Der Aufbau einer maßgeschneiderten XML-basierten Modellierung für ein Wörterbuchnetz

Carolin Müller-Spitzer mueller-spitzer@ids-mannheim.de, Tel.: +49 621 1581-429

1. Einleitung

Im vorliegenden Beitrag soll der Aufbau einer maßgeschneiderten XML-Modellierung für ein Wörterbuchnetz erläutert werden. Diese Schriftfassung beruht auf einem gleichlautenden Vortrag, der auf dem ersten Arbeitstreffen des DFG-Netzwerks „Internetlexikografie“ in Mannheim im Mai 2011 gehalten wurde. Der Beitrag ist als Werkstattbericht zu verstehen, d. h. als praktisch orientierter Blick sowohl darauf, wie wir unsere Modellierung für OWID konzipiert haben, welche Konsequenzen dies für die lexikographische Arbeit sowie für die Recherchemöglichkeiten der Nutzer hat, als auch darauf, welche Vor- und Nachteile wir bei diesem Modellierungsansatz sehen. Der vorliegende Beitrag bietet damit keine umfassende theoretische Auseinandersetzung mit verschiedenen Möglichkeiten der Modellierung. Lediglich im folgenden Kapitel werden die Grundzüge des Modellierungsansatzes kurz erläutert und es wird auf entsprechende weiterführende projektbezogene Literatur verwiesen.

2. Allgemeines zur Modellierung in OWID

Das Online-Wortschatz-Informationssystem Deutsch (OWID) ist das Wörterbuchportal des Instituts für Deutsche Sprache (IDS) in Mannheim (s. www.owid.de). Es beinhaltet wissenschaftliche, korpusbasierte Wörterbücher zum Deutschen mit unterschiedlichen inhaltlichen Schwerpunkten. Dies sind im Moment[27]:

- *elexiko*: *elexiko* verfügt über eine umfangreiche korpusbasiert gewonnene Stichwortliste zum Deutschen mit über 300.000 Einträgen. Zu fast allen Stichwörtern sind in *elexiko* automatisch gewonnene Textbelege sowie orthographische Angaben zu finden. Darüber hinaus sind über 1.500 hochfrequente Stichwörter im „Lexikon zum öffentlichen Sprachgebrauch“ ausführlich lexikographisch beschrieben.
- Feste Wortverbindungen: In diesem Bereich sind lexikographische Ergebnisse der korpusgesteuerten Mehrwortforschung veröffentlicht. Die Wortartikel haben unterschiedliche linguistische Beschreibungstiefen und Darstellungsformate. Derzeit sind etwa 130 Mehrwortartikel in OWID enthalten (weiterführende Informationen s. Wortverbindungen online).
- Neologismenwörterbuch: Das Neologismenwörterbuch präsentiert in über 1.000 Wortartikeln neue Wörter bzw. Wortverbindungen sowie neue Bedeutungen von etablierten Wörtern, die in den 90er Jahren des 20. Jahrhunderts in die Allgemeinsprache eingegangen sind.
- Schulddiskurs 1945-55: In diesem Wörterbuch sind 85 Haupt- sowie über 200 Unterstichwörter zum Schulddiskurs im ersten Nachkriegsjahrzehnt verzeichnet. Dieser Wortschatzbereich ist aus einem breit angelegten Korpus von Texten, die in den Jahren 1945-55 erschienen sind, erarbeitet worden.

[27] Für projektbezogene Literatur s. die Projektseiten unter www.ids-mannheim.de, dort jeweils den Punkt „Publikationen“. Alle genannten Wörterbücher haben außerdem unter OWID eigene Begleittexte.

© 2011 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

Neben Wörterbüchern enthält OWID eine Online-Bibliographie zur elektronischen Lexikographie (OBELEX) sowie eine Datenbank zu Online-Wörterbüchern.

