context, such as "1900 1990" from FAST and "18 53 Russian literature," a term from the Dutch Basic Classification.<sup>49</sup>

The top four subject facet values that result from a search for the keywords "climate change" scoped to consortial records in the University of Oregon's Primo include "Climatic changes" from LCSH, "Climate change" from MeSH and "Climat--Changements" from Répertoire de vedettes-matière. These inconsistencies only multiply when library

- Russian Literature (290)
- Tolstoy Leo Graf 1828 1910 (264)  
Criticism And Interpretation
- Literature (220)
- 1900 1999 (216)
- Authors Russian (196)
- Literatur (173)
- Russisch (169)
- Short Stories (158)
- Soviet Union (151)
- Manners And Customs (127)
- Russian Fiction (124)
- Tolstoj Lev N (123)
- Nouvelles (121)
- 1800 1999 (114)
- Fiction (111)
- 18 53 Russian Literature (107)
- Russian Literature 19th Century (106)  
History And Criticism

**Figure 11.** Heterogeneous list of topical subject facet values resulting from a search for Tolstoy in Primo.

discovery interfaces incorporate data from non-MARC sources. Expanding the search for “climate change” to include external metadata from Ex Libris’s Central Discovery Index, mostly from articles, causes the subject facet list to shift to include many broad categories that are not present as subjects in typical MARC records, such as the new top match of “Science & technology.”



The usability of the Primo topical subject facet, as shown in [Figure 11](#), is impaired because it mixes terms from different vocabularies with different structures, as well as uncontrolled keywords. These vocabularies are in multiple languages and may have different meanings or be at different levels of specificity. This lack of consistency leads to lists of terms within a facet that are not mutually exclusive and perhaps not collectively exhaustive. Lack of consistency reduces both precision and recall and makes the facet list more confusing and less effective for users.

Overlapping and duplicative headings waste prime space at the top of the heading list and require users to select multiple terms if they want comprehensive results. They also reduce variety in the top results and prevent users from easily noticing other aspects or terms that would be helpful. This problem is exacerbated in systems like Ex Libris's Primo that only display a limited number of facet values, commonly twenty. Keeping only a preferred vocabulary and suppressing the others reduces redundancy and increases variety, although it does not solve the recall problem when not all records contain terms from the selected vocabulary.

Populating a single facet with terms from multiple vocabularies can also conceal differences in the meaning of the same term, since terms from different vocabularies may have different definitions. For example, a genre facet could include both the term "Drama" from LCGFT, which is limited to stage drama, as well as "Drama" from FAST, which is used both for stage drama and for film, television, and radio dramas. This conflation means that users are unable to isolate stage drama because it is combined with other forms of drama. Terms may also be used in different ways. The Library of Congress subject heading "children" covers birth through 12 years of age, but the superficially equivalent MeSH term "child" covers only ages six through twelve. The differing structure of the two vocabularies also reduces precision. LCSH is precoordinated and "children" by itself is only used for works that are narrowly focused on children as a topic. MeSH is a post-coordinated vocabulary. For the types of resources where MeSH terms are assigned, children as a concept rarely or never occurs as the main topic of a resource. Rather "child" is used to mark that a resource discusses some disease, treatment, or other topic in relationship to children.

This problem can be mitigated, at the cost of a potential loss of recall, by populating facets with values from only a single controlled vocabulary. In some contexts, libraries may use different vocabularies for different purposes, such as supplementing a general-purpose subject vocabulary like LCSH with additional vocabularies that provide more granular access to certain topics or that support diversity, equity, and inclusion goals. In this case, some sort of mapping or other process to coordinate the vocabularies, so that the union of the vocabularies produces a coherent list of facet

values that are not redundant and do not conflict, is desirable. Metadata remediation to ensure the presence of standardized terms can also improve the situation. For example, Ma describes a project to clean up variations in facet values in a digital collection of oral histories.<sup>50</sup>

### ***Heterogeneity of library resources***

Despite the popularity of collections of tools, board games, or other objects in some libraries, the stock held by libraries is not as diverse as that of a giant ecommerce site like Amazon. Nevertheless, there are many categories of resources held by libraries that could benefit from specialized subsets of facets. In addition to a wide variety of textual resources, most libraries also hold at least some other types of materials, including scores and musical recordings, films and videos, maps, games and objects, images and pictures, or computer software. Textual resources also vary and include not just mainstream published monographs, but journals, articles, and primary sources. Many genres, such as literature, biographies, and dance, also have specialized characteristics.

On many ecommerce websites, this problem is addressed by mapping the items in inventory to categories. When a user searches, their search terms are matched to one or more of these categories. In some cases, the user is able to select a category. For example, a search on Amazon for “Beethoven” brings up facets for departments, such as music and books. Selecting the book department brings up categories appropriate to books, including subcategories like biography and history, formats such as paperback and board books and language. Selecting music brings up some similar categories, including subcategories and formats, but also some different categories like edition and an option to exclude explicit lyrics.

There are many cases where a single search in library catalogs across formats and genres is optimal, but it would be beneficial to be able to also offer users with more specialized needs the ability to hone in on the characteristics of the resources that they are seeking. Many of the categories of library resources discussed above could benefit from focused lists of facets. In addition to developing user interfaces with this capability, it is also necessary to be able to identify the resources that fall into a given category and to ensure that a sufficient percentage of the records for those resources contains adequate metadata to support those facets. For many types of resources, this is challenging. For example, it is not easy using existing MARC metadata to definitively identify records that describe resources that consist of or contain literary works. Much of the metadata that would be useful as facets, such as genre, date of work, original language, and creator demographic characteristics, has not traditionally been recorded and is not consistently added to new records.

Retrospective work is challenging due to the large number of records and inconsistent practices. Other potential categories are easier to identify. Records for some types of resources are more amenable to retrospective data remediation as past and current cataloging practice has been to more consistently record information as structured data that can be used for facets. Perhaps the most promising category of this sort is music. Correctly coded records for scores and audio recordings can be identified by the record type in the leader of the MARC record. Video recordings of musical performances are more challenging, but many of them can probably be identified from subject or genre headings and marked with an additional RDA content type of “performed music.” Incorporating actual musical instruments or resources about music would be more challenging. Below I will discuss some issues with providing faceted access focused only on musical resources themselves.

### ***Music: a case study***

Musical resources are complex and have many unique characteristics. There is also a long history of detailed and relatively consistent cataloging by metadata professionals with deep domain knowledge. Many of the search needs of users of musical resources cannot be effectively met with keyword searching in contemporary discovery interfaces. Music specific facets have the potential to make searching for music resources simultaneously more powerful and easier.

Medium of performance is a prime example of an unmet information need. Musicians and music researchers often wish to search by instrumentation or voices in a piece of music. Historically, this has been difficult to do in library catalogs. With the introduction of the Library of Congress Medium of Performance Terms (LCMPT) vocabulary and the implementation of the 382 (Medium of Performance) field in MARC 21, it has become more common to record this information in a structured manner in library metadata. Catalogers find it easier to enter than the previously available coded fields and the Music Library Association in conjunction with Gary L. Strawn has developed a widely used macro that aids in accurately entering this data into new and existing records.<sup>51</sup> More completely populated data has created incentives for discovery interfaces to provide access to this information. For example, the Orbis Cascade Alliance has created several facets based on LCMPT in the 382 field of the MARC record.<sup>52</sup>

There are three main challenges for using medium of performance from the 382 field in discovery interfaces. One is the complexity of the structure of the field and the need in many cases to manipulate the data before it can be presented to users. The second is the complexities associated with

certain types of medium of performance information that cannot be accurately encoded in the 382 field. The third is the problem of false drops associated with aggregates or compilations.

The Orbis Cascade Alliance has created three facets based on medium of performance information from 382. Creating the values to populate these facets ranges from straightforward pulling of the data from MARC to presenting facet values that are significantly transformed from the data in the underlying MARC record. The medium of performance facet is populated with the individual names of each instrument given in the 382 field (Figure 12). The data is not manipulated for faceting other than mapping repeated subfields that occur in a single instance of 382 independently and double posting instruments and voices that are identified as solos both under the plain instrument name (“piano”) and under the instrument name qualified by the word solo (“piano (solo)”). The number of performers facet is similar (Figure 13). The data is only slightly manipulated for display. The word “performer” or “performers” is added after the bare number given in the MARC data and any number of ensembles is mapped to the generic term “ensemble(s).” The final facet is called medium of performance statement and attempts to represent the complete instrumentation for a piece (Figure 14). This facet requires the most manipulation, which necessitates a system that is capable of transforming data before use and is subject to the constraints of such systems. Alternative and doubled instrumentation is dropped from the medium of performance statement facet because it cannot be presented usefully with the tools available. To promote readability, the number of performers is dropped when it equals one. This also improves consistency by compensating for cases where the cataloger has omitted the “1.” This approach does, however, also conflate single performers with instances where the number of performers is not specified or varies and thus is not recorded. Higher numbers of performers are surrounded with parentheses and each instrument is separated with a semicolon. Some inconsistencies in the underlying data impact performance of this facet. In particular, there is no prescribed order for listing instruments in 382. In Figure 14, “violin; cello; piano” and “piano; violin; cello” are listed separately because they have been entered differently in the MARC record. Compensating for this would require more tools than Primo provides and ideally would be done by making the original metadata more consistent. The medium of performance statement facet also suffers from an occasional mismatch between the atomized terms used in 382 and a collective term in common use. For example, users may be seeking a piece for “string quartet,” while the facet used is “violin (2); viola; cello.”

