CEUR-WS.org/Vol-2836/qurator2021_paper_9.pdf

# Towards Genre Classification in the Library Catalog

Andreas Lüschow[0000−0002−9287−0286] and José Calvo Tello[0000−0002−1129−5604]

Niedersächsische Staats- und Universitätsbibliothek Göttingen
lueschow@sub.uni-goettingen.de

**Abstract.** Library catalogs usually do not contain explicit information about the genre of literary works. However, both for readers and for researchers this information can be useful, e.g, when it comes to creating research corpora based on genre. In this proposal, we offer a first analysis about how genre information has been integrated in catalogs to date, train several algorithms to classify annotated instances and contrast the results to the fields from the library catalogs used as features. Our study is based on data sets from the union catalog (GVK) of the German *Common Library Network* (GBV).

**Keywords:** library catalog · machine learning · genre analysis · digital humanities · bibliographic data science.

## 1 Literary Genre and its Computational Approach

Genres are frequently used as categories to structure literary collections of texts. This is observable in the way bookstores are organized or how publishers market their products. But these categories do not only mediate in the economical process, they also assist in the writing and reading process. Writers often take actively a decision whether they are producing a novel, and if so, if it is, for example, a historical novel [20]. This is also taken into consideration by readers who expect a series of characteristics from the genres [33].

Genres have played a central role in academic circles for many centuries, frequently in a basic three-part schema, most often populated with the categories drama, epic and lyric [14]. For more specific categories, sceptic opinions are frequently hold [9,10]. However, numerous projects use them to define their research object. Let us take three large projects of the past years in different European countries: The Poetry Standardization and Linked Open Data[1] (at the UNED University in Madrid), The Riddle of Literary Quality[2] (Huygens Institute for the History of the Netherlands), and the Distant Reading for European Literary History[3] (coordinated from the University of Trier). None of them had the aim of analyzing genres, but all three use genres to define their research object: poetry in the first case, novels in the latter two.

[1] https://cordis.europa.eu/project/id/679528

[2] https://literaryquality.huygens.knaw.nl/

[3] https://www.distant-reading.net/

Copyright © 2021 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

The ubiquity of genres is also observable in literary corpora. Many corpora use genres as a constant that constitutes a basic criterion for their composition: dramatic texts in DraCor [13], novels in the ELTeC corpus and the Litbank [1]. In the case of the Textbox [27], each subcorpus also contain a single main genre. In more generic corpora such as the German TextGrid Repository [26] or the Spanish CORDE [32] genres are variable and therefore this information is integrated as metadata for each text.

The interest in genres has also been considered by many computational approaches from Computer Science, Computational Linguistics, and Digital Humanities [3,18,28,25,24,17,16,8,34,7]. These works have been applying mainly supervised methods to mono-lingual corpora using linguistic cues from the full text, such as the frequency of the tokens or linguistic annotation. When multi-label classification is applied, the results tend to show relatively high scores, even when the inter-annotator agreement tends to be low [2,8].

Despite this relevance for authors, readers, publishers and researchers, literary genres are seldom found in library catalogs, as we will explain in the following section.

## 2 Genre in Bibliographic Data

Today, library catalogs do not only serve as a means of finding and identifying books held in a single library. Union catalogs such as the GVK (union catalog of the German *Common Library Network* (GBV)) cover stocks of many different libraries, including all types of media. Over the years an immense amount of mostly intellectually generated (i.e., human-annotated) metadata was collected that can be used far beyond its original purpose of registering and describing books for local access [15]. Methods from computer science can be applied to large data sets of bibliographic data to gain valuable insights into the inner construction of the data, allowing for building tools to automatically structure, group, enhance and enrich data. In the age of the rise of Digital Humanities, libraries more and more see the urge and the value of opening their data silos [31] and thinking about making their data more useful for researchers.

Quantitative approaches to using collections of bibliographic records[4] as research material have been studied to a lesser extent than, for example, the analysis of full text resources. All too often, they are considered a “mere retrieval tool” [19], useful in finding, identifying, and gaining access to resources, but not so much as a resource by itself. Although library catalogs allow for—as Bubenhofer and Rothenhäusler put it—new text arrangements and operations that are not possible by solely looking at the texts “themselves” [6].