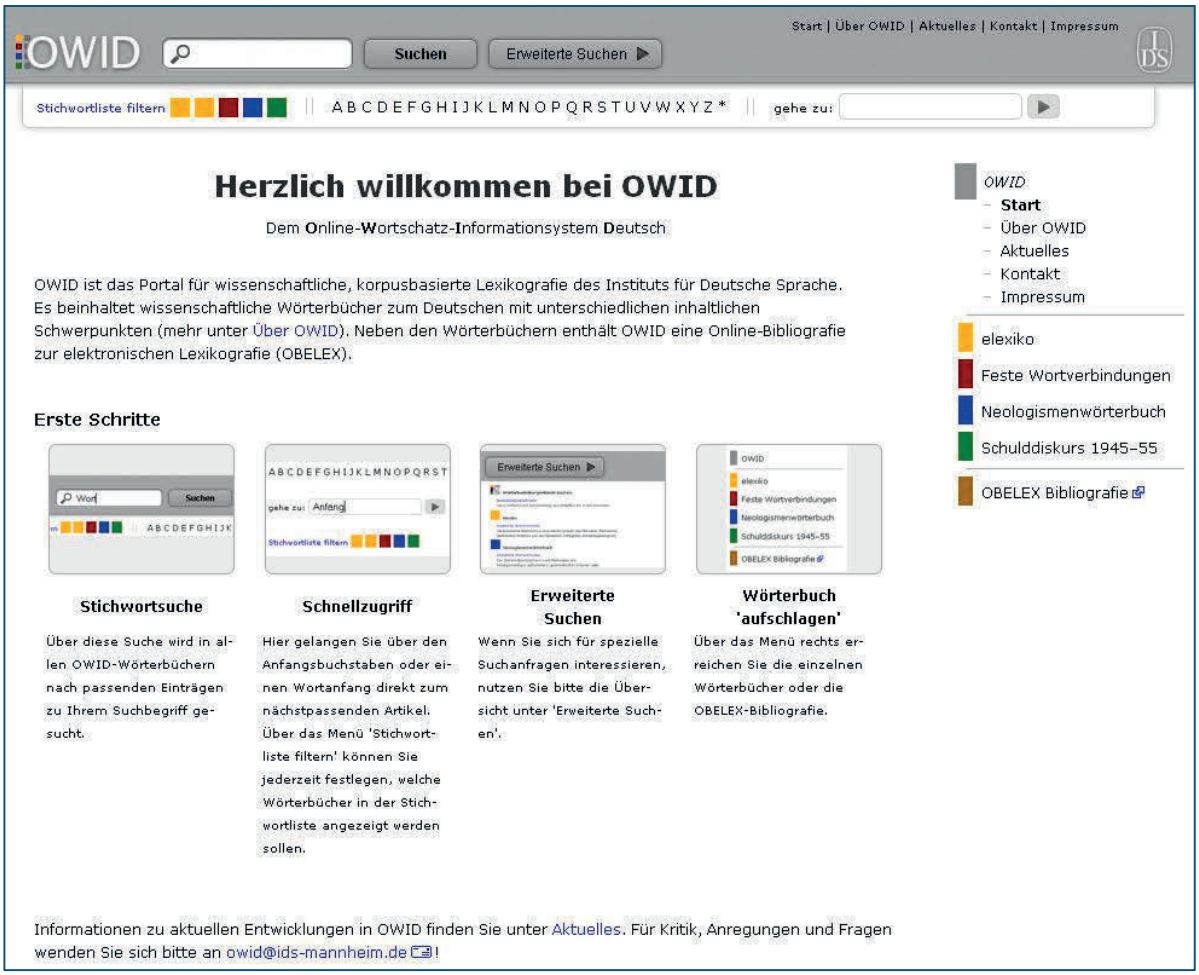

Abb. 1: Startseite von OWID

OWID wird kontinuierlich erweitert. Im Moment wird an der Integration des „E-ValBU“ (http://hypermedia2.ids-mannheim.de/evalbu/index.html), des „Handbuchs deutscher Kommunikationsverben“ (Harras et al. 2004), der „Schlüsselwörter der Wendezeit“ (Herberg et al. 1997) und eines Sprichwörterbuchs gearbeitet. Bei Letzterem handelt es sich um 300 Artikel zu deutschen Sprichwörtern, die im Rahmen des EU-Projekts SprichWort[28] erarbeitet wurden. Alle diese Wörterbücher werden gezielt für eine Publikation im Internet aufbereitet, sowohl was die Art der Datenstrukturierung als auch was die Form der Darstellung betrifft.

OWID hat seine Arbeit als eigenes Projekt erst 2008 aufgenommen. Vorgänger war jedoch das *elexiko*-Portal, welches schon sechs Jahre früher begonnen wurde (vgl. Müller-Spitzer 2008a, b; Klosa 2008). Die Modellierung, die allen Wörterbüchern von OWID zu Grunde liegt, wurde daher bereits in den Jahren 2002ff. entwickelt. Die Leitlinien, die der Modellierung zu Grunde liegen, sind folgende (für eine ausführliche Begründung und Erläuterung der Richtlinien s. Müller-Spitzer 2007a, 2008a, 2008b, 2011):

[28] Siehe dazu auch die IDS-Projektbeschreibung. Die SprichWort-Artikel sind ebenso auf der Sprichwortplattform, in der Sprichwortdatenbank Deutsch, abrufbar (http://www.sprichwort-plattform.org/sp/Sprichwort).

© 2011 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

- Die Modellierung ist XML-basiert, um die Softwareunabhängigkeit und Langlebigkeit der Daten zu gewährleisten.
- Die Modellierung ist maßgeschneidert, um einen genauen Zuschnitt auf die Erfordernisse der einzelnen Wörterbücher im Portal sicherzustellen.
- Die Modellierung ist sehr ‚streng' und genau, um die Lexikographen bei der Einhaltung der formalen Artikelstruktur so gut wie möglich zu unterstützen.
- Die Modellierung ist so granular wie möglich, um eine bestmögliche Flexibilität hinsichtlich der Darstellung und des Zugriffs zu gewährleisten. Diese letzten beiden Aspekte können mit dem Terminus *Inhaltsstrukturmodellierung* zusammengefasst werden (vgl. Müller-Spitzer 2007a, 152ff.).

Um Objekte, die in verschiedenen Wörterbüchern in OWID vorkommen, nur einmal zu modellieren, wurde eine DTD-Bibliothek für OWID angelegt.[29] In dieser DTD-Bibliothek finden sich DTDs mit Bausteinen, die in allen Wörterbüchern vorkommen (wie Belege, Kommentare, Abbildungen etc.), sowie DTDs für objektübergreifende Gruppen (wie Elemente, die sowohl für die Einwortlemmata in *elexiko* sowie für die Neologismen benötigt werden), Bausteinen für einzelne Wörterbücher (z.B. für die Neologismen der Neologismentyp, der sowohl bei Einwort- wie bei Mehrwortlemmata angegeben wird) und zuletzt die Kopf-DTDs für die einzelnen Wörterbücher, die aus der DTD-Bibliothek die Elemente zusammenziehen, die für das jeweilige Wörterbuch relevant sind. So kann man am Aufbau der DTD-Bibliothek bereits erkennen, welche Angaben wörterbuchübergreifend gleich modelliert sind und sich daher z.B. für erweiterte, wörterbuchübergreifende Suchen eignen. Außerdem wird auch deutlich, dass es sich bei OWID nicht um ein Wörterbuchportal mit völlig unterschiedlichen, untereinander nicht verbundenen Ressourcen handelt, sondern um ein Wörterbuchnetz (im Sinne von Engelberg/Müller-Spitzer, im Erscheinen). Im Folgenden sollen die Modellierungsprinzipien anhand eines Beispiels – der Angaben zur Wortbildung – illustriert werden.

DTD-Bibliothek für OWID

Bausteine für alle Wörterbücher		**allg-entities.dtd**	**allg-elemente.dtd**	
Bausteine für übergreifende Objektgruppen	**ewl-objekte.dtd**	**mwl-objekte.dtd**	**ewl_mwl-objekte.dtd**	**ewl-grammatik.dtd**
Bausteine einzelner Wörterbücher		**elexikoBA-allgobj.dtd**	**neo-allgobj.dtd**	
Kopf-DTDs für Wörterbücher	***elexiko*** **elexikoAA-ewl.dtd** **elexikoBA-ewl.dtd**	***Neologismen*** **neo-ewl.dtd** **neo-mwl.dtd**	***Wortverbindungen*** **mwl.dtd** **wv.dtd**	***Schulddiskurs*** **zeitreflektion 1945-55.dtd**

Abb. 2: DTD-Bibliothek für OWID

[29] Da die Modellierung bereits 2002 entwickelt wurde, wurden XML-DTDs verwendet. Intern werden diese mittlerweile zu XML-Schemata konvertiert.