Although the 382 field for medium of performance is complex and includes numerous subfields that enable catalogers to record nuances

## Music: Medium of Performance

piano (5,979)  
orchestra (4,245)  
violin (2,328)  
cello (1,553)  
piano (solo) (1,062)

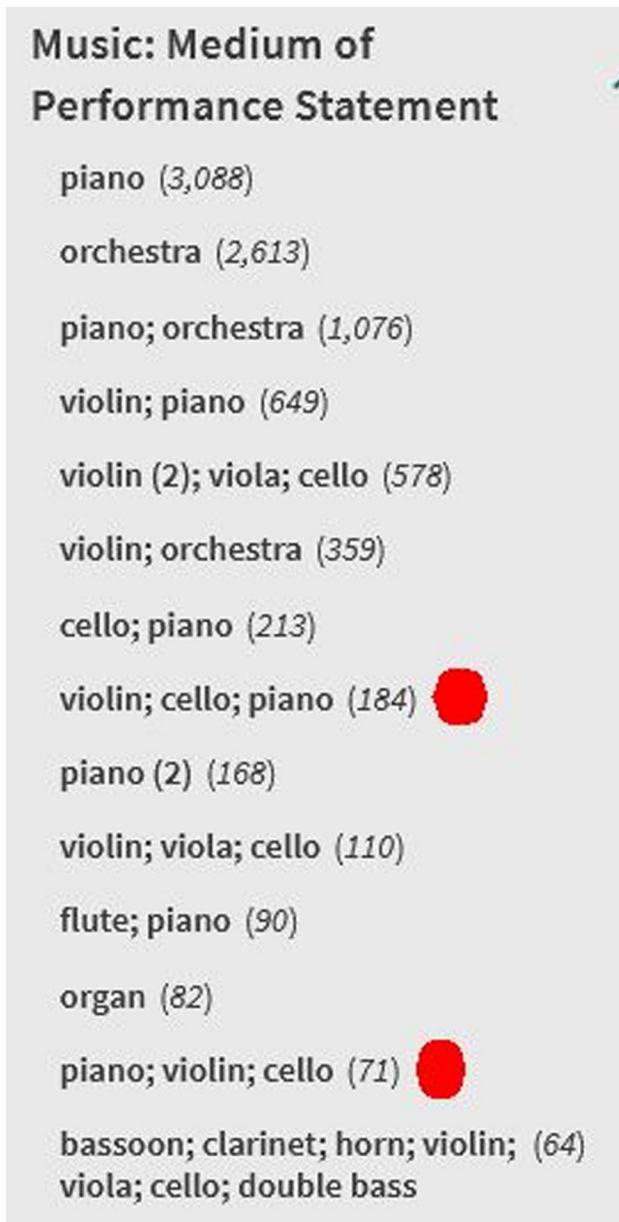
**Figure 12.** The Orbis Cascade Alliance music medium of performance facet.

## Music: Number of Performers

ensemble(s) (4,330)  
1 performer (3,312)  
2 performers (1,587)  
4 performers (750)  
[Show More](#)

**Figure 13.** The Orbis Cascade Alliance music number of performers facet.

of the instrumentation and voices used, it nevertheless falls short of being able to accommodate all situations accurately and with sufficient granularity. One example of this is the number of performers, pianos, and hands involved in the performance of certain piano music. Another situation that presents challenges is the way that a single percussion player may use multiple instruments within a single piece. This can be recorded as either a single percussion player or a



**Figure 14.** The Orbis Cascade Alliance music medium of performance statement facet. If the instruments are listed in different orders in the metadata, they do not collocate, as seen in the two entries for violin, cello and piano.

single performer doubling on multiple specific percussion instruments. The 382 field can be repeated to bring out both aspects, but this is not always done, which leads to data inconsistency. The need to repeat the field also creates extra work for catalogers. Problems with and potential solutions for more complex medium of

Music: Medium of Performance	For violin; harpsichord. Total performers: 2. For piano (2). Total performers: 2. For bassoon; cello. Total performers: 2.
------------------------------	--

**Figure 15.** Selecting facets for violin and piano may find compilations with a piece for piano and a separate piece with a violin part.

performance situations are discussed in Szeto (2022), Lee and Robinson (2018), and Lee (2017).<sup>53</sup>

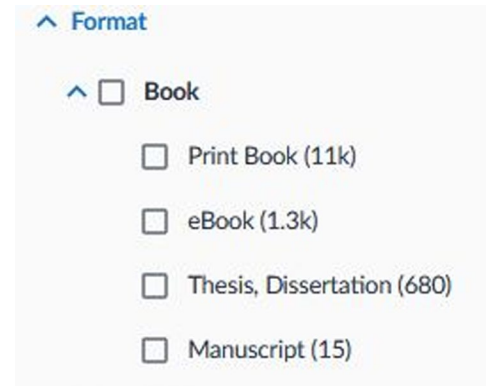
The potential for misleading combinations of facets related to different components of aggregates is particularly common when dealing with musical recordings. Most musical recordings include more than one piece. Each individual piece is commonly described at the level of the piece rather than with a single, broader term that is meant to encompass the shared characteristics of the whole, as is commonly done with literature. For example, a user who searches for violin and piano and then limits with those facets, may still retrieve resources that include some pieces for piano and some different pieces for violin with no overlap (Figure 15).

There are many other types of data unique to music that could be used to create helpful facets. The Orbis Cascade Alliance has created facets for musical key, music number, composer, and performer. The composer and performer facets consist of names from the MARC 1XX and 7XX fields that are associated with relator codes or relator terms for those roles.<sup>54</sup> The composer facet includes additional logic where any 1XX on a score record or any 1XX on a record for a musical recording with a uniform title in a 240 field is mapped to the composer facet. Names in 7XX name-title fields with second indicator 2 for component part or a relationship in \$i indicating that the name and title represent a work within the resource are also mapped to the composer facet on records in the music format. The Music Library Association issued a report on music discovery requirements that makes extensive recommendations on how facets can profitably be used to improve access to music materials, which influenced some of the work of the Orbis Cascade Alliance.<sup>55</sup>

## ***Assorted metadata-related issues***

### ***Hierarchy, nesting, specificity***

Tunkelang points out that hierarchical or nested facets can be an effective way to present facets with long lists of values to users. However, if each top-level facet value has a large number of values immediately under it or, alternatively, the facet hierarchy is many layers deep, this approach potentially introduces usability issues because of increased information load and the complexity of navigating a hierarchical tree.<sup>56</sup> In addition, the vocabulary used to populate the facet must have a hierarchical



**Figure 16.** Nested facets for resource types in WorldCat.

structure and the library discovery interface must be able to interpret and present this structure in a useful way. Most library catalogs make little or no use of hierarchical or nested facets. One example of nested facets is WorldCat's format facet (Figure 16).

McGrath notes that being able to navigate to different layers of a hierarchy in facets gives users the power to adjust their searches depending on their needs and the number of resources in a given category. She gives the example of searches for “everything about communicable diseases (broad) in Kenya (narrow) or AIDS (narrow) in Africa (broad),” which demonstrates the need to support differing levels of specificity based on users' requirements.<sup>57</sup> Supporting multiple levels of specificity also helps users perform more effective searches depending on the number of resources available. The combination of facets for nineteenth century works originally in English will be more effective if it can be combined with specific genres of poetry rather than just the broad category of poetry. However, if the user has additionally selected works by children or if they are looking for nineteenth century poetry originally in Polish, they may be better served by a facet value for poetry that combines all types of poetry, including subgenres. Flexibility for users in navigating hierarchies can either be provided by discovery interfaces that leverage the syndetic structure of controlled vocabularies or by selectively assigning applicable terms from different levels of the hierarchy. For example, if a user is seeking nineteenth century novels written by Americans and set in Mexico, but only the more specific creator demographic group terms such as New Yorkers and Illinoisans are present in the bibliographic record, without system support, the user must manually check for authors from every state and city in the U.S.

