

# Das optimale Datenmodell

## Eine Spurensuche im

## Möglichkeitsfeld der Kodierung

### Saric, Sanja

sanja.saric@uni-graz.at

Institut Zentrum für Informationsmodellierung - Austrian Centre for Digital Humanities, Universität Graz, Österreich; Institut für Sprachwissenschaft, Universität Graz, Österreich

### Steiner, Elisabeth

elisabeth.steiner@uni-graz.at

Institut Zentrum für Informationsmodellierung - Austrian Centre for Digital Humanities, Universität Graz, Österreich

### Vogeltanz, Maximilian

maximilian.vogeltanz@uni-graz.at

Institut für Sprachwissenschaft, Universität Graz, Österreich

## Einleitung

Zu Beginn jedes neuen Projektes in den Digitalen Geisteswissenschaften steht die Frage nach der adäquaten Modellierung der Forschungsdaten. Während im Förderansuchen häufig die Nennung von XML/TEI-Kodierung für Textquellen ausreichend ist, stellt sich die praktische Arbeit meist komplizierter dar: Schon die TEI (TEI Consortium 2021) bietet zahlreiche Möglichkeiten, ähnliche Sachverhalte zu annotieren und die gewählte Strategie muss dabei auf das Material, die Forschungsfrage sowie die Archivierung und Weiterverwendung der Daten Rücksicht nehmen.

Der vorliegende Beitrag stellt die Herausforderungen und Lösungsansätze anhand der Briefkorrespondenz von Hugo Schuchardt vor.

## Herausforderung bestmögliche Kodierung

Ein Modellierungsansatz erfasst die untersuchten Merkmale möglichst genau und standardisiert in einem anerkannten Schema; berücksichtigt Referenzimplementationen und *best practice*-Guidelines; bezieht fachspezifische Vokabularien und Normdaten mit ein; versieht die Daten bereits im Entstehungsprozess mit Metadaten für die Weiterverwendung. Viele dieser Punkte sind ebenfalls Grundpfeiler der FAIR-Datenprinzipien (Wilkinson et al. 2016). In der Projektarbeit begrenzen jedoch oft verfügbare Zeit- und Personalressourcen die Umsetzbarkeit aller Aspekte, was notgedrungen zu Kompromissen führt. Zusätzlich konkurriert das Bedürfnis, projektspezifische Merkmale zu berücksichtigen, mit dem Anspruch an Vergleichbarkeit und Standardisierung. Trotzdem ist gerade die Interoperabilität zentral für die Nachhaltigkeit der Forschungsdaten.

Diese Herausforderungen stellten sich auch im *Hugo Schuchardt Archiv* (Hurch 2007-), einem langjährigen Vorhaben des Instituts für Sprachwissenschaft der Universität Graz. Im Mit-

telpunkt einer Kooperation mit dem Institut Zentrum für Informationsmodellierung – Austrian Centre for Digital Humanities (ZIM-ACDH) steht die Migration aller Ressourcen vom Institut für Sprachwissenschaft in das Repositorium GAMS, um die Korrespondenz und andere Dokumente aus dem Nachlass zu archivieren.<sup>1</sup>

Daher tritt zu den genannten Faktoren ein weiterer hinzu: die Berücksichtigung von Legacy-Daten. Anpassungen an das neue TEI-Modell konnten zwar teilweise automatisch durchgeführt werden, stoßen aber an Grenzen. Manuelle Anpassungen sind jedoch aufgrund des Umfanges von mehreren Tausend Briefen nur eingeschränkt möglich. Dieses Kompatibilitätsproblem tritt einerseits auf technischer Ebene in Erscheinung, wo alle Daten gegen ein einheitliches ODD-Schema validiert werden sollen, andererseits aber ebenso auf inhaltlicher Ebene, was beispielsweise die Strukturierung des Schlagworthesaurus betrifft.

Für die TEI-Annotation der Korrespondenzdaten drängen sich die Empfehlungen der *SIG Correspondence* und die entsprechende Weiterverarbeitung im CMI-Format (Dumont et al. 2019) auf, um den Zugriff über *correspSearch* (Dumont 2016 und Dumont/Grabsch/Müller-Laackman 2021) zu ermöglichen. Als Referenzimplementationen wurden die Briefeditionen der BBAW (insbesondere die beiden Humboldt-Editionen von Ette et al. 2020 und BBAW 2021) konsultiert. Für den Brieftext wurde Kompatibilität mit dem DTA-Basisformat (DTABf) angestrebt. Normdaten und Vokabulare wurden einerseits aus den in Vorarbeiten aufgebauten Thesauri bezogen, andererseits aus *authority files* wie VIAF, GND und *GeoNames*. Das ZIM-ACDH versucht durch interne Kodierungsrichtlinien für den TEI-Header einen Grundstock an Metadaten zu erzeugen, der in weiterer Folge für die Langzeitarchivierung in weitere Standards umgewandelt werden kann. Unter zusätzlicher Berücksichtigung der Legacy-Daten ergab die Zusammenwirkung dieser Anforderungen bereits mehrere Konflikte in der Annotation.