© 2011 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

3. Beispiel: Angaben zur Wortbildung

3.1 Modellierung

Die Angaben zur Wortbildung, die für *elexiko* und das Neologismenwörterbuch (für die Einwortlemmata) einheitlich sind, sind analog zum Modellierungsansatz sehr granular ausgezeichnet. Es werden nicht nur die Wortbildungsarten wie Ableitung, Zusammensetzung etc. codiert, sondern es werden auch die einzelnen gebildeten Teile näher bestimmt. Bei Präverbfügungen wie „weißwaschen“ (*elexiko*) oder „schönrechnen“ (Neologismenwörterbuch) (vgl. Abbildung 3 und 4) werden beispielsweise nicht nur die Art der Wortbildung (Präverbfügung), sondern auch die einzelnen Bestandteile (Präverb, verbale Basis) einzeln codiert und – in *elexiko* – die Wortbildungsbedeutung spezifiziert. Die einzelnen Bestandteile werden – wenn möglich – mit den korrespondierenden Wortartikeln in der Datenbasis verlinkt (ref-ID-Attribute). Online wird nur ein kleiner Teil dieser in der XML-Instanz codierten Informationen angezeigt (vgl. Klosa 2011).

```
<vb-wortbildung>
<praeverbfg>
<praeverbA
     basistyp="adjektiv"
     artikel-refid="137003"
     lesart-refid="0">weiß</praeverbA>
<verb-basisA
     basistyp="verb"
     artikel-refid="136664"
     lesart-refid="0">waschen</verb-basisA>
    </praeverbfg>
<vb-wortblgbedeutungA
     bezeichnung="aktive-zustandsveraenderg"/>
</vb-wortbildung>
```

Abb. 3: Ausschnitt aus der XML-Instanz zum Wortartikel weißwaschen (*elexiko*)

```
<vb-wortbildung>
<praeverbfg>
     <praeverbA
     basistyp="adjektiv"
     artikel-refid="288474"
     lesart-refid="0">schön</praeverbA>
<verb-basisA
     basistyp="verb"
     artikel-refid="283509"
     lesart-refid="0">rechnen</verb-basisA>
</praeverbfg>
</vb-wortbildung>
```

Abb. 4: Ausschnitt aus der XML-Instanz zum Wortartikel schönrechnen (Neologismenwörterbuch)

© 2011 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

3.2 Recherchemöglichkeiten für die Lexikographen

Wichtige Gründe, weshalb die Modellierung bzw. Auszeichnung der lexikographischen Daten in OWID so feingranular ist, ist zum einen die Unterstützung der Lexikographen bei der Einhaltung der formalen Artikelstruktur, aber genauso sind es auch die Möglichkeiten, sehr gezielt auf die erarbeiteten Daten zugreifen zu können. So können z.B. Konsistenzprüfungen über verschiedene Wortartikel hinweg erheblich erleichtert werden.

Alle Daten von OWID werden im „Electronic Database Administration System (EDAS)“ gespeichert, einem Datenbankmanagementsystem basierend auf Oracle 11.[30] Intern können die beteiligten Lexikographen über OWID eine XPath-Suche benutzen, in der die Autoren alles, was XML-basiert ausgezeichnet ist, auch abfragen können. So kann beispielsweise im Bereich der Wortbildung nach allen Präverbfügungen gesucht werden (`//praeverbfg`, vgl. Abbildung 5) oder auch spezifischer nach allen Präverbfügungen mit einem Adjektiv als Präverb (`//praeverbA[@basistyp="adjektiv"]`, vgl. Abbildung 6). Genauso kann auch auf alle Lemmata, bei denen eine bestimmte Wortbildungsbedeutung angegeben wurde, zugegriffen werden.

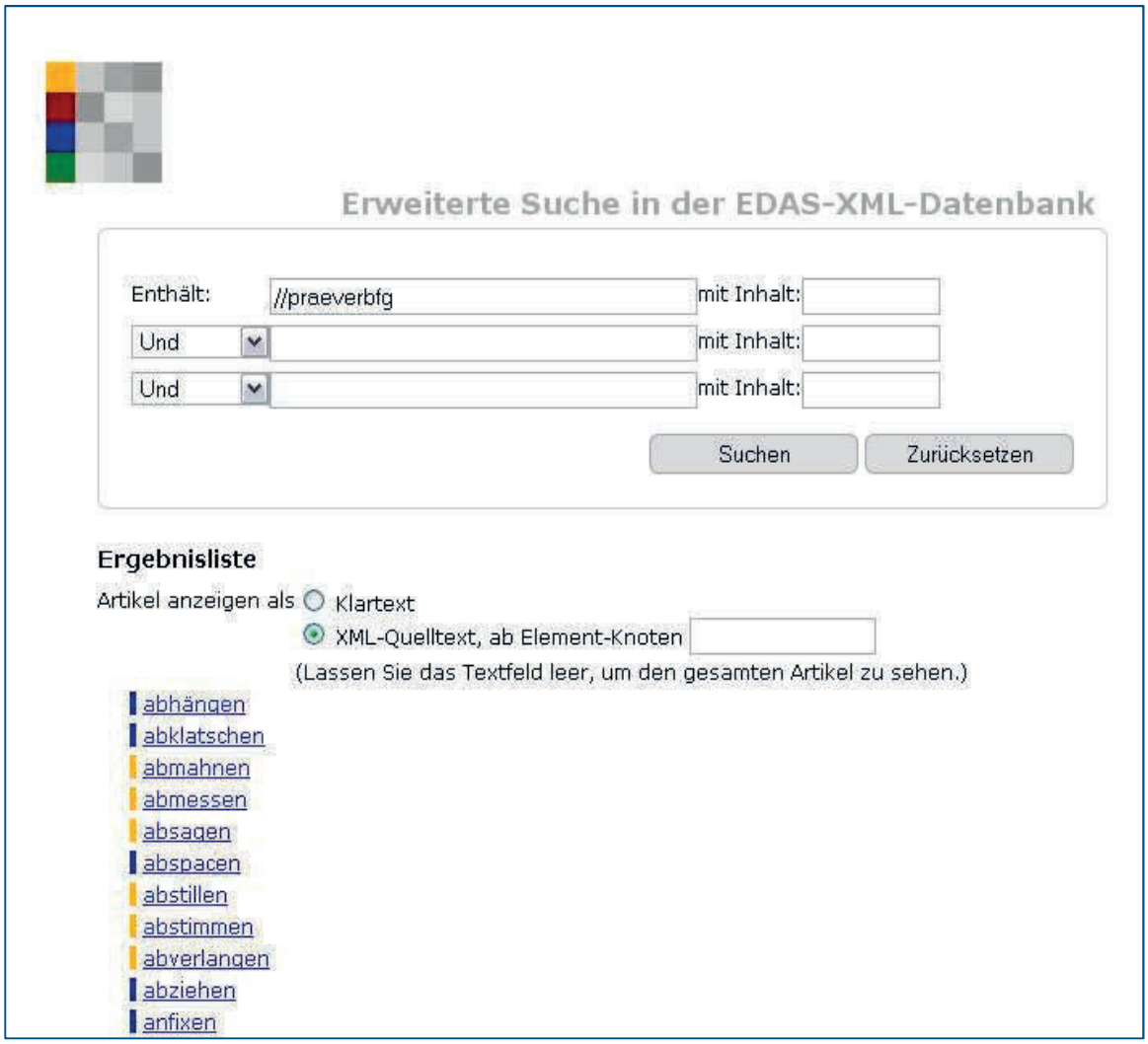

Abb. 5: XPath-basierte interne Suche über OWID nach allen Präverbfügungen

[30] EDAS wurde von Roman Schneider aus der Abteilung Grammatik des IDS entwickelt.

