

## Einleitung

Annette Klosa [klosa@ids-mannheim.de](mailto:klosa@ids-mannheim.de), Tel.: +49 621 1581-411

Carolin Müller-Spitzer [mueller-spitzer@ids-mannheim.de](mailto:mueller-spitzer@ids-mannheim.de), Tel.: +49 621 1581-429

Mittlerweile ist allgemein anerkannt, dass die Datenmodellierung für Wörterbücher in denjenigen lexikographischen Projekten, in denen aus einer Wörterbuchsubstanz verschiedene Wörterbücher in unterschiedlichen Medien hergestellt werden sollen, möglichst medienunabhängig erfolgen sollte. Geeignete Formate für eine solche medienunabhängige Datenmodellierung sind etwa XML-DTDs oder XML-Schemata, aber auch eine netzartige Modellierung in Relationen und Knoten. Neben wörterbuchspezifischen, maßgeschneiderten Modellierungen gibt es Richtlinien bzw. ‚Baukästen‘ für Standardmodellierungen, wie etwa das „Lexical Markup Framework for natural language processing (NLP) lexicons and machine-readable dictionaries (MRD)“ (LMF, ISO-24613:2008)<sup>1</sup> oder die Richtlinien der „Text Encoding Initiative“,<sup>2</sup> deren Einsatz für die Datenmodellierung bei Internetwörterbüchern zu diskutieren ist. Für elektronische Wörterbücher muss die Modellierung aber noch weiteren Anforderungen genügen: Hier bestimmen die erwünschten Zugriffsmöglichkeiten auf die Daten, wie diese modelliert werden müssen (vgl. Gloning/Welter 2001 und Müller-Spitzer 2005). Bei Internetwörterbüchern ist darüber hinaus bei der Modellierung zu berücksichtigen, dass eine flexible Präsentation der Suchergebnisse angestrebt werden soll, und zwar je nach Nutzergruppe oder Nutzersituation (vgl. Storrer 2001). Internetwörterbücher können solche flexiblen Zugriffs- und Präsentationsmöglichkeiten dann realisieren, wenn bei der Datenmodellierung die Funktionalitäten des Computers gleich mitgedacht wurden (vgl. u.a. de Schryver 2003).

Vor dem Hintergrund solcher Überlegungen fand am 5. und 6. Mai 2011 am Institut für Deutsche Sprache in Mannheim das erste Arbeitstreffen des wissenschaftlichen Netzwerks „Internetlexikografie“ (gefördert von der Deutschen Forschungsgemeinschaft) statt.<sup>3</sup> Im Rahmen des Arbeitstreffens wurde die Modellierung verschiedener Internetwörterbücher in Konzeption und Realisierung vorgestellt und mit anderen, von konkreten Wörterbuchprojekten unabhängigen Modellierungsvorschlägen kontrastiert. Im Einzelnen wurde die Modellierung eines semantischen Netzes für lexikographische Anwendungen (am Beispiel der Duden-Ontologie; vgl. den Beitrag von Melina Alexa) präsentiert, die XML-Modellierung für ein Wörterbuchnetz (am Beispiel von [OWID](#); vgl. den Beitrag von Carolin Müller-Spitzer), die TEI-basierte Modellierung von Retrodigitalisaten (am Beispiel des [Trierer Wörterbuchnetzes](#); vgl. den Beitrag von Vera Hildenbrandt), eine FrameNet-basierte Modellierung (am Beispiel des [Kicktionary](#); vgl. den Beitrag von Thomas Schmidt) sowie die Modellierung von Mehrwortverbindungen im [DWDS](#) (vgl. den Beitrag von Alexander Geyken). Außerdem wurde die Datenmodellierung bzw. Architektur für „pluri-monofunctional dictionaries“ von Dennis Spohr vorgestellt.<sup>4</sup>

Dabei wurden Vor- und Nachteile des jeweiligen Vorgehens diskutiert, sodass sich zeigte, welche Formen der Datenmodellierung für welche Form von Internetwörterbüchern bzw. wörterbuchähnlichen Produkten am besten geeignet sind. Mit weiteren Fragen beschäftigten sich die Teilnehmer auch in einer Diskussionsrunde anhand verschiedener, für die Diskussion

<sup>1</sup> Vgl. <http://www.lexicalmarkupframework.org/>.

<sup>2</sup> Vgl. <http://www.tei-c.org/index.xml>.

<sup>3</sup> Zur Arbeit des Netzwerks „Internetlexikografie“ vgl. <http://www.internetlexikografie.de>.

