Metadata Desiderata for Literary Corpora: A survey for starting a conversation between literary studies, libraries and research data repositories

1. Introduction

Defining the criteria for specific corpora is one of the primary tasks for literary scholars. For example, researchers composing literary corpora of female novelists in France from the 20th century, would like to find these metadata for works and authors in catalogs, repositories and authority file resources.

For this reason, we ran a survey on the importance of specific categories for literary studies (cfr. Cox et al., 2019; Swauger and Vision, 2015; Schopf and Newhouse, 2007). This survey was prepared as a joint effort by the Text+ Consortium, the priority program Computational Literary Studies (SPP 2207) in Germany and the European project Computational Literary Studies Infrastructure (CLS INFRA). In its preparation, the feedback of several other scholars unrelated to the previously mentioned projects has been considered.

The results could be used for infrastructure institutions (libraries, institutions curating authority files) and the mentioned consortia to prioritize their efforts and resources such as related to research data repositories (Swauger and Vision, 2015; Strecker, 2022) as well as to search for new methods to provide the missing information that the community requires.

For example, the German consortium Text+ (part of the NFDI) emphasizes the necessity of enriching metadata based on the researchers’ interests (Hinrichs et al., 2022). Authority files play a crucial role, which is the reason why the German National Library (DNB) seeks paths to enhance its Integrated Authority File (GND) through different methods (new GND agency, entityXML; Kett et al., 2022). Within Text+, several features are being developed for the TextGrid Repository for improving the metadata quality and linking it to online resources (Calvo Tello et al., 2023), while researchers in the CLS-Infra project published data from the ELTeC corpus on Wikidata (Nešić et al., 2022).

Furthermore, the results can be also interesting for different research communities, in addition to literary scholars such as library and information scientists or for researchers of bibliographic metadata (Király, 2017; Umerle et al., 2022).