© 2011 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

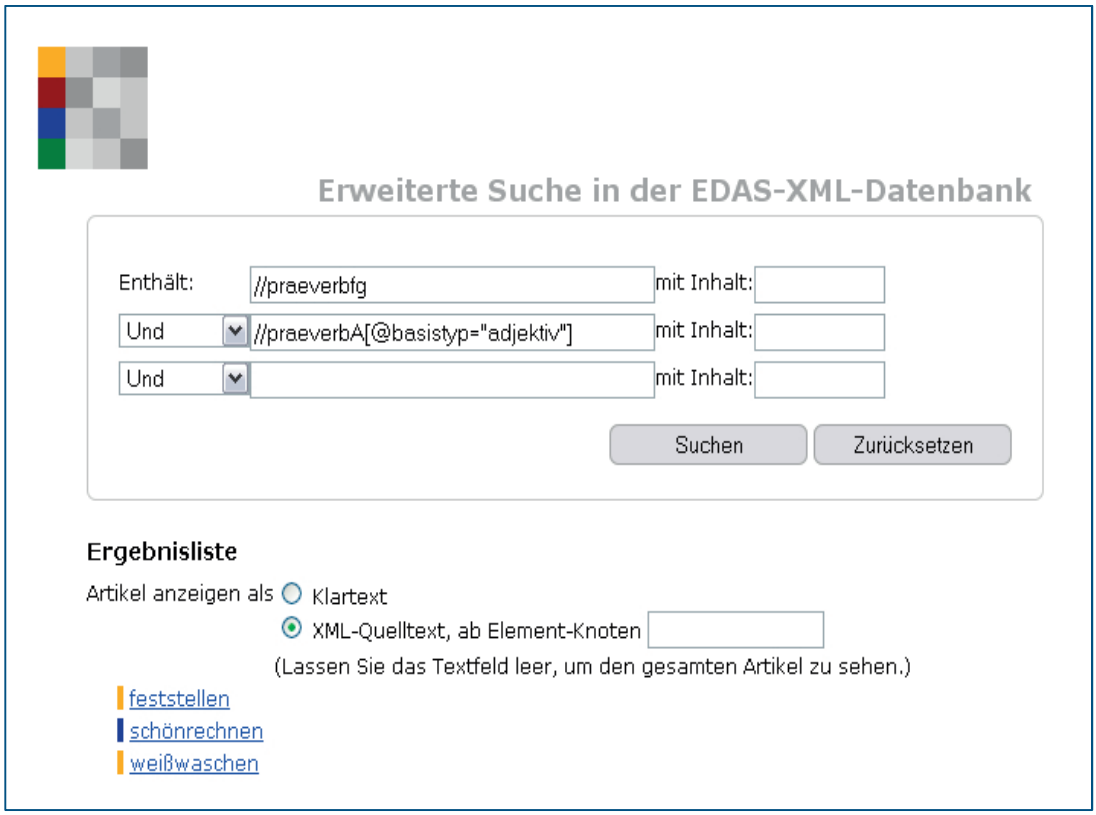

Abb. 6: XPath-basierte interne Suche über OWID nach allen Präverbfügungen mit einem adjektivischen Präverb

3.3 Erweiterte Suchen für die Nutzer

Für Endbenutzer eignen sich diese XPath-basierten Suchen nicht, denn Voraussetzung für eine erfolgreiche Benutzung ist die genaue Kenntnis der XML-Modellierung. Dies würde bedeuten, dass – im Falle von *elexiko* – ein Benutzer über 400 Elemente und über 300 zugehörige Attribute kennen müsste, d.h., die vollständige XML-Struktur müsste nach außen so dokumentiert sein, dass Außenstehende sich einarbeiten könnten. Zusätzlich müsste die XPath-Syntax erlernt werden. Die Hürden, eine solche Suche benutzen zu können, sind demnach zu hoch. Außerdem eröffnet eine solche Suche einen zu offenen Zugriff auf die lexikographischen Daten, der vor Missbrauch (beispielsweise dem Herunterladen ganzer Artikel) schwer zu schützen ist.

Für die Benutzer haben wir deshalb eine andere Form der erweiterten Suche für den Bereich der Wortbildung entwickelt, mit der wir versucht haben, die Gliederung des Angabebereichs graphisch zu veranschaulichen und so den Benutzern einen leichten Zugang zu sehr spezialisierten Suchen zu eröffnen. Diese neue Form der erweiterten Suche ist noch im Pilotstadium und soll voraussichtlich Ende 2011 unter den erweiterten Suchen von OWID online verfügbar sein (vgl. Abbildung 7 und 8).

Die Abbildung kann den interaktiven Aufbau der graphischen Suche nur bedingt verdeutlichen, deshalb wird das Vorgehen kurz erläutert: Klickt ein Benutzer auf den Kasten „Wortbildungsart“, werden folgende Kästen expandiert: „Derivation“, „Komposition“, „Präverbfügung“ und „Kurzwortbildung“. Klickt man wiederum auf eine dieser Arten, expandiert ggf. der nächste relevante Teil des Strukturbaumes; je nach Angabe wiederum als Baum oder als Auswahlmenü (wie im Bereich der Präverbfügung, s. Abbildung 7). Ist schon durch die Aus-

© 2011 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

wahl der Wortbildungsart das Suchergebnis hinreichend überschaubar, wird ein kurzer erläuternder Dialog angezeigt (wie im Bereich der Kontamination, vgl. Abbildung 8). Das Suchergebnis enthält bei dieser erweiterten Suche bereits alle relevanten Angaben als Auszug aus den zugehörigen XML-Instanzen.