In the Orbis Cascade Alliance's former Primo Back Office discovery interface, they compensated for the lack of support for nested facets in Primo in their Medical Subject Headings (MeSH) facet through double posting. Base headings with no subdivisions (e.g., “Kidney Diseases”) were



posted under both “Kidney Diseases” and “Kidney Diseases (general).” Base headings with subdivisions (e.g., “Kidney Diseases \$x therapy”) were posted under both the full string “Kidney Diseases--therapy” and the base heading “Kidney Diseases.” This resulted in a list like the one shown below. The display would have been improved if all of the headings with the base heading of kidney diseases could have been nested under plain “Kidney Diseases” with the heading qualified by general coming first in the list when the hierarchy is expanded.

- Kidney Diseases (2)
- Kidney Diseases (general) (1)
- Kidney Diseases--therapy (1)

Many potentially useful facets in library discovery interfaces would benefit from hierarchical or nested navigation, including formats, topics, genres, geographic areas, time periods, and languages. LCSH is particularly challenging in this regard due to its large size; high number of top-level terms; and idiosyncratic, incomplete syndetic structure. LCSH was not developed in accordance with modern guidelines for developing thesauri and much of its syndetic structure was retrofitted. As Svenonius points out, the Library of Congress took the “quick-fix” approach and used an automated process to convert all of its historically inconsistent see and see also references to broader and narrower terms in one fell swoop, with only “a few ‘Band-Aid’ reparations ... to fix some of the more egregious structural deficiencies.”<sup>58</sup>

#### *Lead-in terms and cross-references*

Facets make it easier for users to take advantage of controlled vocabularies in some ways. For example, facets enable users to recognize relevant terms without having to come up with them themselves. Users may not necessarily know in advance what terms might be used in the library catalog to describe the resources they are seeking. When facets are populated with terms from a consistent controlled vocabulary, they guide users toward search terms that are most likely to lead to their success. As Buckland notes, it is “easier for a person to recognize pertinent terms than to predict them.”<sup>59</sup> However, unless they are performing a search in a system that supports browse without search, users may not retrieve all relevant results if their initial search terms do not include the standardized term in the relevant facet. For instance, a user might search for “World War I” and then select “World War, 1914–1918” from an LCSH-based facet. Without removing the initial keyword search, the user will not retrieve everything with the subject heading “World War, 1914–1918,” since all of those records will not include the string “World War I.” This type of situation can be

mitigated by query expansion where synonyms or terms based on LCSH cross-references are included in the initial result set. There may also be a disconnect between phrases entered by users (e.g., “African-American lesbian poets”) and the atomistic terms used in a faceted vocabulary (e.g., “African Americans” plus “Lesbians” plus “Poets”). Query expansion based on stemming can help with many of these situations. However, sometimes the mismatch is less straightforward. Users may seek musical works for “string quartets,” but LCMPT describes this by listing the instrumentation separately (“violin” plus “viola” plus “cello” or, alternatively as something like “violin (2); viola; cello”). It is theoretically possible to design a search interface that will compensate for known instances of these mismatches, but it is more complex.

### *Coverage, recall, and retrospective enhancement of bibliographic metadata*

As described above, for good coverage in facets, comprehensive structured metadata describing attributes of interest is necessary. Because the MARC format was created in order to print catalog cards in the late 1960s, it originally featured only a small amount of structured metadata mostly in the form of single characters that stand for values rather than in a form that is comprehensible by humans without a key. Some other data is in the form of structured strings that are prone to typos and errors. Over time MARC 21 has shifted away from its original focus on printing and emulating catalog cards and more and more structured elements have been added. Many factors have led to the addition of more fields and subfields to the MARC 21 format that are intended to be machine actionable. These include an increased understanding of the value of machine-actionable metadata, the development of newer and more complex cataloging standards, the creation of a number of new faceted Library of Congress vocabularies and the demands of various user communities. This means that there are many more attributes related to bibliographic resources recorded in MARC records that can potentially be used to populate facets. However, current library discovery interfaces do not make optimal use of them. Presenting users with a large number of choices without overwhelming them is, of course, one challenge. Another frequent reason given for not providing access to these potential facets in discovery interfaces is concern that lack of coverage will mislead users. No large-scale information retrieval system will have perfect recall, but in many catalogs legacy records that do not include data in newer fields and subfields far outnumber newer records that include this metadata. Tunkelang lists two situations where recall is critical for user success.<sup>60</sup> One is when searches would otherwise return few or no results and sparsely-populated facets cause users to believe that the library does not provide access to relevant resources. He notes

that it is also important when “searchers care about aggregate information about the results, such as the total number of results or the distribution of attributes of those results.”

The only practical remedy for this situation is automated or semi-automated methods for adding this data to existing records. As discussed previously, there has been progress in this area for records describing scores and musical recordings through a macro-based approach.<sup>61</sup> However, many of the other enhancements that would be desirable will be much more challenging. The music formats have two significant advantages as targets for this kind of metadata enhancement. First of all, the target set of records can be easily identified. The macro works only on scores and musical sound recordings. These can be reliably identified by the record type in the MARC leader, which is required and is rarely incorrect except for some older records for electronic resources or for scores put on book records. Some spoken recordings and recordings of sounds may be incorrectly coded as musical sound recordings, but the vast majority of records coded as scores and musical sound recordings describe resources that actually fall into those categories. The second factor that makes data remediation for music records a more tractable problem is that music catalogers have traditionally recorded more information in a more consistent fashion, so in most cases the data that the macro needs to generate the new fields is present in the record. Much of the data is in the form and of a type that can be accurately parsed and transformed without human intervention. The new fields generated may be incorrect or incomplete, but they will reflect the existing metadata and be no less accurate. Records may also contain conflicting metadata in different fields, such as genre and medium of performance in Library of Congress subject headings vs. what’s found in 047 (Form of Musical Composition Code) and 048 (Number of Musical Instruments or Voices Code) coded data. An automated process has no way to resolve these discrepancies unless it were sophisticated enough to identify and access external authoritative data.

The American Library Association’s (ALA) Subject Analysis Committee’s (SAC) Subcommittee on Faceted Vocabularies (SSFV) has begun working on the problem of retrospectively enhancing bibliographic records with data to support faceting, such as genre/form terms from LCGFT. They have written a paper providing an overview of the issues and listing many types of information in bibliographic records that would benefit from metadata remediation.<sup>62</sup> SSFV has developed provisional mappings from LCSH form/genre subdivisions to LCGFT and from selected fixed fields to both LCGFT and LCDGT.<sup>63</sup>

The subcommittee has begun work on mapping LCSH for literature to LCGFT. This is more challenging for several reasons. It is harder to reliably identify records that consist of or contain literary works. Records

for literature, especially older records, are also less likely to contain genre or form information in any form than records for music, which means that there is no data for an automated process to work with. The *Subject Headings Manual* tells catalogers not to record genre/form terms in MARC field 650 (topical subject headings) for individual works of fiction and literature, so if LCSH is correctly applied, these terms are only recorded for anthologies.<sup>64</sup> Even when a record contains an LCSH term for genre or form, it may be difficult to reliably determine whether the resource consists of or contains that genre or form or if the resource only contains criticism. For both users and catalogers, the difference can be very subtle. In LCSH, “Symphonies” is used to describe resources that contain symphonies while “Symphony” is used for resources about symphonies. Often the same base term is used with some sort of subdivision appended to indicate that something is or contains criticism, for example “Poetry” for poetry and “Poetry--History and criticism” for works about poetry. This distinction is not always reliably made in practice. In particular, some older records describing critical resources may lack the subdivision while records that describe resources that contain both poetry and criticism of poetry do not necessarily include both subject headings.

Although this method is more error-prone, data is sometimes available in the record in free text fields, such as notes, and work is being done to find automated methods to map it to structured data. Progress in machine learning and natural language processing in combination with the ability to match entities described in bibliographic records with external, trusted descriptions of those entities elsewhere also represents a possible approach for enhancing library metadata.

### ***Structured metadata and metadata preparation***

Even when structured fields exist to record attributes of interest and data in these fields is sufficiently populated within the dataset, the data in some fields requires additional processing and transformation before it can be presented to end users. For example, the 008/22 target audience fixed field in the book format contains data like “a,” which must be transformed into its meaning of “preschool” before it can be used in a facet. Some fields, such as 382 medium of performance, may require even more transformation before they are suitable for display or faceting. Some vocabularies would also benefit from more complex mapping. For example, the Program for Cooperative Cataloging (PCC) has recently issued guidelines for recording ISO 639-3 language codes in MARC records.<sup>65</sup> However, few or no library discovery systems are capable of mapping these codes to words out-of-the-box. For an optimal user experience, some of these codes

should be mapped to more than one facet value. For example, ISO 639-3 includes both a code for Chinese as a collective macro language for all varieties of Chinese and codes for individual varieties, such as Mandarin and Cantonese. In order to have good recall for users who select the facet value for Chinese that facet value must also bring up the records that are coded for Mandarin and Cantonese.