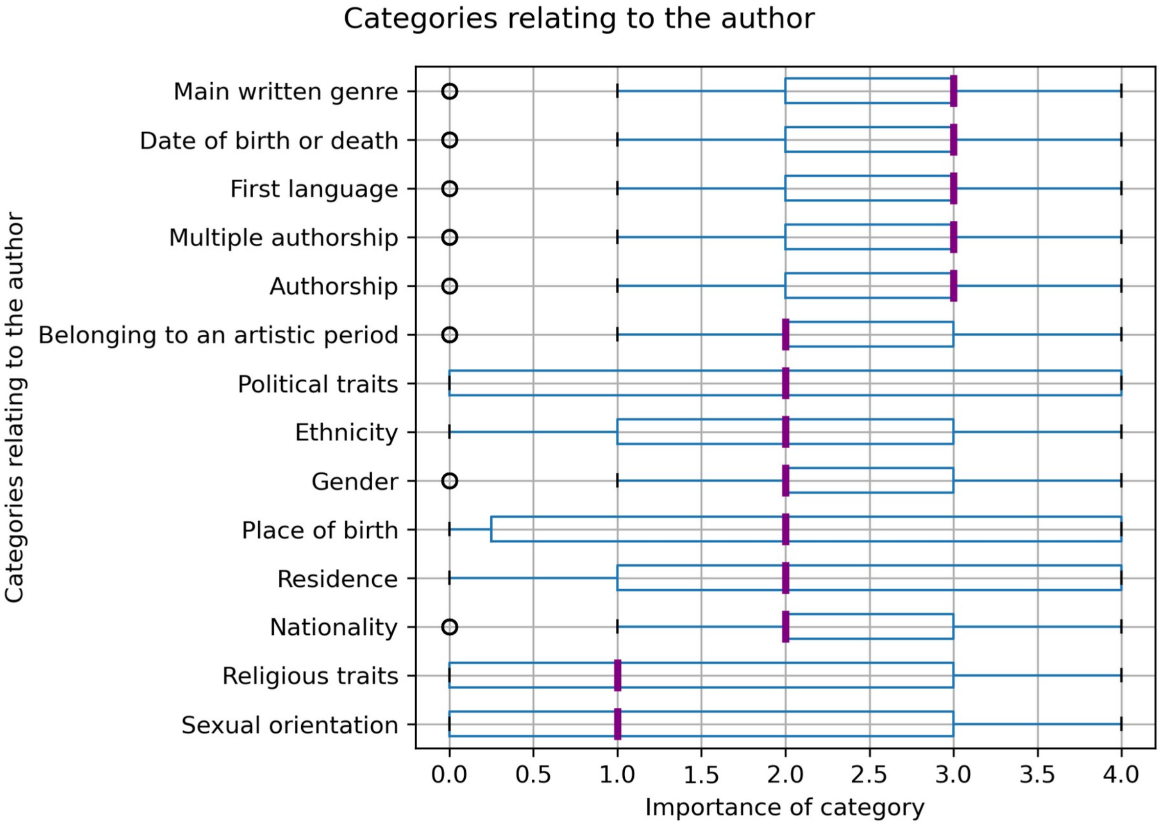
2. Survey Design

The survey is structured in three main sections: questions about author, work and text. This resembles a simplified version of the librarianship model FRBR (Taylor, 2007) and enables the differentiation between characteristics of the work (first publication, genre, etc.) and the characteristics of the text (actual language, format, etc.). This structure allows us to more easily associate certain categories to certain resources (authority files with work and author; catalogs and repositories with text). The survey also contains a personal information