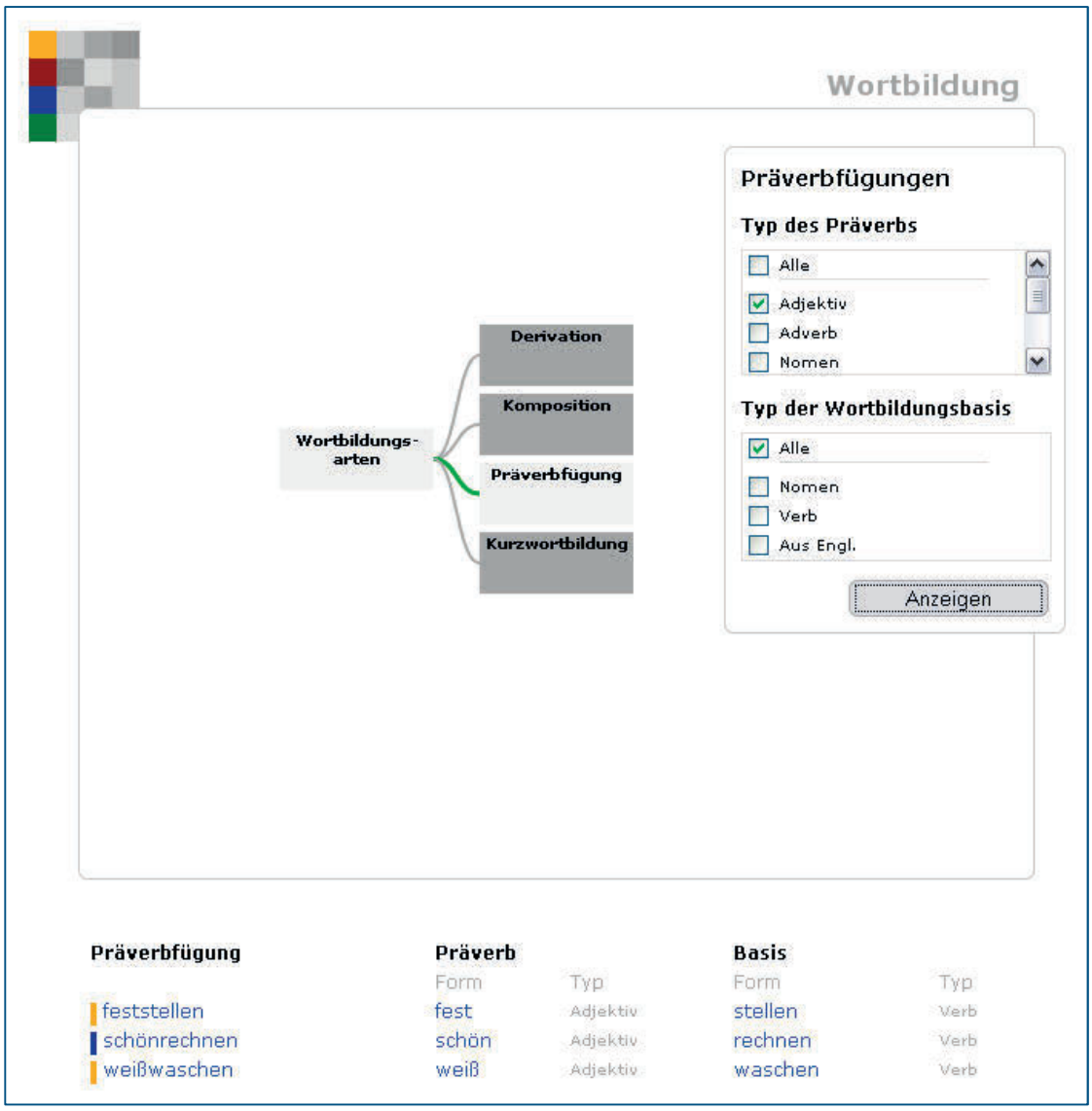

Abb. 7: Erweiterte OWID-Suche nach Präverbfügungen

Mit dieser neuartigen Form von erweiterten Suchen soll OWID als Portal für wissenschaftliche Lexikographie wie auch als Experimentierplattform dienen, in der neue Darstellungsmöglichkeiten erprobt und anhand empirischer Wörterbuchbenutzungsforschung (s. www.benutzungsforschung.de) auf ihre Anwendbarkeit überprüft werden können.

© 2011 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

Abb. 8: Erweiterte OWID-Suche nach Kontaminationen

4. Unterstützung beim redaktionellen Arbeiten

Auch bei der redaktionellen Arbeit sollte die granulare, strenge Modellierung eine Unterstützung sein; so war der Anspruch bei der Entwicklung der DTDs. Ein wichtiger Bereich dabei ist die Erarbeitung, Verwaltung und Konsistenzsicherung der Verweisstrukturen lexikographischer Daten (vgl. Müller-Spitzer 2007b). Um zu verdeutlichen, wie in OWID, in diesem Fall genauer in *elexiko*, die Lexikographen dabei formal unterstützt werden, soll hier als ein Beispiel die Verwaltung der sinn- und sachverwandten Wörter demonstriert werden.

In *elexiko* werden zu allen Stichwörtern bzw. ihren Einzelbedeutungen (Lesarten) möglichst exhaustiv sinnverwandte Stichwörter (ggf. mit zugehörigen Einzelbedeutungen) korpusbasiert erarbeitet und im Wortartikel verzeichnet (vgl. u. a. Storjohann 2006 und 2011). Abbildung 9

© 2011 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

zeigt als ein Beispiel einen Ausschnitt der sinnverwandten Wörter der Lesart ‚Material' im Wortartikel Holz.

Abb. 9: Ausschnitt aus dem *elexiko*-Wortartikel Holz, Lesart ‚Material', Bereich ‚Sinnverwandte Wörter'

Die Vernetzungsstruktur in *elexiko* ist demnach sehr umfangreich und ausführlich. Zur Verdeutlichung der Dimension: 1.250 ausgearbeitete Wortartikel enthalten 26.488 Relationspartner insgesamt, d.h., im Durchschnitt werden in *elexiko* 21 Relationspartner pro Stichwort verzeichnet. In den XML-Instanzen werden im Bereich der sinnverwandten Wörter die Typen der Vernetzung genau codiert, also ob es sich um Synonyme, inkompatible Partner oder komplementäre Partner etc. handelt, evtl. Kommentare zu Vernetzungen werden skopusgenau abgespeichert, und die ID des Zielstichworts bzw. der zugehörigen Lesart wird festgehalten. Letzteres ist besonders schwierig konsistent zu halten. *elexiko* ist ein im Aufbau befindliches Wörterbuch, d.h., Wortartikel werden kontinuierlich erarbeitet. Die Vernetzungen im Bereich der sinnverwandten Wörter sollen möglichst lesartenbezogen sein. Wenn nun aber beispielsweise der Artikel „Holz" erarbeitet und „Beton" als inkompatibler Partner verzeichnet wird, „Beton" aber noch nicht bearbeitet ist, dann kann auch keine Einzelbedeutung als Zieladresse angegeben werden. Wenn der Wortartikel „Beton" allerdings zu einem späteren Zeitpunkt bearbeitet wird, wünscht man sich als Lexikographin eine Information darüber, ob das Stichwort „Beton" bereits als Ziel in anderen Artikeln genannt wird. Genauso sollte überprüft werden können, ob Synonymverweise immer in beide Richtungen angelegt sind, wie dies konzeptionell gewünscht ist.