This means that effective implementation of facets in library discovery interfaces depends on the capabilities of the software being used in combination with deep understanding of the metadata and likely user needs. Local control over the data being displayed, indexed, and faceted in library discovery interfaces varies greatly. With some online catalogs, the library may have no control over display and facets or the library may only be able to choose which fields and subfields to use without being able to manipulate them in any way. Other products, such as Ex Libris's Primo, support much more powerful manipulation of the metadata by local institutions for use in their discovery interface. Open-source discovery interfaces, such as Blacklight, with sufficient investment offer even greater flexibility and power. Even in these cases, there are limits that may prevent some desired transformations. An alternative approach would be to alter the underlying metadata in some way, but this may lead to problems with nonstandard metadata.

One question that is not asked enough in the library metadata world is what question or questions is this metadata trying to answer? Related to this, does the way that the field for this metadata is defined and structured, as well as the way that the metadata is entered in practice, enable it to answer that question or questions? Evolving needs and historical contingencies sometimes mean that the answer is no. For example, the path of evolution of subfields associated with dates of creation in 046 means that data in some subfields cannot be accurately interpreted without evaluating it in association with what other subfields are present. This makes the field much more difficult for catalogers to understand how to use correctly while also making it more complicated for systems to use. Useful facets require not only consistent, structured metadata, but also that the metadata be structured in such a way that it can either be used in facets in its raw form or be accurately transformed in as straightforward a way as possible.

### *The unknown, the unknowable, the vague, and the inconsistent*

Many challenges for incorporating facets into library discovery interfaces are technical or could be solved with sufficient time, trained personnel, and funding, but some issues are less tractable. These include missing values, imprecise values, and inconsistent values.

Certainly, no large database will ever be clean enough and complete enough to have perfect recall. Older bibliographic records are often missing data that would be common to find in more recent records. Even in contemporary records, data may be missing because it is impractical to spend the time and effort to identify it. In some cases, no amount of time or effort will uncover the correct value. There are other reasons for incomplete coverage, such as cases where an appropriate subject heading has not yet been or cannot be established at the time of cataloging. Jahnke gives the example of Eve Sedgwick's *Epistemology of the Closet*, which was written by a founder of queer theory, but was published before there was literary warrant or the idea had coalesced into a namable entity.<sup>66</sup> With currently available resources and technology, it is impractical at best to later identify all these cases and go back to add the new information. However, fixing sizable gaps that are of clear interest to users, such as reliably identifying fiction, should be a priority.

The messiness and fuzzy boundaries of the real world are often at odds with the clearly defined categories required for optimal facets. There has been much recent criticism in the library world of the practice of categorizing gender as binary with sharp boundaries, but the problem of what philosophers call vague predicates pervades the categories used in bibliographic metadata. As previously discussed, many named time periods have fuzzy boundaries. This is also true of places, classes of persons, languages and most other topics described by library metadata. Based on a project to map statements about responsibility in moving image records to standardized role designations, McGrath discusses additional situations where it is difficult to map information provided by a resource to clearly defined categories. For example, it may be difficult to interpret cases where language use has changed over time, such as the earlier use of the credit "art director" for what is now called "production designer."<sup>67</sup>

Inconsistent cataloging practices or lack of inter-indexer consistency potentially has negative effects on facet usefulness. Cataloging practices have changed significantly over time. Newer records are often fuller and contain more structured metadata. Older records, vendor-created records, or records created according to minimal standards may lack useful metadata. Metadata values reflect the information available at the time, as well as the prejudices and perspectives of the era. Different catalogers bring differing amounts of expertise, time, and inclinations to their work. For example, there is often a conflict between catalogers who take a maximalist approach and prefer to add any potentially relevant value and catalogers who emphasize precision. To take one example, the RDA content type "still image" is added inconsistently to records for books. Some catalogers add "still image" even if there is just one portrait on a frontispiece on the principle that images are present. Others include "still image" only if

the book contains a significant number of images that are topically relevant, such as in a graphic novel or book of photographs. When catalogers do not agree, the metadata, and any facets it generates, will answer neither the question “Does the book contain any images at all?” nor the question “Does the resource contain interesting, useful images related to the topic of the book?” Variation in cataloger judgment and practice is exacerbated by the fact that most values in controlled vocabularies used by libraries are subject to the paradox of the heap<sup>68</sup> where there are inevitably situations where there is not consensus about what value should be recorded in the metadata.

### ***LCSH and topics: a case study***

One of the most important and challenging types of information to incorporate into facets in library discovery interfaces is topical or subject information. The largest and most widely used controlled vocabulary for subjects in library catalogs is LCSH. A number of characteristics of LCSH make it challenging to present as facet values. These include its origin as a pre-coordinated and nonsystematic vocabulary combined with an inability to easily deconstruct it into more granular facets and its historical use for recording information about both topical and non-topical aspects of a resource.

One fundamental challenge is that LCSH was designed to be used in multifaceted, precoordinated strings that attempt to collocate all the significant aspects of what a resource is about. It was also designed to be browsed in a left-anchored, alphabetical list. Both the LCSH strings as a whole and the individual parts of the strings combine different types of information. Some of this is distinguished in MARC records by subfield coding, such as chronological aspects in \$y and geographical aspects in \$z and can be easily separated. In other cases, different kinds of information may be encoded in a single subfield, which makes it harder to separate. For example, “Waterloo, Battle of, Waterloo, Belgium, 1815” is recorded in a single topical 650 \$a, but includes chronological and geographical information that is not separately subfielded. Some topical subject headings in LCSH include prepositions that relate more than one term and conflict with the atomistic concepts that are optimal for faceted search. In some cases, compound terms seem to merely present synonyms, near synonyms or opposites (e.g., “Ambushes and surprises,” “Belief and doubt”). In other cases, they combine related terms that are different types of things, which should properly be separated for faceting (e.g., “Boats and boating,” “Collectors and collecting”). Structurally similar headings may also be used to present the relationship between two distinct things (e.g., “Age and sports,” “Artists and architects”). Conversely, parallel meanings

may be represented by different structures (e.g., “Africa \$x In motion pictures” vs. “African American cowboys in motion pictures”). This last example actually represents three concepts: African Americans, cowboys, and portrayal in movies. Young notes that a given structural pattern in LCSH does not always have the same meaning, which, in addition to potentially confusing users, impedes mapping to a more faceted presentation. She gives the example of “Children’s diaries,” which is used for diaries written by children and “Children’s films,” which is used for films made for children.<sup>69</sup> “African Americans in motion pictures” (the portrayal of African Americans in movies) and “African Americans in the motion picture industry” (African Americans working in the movie industry) do not imply the same relationship between the two nouns. These concepts cannot be further combined into one long string, such as “African Americans in the motion picture industry in motion pictures,” but a thoroughly faceted vocabulary should enable these sorts of novel combinations without requiring that they be precomposed. McGrath points out that “terms, such as ‘Cookery, Japanese’ or ‘Adult children of alcoholics, Writings of,’ that incorporate more than one facet or aspect of a concept reduce the power and flexibility of faceting by preventing users from limiting by the individual aspects separately.”<sup>70</sup>

This is further complicated by the fact that LCSH has historically been used to encode some non-topical information as well, such as genre and audience. Young writes that

LCSH combines the topical, genre/form, creator, audience, and medium of performance facets in contradictory and sometimes unpredictable ways, and even headings that are similarly formatted may denote quite different facets. Those problems are only exacerbated by the fact that form headings can also usually be used as topics.<sup>71</sup>

There have been two main attempts to improve the suitability of LCSH for faceting. Both have as goals simplifying metadata creation and making it easier for users to discover library resources.<sup>72</sup> One is OCLC’s development of FAST, which is derived from LCSH, but breaks it down into more post-coordinate categories, such as topical, geographic, chronological, and form/genre aspects. The other is the Library of Congress’s creation of several new vocabularies to accommodate non-topical information currently contained in LCSH.