Bibliographic records can be seen as big data [23]. Because sophisticated cataloging standards and conventions exist in librarianship, one may come to the assumption that these data are consistently structured and largely homogeneous. However, this is not the case; bibliographic data are the results of multi-layered

[4] Hereafter, we refer to catalog data sets as *bibliographic data* or *bibliographic records*, respectively.

historical processes [19] and shaped mainly due to different cataloging standards applied over time [23,36].

Lahti et al. [19] refer to the kind of research that is based on catalog data as *bibliographic data science* (BDS). BDS can easily be seen as part of a larger process called *Bibliomining* [21] in which libraries use their large data corpora to generate findings and new knowledge about their central field of competence. Recently, Wallbank et al. [35] illustrated the complexity of bibliographic data science by pointing out the need for extensive data cleaning—even if the data extraction from the catalog follows very clear and distinct rules. Or, as Suarez [30] phrased it: "it is salutary to calibrate the instrument we are using".

Studies focusing on genre information in bibliographic data are sparse and allude to different aspects of the representation of genre, ranging from the possibilities to use genre facets in user interfaces [12] to the significance of genre/form aspects in American cataloging standards over time [36]. In the *Functional Requirements for Bibliographic Records* (FRBR), on which modern cataloging standards such as *Resource Description and Access* (RDA) are based, the concept of genre is one of the primary characteristics in distinguishing one work from another. However, genre was not clearly defined in the library sector to date [12] and its recording is not obligatory in most cases.

Most of the studies related to BDS emphasize the importance of genre, especially for making access to collections easier and (re)use of bibliographic data more common. This point of view is already reflected in relevant documents, for example in version 5.0 of the *Metadata Application Profile* (MAP) of the *Digital Public Library of America* (DPLA), where a list of values for the *edm:hasType* property was compiled based on "areas of interest for researchers", amongst others [11]. The values from this list are used in cataloging practice, thus adding to a more uniform and widespread usage of genre attributes in bibliographic data.

## 3 Data Set, Genre Labels and Features

In this section we describe our approach to enriching library catalogs with genre information based on machine learning using already labeled data sets for classification.[5]

To compile our data set for classification we use keywords inside bibliographic records and their relations to authority records. We extract all subject headings from the German national authority file *Gemeinsame Normdatei* (GND)[6] that are classified in GND's own taxonomy [29] as "Literaturgattung" (literary genre). This yields a total of 1319 different genre labels. For each genre, we save its label, GND identifier and union catalog identifier. The union catalog identifier is then used to retrieve all catalog records from the GVK associated with this keyword using the *Search/Retrieve via URL* (SRU) protocol[7] via the SRU-API offered by

[5] We also conducted experiments with unsupervised learning (clustering) but for this short paper we focus only on the classification results.

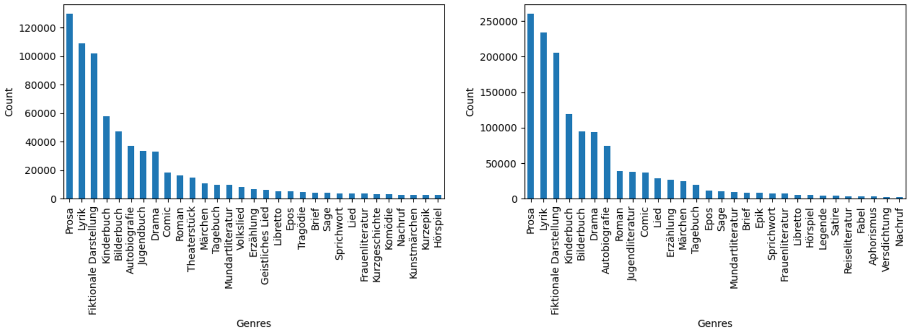
[6] https://www.dnb.de/DE/Professionell/Standardisierung/GND/gnd.html

[7] http://www.loc.gov/standards/sru/

the GBV[8]. In total, we are able to download 746,786 records. All these records are saved into a single data table, where each catalog field (i.e., author, title, keywords etc.) is represented by a single column.

For this first attempt—and to limit both the difficulties of the heterogeneity of the data and the computational costs—, instead of using the precise values in each catalog field as features we decide to use only binary features. This leads to a table where for each record and each field either a “1” or a “0” denotes the presence of information in this field for this record. Our hypothesis is that the structure of a record can lead to indirect conclusions about the media type, the publication format, the genre, or at least the cataloging standard used. To a certain degree, the algorithm should be able to “re-create” the cataloging rules based on the data fields used for a record, but also identify implicit rules for the genre assignment.

