Three Use Cases for the ARK Metadata

User 1: Study on format, language and genre of books, 1500-1950

As a researcher, I would like to investigate the relationship between book size (the format), the language of publication and the genre of these books in the period 1500 to 1950. I am primarily interested in whether there are statistically significant connections (correlations) between the format and the language or between the format and the genre of the books. I would like to analyse the number of publications over time using the three main criteria and place them in the wider historical context. My working hypothesis is that at the beginning of the period under consideration, large-format publications in folio format in Latin and in academic genres (e.g. dissertations, treatises) dominate, while the quantitative focus shifts in the course of the 19th century towards small-format German-language publications (octavo format), which are more entertaining in character (German literature: novels, drama, poetry). I am more interested in general trends in book history. The shift away from Latin as the language of publication and the rise of High German and regional dialects can be read in the context of nation-building. It is therefore not so important to look exhaustively at all publications in a given period and to verify or standardise the identity of original works on the basis of author and title details, but rather to identify trends and patterns.

Basically, I need the usual metadata fields author (*3000 Person Familie als 1. geistiger Schöpfer; 3001 + 3002 2. und weitere Verfasser; 3010 Person/Familie als 2. und weiterer geistiger Schöpfer*), title (*3210 Werktitel und sonstige unterscheidende Merkmale des Werks; 3211 Weiterer Werktitel und sonstige unterscheidende Merkmale; 4000 Haupttitel, Titelzusatz, Verantwortlichkeitsangabe; 4150 Gesamttitel der mehrteiligen Monografie; 4213 Frühere/frühester Haupttitel*), subtitle (*3260 Abweichender Titel (Sucheinstieg); 4002 Paralleltitel, paralleler Titelzusatz, parallele Verantwortlichkeitsangabe; 4010 Weitere Titel etc. bei Zusammenstellungen; 4011 Titelzusätze und Verantwortlichkeitsangabe zur gesamten Vorlage; 4212 Abweichender Titel*), place of publication (*1700 Code für Erscheinungsland, 4035 Weiterer und oder früherer Verlagsort und Verleger; 4040 Normierter Ort; 4043 Drucker, Verleger oder Buchhändler (bei Alten Drucken*); *4050 Verbreitungsort in normierter Form*) and date of publication (*1100 Erscheinungsdatum / Entstehungsdatum; 1108 Copyright-Datum, Vertriebsdatum, Herstellungsdatum*), but also information on the format (*4062 Format, Maße*) and scope (*4060 Umfang*), the language of the text and the language of the original (in the case of a translation) (*1500 Sprachcodes; 4221 Angaben über Sprache und Schrift der Expression; 4248 Beziehungen zwischen Sprachausgaben (Expressionsebene)*), as well as the genre (*1131 Art des Inhalts; 4204 Hochschulschriftenvermerk; 4207 Inhaltliche Zusammenfassung; 5010 DDC Notation; 5025 Gattungsbegriffe (DNB); 5030 LCC-Notation; 5050 Sachgruppen der Deutschen Nationalbibliografie bis 2003; 5090 Regensburger Verbundklassifikation (RVK); 5100-5199 Schlagwortfolgen (DNB und Verbünde); 5200-5210 STW-Schlagwörter; 5249 ZBW-Schlagwörter -Veröffentlichungsart; 5301 Basisklassifikation; 5400-5410 DDC-Notation; 5500 LoC Subject Headings; 5550-5559 Schlagwortfolgen (GBV, SWB, K10plus); 5570 Gattungsbegriffe bei Alten Drucken; 5580-5588 Einzelschlagwörter; 6000-6099 Lokale Notationen (lokal verwendete Systematiken); 6300-6399 Lokale Notationen (auf bibliographischer Ebene); 6500-6599 Lokale Schlagwörter (lokal verwendete Regelwerke); 6700-6799 Lokale Notationen; 6800-6899 Lokale Schlagwörter*), which may be derived via fields on the type of content and keywords if necessary. Title details are important because their length correlates with the format of the books (large book => long title is possible; small books => short titles). I would also be interested in information on the length of the books (number of pages) (*4060 Umfang*), which could indicate the development of standardised book lengths ("a novel comprises 400 pages towards the end of the 19th century"); and whether the book was printed by hand (which was common until 1830) or by a mechanical press (*4043 Drucker, Verleger oder Buchhändler (bei Alten Drucken)*), because the connections between nation-building and the emergence of mass literature, especially after the introduction of paper rolls and rotary printing presses in the 1860s and 1870s, are of great importance for the study.