FAST is a largely post-coordinate vocabulary that can be assigned independently or automatically derived from LCSH strings. FAST takes advantage of MARC field and subfield codes to create nine separate facets for topics, personal names, corporate names, meetings, named events, uniform titles, chronological information, geographic areas, and form and genre.<sup>73</sup> However, it does not merely perform a naïve mapping of the components



of LCSH strings to these categories, but rather transforms the data in a number of ways that make it clearer and more amenable for use in faceted search. The FAST Quick Start Guide notes that FAST introduces useful distinctions that are not made in LCSH, such as named events, which are described as “events associated with a particular date, and possibly a particular geographic location, and that are well known by a recognized name,” such as particular battles or earthquakes.<sup>74</sup> In some areas, such as its chronological facet, FAST introduces flexibility that is not available in LCSH. Time spans are recorded in FAST using explicitly coded beginning and end years. Chronological information can therefore be coextensive with the coverage of the resource, although in practice, much chronological information is automatically derived from more generic information found in time spans given in LCSH topical headings. Some relationships are more explicitly encoded in FAST. Geographic headings are given in indirect order and include the relationship to the larger place up to the country or state level. Geographic headings are given in a consistent form without abbreviations or inversion. FAST uses “Illinois--Chicago” rather than LCSH’s use of both “Illinois--Chicago” and “Chicago (Ill.)” This makes searching by place names more predictable. It also supports a certain amount of hierarchical access, but does not always include the country name and does not include the continent level. However, FAST authority records for geographical areas do include the MARC geographical area code that does map to higher levels and could be employed to generate a fuller hierarchy.

FAST is not a completely post-coordinated vocabulary since it allows combination of terms from the same facet category. In particular, main topics and topical subdivisions continued to be pre-coordinated in FAST. This improves precision in some cases (e.g., “History--Philosophy” vs. “Philosophy--History”) while reducing flexibility in others. For example, the topical subdivision “Economic aspects” that follows topics remains pre-coordinated while the topical subdivision “Economic conditions” that follows places is given separately. In a more fully faceted system, these two topical subdivisions could probably be profitably combined into a single term. FAST also makes no effort to create separate categories for types of topical headings. For example, one can imagine that a separate facet for classes of persons would better support browsing for biographies or types of characters in literary works.

In 2007, the Library of Congress began a project to develop LCGFT, a vocabulary to separately describe the form or genre of resources. Their intent is to remove form and genre from LCSH and record it only using LCGFT in a separate field designated for form and genre information. Form and genre terms remaining in LCSH would only be used for works about those forms and genres. As the Library of Congress worked on this

project, they realized that other non-topical aspects of resources are recorded in LCSH that need to be moved elsewhere in order to limit LCSH and topical fields to topical content. The Library of Congress has since developed vocabularies for musical medium of performance and for demographic group terms that can be used to describe creators and intended audiences.

McGrath describes a number of problems that arise when trying to use LCSH in a faceted search interface that are not necessarily resolved by FAST or the new Library of Congress vocabularies. For example, there are situations where aspects of an LCSH string are implicit, so there is no data to populate facets.<sup>75</sup> Practices around implicit information in LCSH often have their origin in its roots as a vocabulary designed for left-anchored, alphabetical browsing where shorter strings may be desirable and lead to less fragmentation. For example, the subject heading “National socialism” is used both for national socialism in general and national socialism in Germany as a whole. The subject heading is only geographically subdivided for works about Nazism in smaller places within Germany (e.g., Berlin) or for works about allied countries, such as Austria. This means that even if the geographic subdivision is presented in a separate facet in a library discovery interface, not all the works about Nazism in Germany will include the term for Germany in this facet, resulting in incomplete recall. “African Americans--United States” is a cross-reference for “African Americans,” so a naïve automated system for splitting LCSH strings into multiple facets based on subfield data will not create a geographic facet value for United States from “African Americans--History.” This means that a user who searches for a historical topic and then limits by United States in the geographic facet will not retrieve a resource based on this subject heading. Some topical subdivisions are also omitted on the basis that the topic is clear from the main heading. For example, the topical subdivision “Law and legislation” is not used under topics such as “Human rights.”

In addition to metadata values that are not explicitly recorded because they were deemed not useful in left-anchored, alphabetical browsing, LCSH use in practice reduces recall in some areas. A prime example is a genre facet that includes “Fiction” from the \$v genre/form subdivision that follows headings for what a novel or short story is about. These kinds of headings are not found on most older records or on records where the work does not have an easily identifiable topical aspect. Even when the facet value is supplemented with information from the literary form fixed field in MARC 008, many older records or minimally coded records lack the correct fixed field coding.

Chronological information can be difficult to extract from LCSH or may describe a less precise time span than the resource covers. McGrath

describes a number of situations where chronological information is not given explicitly in LCSH strings or where it may not be coextensive with the time period covered by the work.<sup>76</sup> FAST is a significant improvement on LCSH here because time periods are always encoded separately as numbers, either as a single year or as the beginning and ending date of a time span. Some chronological information in LCSH is not subfielded separately and is not easily accessible to populate a facet. This situation is improved in FAST, where dates contained within topical headings have been mapped to explicit dates or date ranges, although FAST does not provide access to time spans shorter than a year. For example, “Chile Earthquake, Chile, 2010 (February 27)” maps to 2010. This mapping does not work when the date is not recorded or implied somewhere in an existing LCSH string, but could be recorded proactively at the time of metadata creation. It also remains difficult to map named or imprecise time periods with fuzzy boundaries to specific beginning and ending dates in a way that works for all search queries and resources.

Nested, hierarchical facets are a good way to present multiple levels of geographic information. However, it is not always possible to identify the type and level of geographic entity being described in LCSH in an automated way. In some cases, authority records could be used to identify the type of geographic entity and the related broader and narrower places to create a hierarchy. It might also be desirable for a geographic facet to include certain places that are not currently marked as places in LCSH. These are generally names that identify something that can both be a place and act as an agent. The *Subject Headings Manual* section H405, often colloquially known as “the division of the world,” provides guidance on how to treat these entities.<sup>77</sup> Some of these, such as “Buckingham Palace (London, England),” do have broader terms that indicate the type of entity and its location. In this case, the broader term is “Palaces--England.” Other examples do not have broader terms, but in some cases, such as “Museo Guggenheim Bilbao,” the type of entity and its location are encoded elsewhere in the authority record. Increasing the number of cases where entity types can be explicitly modeled could help improve this situation.

As mentioned above, there are places where LCSH uses the exact same construction to mean more than one thing. Some of these have to do with geographic information and often reflect a conflation of nationality and place. For example, the subject heading “Prisoners of war” may be geographically subdivided. Unfortunately, the authority record instructs catalogers to use a place both to designate the current location of prisoners of war and the place of origin of the prisoners. Subject heading strings make no distinction between prisoners of war being held in France and French prisoners of war being held anywhere.

Language, nationality, and ethnicity are also entangled in many literary headings in such a way that it is difficult for users to build a coherent mental model. For example, all of the following are legitimate LCSH strings: “Nigerian drama (English),” “English drama--Irish authors,” “American drama,” and “Hispanic American drama (Spanish).” For literary works themselves, the introduction of LCDGT and an associated MARC field for creator demographic terms promises the ability to clearly identify the nationality and ethnicity of authors. In combination with the expanded definition of MARC 041 \$h to allow the recording of the original language of the work regardless of whether or not there is a translation involved, use of LCDGT will enable the disentangling of these concepts. However, it is less clear how to resolve the problem of disentangling and clarifying headings for resources consisting of criticism and other types of works about literature that will remain in LCSH.

There are several decisions that must be made when incorporating LCSH facets into a library discovery system. First is the question of whether each heading should be displayed as a complete string or whether the headings should be split based on the subfield markers. If the headings are split, a further question is whether to combine them in a single topical LCSH facet or to divide them into separate facets for topic, time period, geographic focus, and genre/form. Although full LCSH strings are not designed for faceting, they may potentially increase precision. However, a significant drawback of presenting full LCSH strings is that it greatly increases the number of facet values while decreasing the number of records associated with each individual facet value. This means that the number of records that a user is exposed to through the top ten or twenty facet values is greatly reduced. On the other hand, if every subfield is mapped to a separate facet value, there is loss of accuracy. “Philosophy--History” (history of philosophy) is not the same as “History--Philosophy” (philosophy of history). McGrath suggested that many headings would be clearer to users (and catalogers) if the relationship between the parts of an LCSH string were made more explicit than the use of double dashes.<sup>78</sup> Thus “United States \$x Geography” could be displayed as “Geography of the United States” rather than “United States--Geography” and “Geography \$z United States,” could be displayed as “Geography (discipline) in the United States” rather than “Geography--United States.” This is more difficult to do in situations where an identical string can mean more than one thing, but introducing a way to explicitly encode prepositions to show a relationship is a potential way to reduce ambiguity. However, it is not obvious how to make these relationships clear to users when the individual terms are in different facets. When placed into separate topical and geographic facets, “United States \$x Geography” and “Geography \$z United States” look the same.

