Accuracy vs. Consistency: A Case Study Assessing Data Quality in Metadata of Early Modern Dissertations

Heßbrüggen-Walter, Stefan
early.modern.thought.online@gmail.com
Universität Münster, Deutschland

As I have discussed in previous work (Heßbrüggen-Walter 2022), the application of digital methods to the analysis of bibliographical metadata must include reflections on the quality of the data we analyse.[1] It can be conceded that in its general form the insight is not really revolutionary: NFDI4Memory, the German research data infrastructure for historical data, has created a whole task area dedicated to questions of data quality (Mainz, Johannes Gutenberg University. n.d.). Still, it might be helpful to contribute to this ongoing discourse using concrete examples of how problems of data quality influence DH analyses in order to contribute to the creation and evaluation of "domain-relevant community standards" (Axton et al. 2016). In the present case this concerns the assessment of the quality of bibliographical metadata, in particular when used as research data in the digital humanities.

This last qualification is essential, because the literature on metadata quality understands it primarily as 'fitness for a purpose' (Király / Brase 2021, 357): the analysis of bibliographical metadata at scale to be undertaken by a digital humanist pursues goals that may differ from other use cases (Király / Brase 2021, 358f). But criteria for the assessment of data quality do not only depend on the goal to be pursued, we should also note that different criteria may at times be in conflict. In the case study I present here this concerns the tension between the 'consistency' of metadata, i. e. the uniformity of mappings between features of a resource and metadata fields, and 'accuracy', the description of the resource in question according to what we can and cannot know about it.

In order to substantiate this claim, I will discuss the question how to characterise the authorship of early modern dissertations within a contemporary metadata schema. The problems I describe arose during the compilation of a dataset if early modern dissertations across disciplines which were published between approx. 1550 and 1800 and preserved in French university libraries and the national library of France. Both the union catalog of French university libraries (SUDOC) and the general catalog of the BNF provide SRU interfaces allowing for the retrieval of limited datasets. Relevant data were retrieved and parsed using Python scripts by the author. The dataset itself still needs to undergo peer review and will hopefully be published in a separate data paper.

Early modern dissertations usually required the participation of contributors fulfilling two different functions, the *praeses* and one or more *respondentes*. While the *praeses* was in most cases a university teacher, e. g. a professor, the *respondens* was in the majority of cases a student. The crux of identifying the authors of early modern dissertations has been succinctly summarised by Joseph S. Freedman:

"The question of determining authorship of these disputations has eluded – and continues to elude – a clear and universally accepted answer. Towards the end of the 19th century, German catalogers began to intensively discuss the question of whether the presider or the respondent should be listed as the author of any given disputation; since then, a number of scholars have addressed this same issue. The following conjecture will be ventured here: there is no simple correct answer, […]." (Freedman 2005, 32f)

From a bibliographical point of view this is a challenging situation, because the contribution of persons listed in a catalog record should in principle be obvious to the user. To this purpose, the UNIMARC standard used in French libraries lists 'relators', codes that clarify the relation between the work in question and the persons listed in the catalog record. But neither *praes id es* nor *respondens* are categories to be found in this list (Willer 2009, 600–12). So how did French librarians deal with this situation?

The preliminary dataset of catalog records for early modern dissertations contains 56316 catalog records, 38958 from the SUDOC catalog, 17358 from the general catalog of the BNF. We find that 12274 SUDOC records (31.51%) and 13796 BNF records (79.48%) list only one person as a contributor. While it is possible to find dissertations without an explicit designation of *praeses* and *respondens*, this is certainly not the rule for early modern dissertations. It seems that in these cases cataloguers simply picked one or the other as the only author of the work.

Limiting our investigation to SUDOC records the summary of how librarians used relator codes for the attribution of contributor roles in the remaining catalog records can be found in table 1.

Table 1: 10 most frequent pairs of 'relator codes' for first and second person in SUDOC catalog records.

Code 1	Code 2	Code Term 1	Code Term 2	Count
70	727	Author	Thesis advisor	18576
70	610	Author	Printer	2574
70	70	Author	Author	2031
70	0	Author	No coding.	1231
70	340	Author	Editor	665
70	555	Author	Opponent	185
70	956	Author	Not resolvable.[2]	177
70	60	Author	Related Individual	164
727	70	Thesis advisor	Author	158
70	280	Author	Dedicatee	125

The coding of persons in catalog records of dissertations can be interpreted as putting forward different implicit theories about who wrote these texts. Two pairings follow the model of the 'sole author', since neither dedicatees nor printers are involved in the production of the original text. The most common solution contrasts an 'author', presumably the *respondens*, and the 'thesis advisor', mirroring the role of the *praeses*, with the one significant difference that modern advisors might not write the thesis they supervise themselves. Some librarians espouse the model of 'co-authorship', naming both contributors as authors. Others contrast an 'author' and an 'opponent', asserting implicitly the authorship of the *praeses*. Finally, a *praeses* (or a *respondens*?) is taken to be an editor of a text prepared by an author. The use of the code '0' or '60' indicates neutrality or lack of verifiable knowledge with regard to the exact contribution of each named person.

In sum, 18919 SUDOC records, namely records featuring author and thesis advisor or author and opponent, allow us to at least infer *praeses* and *respondens*. For the remaining 20039 records this is not possible. This means that only 48.56% of SUDOC records allow for the tentative identification of *praeses* and *respondens*.