Out of 1319 literary genre labels from the authority file, only 1150 have at least one associated bibliographic record in the union catalog; most genres are related to very few records (mean: 612 records; median: 10; standard deviation: 6114.8; IQR: 49), and only 31 genres (2.6 %) are used in more than 2500 records (cf. Fig. 1a).



(a) original data

(b) main genres

Fig. 1: Number of catalog records by genre (considering only genres with more than 2500 assigned records)

Genre information can be found in the catalog in a very scattered manner. In fact, over 15 fields in the catalog records contain information related to genre so it is not a trivial task to determine the genre a record belongs to. After identifying these 15 fields we find that 169,610 out of 746,786 records have multiple genres assigned, i.e., 77.29 % of the records have a single genre assigned. Even if other more flexible theoretical models have been applied to explain the relations between instances and genres [25,16,8], in this work we remain in the more sim-

[8] https://verbundwiki.gbv.de/display/VZG/SRU

ple model that each record is associated with only one genre. Thus, we decide to keep only the first occurring genre in the data if multiple genres are available for a record. Additionally, many genres from the authority file are rather too specific which leads to very few records assigned to them, e.g., "Universitätsroman" (college novel). We thus add a column to the data table in which each record's original genre is mapped to a manually compiled list of 396 main genres (i.e., "college novel" will just become "novel"). However, the general distibution of records across these main genres remains stable, cf. Fig. 1b, showing now 30 genres with more than 2500 records.

Finally, to keep our training data manageable, we further observe only those records that belong to the top 25 genres.[9] We also ignore all catalog fields (i.e., columns in our data table) that are present in less than 0.1 % of the records. Out of 1944 fields available in total, 870 remained (44.75 %). Based on these 870 features, we try to predict the one single genre label for each.

## 4 Results and Evaluation

Using the Python programming language and its machine learning library *scikit-learn* [22] we apply three classification algorithms to our data: Logistic Regression, Random Forest [5] and Decision Tree [4]. For all runs the same 10 % of the data are used as a test set to evaluate the models.[10] We evaluate the model performance on both the original genres ("all") and our manually mapped main genres ("main").

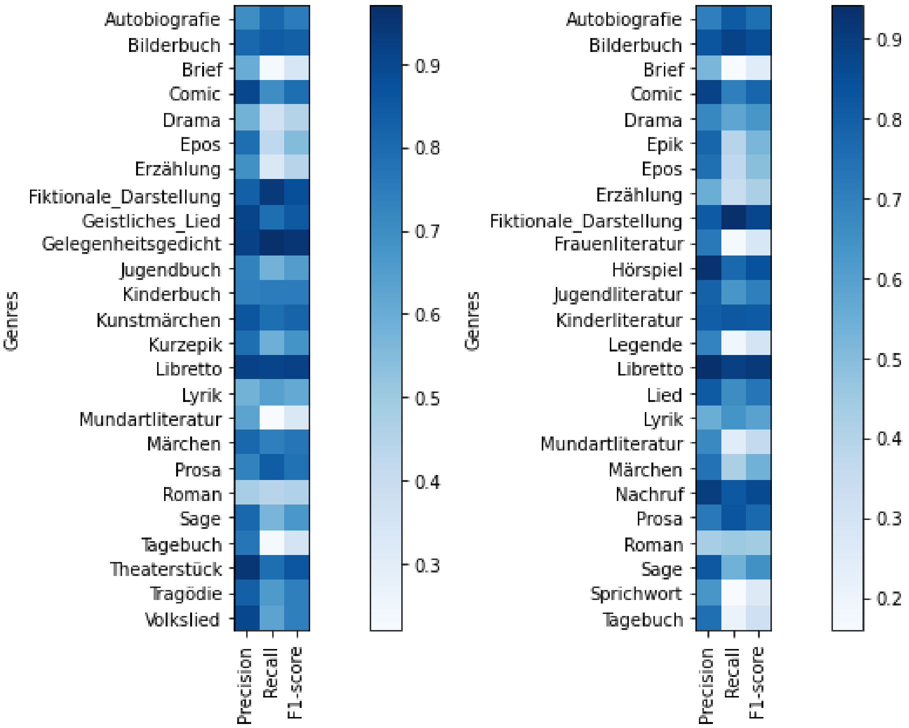
Table 1: Classification evaluation results