section for knowing more about respondents and to better understand their preferences. The questionnaire has been published in Zenodo (Calvo Tello et al., 2023).

3. Results

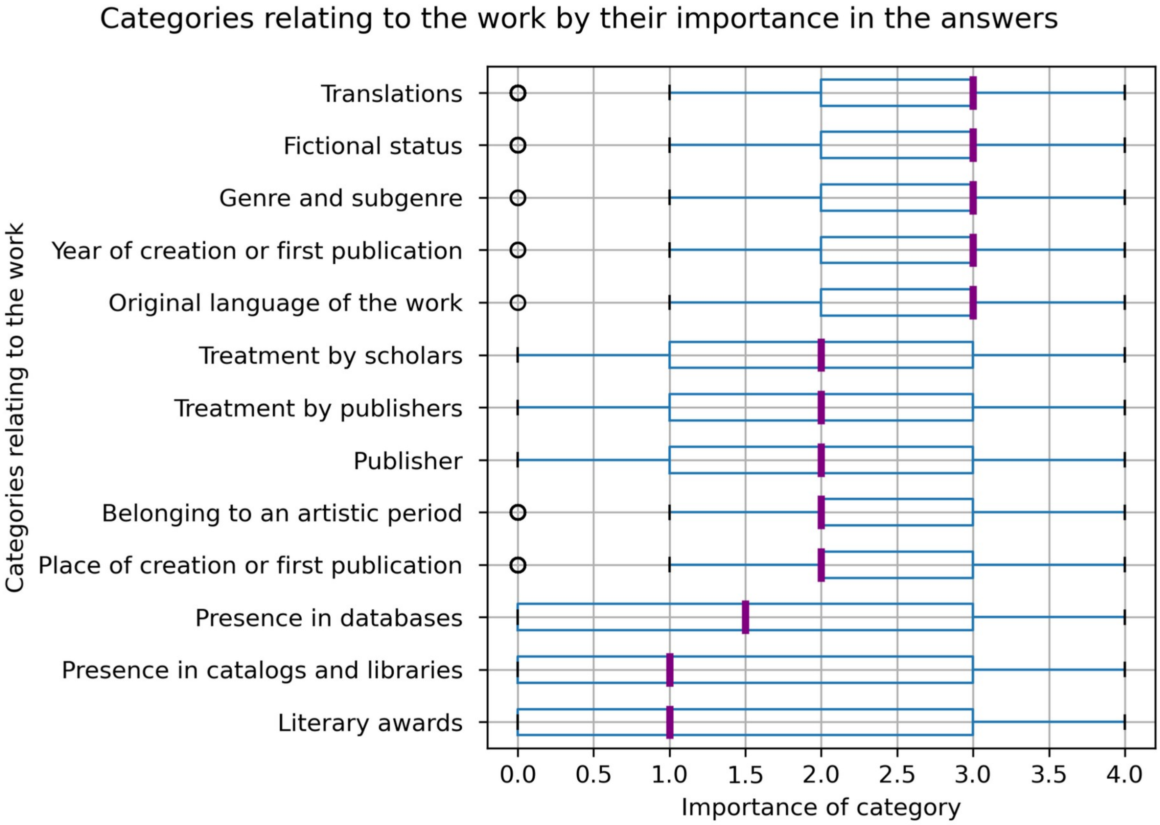
In total, the survey was accessed 317 times with 111 complete answers. In this analysis, we only used the fully completed questionnaires as a basis. Responses came from participants working in 19 different countries, 52% in Germany. The median age of participants was 42; 45% of the participants were female and 53% were above the PhD level (plus 25% with other positions within Academia).