Die Modellierung kann für diese Anfragen die Basis liefern, indem alle relevanten Informationen in der XML-Struktur codiert sind. Allerdings ist zur Auswertung für die Lexikographen eine gesonderte Software nötig. Ein solcher Vernetzungsmanager wurde für *elexiko* im Rah-

© 2011 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

men des Projekts BZV*elexiko* entwickelt.[31] Die Arbeit mit diesem Vernetzungsmanager kann grob folgendermaßen skizziert werden: Arbeitet ein Lexikograph im XML-Editor an einem Wortartikel (wie hier an „Holz“) und öffnet dann den Vernetzungsmanager, werden ihm alle eingehenden sowie alle aus dem Wortartikel ausgehenden Vernetzungen angezeigt. In der Spalte „Status“ wird zu jeder dieser Vernetzungen vermerkt, ob sie korrekt ist oder ob beispielsweise die Ziel-ID nicht korrekt ist oder die Vernetzung in der einen Richtung lesartenbezogen, in der anderen aber lesartenübergreifend ist etc. (vgl. Abbildung 10). Möchte der Lexikograph fehlerhafte Vernetzungen korrigieren, unterstützt der Vernetzungsmanager die Arbeit dahingehend, dass die relevanten Teile der Zielinstanz über den Manager geladen, korrigiert und wieder in die Datenbank eingecheckt werden können. Basis dafür, dass diese Abfragen trotz hoher Komplexität sehr performant sind, ist eine Linkdatenbank, die alle relevanten Extrakte aus den XML-Instanzen beinhaltet und über die die Abfragen des Vernetzungsmanagers laufen (vgl. Meyer/Müller-Spitzer 2010).

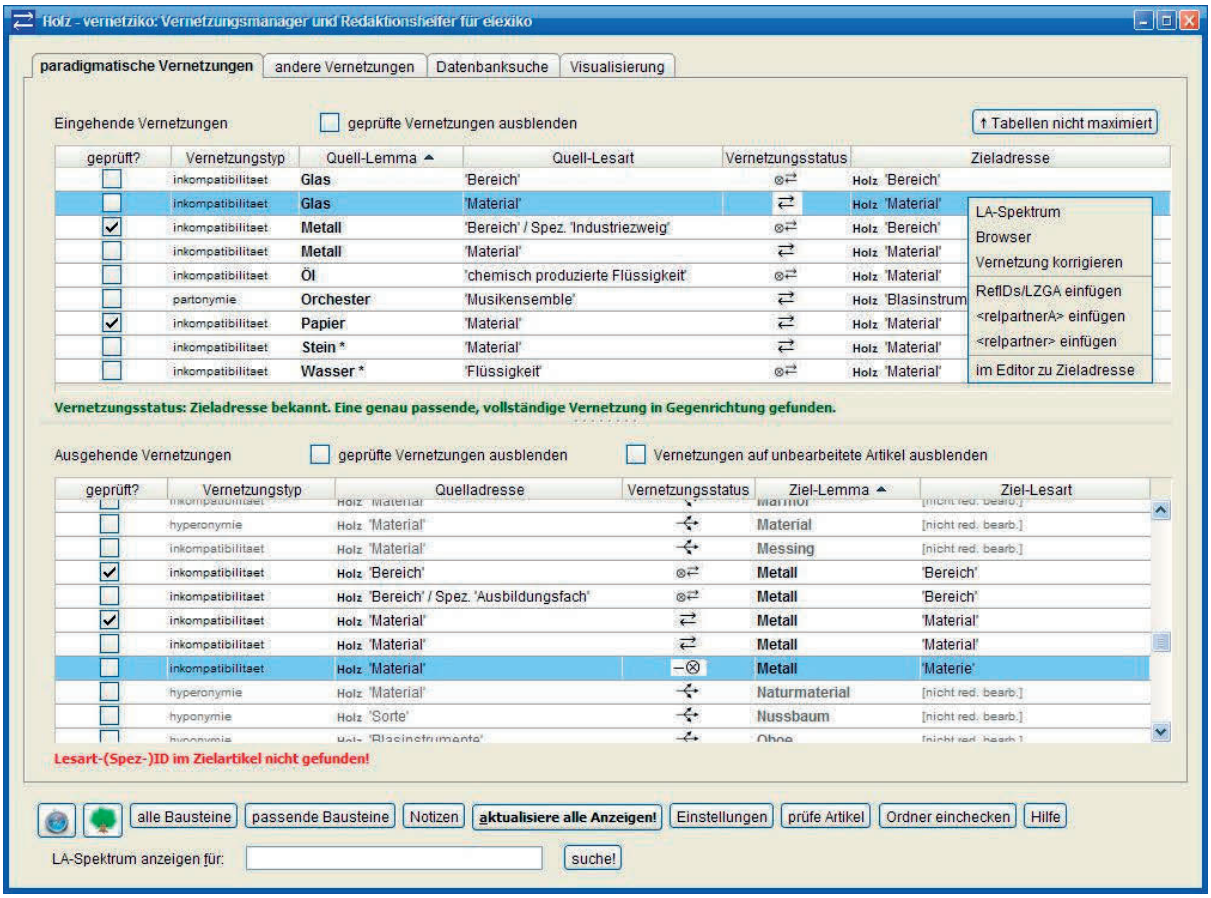

Abb. 10: Bildschirmansicht des Vernetzungsmanagers (Übersicht über ein- und ausgehende Vernetzungen)

Der hier dargestellte Vernetzungsmanager bietet darüber hinaus erweiterte Datenbanksuchen, die vor allem für die leitenden Lexikographen von Interesse sind. So können inhaltliche XPath-Abfragen mit weiteren Informationen aus der Datenbank wie Bearbeitungsstatus oder Bearbeitungszeit kombiniert werden, Gruppen von Instanzen können zusammen ein- und ausgecheckt werden oder Suchen mit selbstgewählten Extrakten aus den XML-Instanzen können spezifiziert werden (vgl. Abbildung 11). Seit der Einführung dieses Vernetzungsmanagers hat sich die redaktionelle Arbeit in *elexiko* daher deutlich vereinfacht und die Konsistenz der Da-

[31] Peter Meyer hat sowohl den Vernetzungsmanager als auch das Tool zur graphischen Visualisierung von Vernetzungen entwickelt.

© 2011 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

ten hat sich erheblich – auch durch intensive redaktionelle Nacharbeiten – verbessert. Die Basis durch die genaue Modellierung war von Anfang an gegeben, allerdings fehlte bis vor kurzem eine entsprechende Softwareunterstützung.