| **Algorithm** | **Accuracy** | **Precision** | **Recall** | F1 **score** |
|---|---|---|---|---|
| Logistic Regression (all) | 0.62 | 0.60 | 0.62 | 0.60 |
| Logistic Regression (main) | 0.60 | 0.57 | 0.60 | 0.57 |
| Random Forest (all) | **0.73** | **0.73** | **0.73** | **0.72** |
| Random Forest (main) | **0.71** | **0.72** | **0.72** | **0.71** |
| Decision Tree (all) | 0.66 | 0.65 | 0.66 | 0.65 |
| Decision Tree (main) | 0.65 | 0.65 | 0.65 | 0.65 |

[9] This is an artificial limitation that allows for reasonable computational costs while keeping enough analyzable data. At this point, we also have to place emphasis on the fundamental problem that these top 25 genres are not on the same conceptual level, containing for example "Roman" (novel) as well as "Prosa" (prose) and "Fiktionale Darstellung" (fiction). This makes especially these broad categories very heterogeneous and their constitution nondistinctive.

[10] All data and source code are available at https://github.com/alueschow/qurator21_towards_genre.

As Table 1 shows, the less detailed main genres yield slightly worse predicitions compared to when no genre mapping is used, maybe due to the more heterogeneous and larger genre classes created by the mapping. In both cases, Random Forest models perform notably better.



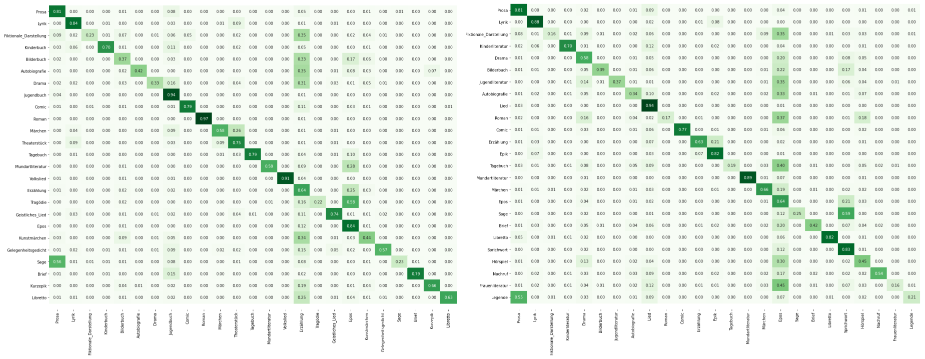
(a) top 25 original genres

(b) top 25 main genres

Fig. 2: Precision, Recall, and F1-Score (Random Forest)

Figure 2 illustrates the model performance for the 25 genres observed, clearly showing that the correct prediction rate highly depends on the individual genres. Generally, specific genres such as “Libretto” have a considerably better outcome than generic genres, e.g., “Frauenliteratur” (women’s literature). However, the genre “Fiktionale Darstellung” (fiction) scores surprisingly well, considering its very generic scope.

The confusion matrix for the winning models visualized in Fig. 3 is based on accuracy values for genre predictions. In both cases—original genres and main genres—the false positives generated by the model seem to be placed predominantly in one to two rather broad genres: “Erzählung” (narrative) and “Epos” (epic) for original genres, “Epos” and “Sprichwort” (proverb) for the main genres. A more detailed analysis on the single genre level may be helpful in determining the reasons behind these assignments, but is beyond the scope of this short paper.



(a) Random Forest (all) (b) Random Forest (main)

Fig. 3: Accuracy of best algorithm for top 25 genres

## 5 Conclusion

In this paper we have identified advancements in library cataloging—understood as enhancements in data quality and access, but primarily their coverage—as being fundamental for the widespread use of bibliographic records as a research source by itself. The use of library catalog data sets is hitherto uncommon, albeit first examples of *bibliographic data science* (BDS) start to take soundings on how this data can be most effectively obtained, processed, and analyzed.