## **Recommendations for future work**

From the discussion above, it is possible to extract a number of recommendations for improvements to faceted search in library catalogs. These are listed below. Note that many of these recommendations could potentially be listed under more than one category.

### ***Computational efficiency***

Library discovery interfaces should maximize the efficiency of the underlying systems that index and retrieve their metadata and related facets in order to increase the speed of response and expand the number of facets and facet values that can be presented to users.

### ***User interface design***

Usability testing and user needs analysis should be done to better determine what facets users are interested in and what challenges they face when using facets. There should be more experimentation with and assessment of number and order of facets and facet values. Functionality should be developed to allow users to customize the order and number of facet values displayed. User studies should be performed to identify the most easily understood labels for facets and facet values. It would also be useful to investigate whether there are ways that Boolean operations with facets can be made more transparent and understandable for users. Where possible, methods should be developed to present facets most likely to be relevant to the user's search in the way that Amazon presents book-related facets if it detects that a user appears to be searching for a book title. For example, a search for Beethoven could highlight music-related facets. Interfaces should be developed that allow users to take advantage of the hierarchical structure of many controlled vocabularies.

### ***Exploratory search and recall***

Library discovery interfaces could better support exploratory search and browsing by allowing users to select facet values without first doing a keyword search, i.e., search-free browsing. It should be possible for users to obtain a complete list of facet values in a particular facet that are related to their search. A user looking for novels should be able to get a comprehensive list of genres or creator demographics not just the top twenty results. Users should be able to remove keywords while retaining the facet values that they have selected. This will enable them to do a keyword search to find relevant facet values and then remove the keywords

in order to get complete recall for the facet value. There should be investigation into what facets and vocabularies are most useful for search-free browsing both for a general-purpose interface and for specialized views focused on particular types of resources, such as music, moving images, or literature.

### ***Multiple entity types***

This is a thorny problem. First of all, the metadata must be structured in such a way that information is clearly associated with specific entities in a consistent, machine-actionable way. Although existing systems take little advantage of them, authority records for persons and corporate bodies could be leveraged for this purpose. However, much work remains to be done before information related to the parts of the WEMI stack or for aggregated and aggregating works can be cleanly specified in this way. When properly structured metadata is available, faceted search interfaces that utilize this information in a more accurate manner should be developed.

### ***Heterogeneity of library metadata and resources***

Work should be done to integrate values in facets that will be drawing on multiple vocabularies in order to present users with a more coherent list that minimizes redundancy and maximizes variety and relevance. This may include both vocabulary mapping work and system design. Experimentation should be done with customized interfaces for subsets of library resources, as well as developing systems that select the most relevant facets for display based on a user's search.

### ***Cross-references and facets***

Systems should allow users to remove their initial keyword search term or terms once they have identified the relevant controlled vocabulary term as a facet. There should be user studies and system design experimentation to identify ways to handle mismatches between the atomistic nature of facets (e.g., a list of individual instruments) and the phrases that users might be seeking (e.g., string quartets, piano trios).

### ***Coverage, recall, and retrospective metadata enhancement***

Tools that make it easier for catalogers to add structured data should be developed and improved. Cooperative projects, such as the work of ALA's SAC Subcommittee on Faceted Vocabularies, to identify and implement

strategies and processes to retrospectively enhance bibliographic metadata should be undertaken.

### **Other metadata issues**

Fields intended to populate facets should be examined to make sure that they meet the conditions of being clearly defined, mutually exclusive, and collectively exhaustive inasmuch as is possible. The fields and facet values should have clearly defined operational definitions that lead to consistent application. The definitions and metadata structure should unambiguously answer the question or questions that they are intended to answer. Facets and facet values should undergo user testing to make sure that they are easily understood and meet relevant information needs. Values should be recorded in a way that makes them easy to use as facets without complex processing.

### **Acknowledgments**

The author is grateful to Chew Chiat Naun, Casey Mullin, and Adam Schiff for helpful and insightful comments on a draft of this article.

### **ORCID**

Kelley McGrath  <http://orcid.org/0000-0002-5524-6417>

### **Notes**

1. Daniel Tunkelang, *Faceted Search* (San Rafael, CA: Morgan & Claypool Publishers, 2009).
2. Kate Moran, "The State of Ecommerce Search," *Nielsen Norman Group*, June 24, 2018, <https://www.nngroup.com/articles/state-ecommerce-search>.
3. Karen Coyle, "KO is KO'd," *Coyle's Information*, January 10, 2023, <https://kcoyle.blogspot.com/2023/01/ko-is-kod.html>.
4. Mia Massicotte, "Improved Browsable Displays for Online Subject Access," *Information Technology and Libraries* 7, no. 4 (1988): 373–80.
5. Marek Nahotko, "Knowledge Organization Affordances in a Faceted Online Public Access Catalog (OPAC)," *Cataloging & Classification Quarterly* 60, no. 1 (2022): 86–111, doi: [10.1080/01639374.2021.2015734](https://doi.org/10.1080/01639374.2021.2015734).
6. Tunkelang, *Faceted Search*.
7. Kathryn Whitenton, "Filters vs. Facets: Definitions," *Nielsen Norman Group*, March 16, 2014, <https://www.nngroup.com/articles/filters-vs-facets>.
8. "Library of Congress Genre/Form Terms," Library of Congress, accessed March 1, 2023, <https://www.loc.gov/aba/publications/FreeLCGFT/freelcgft.html>.
9. "Library of Congress Medium of Performance Thesaurus for Music," Library of Congress, accessed March 1, 2023, <https://www.loc.gov/aba/publications/FreeLCMPT/freelcmpt.html>.

10. "Library of Congress Demographic Group Terms," Library of Congress, accessed March 1, 2023, <https://www.loc.gov/aba/publications/FreeLCDGT/freelcdgt.html>.
11. Tunkelang, *Faceted Search*, 48.
12. Ibid.
13. Ibid.
14. Daniel Tunkelang, "Facets of Faceted Search," *Query Understanding*, November 23, 2020, <https://medium.com/@dtunkelang/facets-of-faceted-search-38c3e1043592>.
15. Tunkelang, *Faceted Search*.
16. Ibid.
17. "Facets," Ex Libris Knowledge Center, accessed March 1, 2023, [https://knowledge.exlibrisgroup.com/Primo/Product\\_Documentation/Primo/Back\\_Office\\_Guide/100Facets](https://knowledge.exlibrisgroup.com/Primo/Product_Documentation/Primo/Back_Office_Guide/100Facets).
18. Tunkelang, *Faceted Search*, 52.
19. Chels Upton, "The Backlash Against America's Most Popular Novelist Is Way Less Satisfying Than I'd Hoped," *Slate*, February 2, 2023, <https://slate.com/culture/2023/02/colleen-hoover-domestic-violence-ends-with-us.html>.
20. Tunkelang, *Faceted Search*, 65.
21. Ibid., 66.
22. Dana McKay, George Buchanan, and Shanton Chang, "It Ain't What You Do, It's the Way That You Do It: Design Guidelines to Better Support Online Browsing," *Proceedings of the Association for Information Science and Technology* 55, no. 1 (2018): 347–56, doi: [10.1002/pr2.2018.14505501038](https://doi.org/10.1002/pr2.2018.14505501038).
23. Bill Kules and Robert Capra, "Creating Exploratory Tasks for a Faceted Search Interface," in *Proceedings of 2nd Workshop on Human-Computer Interaction*, 2008, 18–21.
24. McKay, Buchanan, and Chang, "It Ain't What You Do, It's the Way That You Do It," 355.
25. Kelley McGrath, "Facet-Based Search and Navigation with LCSH: Problems and Opportunities," *The Code4Lib Journal*, no. 1 (2007).
26. McKay, Buchanan, and Chang, "It Ain't What You Do, It's the Way That You Do It"
27. Kelley McGrath, Bill Kules, and Chris Fitzpatrick, "FRBR and Facets Provide Flexible, Work-Centric Access to Items in Library Collections," in *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries*, 2011, 49–52, doi: [10.1145/1998076.1998085](https://doi.org/10.1145/1998076.1998085).
28. Tunkelang, *Faceted Search*, 54.
29. IFLA Study Group on the Functional Requirements for Bibliographic Records, "Functional Requirements for Bibliographic Records: Final Report, As Amended and Corrected through February 2009" (International Federation of Library Associations and Institutions, February 2009), <https://repository.ifla.org/handle/123456789/811>.
30. Pat Riva, Patrick Le Bœuf, and Maja Žumer, "IFLA Library Reference Model: A Conceptual Model for Bibliographic Information, As Amended and Corrected through December 2017" (Den Haag, IFLA, January 2018), <https://repository.ifla.org/handle/123456789/40>.
31. Tunkelang, *Faceted Search*, 54.
32. Music Discovery Requirements Update Task Force, "Music Discovery Requirements," Version 2 (Music Library Association, August 2017), <https://www.musiclibraryassoc.org/resource/resmgr/mdr/MusicDiscoveryRequirements2.pdf>.
33. Jerry L. McBride, "Faceted Subject Access for Music through USMARC: A Case for Linked Fields," *Cataloging & Classification Quarterly* 31, no. 1 (2000): 15–30, doi: [10.1300/J104v31n01\\_03](https://doi.org/10.1300/J104v31n01_03).