3.1. Author



The results show that authorship (e.g. one-author-corpus) or multiple authorship are important categories, together with basic information such as language or date of birth, and their main genre (i.e. novelists, poets). We expected gender to have a higher score, but it is in the middle with other characteristics such as ethnicity or artistic period. In terms of cultural affiliation, language is more important than nationality, place of residence or birth. Religious characteristics or sexual orientation received the lowest scores.

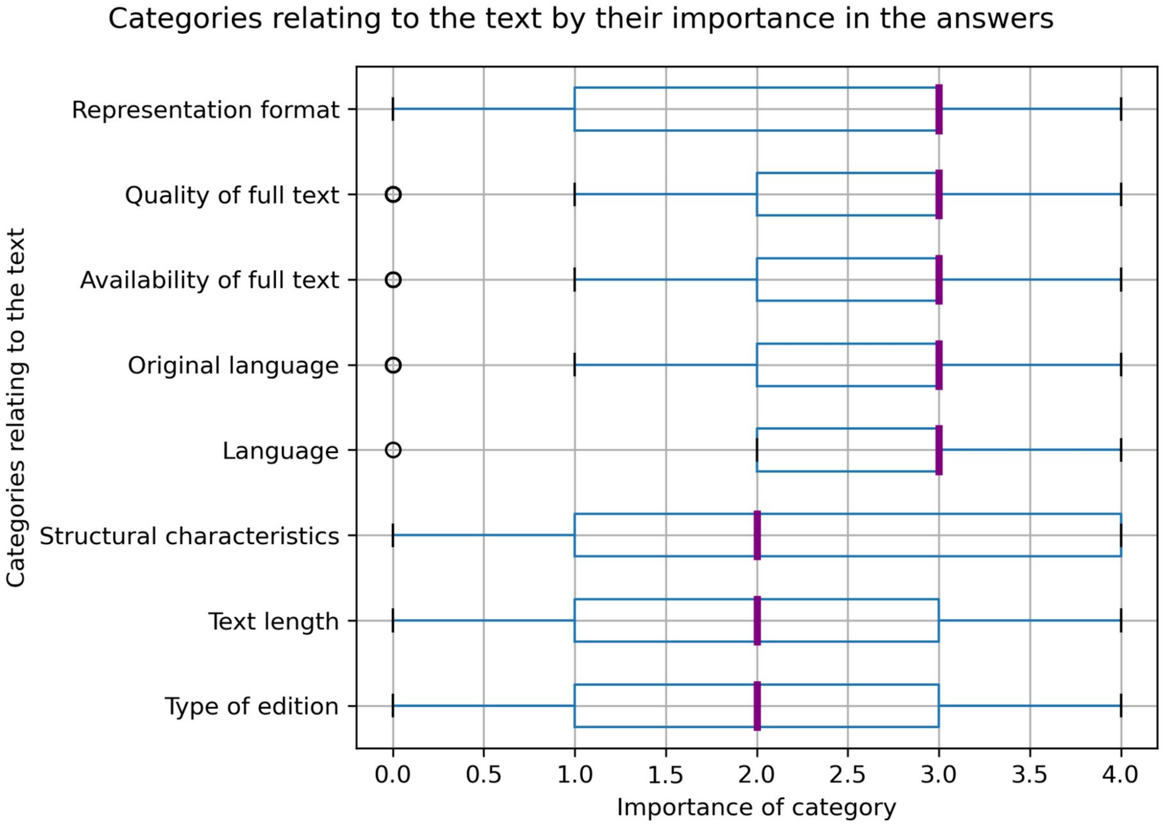
3.2. Work



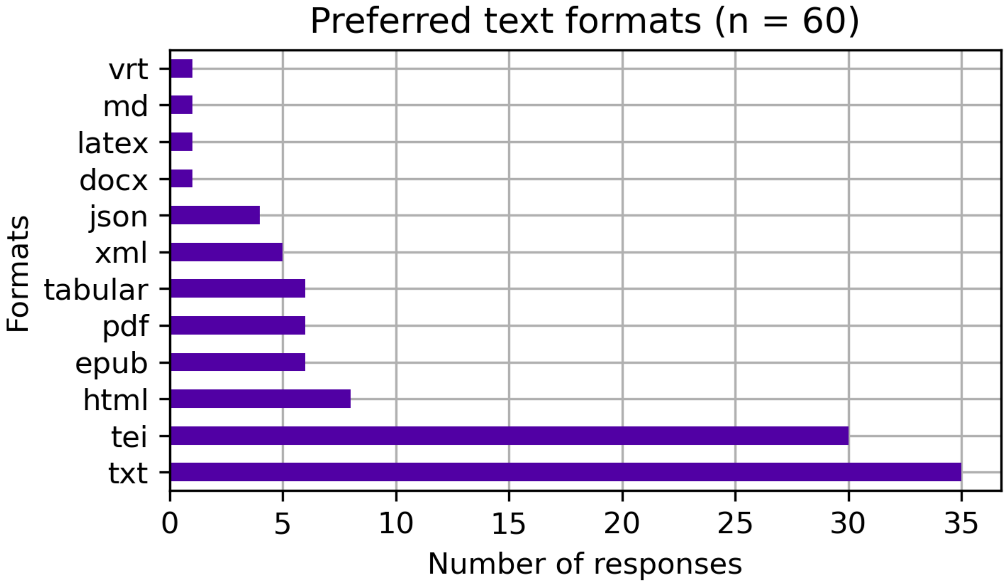
The most important categories relating to the work are associated with basic information (original language, year of creation), translation, or internal aspects such as its fictionality and genre. Treatment by scholars and information about publishers receive higher scores than presence in databases, catalogs, or literary awards.

The high scores in many of these categories are particularly important because library catalogs, authority files, and repositories tend to neglect work level information. For example, the genre of the work is very important for the community, but it is frequently missing in resources.

3.3. Text



Relating to the text, the categories with the highest scores are related to language (first language and actual language), and to the availability, quality and format of the text. Other aspects such as structural characteristics, length or type of edition (critical, philological, non-professional, etc.) are less important, but still relevant.



In an open field, we asked about the preferred format of the text. Two formats show a clear advantage over the rest: plain text (txt) and TEI, while other structured (HTML, ePUB, XML, JSON) and unstructured (PDF, tabular) formats receive less but still notable support. TEI seems to be the format that would satisfy a larger number of researchers. Nevertheless, providing a variety of standard formats would be the aim for digital collections. Similar results were observed in a comprehensive research data management survey in the priority program SPP 2207 (Helling et al., 2022).

4. Outreach - Future Plans

With this survey, the community had the opportunity of expressing their interest in categories of metadata for literary studies. The conversation now moves to the side of research infrastructure institutions and projects, which can use these results to evaluate their resources, metadata models, and actual practices. We are going to discuss questions such as: Of the highest ranked categories, which should be supplied by authority files, by library catalogs, by repositories? Are some categories (such as religious traits or sexual orientation) too sensitive and too personal information about the authors? How can the requested information be created and provided in the first place?