Studying the example of classifying genre information based on bibliographic data we were able to show that the application of well-established algorithms (Logistic Regression, Random Forest, Decision Tree) already leads to a robust performance. We used the existence of single data fields as binary features for 746,786 data sets and observed the most frequent 25 genre labels for both the original genre assignments and a manually generated mapping to more broadly defined genres. Our results indicate that keeping the detailed genre information may be beneficial for classification performance.

In the near future, we would like to take into consideration other aspects related to the genre labels (e.g., different conceptual levels of genres, deeper analysis of individual genres, multi-label classification, artificially created distinctive genres), the features (feature engineering, feature analysis, multilinguality of the features), resources used for training and the methods used. In any case, in this first step we have systematically analyzed genre labels in bibliographic data. Our hope is that we can awaken the interest in libraries to make their catalogs more useful for (literary) research through the categories of literary genre.

## References

1. Bamman, D.: LitBank: Born-Literary Natural Language Processing. In: Johnson, J.M., Mimno, D., Tilton, L. (eds.) Computational Humanites, Debates in Digital Humanities (2020, preprint), https://people.ischool.berkeley.edu/~dbamman/pubs/pdf/Bamman_DH_Debates_CompHum.pdf
2. Berninger, V., Kim, Y., Ross, S.: Building a Document Genre Corpus: a Profile of the KRYS I corpus. In: BCS-IRSG Workshop on Corpus Profiling. London (2008), http://eprints.gla.ac.uk/47699/1/ewic2_ir08_s1paper2.pdf
3. Biber, D.: The multidimensional approach to linguistic analyses of genre variation: An overview of methodology and finding. Computers in the Humanities **26**(5-6), 331–347 (1992)
4. Breiman, L., Friedman, J., Olshen, R., Stone, C.J.: Classification and Regression Trees. Wadsworth, Belmont, CA (1983). https://doi.org/10.2307/2530946
5. Breiman, L.: Random Forests. Machine Learning **45**(1), 5–32 (2001). https://doi.org/10.1023/A:1010933404324
6. Bubenhofer, N., Rothenhäusler, K.: "Korporatheken": Die digitale und verdatete Bibliothek. 027.7 Zeitschrift für Bibliothekskultur / Journal for Library Culture **4**(2), 60–71 (2016). https://doi.org/10.12685/027.7-4-2-154
7. Calvo Tello, J.: The Novel in the Spanish Silver Age: A Digital Analysis of Genre through Machine Learning. University of Würzburg, PhD Thesis (unpublished)
8. Calvo Tello, J.: Genre Classification in Spanish Novels: A Hard Task for Humans and Machines? In: Data in Digital Humanities. EADH, Galway (2018), https://eadh2018.exordo.com/programme/presentation/82
9. Croce, B.: Estetica come scienza dell'espressione e linguistica generale. Sandron, Milano, Italy (1902)
10. Derrida, J.: The Law of Genre. Critical Inquiry **7**(1), 55–81 (1980)
11. Digital Public Library of America: Metadata Application Profile, version 5.0 (2017), https://drive.google.com/file/d/1fJEWhnYy5Ch7_ef_-V48-FAViA72OieG/view
12. Dragon, P.M.: Form and Genre Access to Academic Library Digital Collections. Journal of Library Metadata **20**(1), 29–49 (2020). https://doi.org/10.1080/19386389.2020.1723203
13. Fischer, F., Trilcke, P., Jennifer Beine, J., Orekhov, B.: Dracor (2018), https://dracor.org/
14. Genette, G.: Genres, types, modes. Poetique **32**, 389–421 (1977)
15. Gorman, J.: A Systems Librarian's Cataloging Daydream. In: Sanchez, E. (ed.) Conversations With Catalogers in the 21st Century, pp. 73–94. Libraries Unlimited, Santa Barbara, California (2011)
16. Henny-Krahmer, U., Betz, K., Schlör, D., Hotho, A.: Alternative Gattungstheorien: Das Prototypenmodell am Beispiel hispanoamerikanischer Romane. In: DHd 2018: Kritik der digitalen Vernunft. Konferenzabstracts. pp. 105–112. Köln (2018), http://dhd2018.uni-koeln.de/wp-content/uploads/boa-DHd2018-web-ISBN.pdf
17. Hettinger, L., Reger, I., Jannidis, F., Hotho, A.: Classification of Literary Subgenres. In: DHd 2016: Modellierung–Vernetzung–Visualisierung. Die Digital Humanities als fächerübergreifendes Forschungsparadigma. Konferenzabstracts. pp. 154–158. Leipzig (2016), http://dhd2016.de/boa.pdf
18. Kessler, B., Numberg, G., Schütze, H.: Automatic detection of text genre. In: Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for