34. Kelley McGrath, "Will RDA Kill MARC?" (American Library Association Midwinter Meeting, San Diego, CA, January 8, 2011), <http://hdl.handle.net/1794/23939>.
35. Riva, Le Boëuf, and Žumer, "IFLA Library Reference Model."
36. IFLA Study Group on the Functional Requirements for Bibliographic Records, "Functional Requirements for Bibliographic Records."
37. Riva, Le Boëuf, and Žumer, "IFLA Library Reference Model," 21-27.
38. "RDA Toolkit," ALA Publishing, accessed March 1, 2023, <https://access.rdatoolkit.org>.
39. "BIBFRAME 2 List View," Library of Congress, accessed March 1, 2023, <https://id.loc.gov/ontologies/bibframe.html>.
40. "Share-VDE Model (Simplified Version)," Casalini Libri, accessed March 1, 2023, [https://docs.google.com/presentation/d/1cTf6UC\\_wSj-C43OxGj0du47HwOGl8FQbVLVF3goNUG8/edit#slide=id.g18c8243e708\\_0\\_0](https://docs.google.com/presentation/d/1cTf6UC_wSj-C43OxGj0du47HwOGl8FQbVLVF3goNUG8/edit#slide=id.g18c8243e708_0_0).
41. McGrath, Kules, and Fitzpatrick, "FRBR and Facets Provide Flexible, Work-Centric Access to Items in Library Collections."
42. Karen Coyle, "Works, Expressions, Manifestations, Items: An Ontology," *The Code4Lib Journal*, no. 53 (2022), <https://journal.code4lib.org/articles/16491>.
43. Ibid.
44. "Extended Date/Time Format (EDTF) Specification," Library of Congress, accessed March 1, 2023, <https://www.loc.gov/standards/datetime>.
45. "MARC Proposal No.: 2014-06: Defining New Field 388 for Time Period of Creation Terms in the MARC 21 Authority and Bibliographic Formats," Library of Congress, accessed March 1, 2023, <https://www.loc.gov/marc/mac/2014/2014-06.html>.
46. Joan Neuberger, "Not a Film but a Nightmare: Revisiting Stalin's Response to Eisenstein's Ivan the Terrible, Part II," *Kritika* 19, no. 1 (2018): 115-142, doi: 10.1353/kri.2018.0005.
47. Daniel N. Joudrey, Arlene G. Taylor, and David P. Miller, *Introduction to Cataloging and Classification*, 11th ed. (Santa Barbara, CA: Libraries Unlimited, 2015), 679.
48. Jung-Ran Park, "Metadata Quality in Digital Repositories: A Survey of the Current State of the Art," *Cataloging & Classification Quarterly* 47, no. 3-4 (2009): 213-28, doi: 10.1080/01639370902737240.
49. "Dutch Basic Classification," BARTOC, last modified July 14, 2021 <https://bartoc.org/en/node/745>.
50. Xiaoli Ma, "One Concept, One Term, Good Practice but How to Achieve? – Improving Facet Values Quality for Samuel Proctor Oral History Collection, Hosted by the University of Florida Digital Collections," *Journal of Library Metadata* 22, no. 3-4 (2022): 167-83, doi: 10.1080/19386389.2022.2096385.
51. "New OCLC Music Toolkit for Generating Faceted Music Data," Music Library Association Cataloging and Metadata Committee, April 20, 2018, <https://cmc.wp.musiclibraryassoc.org/2018/04/20/new-oclc-music-toolkit-for-generating-faceted-music-data>.
52. Kelley McGrath and Lesley Lowery, "Getting More out of MARC with Primo: Strategies for Display, Search and Faceting," *The Code4Lib Journal*, no. 41 (2018), <https://journal.code4lib.org/articles/13600>.
53. Kimmy Szeto, "Ontology for Voice, Instruments, and Ensembles (OnVIE): Revisiting the Medium of Performance Concept for Enhanced Discoverability," *The Code4Lib Journal*, no. 54 (August 29, 2022), <https://journal.code4lib.org/articles/16608>; Deborah Lee and Lyn Robinson, "The Heart of Music Classification: Toward a Model of Classifying Musical Medium," *Journal of Documentation* 74, no. 2 (March 12, 2018): 258-77, <https://doi.org/10.1108/JD-08-2017-0120>; Deborah Lee, "Numbers, Instruments and Hands: The Impact of Faceted Analytical Theory on Classifying Music Ensembles," *Knowledge Organization* 44, no. 6 (2017): 405-15.

54. McGrath and Lowery, "Getting More out of MARC with Primo."
55. Music Discovery Requirements Update Task Force, "Music Discovery Requirements."
56. Tunkelang, *Faceted Search*.
57. McGrath, "Facet-Based Search and Navigation with LCSH."
58. Elaine Svenonius, "LCSH: Semantics, Syntax and Specificity," *Cataloging & Classification Quarterly* 29, no. 1–2 (2000): 17–30, doi: [10.1300/J104v29n01\\_02,22](https://doi.org/10.1300/J104v29n01_02,22).
59. Michael Buckland et al., "Mapping Entry Vocabulary to Unfamiliar Metadata Vocabularies," *D-Lib Magazine* 5, no. 1 (January 1999), <https://doi.org/10.1045/january99-buckland>.
60. Daniel Tunkelang, "The 3 Rs of Search: Relevance, Recall, and Ranking," *Query Understanding*, December 21, 2020, <https://dtunkelang.medium.com/the-3-rs-of-search-relevance-recall-and-ranking-c9a785578653>.
61. "New OCLC Music Toolkit for Generating Faceted Music Data."
62. ALA Core Subject Analysis Committee, Subcommittee on Faceted Vocabularies, "Retrospective Implementation of Library of Congress Faceted Vocabularies: Best Practices for Librarians and Programmers," last updated March 25, 2022, <http://hdl.handle.net/11213/17998>.
63. Casey A. Mullin, "Iteration, Not Perfection: The 'Long Game' of Retrospective Implementation of Faceted Vocabularies" (IFLA Subject Analysis and Access webinar: "Fascinating Facets," May 19, 2022), [https://cdn.ifla.org/wp-content/uploads/2CaseyMullin\\_IFLA-Webinar-220519-Mullin.pdf](https://cdn.ifla.org/wp-content/uploads/2CaseyMullin_IFLA-Webinar-220519-Mullin.pdf)
64. "H 1775: Literature: General," in *Subject Headings Manual* (Washington, DC: Library of Congress, 2015), <https://www.loc.gov/aba/publications/FreeSHM/H1775.pdf>
65. "Guidelines for the Use of ISO 639-3 Language Codes in MARC Records," (Program for Cooperative Cataloging, January 12, 2023), <https://loc.gov/aba/pcc/scs/documents/ISO-639-3-guidelines.pdf>.
66. Lori M. Jahnke, Kyle Tanaka, and Christopher A. Palazzolo, "Ideology, Policy, and Practice: Structural Barriers to Collections Diversity in Research and College Libraries," *College & Research Libraries* 83, no. 2 (March 3, 2022): 166, doi: [10.5860/crl.83.2.166](https://doi.org/10.5860/crl.83.2.166).
67. Kelley McGrath, "Ostriches, Minotaurs, Ghosts and Fossils in the Brave New Metadata World: Categorization & Linked Data" (Online Northwest, Portland, OR, May 31, 2017), <http://hdl.handle.net/1794/23941>.
68. "Sorites paradox," Stanford Encyclopedia of Philosophy, last modified March 26, 2018, <https://plato.stanford.edu/entries/sorites-paradox/>.
69. Janis L. Young, "Unlimited Opportunities for Enhanced Access to Resources: The Library of Congress' Faceted Vocabularies" (Subject Access: Unlimited Opportunities, Columbus, Ohio, USA, 2017), <http://library.ifla.org/2074/>.
70. McGrath, "Facet-Based Search and Navigation with LCSH."
71. Young, "Unlimited Opportunities for Enhanced Access to Resources," 3.
72. Rebecca J. Dean, "FAST: Development of Simplified Headings for Metadata," *Cataloging & Classification Quarterly* 39, no. 1/2 (2004): 331–52, doi: [10.1300/J104v39n01\\_03](https://doi.org/10.1300/J104v39n01_03); Young, "Unlimited Opportunities for Enhanced Access to Resources."
73. Chew Chiat Naun, Kerre Kammerer, Kim Mumbower, and Dean Seeman of the FAST Policy and Outreach Committee, "FAST Quick Start Guide," (OCLC, April 2022), <https://www.oclc.org/content/dam/oclc/fast/FAST-quick-start-guide-2022.pdf>
74. *Ibid.*, 13.
75. McGrath, "Facet-Based Search and Navigation with LCSH."
76. *Ibid.*