I am familiar with common techniques of data extraction and transformation, otherwise I would not dare to tackle a very extensive data set. I am therefore prepared to prepare a dataset that serves my research interests (data cleaning), while I probably need information on the content of the various existing fields in order to be able to produce derived information if necessary. A perspective that goes beyond the specific research project is the use of digital copies of books from the period under consideration in order to be able to make well-founded statements about the content of certain genres using the existing OCR.

User 2: Study on the paratextual formatting of scientific disciplines, 1800-1950

As a researcher, I would like to write a study on the formatting of scientific publications: The history of science and the formation of presentation standards becomes tangible here in the way scientific books are laid out, and in particular in the paratexts, i.e. the 'companions' to the works. The starting hypothesis is that between 1800 and 1950, presentation routines were established that were normative for each discipline (*6200-6299 Lesesaalsystematik der SBB, containing the ARK-subject index*). According to Gérard Genette, paratexts are all the information that surrounds a text, presents it and controls its reception and consumption. For him, this includes the author's name(s) (*3000 Person Familie als 1. geistiger Schöpfer; 3001 + 3002 2. und weitere Verfasser; 3010 Person/Familie als 2. und weiterer geistiger Schöpfer*), the title (*3210 Werktitel und sonstige unterscheidende Merkmale des Werks; 3211 Weiterer Werktitel und sonstige unterscheidende Merkmale; 4000 Haupttitel, Titelzusatz, Verantwortlichkeitsangabe; 4150 Gesamttitel der mehrteiligen Monografie; 4213 Frühere/frühester Haupttitel*), tables of contents, dedications, mottos, forewords, intertitles, notes, bibliographies and appendices (*0501 Inhaltstyp; 1131 Art des Inhalts; 1140 Veröffentlichungsart und Inhalt; 3433 URL für Volltext und Kataloganreicherung; 4060 Umfang; 4201 Sonstige Anmerkungen; 4203 Zusammenfassende Register; 4207 Inhaltliche Zusammenfassung; 4222 Angaben zu enthaltenen unselbstständigen Werken; 4950 URL zum Volltext; 7124 Einleitender Text, 9000 Inhaltliche Zusammenfassung*). With regard to non-fictional ('scientific') works–Genette was essentially concerned with fiction–illustrations, photographs, tables, charts, maps and indices can also be understood (*3433 URL für Volltext und Kataloganreicherung; 4061 Illustrationsangabe bzw. sonstige physische und technische Angaben; 4063 Begleitmaterial; 4278 Beschreibung des Einbands; 4803 Bestandsschutzmaßnahmen*).

I hope to find references to a large part of these paratexts in the metadata provided by the Staatsbibliothek zu Berlin – Berlin State Library. If not all, then most of the information will be found in this data provided by one of the largest academic research libraries. I would prepare the data myself: Firstly, filter by academic discipline and create separate data sets for each discipline; then convert or quantify the presence or absence of paratexts into numerical values, and finally cluster these combinations of values in order to identify groups of texts within the individual disciplines and thus also to be able to infer a development over time, i.e. to empirically prove the emergence of presentation standards.

Authors do not publish in a vacuum and do not design their texts alone. In addition to the traditions of their discipline, publishers are also involved in the design of the books (*3010 Person/Familie als 2. und weiterer geistiger Schöpfer, sonstige Personen/Familien, die mit dem Werk in Verbindung stehen, Mitwirkende, Hersteller, Verlage, Vertriebe; 3110 Körperschaft als 2. und weiterer geistiger Schöpfer, sonstige Körperschaften, die mit dem Werk in Verbindung stehen, Mitwirkende, Hersteller, Verlage, Vertriebe; 4035 Weiterer und oder früherer Verlagsort und Verleger; 4043 Drucker, Verleger oder Buchhändler (bei Alten Drucken)*). Therefore, basic socio-demographic information on the authors is just as relevant as data on the publishers into whose format specifications the content is pressed. It is therefore also important here to be able to identify the identical presentations within a published series and to track their changes over time. Publishers add genre information, tables of contents or indices, for example, and use their symbolic capital as renowned academic publishers to establish the credibility that academics need in order to be able to convey autobiographical content as truthful, for example. The focus on publishers will hopefully make it possible to carry out correlation analyses (identification of presentation patterns typical of publishers) on the one hand, and to apply time-space analysis techniques from ecology and related fields to the available data on the other.