We may speculate on what a satisfactory solution of the problem within the confines of UNIMARC could consist in. If we accept Freedman's diagnosis that the contribution of *praeses* and *respondens* to the content of the dissertation is is in the majority of cases not clear, the most accurate strategy seems to consist in coding both roles as 'individuals related to the work in an unspecified manner'. I anticipate that this might not be a feasible solution, since a work should have an author (even an anonymous one). However, the attribution of author roles to both *praeses* and *respondens* or the distinction between 'thesis supervisor' and 'author' or 'author' and 'opponent' may not be correct for some works either. Thinking outside the box, the most desirable outcome would be a change of the standard acommodating these particular roles in early modern dissertations, although this is probably not a realistic expectation and would still leave us with legacy data not taking the distinction into account.

However, from the point of view of a digital humanist who wishes to turn these catalog records into a dataset allowing for further investigation of the genre, accuracy with regard to the intellectual contributions of *praeses* and *respondens* in individual works is of secondary importance anyway.[3] If it is possible to map the roles of *praeses* and *respondens* on some categories of the metadata standard, we can reconstruct or infer the historical roles from the corresponding fields of the metadata standard. So we would have to pick one asymmetrical pair of categories – be it 'author vs. supervisor' or 'author vs. opponent' – which then would have to be applied consistently across the whole dataset.

As we have seen, this is not the case here, because most records either lack this information or provide it in an inconsistent manner. This means that for the purpose of digital investigations of these data, the role of *praeses* and *respondens* cannot be analysed with the required degree of representativity: if less than half of the records in one dataset and about 80% of records in the second dataset do not provide reliable information about these roles, even the information we have is of no value, because we cannot be sure to which extent the records that do contain this information are representative for the dataset as a whole. In other words, the existing data about the authorial roles of *praeses* and *respondens* are unusable not because they may not be accurate, but because the corresponding UNIMARC relators are applied in an inconsistent manner: if in some cases a *praeses* appears as 'thesis advisor' in the record, although he was in fact the main author of the text, this misattribution is inaccurate with regard to the individual work. But if the role of 'thesis advisor' for the *praeses* were applied consistently across all available data, we could translate it back into the category ' *praeses*' without caring about the correctness or incorrectness of this modern attribution in some cases: consistency trumps accuracy.

In closing, I would like to add two short remarks regarding the broader significance of these findings for scholarly practices in the digital humanities. First, if we accept provisionally the idea that research in the digital humanities is at least to a significant extent based on the analysis of digital objects 'at scale', uniformity of data that makes it possible to compare subsets of data in an instructive manner may often be a more pressing concern than the philological exactitude (*Genauigkeit*) with regard to the individual datum. Not just in the case presented here we may be tempted to value consistency higher than accuracy, so there may be reason to assume that it is precisely the digital humanities that are 'inexact' disciplines (for a different point of view see Lauer 2019). Second, if we leave aside the methodological dimension of this contribution, i. e. the role of different criteria for the evaluation of bibliographical metadata as research data, it reports, strictly speaking, a null result: the data made available in French library catalogs are not sufficient for the identification of *praeses* and *respondens* in early modern dissertations. Usually, this would have been a footnote in a data paper that may only have been inserted after the intervention of a reviewer. How the reporting of such results can contribute to the further development of the digital humanities is, however, probably a topic left best for another occasion.

Fußnoten

[1] The work described in this abstract was funded by the European Union through a FIAS research professorship under the Marie Skłodowska-Curie grant agreement No 945408.
[2] The relator code '956' is not referenced in Willer 2009. I was therefore unable to resolve it.
[3] Cf. Freedman 200, 34: "[…] there is no simple correct answer [to the question of authorship of a given dissertation], and it might in many cases to be best to list the presider and the respondent(s) as joint author."

Bibliographie

Axton, Arie Baak, Niklas Blomberg, et al. 2016. “The FAIR Guiding Principles for Scientific Data Management and Stewardship.” *Scientific Data* 3 (1): 160018. https://doi.org/10.1038/sdata.2016.18.

Freedman, Joseph S. 2005. “Disputations in Europe in the early modern period”. In Breimer, Douwe Durk, ed., *Hora Est!: On Dissertations*, 30–50. Leiden: Universiteitsbibliotheek Leiden.

Heßbrüggen-Walter, Stefan 2022. “Data Cleaning als digitale Quellenkritik - VD17 und das Genre der katholischen Dissertation im Alten Reich.” In *Dhd-2022 Book of Abstracts*. Potsdam. https://zenodo.org/record/6328027.

Király, Péter, and Jan Brase. 2021. “4.3 Qualitätsmanagement.” In *Praxishandbuch Forschungsdatenmanagement*, 357–80. De Gruyter Saur. https://doi.org/10.1515/9783110657807-020.

Lauer, Gerhard. 2019. “Über den Wert der exakten Geisteswissenschaften.” In *Geisteswissenschaft – Was Bleibt? Zwischen Theorie, Tradition Und Transformation*, edited by Hans Joas and Jörg Noller, 152–73. Freiburg: Alber.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles

Willer, Mirna, ed. 2009. *UNIMARC Manual: Authorities Format*. 3rd ed. IFLA Series on Bibliographic Control, Vol. 38. München: Saur.

Mainz, Johannes Gutenberg University. n.d. “Task Area 1: Data Quality | 4Memory/Nationale Forschungsdaten Infrastruktur (NFDI).” Johannes Gutenberg University Mainz. Accessed July 24, 2024. https://4memory.de/task-areas/task-area-1-data-quality/.