77. “H 405: Establishing Certain Entities in the Name or Subject Authority File,” in *Subject Headings Manual* (Washington, DC: Library of Congress, 2021), <https://www.loc.gov/aba/publications/FreeSHM/H0405.pdf>.
78. McGrath, “Facet-Based Search and Navigation with LCSH.”
79. “046 – Special Coded Dates,” Library of Congress, accessed March 1, 2023, <https://www.loc.gov/marc/bibliographic/bd046.html>.
80. “Proposal No.: 2002-03: Expanding Field 046 for Other Dates in the MARC 21 Bibliographic Format,” Library of Congress, accessed March 1, 2023, <https://www.loc.gov/marc/marbi/2002/2002-03.html>.
81. “Best Practices for Cataloging DVD-Video and Blu-ray Discs Using RDA and MARC21,” Version 1.1, OLAC, accessed March 1, 2023, <https://cornerstone.lib.mnsu.edu/olac-publications/4>.
82. “MARC Proposal No.: 2013-07: Defining Encoding Elements to Record Chronological Categories and Dates of Works and Expressions in the MARC 21 Bibliographic and Authority Formats,” Library of Congress, accessed March 1, 2023, <https://www.loc.gov/marc/marbi/2013/2013-07.html>.
83. “MARC Proposal No.: 2016-03: Clarify the Definition of Subfield \$k and Expand the Scope of Field 046 in the MARC 21 Bibliographic Format,” Library of Congress, accessed March 1, 2023, <https://www.loc.gov/marc/mac/2016/2016-03.html>.
84. “MARC Proposal No.: 2021-06: Accommodating Work and Expression Dates, and Related Elements, in Bibliographic and Authority Field 046,” Library of Congress, accessed March 1, 2023, <https://www.loc.gov/marc/mac/2021/2021-06.html>.
85. Best Practices for Recording Faceted Chronological Data in Bibliographic Records, Version 1.0 <http://hdl.handle.net/11213/16710>

## Appendix The tangled history of MARC field 046 and work creation dates

The 046 field was originally created to record dates that could not be accommodated in the date fixed fields, such as BCE dates.<sup>79</sup> Additional subfields were added in 2002 for the purpose of recording data about internet resources.<sup>80</sup> Moving image catalogers long wanted a place to unambiguously record the original release date of a film, which is important to users and often unrelated to the date of publication. The existing subfield 046 \$k (Beginning or single date created) was repurposed to meet this need. For some time, it was informally recommended in the audiovisual cataloging community. It was first officially recommended in 2017 in OLAC’s “Best Practices for Cataloging DVD-Video and Blu-ray Discs Using RDA and MARC21.”<sup>81</sup> Meanwhile in 2013, ALA’s SAC Subcommittee on Genre/Form Implementation proposed two new subfields for a related but different use case.<sup>82</sup> These subfields, \$o and \$p, are intended to encode the beginning and ending dates of aggregated content where the genre or form is described using LCSH strings, such as “Operas \$y Eighteenth century.” This supports the Library of Congress’s plan for disaggregating non-topical information traditionally found in LCSH. Later, the MARC documentation was modified to adjust some wording that seemed to conflict with effectively using \$k (date created) to record the date of the work.<sup>83</sup> The original definition stated that dates recorded in \$k could not be recorded elsewhere in the same record. This worked for the original use case for “a data element for creation date not recorded elsewhere” but is incompatible with unambiguously recording the original date of the work when it is the same as the date of the manifestation. Finally, in 2021 the 046 was again modified to incorporate new indicators that distinguish between dates associated with a work and dates associated with an expression.<sup>84</sup>

When aggregates are introduced, dates become even more complicated. In addition to the publication date of the manifestation, there are expression and work dates associated with both the aggregating work and expression and the individual works and expressions that are being aggregated. Although the 046 field has been modified and expanded in order to support distinctions between work and expression dates, as well as between dates associated with the aggregating work and the works being aggregated, this was not designed into the field as it was constructed. The result is not intuitive for catalogers to apply. It has also led to recommended practices changing over time. This evolutionary legacy complicates and compromises the reliability of machine interpretation of this data.

At the time that OLAC began recommending the use of 046 \$k to record the date of the work for moving images, they only anticipated using this data for the date of work of the movie or movies contained in the resource and not for the aggregating work. The OLAC documentation recommended, and continues to recommend, coding dates for individual works that are part of compilations either individually in multiple \$k or as a range in \$o and \$p. This is not a problem, so long as only the dates of the individual works are recorded. However, as interest in the 046 field expanded for other uses, such as music or literature, some catalogers began to record the date of the aggregating work in \$k and the dates of the aggregated works in \$o and \$p. This meant that \$k began to be used for two purposes. If there is only a single work, the date of the work is recorded in 046 \$k. However, in the case of an aggregate if a cataloger wants to record the original date of the aggregating work, 046 \$k will contain the date of the aggregating work. Users are likely to find it confusing to have these two types of dates mixed up in a single facet. A user seeking twentieth century poetry will probably be unhappy if an anthology of seventeenth century poetry published in 1995 comes up in their results.

In response to this problem, the ALA SAC Subcommittee on Faceted Vocabularies recommended always coding the dates of aggregated works and expressions in \$o and \$p, even if only a series of single dates rather than a range is being recorded and even if the date of the aggregating work is not being recorded.<sup>85</sup> However, two unresolved problems remain. For a date of creation of the work facet that only includes the contents of resources and not aggregating works, \$k should only be included conditionally, which requires preprocessing tools that not all systems have. In addition, if catalogers ever use \$k for an aggregating work without corresponding \$o for the aggregated works, there is no possible logic to distinguish it from the date of creation of a single work. Specific systems may have additional limitations. For example, it is not possible to make the necessary logic work in Primo VE if \$k and \$o are recorded in different instances of field 046.

Table A1 shows common practices and recommendations. Although the dates of aggregated works (\$o and \$p) can be consistently interpreted, dates in \$k and \$l cannot. This greatly increases the preprocessing required to generate a coherent facet. In order to exclude the dates of aggregating works, it is necessary to only include single dates or ranges of dates of creation (\$k and \$l) when the dates of any aggregated works (\$o and \$p) are not present. The ability to do this depends on the affordances of a particular system. For example, in Primo VE, this is possible if all the relevant subfields are in the same instance of field 046. However, it turns out to be impossible to do this in the situation given on line four of the table where \$k and \$o are in separate instances of field 046. It is also impossible to distinguish the situation where only a single date of creation is reported in 046 \$k (line 2 of the table) and the situation where only the date of the aggregating work is recorded and thus only 046 \$k exists (line 6 of the table). It might almost be better to abandon field 046 for this purpose and record numeric dates in 388 where the distinction between aggregating and aggregated works is made more clearly. It would be necessary to add an indication of whether the date or dates apply to a work or expression to field 388, though.

**Table A1.** Comparison of 046 original date coding practices for a single work and an aggregate.

	OLAC best practices	Common practice when including aggregating work	SSFV recommendations
Single work			
Date of work: 1995	\$k 1995	\$k 1995	\$k 1995
Aggregating work (compilation): 2023			
Aggregated work: 1983			
Aggregated work: 1995			
Aggregated work: 2008			
Work dates in an aggregate recorded separately			
	\$k 1983	\$k 2023	\$k 2023
	\$k 1995	\$o 1983	\$o 1983
	\$k 2008	\$o 1995	\$o 1995
		\$o 2008	\$o 2008
Work dates in an aggregate recorded as a range			
	\$o 1983	\$k 2023	\$k 2023
	\$p 2008	\$o 1983	\$o 1983
		\$p 2008	\$p 2008
Aggregated works only			
	\$k 1983	\$k 1983	\$o 1983
	\$k 1995	\$k 1995	\$o 1995
	\$k 2008	\$k 2008	\$o 2008
Aggregating work only			
		\$k 2023	\$k 2023