The challenge of contextualization and data transparency in structured research data!

Katrin Moeller*[1]

[1]Historical Data Center Saxony-Anhalt – Germany

Abstract

The conference aims to discuss questions of data modelling and the development of ontological approaches. With the development or expansion of various classification systems to build up an ontology of "Historischen deutschsprachigen Amts- und Berufsbezeichnungen" and with the establishment of the Historical Data Centre of Saxony-Anhalt (https://opendata.uni-halle.de/handle/1981185920/66) I have been dealing with the challenges of data modelling of historical data for many years.
In the meantime, the humanities disciplines are also following the path of the FAIR principles of data management. A very important requirement of these principles is the interoperability of research data, which means on the one hand the use of controlled vocabularies/ontologies for the exploitation of data, and on the other hand the transparency of the creation and decision-making process as well as the traceability of data collection. These measures, as well as the consideration of community specific principles, should ultimately increase the data quality and the reusability of research data. It can be observed in the data community that although the use of standard data is already being intensively discussed and thought about, the aspect of transparency and traceability of data is relatively seldom discussed in depth outside the discussion of metadata. Yet it is precisely this claim that also in the field of historical scholarship places considerable new methodological demands on the production, documentation and curation of data. For this is associated with very considerable demands, which are also being made in the research community of the disciplines working in the field of history, and which have rarely been met to date.

Especially when modelling historical data from heterogeneous sources, often under the conditions of an analogue source situation and its transformation into the digital, it is not at all easy to actually adequately represent even previous community-specific requirements of the historical sciences (citation and contextualisation) in structured data. In the process, mostly complex decisions and information usually require very reduced entities and summaries to be documented. This does not apply in the same way to textual sources, whose form of representation ultimately allows context and annotation much more easily and in a technically more standardised way than is possible for structured data today.

If one tries to analyze these problem layers, it is not so much a problem of data modeling that is revealed here, but primarily the challenges of heterogeneous information sources and the limited possibilities of representation in conventional, but also widely used tools. Using the example of a research data set for recording the age at death of the population of Halle between 1670 and 2018, I would like to make the challenges and individual problem

*Speaker

layers of this problem visible. In doing so, both the substantive aspects of the temporal comparison in the long-term perspective and the dimensions of the changing space and the changing group of people will be discussed, which have to be represented in many different ways and concern central dimensions of the modelling paradigm discussed at the conference. At the same time, the handling of data errors, multiple representations, missing data and changing categories and feature dimensions must be considered.

At the same time, the influence of these changes on the result of the analysis is to be discussed and methodical approaches to solutions that have been realized in this project are to be shown. At the same time, the outline of decision problems and the lack of representability at the interface to information will be discussed. In this context, it is also necessary to weigh up between reusability on the one hand and effectiveness and usability for the data producer on the other. All in all, I would like to argue more strongly in favour of integrating the documentation process of research data and its significance for the methodological and technical repositioning of the historical sciences and the processing of such knowledge in various forms of quality assurance, which is as important for the traceability and understanding of data as the use of standardised norm data.

Die Herausfoderung von Kontextualisierung und Datentransparenz in strukturierten Forschungsdaten!

Dr. Katrin Moller (Historisches Datenzentrum Sachsen-Anhalt)

Die Tagung setzt sich zum Ziel, Fragen der Datenmodellierung und Ausbildung von ontologischen Ansätzen zu diskutieren. Mit der Entwicklung bzw. dem Ausbau verschiedener Klassifikationssysteme zum Aufbau einer Ontologie der Historischen deutschsprachigen Amts- und Berufsbezeichnungen" sowie mit dem Aufbau des Historischen Datenzentrums Sachsen-Anhalt (https://opendata.uni-halle.de/handle/1981185920/66) beschäftige ich mich bereits seit vielen Jahren mit den Herausforderungen der Datenmodellierung historischer Daten.