With regard to the metadata provided by the Staatsbibliothek zu Berlin – Berlin State Library, I do not assume that it contains all the information relevant to the planned study. However, since an extensive corpus of scholarly works can already be found in the digitised collections, the manual collection of paratextual data (e.g. mottos, dedications) from groups of exemplarily selected digitised works should lead to results. In this way, the characteristic patterns brought to light by the cluster analysis can be substantiated, which is conducive to a deeper interpretation.

User 3: Machine learning based on metadata

As a library employee, I would like to develop a recommendation system that, on the one hand, helps library users to find relevant and interesting content based on keywords and keywords entered, and on the other hand, the system generates suggestions for librarians based on the patterns and relationships present in the data for the indexing of historical works from the period 1500 to 1950. The recommendation system should generate personalised recommendations for the users and recommend similar and related works based on the metadata that the users do not yet know. Particularly relevant metadata is author information, which leads to the recommendation of authors from the same time period and with similar characteristics (language, genre, scope of the works, etc.), the titles of the works in order to be able to recommend thematically similar works, and the combinations of different keywords. Overall, the system should lead to increased user satisfaction and a broadening of the diversity of the offer by enabling library users to select works that would be difficult to find without a precise recommendation. A further option is to use the full texts available in the digitised collections to identify and search for thematically similar content, for example via topic modelling.

The other task mentioned above aims to support librarians of other German libraries with historical collections in the indexing of their holdings. A special feature of the metadata provided by the Staatsbibliothek zu Berlin – Berlin State Library is that it not only uses keywords that are generally used (*5010 DDC Notation; 5025 Gattungsbegriffe (DNB); 5030 LCC-Notation; 5050 Sachgruppen der Deutschen Nationalbibliografie bis 2003; 5090 Regensburger Verbundklassifikation (RVK); 5100-5199 Schlagwortfolgen (DNB und Verbünde); 5200-5210 STW-Schlagwörter; 5249 ZBW-Schlagwörter -Veröffentlichungsart; 5301 Basisklassifikation; 5400-5410 DDC-Notation; 5500 LoC Subject Headings*), but also internal markers that were used for the old real-world catalogue (ARK) and that are contained as local notations (*5550-5559 Schlagwortfolgen (GBV, SWB, K10plus); 5570 Gattungsbegriffe bei Alten Drucken; 5580-5588 Einzelschlagwörter; 6000-6099 Lokale Notationen (lokal verwendete Systematiken); 6300-6399 Lokale Notationen (auf bibliographischer Ebene); 6500-6599 Lokale Schlagwörter (lokal verwendete Regelwerke); 6700-6799 Lokale Notationen; 6800-6899 Lokale Schlagwörter*). These subject-specific keyword combinations (law, Eastern Europe, languages, literatures, etc., see also *6200-6299 Lesesaalsystematik der SBB, containing the ARK subject index*) can be very valuable for other libraries so that they can index their historical prints in a similar way as the Staatsbibliothek zu Berlin – Berlin State Library has already done with its holdings. The recommendation system should therefore make it possible to enter a few keywords (e.g. author, title, year) and have relevant keyword combinations displayed.

With regard to the data provided by the Staatsbibliothek zu Berlin – Berlin State Library, I am aware of the limitations that result from the exclusive use of metadata, in particular the lack of context, from which no information about the content, quality or relevance of the work can be derived. In order to counteract this at least to some extent, users should be shown the GND links contained in the metadata of the recommended works so that they can draw their own conclusions about the context. In addition, I will endeavour to make the system's recommendations comprehensible to users or, ideally, to explain them ("explainable AI"). I therefore endeavour to make the recommendation system as transparent as possible so that users can understand how the recommendations were generated.