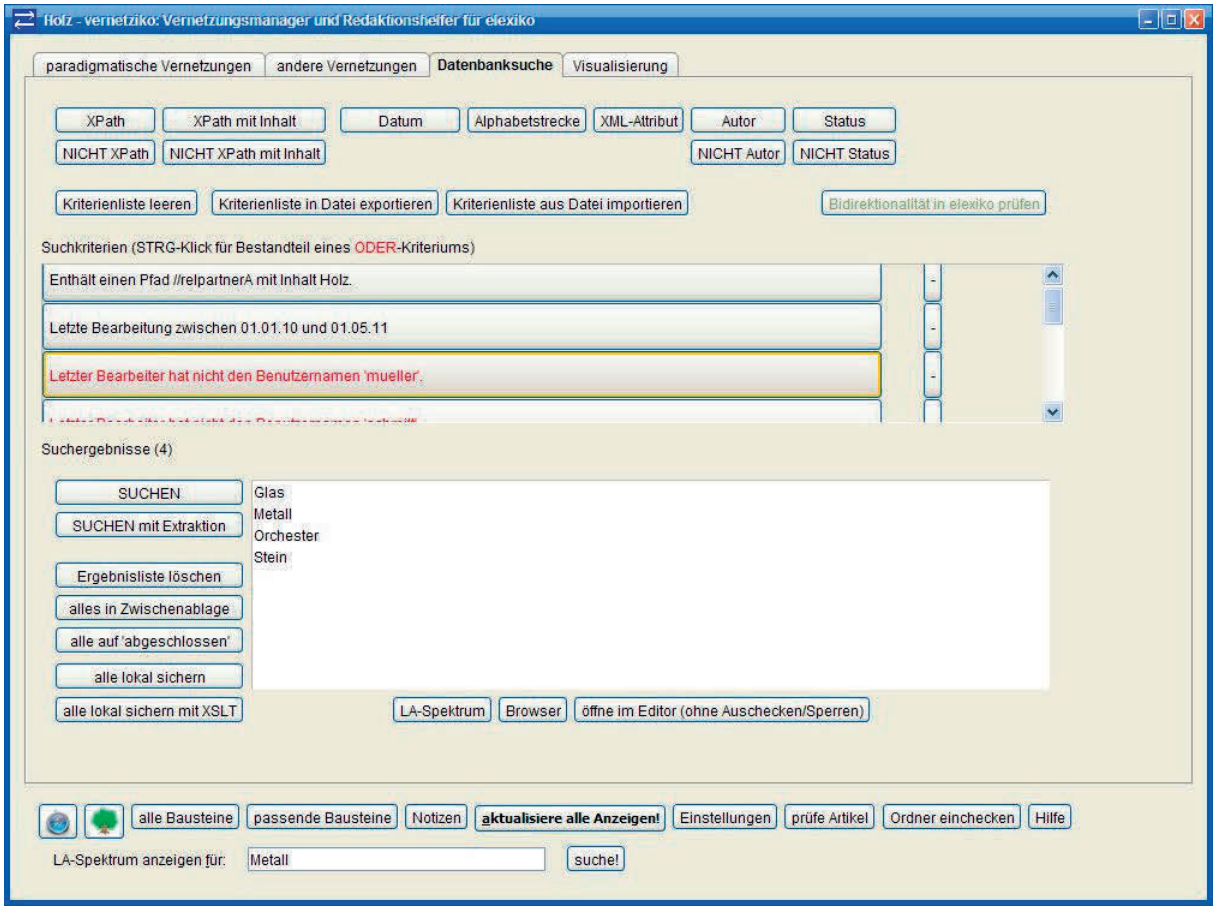

Abb. 11: Bildschirmansicht des Vernetzungsmanagers (Erweiterte Datenbanksuche)

Durch die Speicherung aller vernetzungsrelevanten Daten in einer separaten Linkdatenbank sind auch andere Darstellungen der Vernetzungen gut zu entwickeln und performant abfragbar. So haben wir beispielsweise eine graphische Visualisierung der sinnverwandten Wörter als Experimentierplattform für interne Zwecke (vgl. Abbildung 12) erarbeitet. Diese wird online noch keinem Benutzer zur Verfügung gestellt, weil noch nicht klar ist, für welche Benutzungssituation das Tool gewinnbringend einzusetzen ist.

© 2011 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

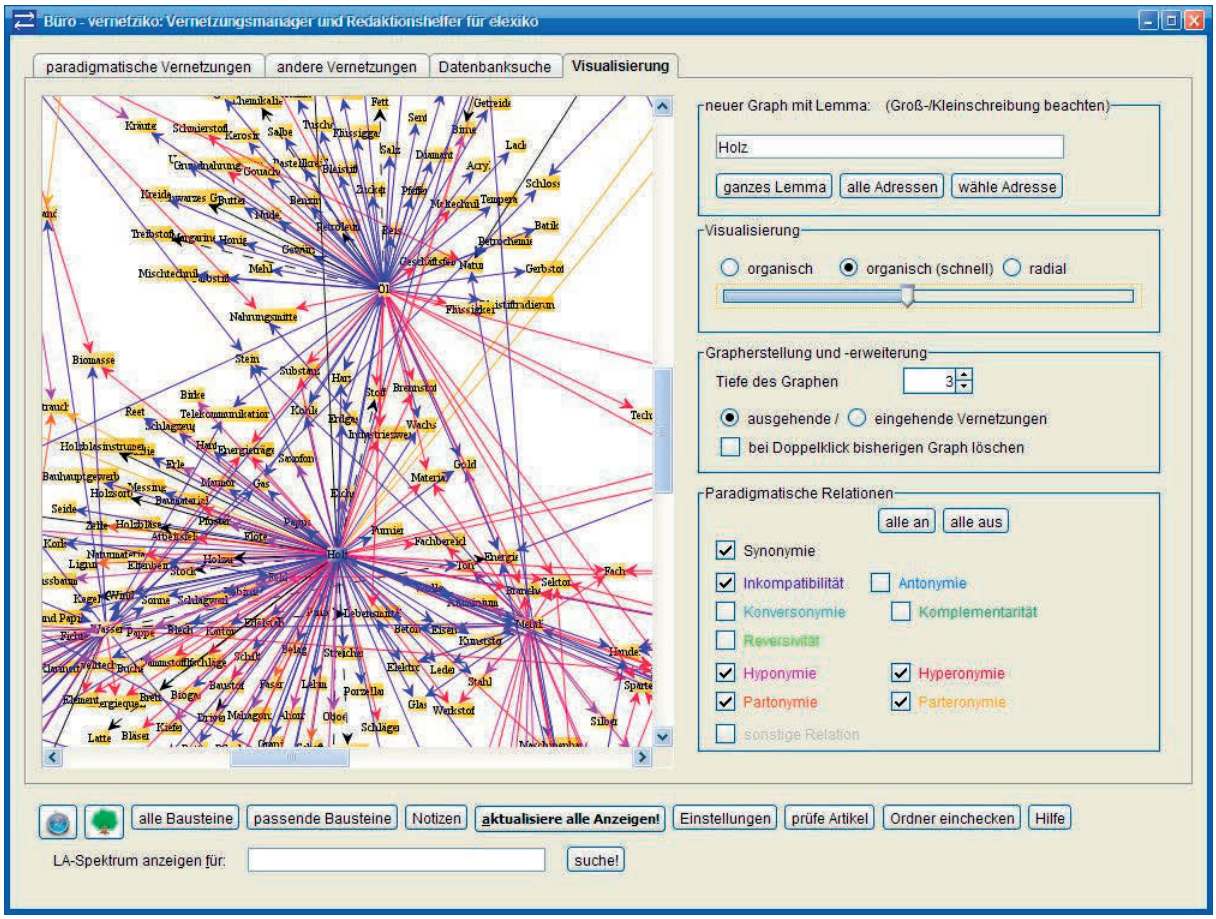

Abb. 12: Ausschnitt aus einer Bildschirmansicht des elexPlorer (Tool zur graphischen Visualisierung der sinnverwandten Wörter)

5. Vor- und Nachteile des Modellierungsansatzes

Nachdem in den vorangegangenen Abschnitten die Arbeit mit der Modellierung in OWID verdeutlicht wurde und damit auch mehr die Vorteile des gewählten Modellierungsansatzes im Vordergrund standen, sollen abschließend kurz Vor-, aber auch Nachteile des hier vorgestellten Ansatzes gegenübergestellt werden. Dabei werden zwei wesentliche Aspekte herausgegriffen: die maßgeschneiderte Modellierung und die Ausrichtung als Inhaltsstrukturmodellierung.