<sup>4</sup> Zu den Vorschlägen von Dennis Spohr für eine solche Modellierung vgl. detailliert Spohr (2011).

vorgegebener Fragen, deren Ergebnisse in den folgenden Abschnitten zusammengefasst werden.<sup>5</sup>

## 1. Welche Formate eignen sich am besten für die Modellierung von Internetwörterbüchern (DTDs, Schemata, Netze etc.)?

Hinsichtlich der Eignung verschiedener Formate für die Modellierung von Internetwörterbüchern war die einhellige Meinung der Diskussionsrunde: Komplexere Schemata sind im Hinblick auf ihre Mächtigkeit den DTDs überlegen. Außerdem können bestimmte funktionale Abhängigkeiten zwischen Attributen in ihnen expliziter formuliert werden, mögliche Verstöße gegen die formale Inhaltsstruktur sind somit besser zu kontrollieren. Dieser Vorteil ist jedoch weniger relevant speziell für die Präsentation von Internetwörterbüchern, vielmehr ist er für die Erstellung aller Arten von Wörterbüchern von Bedeutung. Wenn eine DTD für eine bestimmte Zielvorstellung nicht ausreichend ist, bietet sich daher der Einsatz von XML-Schemata an.

Zu bedenken ist auch, dass DTDs in Bezug auf die Definition von Datenformaten veraltet erscheinen, exakte Bestimmungen (wie etwa die Beschränkung auf ein Datumsformat) sind nicht möglich. Jedoch zeichnen sich DTDs durch den nicht unerheblichen Vorteil aus, dass sie für den Lexikographen lesbar sind. Man muss sich nicht auszugsweise über Probeartikel einen Überblick über die Modellierung verschaffen (den man auf diesem Weg zumindest bei einer komplexen Modellierung auch kaum bekommt), sondern kann die Modellierung in ihrer Gesamtheit in der DTD nachvollziehen; Entwürfe von DTDs können zwischen den Lexikographen und der Person, die sie modelliert, besprochen werden. Zudem sind DTDs für neue Wörterbuchprojekte automatisch in Schemata zu konvertieren. Die Beantwortung der Frage, ob Schemata oder DTDs herangezogen werden sollen, hängt somit vor allem davon ab, was in der Modellierung festgehalten werden soll, wie das lexikographische Team zusammengesetzt ist, wer die Modellierung lesen können soll und welche Rolle sie im Prozess der Wörterbucherstellung spielt.

Für bestimmte Ansätze sind DTDs als Grundlage möglich und ausreichend und insbesondere einfach zu handhaben. In einem Projekt wie *elexiko* beispielsweise ist die DTD sehr narrativ und auch für Uneingeweihte und Außenstehende nachzuvollziehen.<sup>6</sup> Andere Projekte, wie die Duden-Ontologie, benötigen eine netzbasierte Modellierung.<sup>7</sup> Die entscheidende Überlegung für die Wahl eines Modellierungsformates muss also – so der Tenor der Teilnehmer – sein: Welcher Formalismus kommt welchem Zweck zugute?

Es gilt auch zu bedenken, ob die modellierten und angehäuften Daten auf lange Sicht maschinell lesbar bleiben sollen und in welche Formate sie jeweils exportiert werden können. Generell kann festgehalten werden, dass das Verhältnis zwischen dem darzustellenden Inhalt und dem nötigen Aufwand bedacht werden muss, um die Entscheidung für die richtige Modellierungsmethode zu erleichtern. Schließlich muss man sich stets darüber klar sein, was und wie viel man investieren will, um die Daten langfristig kohärent zu halten. Neue Wörterbuchprojekte sind vor diesem Hintergrund auf jeden Fall gut beraten, die Frage der Datenmodellie-

---

<sup>5</sup> Die folgende Zusammenfassung beruht auf den Protokollen unserer Hilfskräfte Martin Loder, Bianca Pargner und Sandra Zimmermann, denen wir an dieser Stelle herzlich für ihre Unterstützung danken.

<sup>6</sup> Vgl. den Beitrag von Carolin Müller-Spitzer in diesem Band sowie Müller-Spitzer (2011).

<sup>7</sup> Vgl. den Beitrag von Melina Alexa in diesem Band.

rung ausreichend zu diskutieren, verschiedene Modellierungsansätze zu prüfen und genügend Zeit für die Umsetzung einzuplanen.