One of the clearest results of the survey is the importance of categories at the work level, such as the first language, the first year of publication, genre or fictionality. In authority files (such as the GND), the coverage of literary works tends to be much worse than authors. Moreover, many resources (like repositories or catalogs) tend not to be connected to work entities in authority files. This poor treatment of works in research infrastructures causes a systematic lack of information for literary studies.

Future actions include the evaluation of resources such as library catalogs, research data repositories (TextGrid-Repository) and authority files (GND) but also starting with testing workflows for data enrichment using Machine Learning methods (cfr. Kokash et al. 2023). The key to solve this problem is to integrate metadata created directly in research projects. Accordingly, further steps will be taken together with infrastructure and other research stakeholders.

References

Calvo Tello, J., Funk, S. E., Göbel, M., Kurzawe, D., Rißler-Pipika, N. and Veentjer, U. (2023). Between Corpora, Tools, and Authority Files: TextGrid Repository for Hispanic Studies. , **8**. (Revista de Humanidades Digitales): 90–108 doi:10.5944/rhd.vol.8.2023.37994.

Calvo Tello, J., Rißler-Pipika, N., Barth, F., Kerstin, J. and Schöch, C. (2023). Questionnaire of the survey: How do you Compose your Literary Corpus or Literary Collection? Questionnaire Zenodo doi:10.5281/zenodo.10203624.

Cox, A. M., Kennan, M. A., Lyon, E. J., Pinfield, S. and Sbaffi, L. (2019). Progress in Research Data Services. *International Journal of Digital Curation*, **14**(1): 126–35 doi:10.2218/ijdc.v14i1.595.

Hinrichs, Erhard, Alexander Geyken, Peter Leinen, Andreas Speer, Regine Stein, Jonathan Blumtritt, Luise Borek, et al. (2022). Text+: Language- and Text-Based Research Data Infrastructure. doi:10.5281/zenodo.6452002.

Kett, J., Kudella, C., Rapp, A., Stein, R. and Trippel, T. (2022). Text+ und die GND – Community-Hub und Wissensgraph. *Zeitschrift Für Bibliothekswesen Und Bibliographie*, **69**(1–2): 37–47 doi:10.3196/1864295020691262.

Király, P. (2017). Measuring completeness as metadata quality metric in Europeana. https://dh-abstracts.library.cmu.edu/works/4058.

Kokash, N., Romanello, M., Suyver, E. and Colavizza, G. (2023). From Books to Knowledge Graphs. *Journal of Data Mining & Digital Humanities*, 2023: 9380 doi:10.46298/jdmdh.9380.

Nešić, M. I., Stanković, R., Schöch, C., and Mihailo Skoric (2022). From ELTeC Text Collection Metadata and Named Entities to Linked-data (and Back). *8th Workshop on Linked Data in Linguistics*. Marseille: LREC, pp. 7–16 http://www.lrec-conf.org/proceedings/lrec2022/workshops/LDL/pdf/2022.ldl2022-1.2.pdf.

Schopf, J. M. and Newhouse, S. (2007). User Priorities for Data: Results from SUPER. *International Journal of Digital Curation*, **2**(1): 149–55 doi:10.2218/ijdc.v2i1.23.

Strecker, D. (2022). Quality of metadata describing research data and the influence of repository characteristics. *Young Information Scientist*, **7** doi:10.25365/yis-2022-7-2. https://yis.univie.ac.at/index.php/yis/article/view/7595.

Swauger, S. and Vision, T. J. (2015). What Factors Influence Where Researchers Deposit their Data? A Survey of Researchers Submitting to Data Repositories. *International Journal of Digital Curation*, **10**(1): 68–81 doi:10.2218/ijdc.v10i1.289.

Taylor, A. G. (ed). (2007). *Understanding FRBR: What It Is and How It Will Affect Our Retrieval Tools*. Westport, Conn: Libraries Unlimited.

Umerle, T., Colavizza, G., Herden, E., Jagersma, R., Király, P., Koper, B., Lahti, L., et al. (2022). *An Analysis of the Current Bibliographical Data Landscape in the Humanities. A Case for the Joint Bibliodata Agendas of Public Stakeholders*. Zenodo doi:10.5281/zenodo.6559857.