Mittlerweile verfolgen auch die geisteswissenschaftlichen Disziplinen den Pfad der FAIR-Prinzipien des Datenmanagements. Eine ganz wesentliche Forderung dieser Prinzipien ist die Interoperabilität von Forschungsdaten, worunter einerseits die Verwendung von kontrollierten Vokabularen/Ontologien zur Erschließung von Daten, andererseits die Transparenz des Entstehungs- und Entscheidungsprozesses sowie die Nachvollziehbarkeit der Datenerhebung verstanden werden. Diese Maßnahmen wie ebenso die Beachtung community-spezifischer Prinzipien soll letztlich die Datenqualität und die Nachnutzbarkeit von Forschungsdaten steigern. Dabei kann man in der Datencommunity beobachten, dass zwar über die Anwendung von Normdaten bereits intensiv diskutiert und nachgedacht wird, der Aspekt der Transparenz und Nachvollziehbarkeit von Daten aber außerhalb der Diskussion um die Metadaten relativ selten vertieft wird. Dabei ist es genau dieser Anspruch, der auch in der Geschichtswissenschaft erhebliche neue methodische Ansprüche an die Produktion, Dokumentation und Kuration von Daten stellt. Denn damit verbindet sich auch in der Forschungscommunity der historisch arbeitenden Disziplinen ganz erhebliche Anforderungen, die bisher nur selten erfüllt werden.

Gerade bei der Modellierung von historischen Daten aus heterogene Quellen, oft unter den Bedingungen einer analogen Quellensituation und ihrer Transformation in das Digitale ist es gar nicht so einfach, selbst bisherige communityspezifische Anforderungen der Geschichtswissenschaften (Zitation und Kontextualisierung) auch in strukturierten Daten tatsächlich adäquat abzubilden. Dabei sind meist komplexe Entscheidungen und Informationen meist sehr reduzierte Entitäten und Zusammenfassungen zu dokumentieren. Dies gilt in gleicher Weise nicht für textuelle Quellen, die über ihre Repräsentationsform letztlich Kontext und Annotation wesentlich einfacher und fachlich standardisierter ermöglichen, als dies für strukturierte Daten heute möglich wird.

Versucht man diese Problemschichten zu analysieren, offenbart sich hier gar nicht so sehr ein Problem der Datenmodellierung, sondern in erster Linie vor allem Herausforderungen heterogener Informationsquellen und den bisher begrenzten Abbildungsmöglichkeiten in konventionellen, aber eben auch weit verbreiteten Arbeitswerkzeugen. Am Beispiel eines Forschungsdatensatzes zur Erfassung des Sterbealters der Bevölkerung Halles zwischen 1670 und 2018 möchte ich die Herausforderungen und einzelnen Problemschichten dieses Problems sichtbar machen. Dabei werden sowohl die inhaltlichen Aspekte des zeitlichen Vergleichs in der Langzeitperspektive wie ebenso die Dimensionen des sich verändernden Raumes und des sich veränderten Personenkreises diskutiert, die in vielfältiger Weise abgebildet werden müssen und zentrale Dimensionen des auf der Tagung diskutierten Modellierungsparadigmas betreffen. Gleichzeitig ist der Umgang mit Datenfehlern, mehrfachen Repräsentationen, fehlenden Daten und wechselnden Kategorien sowie Merkmalsdimensionen zu betrachten.
Gleichzeitig soll der Einfluss dieser Veränderungen auf das Ergebnis der Analyse diskutiert und methodische Lösungsansätze aufgezeigt werden, die in diesem Vorhaben realisiert wurden. Gleichzeitig soll es um die Skizzierung von Entscheidungsproblemen gehen und die mangelnde Abbildbarkeit an der Schnittstelle zur Information. Dabei ist zudem zwischen Nachnutzbarkeit auf der einen sowie Effektivität und Usability für den Datenproduzenten auf der anderen Seite abzuwägen. Insgesamt möchte ich damit stärker für die Einbindung des Dokumentationsprozesses von Forschungsdaten und seine Bedeutung für die methodisch-fachliche Neuaufstellung der Geschichtswissenschaften und eine Verarbeitung solchen Wissens in verschiedene Formen der Qualitätssicherung plädieren, welche für die Nachvollziehbarkeit und das Verständnis von Daten ähnlich hohe Bedeutung besitzt wie die Verwendung standardisierter Normdaten.

Keywords: data transparency