## **2. Was sind die Vor- und Nachteile projektspezifischer Modellierungen und allgemeiner Modellierungen wie TEI oder „Lexical Markup Framework“?**

Entsprechend den unterschiedlichen Projekten, die die Diskussionsteilnehmer vertraten, waren sowohl Anwender von standardbasierten als auch von maßgeschneiderten Modellierungen vertreten. Allgemeine Übereinkunft bestand darin, dass eine vollkommene Austauschbarkeit von Daten zwischen verschiedenen elektronischen Wörterbüchern und Projekten auch bei einer Standardmodellierung nicht gewährleistet ist. Dafür ist die Bandbreite, die die Standardmodellierungen als Baukästen bieten, zu groß. Anders verhält es sich bei Retrodigitalisaten: Wenn in einer Institution (wie z.B. im [Trierer Wörterbuchnetz](#)) verschiedene Printwörterbücher in ein und demselben Modell für die elektronische Präsentation erfasst werden, bietet sich die Anwendung der TEI an.

Die TEI ist somit für die Modellierung neuer Wörterbücher eine gute Quelle, aus der man lernen kann. Ob allerdings die Anwendung einer TEI-konformen Modellierung generell als sinnvoller erachtet werden muss als eine genau auf die Bedürfnisse des Projekts angepasste Modellierung, wurde bezweifelt. Letztere kann in einem solchen Fall die bessere Alternative sein, auch weil ihre Anwendung möglicherweise weniger komplex und besser überschaubar ist. Entscheidend für diese Frage ist auch, wie das lexikographische Team zusammengesetzt ist. Ist z.B. überhaupt jemand vorhanden, der von Grund auf eine neue Modellierung entwickeln kann oder muss sich der- bzw. diejenige an Standards orientieren?

## **3. Ist bei Standardmodellierungen wirklich Austauschbarkeit der Daten gewährleistet?**

Die Weitergabe und der Austausch von größeren Datenmengen kann für verschiedene Wörterbuchprojekte von gegenseitigem Nutzen sein, wie allgemein konstatiert wurde. Vollkommene Austauschbarkeit ist dabei natürlich nie gewährleistet. Doch ist das Minimieren des dafür nötigen Aufwandes sehr erwünscht und durch Standards ermöglicht, wobei sich allerdings auch kleinere, nicht standardmäßig modellierte Datenmengen, wenn sie denn klar definiert sind, prinzipiell ohne großen Aufwand in ein allgemeineres Standardformat migrieren und somit austauschen lassen. Jedenfalls sind in der TEI viele philologische Prinzipien und Richtlinien repräsentiert, auf die man gut zurückgreifen kann, wenn es um flexible Modelle für eine je individuelle Modellierung geht.

## **4. Wie können/müssen bei der Modellierung verschiedene Zugriffsmöglichkeiten berücksichtigt werden?**

Wenn man ein Wörterbuch schreibt, weiß man nicht von Anfang an, was potentielle Nutzer alles wissen wollen. Die Frage, die in der Runde diskutiert wurde, war daher u.a., ob man wirklich eine vollkommen nutzerunabhängige Datenmodellierung aufbauen kann bzw. wo Restriktionen hierfür liegen. Zum Beispiel ist die Formulierung einer Bedeutungsparaphrase in den meisten Fällen nicht nutzerunabhängig, sondern auf eine bestimmte Zielgruppe zugeschnitten. Man sollte sich auch fragen, ob sehr komplexe und/oder spezifische Daten letztlich

tatsächlich genutzt werden, d.h., ob der Aufwand für eine möglichst nutzerunabhängige Modellierung gerechtfertigt ist.

Andererseits geht es bei der Digitalisierung von Daten zumindest in wissenschaftlichen Kontexten ja auch darum, eine Forschungsgrundlage für die spätere wissenschaftliche Beschäftigung mit einer Datenmenge zu schaffen, deren Ergebnisse und Methoden zunächst noch nicht abzusehen sind. Folglich sind Datenkomplexität und ein spezifisches Informationsangebot durchaus auch von linguistischer Seite erwünscht. Der wissenschaftlichen Lexikographie bleibt vor diesem Hintergrund deshalb vor allem der Auftrag, neue Arten der Datenmodellierung und des Informationsangebots auszuprobieren und auch neue, bisher unbekannte Arten des Zugriffs anzubieten. Eine Modellierung sollte möglichst ein Maximum des Darzustellenden vorsehen und anstreben. Die Internetlexikographie entwickelt sich somit – zumindest was die Ebene der Datenbasis angeht – in Richtung der lexikalischen Datenbank.