Computational Linguistics. pp. 32–38. ACL '98, Association for Computational Linguistics, Stroudsburg, PA, USA (1997). https://doi.org/10.3115/976909.979622

19. Lahti, L., Marjanen, J., Roivainen, H., Tolonen, M.: Bibliographic Data Science and the History of the Book (c. 1500–1800). Cataloging & Classification Quarterly **57**(1), 5–23 (2019). https://doi.org/10.1080/01639374.2018.1543747
20. Lukács, G.: Der Historische Roman. Aufbau-Verlag, Berlin (1955)
21. Nicholson, S.: The basis for bibliomining: Frameworks for bringing together usage-based data mining and bibliometrics through data warehousing in digital library services. Information Processing  Management **42**(3), 785–804 (2006). https://doi.org/10.1016/j.ipm.2005.05.008
22. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al.: Scikit-learn: Machine learning in Python **12**, 2825–2830 (2011)
23. Prescott, A.: Bibliographic Records as Humanities Big Data. In: Hu, X., et al. (eds.) 2013 IEEE International Conference on Big Data. pp. 55–58 (2013). https://doi.org/10.1109/BigData.2013.6691670
24. Riddell, A., Schöch, C.: Progress through Regression. In: Digital Humanities 2014: Conference Abstracts. UNIL/EPFL, Lausanne (2014), http://dharchive.org/paper/DH2014/Paper-60.xml
25. Santini, M.: Automatic Identification of Genre in Web Pages: A new perspective. LAP Lambert Academic Publishing, Saarbrücken (2011)
26. Schmunk, S., Funk, S.E.: Das DARIAH-DE- und das TextGrid-Repositorium: Geistes- und kulturwissenschaftliche Forschungsdaten persistent und referenzierbar langzeitspeichern. Bibliothek Forschung und Praxis **40**(2) (2016). https://doi.org/10.1515/bfp-2016-0020
27. Schöch, C., Calvo Tello, J., Henny-Krahmer, U., Popp, S.: The CLiGS textbox: Building and Using Collections of Literary Texts in Romance Languages Encoded in XML-TEI. Journal of the Text Encoding Initiative (Rolling Issue) (2019). https://doi.org/10.4000/jtei.2085
28. Stamatatos, E., Fakotakis, N., Kokkinakis, G.: Automatic Text Categorization in Terms of Genre and Author. Computational Linguistics **26**(4), 471–497 (2001)
29. Stella, K., Scheven, E. (eds.): GND-Systematik: Leitfaden zu ihrer Vergabe. Deutsche Nationalbibliothek, Leipzig (2011)
30. Suarez, M.: Towards a bibliometric analysis of the surviving record, 1701–1800. In: Suarez, Michael F., S., Turner, M.L. (eds.) The Cambridge History of the Book in Britain: Volume 5: 1695–1830, pp. 37–65. Cambridge University Press (2009). https://doi.org/10.1017/CHOL9780521810173.003
31. Suominen, O., Hyvönen, N.: From MARC silos to Linked Data silos? o-bib **4**, 1–13 (2017). https://doi.org/10.5282/O-BIB/2017H2S1-13
32. Sánchez Sánchez, M., Domínguez Cintas, C.: El banco de datos de la RAE: CREA y CORDE. Per Abbat: boletín filológico de actualización académica y didáctica (2), 137–148 (2007)
33. Todorov, T.: The Origin of Genres. New Literary History **8**(1), 159–170 (1976)
34. Underwood, T.: Distant Horizons: Digital Evidence and Literary Change. University of Chicago Press, Chicago (2019). https://doi.org/10.7208/9780226612973
35. Wallbank, S., Kane, D.A., Dickerson, M., Hutchinson, J.: Exploring Bibliographic Records as Research Data. Catalogue and Index **197**,  3–9 (2019)
36. Zhang, L., Lee, H.L.: The Role of Genre in the Bibliographic Universe. Advances in Classification Research Online **23**(1), 38–45 (2012). https://doi.org/10.7152/acro.v23i1.14236