## Lösungsansätze

Das Abstimmen verschiedener Anforderungen und der Abgleich mit den vorhandenen Daten nahm erhebliche Zeit in Anspruch. Danach wurde evaluiert, wo eine automatische Anpassung der Altdaten vertretbar ist und wo sich das ODD-Schema den Daten anpassen muss. Schließlich wurden die Konflikte in den unterschiedlichen Kodierungsvarianten besprochen und versucht, die passendste Lösung zu finden. Dieser Prozess mündete in einem Schema, das die gewünschten Eigenschaften vereinte, aber notgedrungen Abweichungen zu den Ausgangsschemata enthielt. Um diesen Schwachpunkt zu entschärfen, wurde versucht, sowohl die Abweichungen wie auch die Gründe dafür innerhalb und außerhalb des Schemas zu dokumentieren. Dies dient nicht nur in der Erfassung als Referenz für unterschiedliche BearbeiterInnen, sondern gewinnt vor allem in der erhofften Weiternutzung der Daten durch Dritte an Bedeutung.

## Zusammenfassung und Ergebnisse

Der Prozess der Datenmodellierung muss zahlreiche Einflussgrößen berücksichtigen, wie am Beispiel des *Hugo Schuchardt Archivs* illustriert wurde. Dieser Prozess beinhaltet immer kritische Entscheidungen und Kompromisse, die sich aus dem Material, aber auch durch eingeschränkte Zeit- und Personalressourcen ergeben. Die Frage nach der bestmöglichen Kodierung kann da-

her nicht allgemeingültig beantwortet werden, vielmehr muss sie individualisiert betrachtet werden. Trotzdem können konstituierende Eigenschaften für eine *gute* Annotationspraxis beobachtet werden: sie sollte gut dokumentiert sein und unter der Berücksichtigung von FAIR-Data die Weiterverwendung erlauben. Gerade dem Aspekt der Metadaten, die für die Archivierung und Aggregation nach dem Ende des befristeten Projektes zentral sind, wird zu Beginn oft wenig Beachtung geschenkt. Für die nachhaltige Weiternutzung der Daten im wissenschaftlichen Kontext stellt dies jedoch einen essenziellen Bestandteil dar. In diesem Sinne sollte auch die Interoperabilität mit ähnlichen Ressourcen als Faktor bei Kodierungsentscheidungen in Betracht gezogen werden.

## Fußnoten

1. Das Hugo Schuchardt Archiv wurde von zahlreichen Fördergebern berücksichtigt, hervorzuheben sind die letzten FWF-Projekte „Netzwerk des Wissens“ (P 24400, Bernhard Hurch 2012–2016) und „Philingk: Verlinktes Wissen zur Fachgeschichte“ (I 5076, Ursula Bähler und Bernhard Hurch 2021–2023).

## Bibliographie

**Berlin-Brandenburgischen Akademie der Wissenschaften** (2011–2020): *DTABf. Deutsches Textarchiv – Basisformat*. <http://deutschestextarchiv.de/doku/basisformat> [letzter Zugriff 15. Juli 2021].

**Berlin-Brandenburgischen Akademie der Wissenschaften** (2021): *Wilhelm von Humboldt: Sprachwissenschaftliche Korrespondenz*. <https://wvh-briefe.bbaw.de> [letzter Zugriff 15. Juli 2021].

**Dumont, Stefan** (2016): "correspSearch – Connecting Scholarly Editions of Letters" in: *Journal of the Text Encoding Initiative* 10. <https://doi.org/10.4000/jtei.1742> [letzter Zugriff 15. Juli 2021].

**Dumont, Stefan / Börner, Ingo / Müller-Laackmann, Jonas / Leipold, Dominik / Schneider, Gerlinde** (2019): "Correspondence Metadata Interchange Format (CMIF)" in: Dumont, Stefan / Haaf, Susanne / Seifert Sabine (eds.): *Encoding Correspondence. A Manual for Encoding Letters and Postcards in TEI-XML and DTABf*. Berlin. <https://encoding-correspondence.bbaw.de/v1/CMIF.html> [letzter Zugriff 15. Juli 2021].

**Dumont, Stefan / Grabsch, Sascha / Müller-Laackman, Jonas** (2021): *correspSearch – Briefeditionen vernetzen*. Version 2.0.0. Berlin-Brandenburgische Akademie der Wissenschaften. <https://correspSearch.net> [letzter Zugriff 15. Juli 2021].

**Ette, Ottmar** et al. (2020): *edition humboldt digital*. Berlin-Brandenburgische Akademie der Wissenschaften. Version 6. <https://edition-humboldt.de> [letzter Zugriff 15. Juli 2021].

**Hurch, Bernhard** (2007–): *Hugo Schuchardt Archiv*. <http://schuchardt.uni-graz.at> [letzter Zugriff 15. Juli 2021].

**Institut Zentrum für Informationsmodellierung – Austrian Centre for Digital Humanities** (2021): *Geisteswissenschaftliches Asset Management System (GAMS)*. <https://gams.uni-graz.at> [letzter Zugriff 15. Juli 2021].

**Steiner, Elisabeth / Stigler, Johannes** (2018): "GAMS – Eine Infrastruktur zur Langzeitarchivierung und Publikation geisteswissenschaftlicher Forschungsdaten". In: *Mitteilungen der Vereinigung Österreichischer Bibliothekarinnen und Bibliothekare* 71: 207–216.

**TEI Consortium** (2021): *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Version 4.2.2. <https://tei-c.org> [letzter Zugriff 15. Juli 2021].

**Wilkinson, Mark D.** et al. (2016): "The FAIR Guiding Principles for scientific data management and stewardship" in: *Sci. Data* 3:160018. <https://doi.org/10.1038/sdata.2016.18> [letzter Zugriff 15. Juli 2021].