### **5. Hat die Art der lexikographischen Primärquellen (Belegarchiv, elektronisches Textkorpus) Einfluss auf die Wahl und Art der Modellierung?**

Unterschiedliche Quellen erfordern selbstverständlich unterschiedliche Behandlungen; hier ist die Effizienzfrage entscheidend. Belege aus einem elektronischen Textkorpus sind (je nach Korpus in verschiedenem Maße) anderer Natur als solche aus dem lexikographischen Belegarchiv im Zettelkasten. Vor allem bei der Verlinkung auf die Belege im Internetwörterbuch ist dies relevant. Für die Modellierung von Internetwörterbüchern ist also nicht so sehr die Frage nach der Art der Primärquellen von Belang, sondern diejenige danach, wie zugegriffen bzw. verlinkt werden und was zur Darstellung kommen soll.

### **6. Welche Rolle spielt die Modellierung im lexikographischen Prozess von Internetwörterbüchern?**

Mittlerweile sind sich sowohl die Verantwortlichen in Forschungseinrichtungen als auch kommerzielle Geldgeber darüber im Klaren, dass Internetlexikographie nicht ohne eine stärkere Gewichtung der technischen Aufgaben auskommt, was sich insbesondere am Beispiel der Datenmodellierung zeigt. Trotzdem ist es häufig noch schwierig, kompetente technische Mitarbeiter für die wissenschaftliche Lexikographie zu gewinnen, da oftmals nur geteilte oder befristete, zuweilen auch unterbezahlte Stellen für technisches Personal angeboten werden, das sich deswegen eher an die freie Wirtschaft bindet. Eine Veränderung dieser Situation wurde von allen Diskussionsteilnehmern als äußerst wünschenswert erachtet.

### **7. Literatur**

- Gloning, Thomas/Welter, Rüdiger (2001): Wortschatzarchitektur und elektronische Wörterbücher: Goethes Wortschatz und das Goethe-Wörterbuch. In: Lemberg, Ingrid/Schröder, Bernhard/Storrer, Angelika (Hg.): Chancen und Perspektiven computergestützter Lexikographie: Hypertext, Internet und SGML/XML für die Produktion und Publikation digitaler Wörterbücher. Tübingen, S. 117-132.
- Lexical Markup Framework for natural language processing (NLP) lexicons and machine-readable dictionaries (MRD). Internet: <http://www.lexicalmarkupframework.org/>. (Stand: Oktober 2011).
- Müller-Spitzer, Carolin (2005): Die Modellierung lexikografischer Daten und ihre Rolle im lexikografischen Prozess. In: Haß, Ulrike (Hg.): Grundfragen der elektronischen Lexikographie. *elexiko* – das Online-Informationssystem zum deutschen Wortschatz. Berlin/New York, S. 21-54. (Schriften des Instituts für Deutsche Sprache 12).

- Müller-Spitzer, Carolin (2011): Der Einsatz einer maßgeschneiderten, feingranularen XML-Modellierung im lexikographischen Prozess. In: Klosa, Annette (Hg.): *ellexiko*. Erfahrungsberichte aus der lexikographischen Praxis eines Internetwörterbuchs. Tübingen: Narr, S. 173-191. (Studien zur deutschen Sprache 55).
- de Schryver, Gilles-Maurice (2003): Lexicographers' Dreams in the Electronic-Dictionary Age. In: *International Journal of Lexicography* 16/2, S. 143-199.
- Spohr, Dennis (2011): A Multi-layer Architecture for „Pluri-monofunctional“ Dictionaries. in: Fuertes-Olivera, Pedro A./Bergenholtz, Henning: *E-Lexicography – The Internet, Digital Initiatives and Lexicography*. London/New York, S. 103-120.
- Storrer, Angelika (2001): Digitale Wörterbücher als Hypertexte: Zur Nutzung des Hypertextkonzepts in der Lexikographie. In: Lemberg, Ingrid/Schröder, Bernhard/Storrer, Angelika (Hg.): *Chancen und Perspektiven computergestützter Lexikographie: Hypertext, Internet und SGML/XML für die Produktion und Publikation digitaler Wörterbücher*. Max Niemeyer: Tübingen, S. 53-69.
- Text Encoding Initiative. Internet: <http://www.tei-c.org/index.xml>. (Stand: Oktober 2011).
- Wissenschaftliches Netzwerk „Internetlexikografie“. Internet: <http://www.internetlexikografie.de>. (Stand: Oktober 2011).