5.1 Maßgeschneiderte Modellierung

Eine maßgeschneiderte Modellierung zu entwickeln ist in einer Zeit zunehmender Standardisierungsbemühungen und Gründung von Infrastrukturprojekten mit dem Ziel, ein möglichst hohes Maß an Austauschbarkeit von Daten zu gewährleisten (wie TextGrid, CLARIN, D-Spin etc.[32]), in gewissem Sinne unmodern. Trotzdem haben wir mit dieser Entscheidung nur gute Erfahrungen gemacht. Eine maßgeschneiderte Modellierung ermöglicht einen passgenauen Zuschnitt auf die Erfordernisse der einzelnen Wörterbücher in einem Portal. Gerade wenn die lexikographischen Daten kontinuierlich erarbeitet werden, ist dies ein enormer Vorteil. Wollte

[32] Siehe www.textgrid.de, www.clarin.eu, http://weblicht.sfs.uni-tuebingen.de/.

© 2011 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

man eine Standard-Modellierung so genau auf die individuellen Bedürfnisse der einzelnen lexikographischen Ressourcen anpassen, wäre auch eine ursprünglich standardbasierte Modellierung von einer maßgeschneiderten nicht weit entfernt. Sehr anders sieht die Situation bei der Strukturierung retrodigitalisierter Wörterbücher aus (vgl. den Beitrag von Hildenbrandt in diesem Band). Allerdings muss man zu allen Standard-Modellierungen wie der TEI (www.tei.org) bemerken: Der Vorteil einer Standard-Modellierung soll die Austauschbarkeit von Daten sein; gleichzeitig muss eine Standard-Modellierung hinreichend offen sein, um von unterschiedlichsten Projekten angewendet werden zu können. Diese beiden Pole stehen oft im Widerstreit zueinander. Man kann davon ausgehen, dass zwei lexikographische Projekte, die unabhängig voneinander beispielsweise die P5-Richtlinien der TEI anwenden, ihre Daten trotzdem nicht ohne Weiteres austauschen können, da diese Richtlinien eben sehr unterschiedlich angewendet werden können. Trotzdem sollte man sich immer über die Standards informieren (so wie es am IDS auch getan wurde), da meist viele Fachleute an der Entwicklung beteiligt sind und man diese Richtlinie als Orientierung benutzen kann, auch wenn die gesamte Modellierung maßgeschneidert ist.

Da die für OWID gewählte Modellierung allerdings möglichst feingranular und genau ist, lässt sich diese maßgeschneiderte Modellierung jederzeit in eine standardbasierte Modellierung z.B. analog zu den TEI-Richtlinien überführen, da diese Standard-Modellierungen immer sehr viel allgemeiner gehalten sind. Für die Beteiligung an Infrastrukturprojekten wurde eine solche Migration bereits in der Praxis erprobt.

5.2 Inhaltsstrukturmodellierung

Der Ansatz einer Inhaltsstrukturmodellierung, d.h. die möglichst granulare Strukturierung aller lexikographischen Angaben, strikt orientiert am inhaltlichen Gehalt der Daten, hat wie oben ausgeführt die Vorteile, die Lexikographen bei der Einhaltung der formalen Artikelstruktur bestmöglich zu unterstützen sowie die Basis für sehr flexible Zugriffsmöglichkeiten für Lexikographen und Endbenutzer zu legen. Dieser Ansatz birgt allerdings auch Nachteile: Der Aufwand, aus solchen granular gegliederten Daten, die von ihrem Aufbau her mehr an den linguistischen Strukturen als an der Gliederung des Wortartikels im Online-Wörterbuch orientiert sind, eine Präsentation (z.B. über XSLT-Stylesheets) zu entwickeln, ist sehr hoch. Beispielsweise muss für jedes der 400 Elemente und zugehörigen Attribute von *elexiko* festgelegt werden, wie die entsprechende Information in einer Online-Ansicht dargestellt werden soll. Außerdem ist die Modellierung rein am Inhalt orientiert, d.h., viele Elemente müssen in eine andere Reihenfolge etc. gebracht werden, um im Wortartikel in der gewünschten Form zu erscheinen. Diese grundlegende Schwierigkeit wird für *elexiko* verschärft dadurch, dass zunächst die Modellierung der Daten entwickelt wurde (an der inhaltlichen Struktur orientiert) und erst danach die Gliederung der Wortartikel für die Online-Darstellung festgelegt wurde. Die Schere zwischen der Gliederung der Daten in der Datenbasis und der Online-Ansicht klafft daher an manchen Stellen weit auseinander.

Ein weiterer, etwas diffiziler zu erklärender Nachteil kommt hinzu: Der Anspruch der Inhaltsstrukturmodellierung ist es, Inhalte zu modellieren und die Skopusbeziehungen beispielsweise von Angaben und dazugehörigen Kommentaren so genau wie möglich abzubilden, losgelöst von Aspekten der Präsentation. Die granulare Strukturierung erlaubt es, aus einer so strukturierten Datenbasis für eine Präsentation einen lexikographischen Text ‚zusammenzubauen'. Wenn nun aber ein Wörterbuch wie *elexiko* oder das Neologismenwörterbuch über Jahre erarbeitet wird und die Wortartikel in der Online-Ansicht eine fest definierte Gestalt haben, kann über die Jahre Folgendes passieren: Die Lexikographen denken bei der Datenerarbeitung

© 2011 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

und -strukturierung mehr daran, wie die Daten im Wortartikel online aussehen sollen, als daran, wo sie eigentlich inhaltlich hingehören. Ein Beispiel: Zu einem Relationspartner aus dem Bereich der sinnverwandten Wörter soll ein Kommentar gegeben werden. Fiktiv sei es so, dass online sowohl ein Kommentar zu allen Relationspartnern dieses Typs wie auch die Kommentare zu einzelnen Partnern an der derselben Stelle erscheinen können. In einem solchen Fall kann es passieren, dass ein Lexikograph den Kommentar an der falschen Stelle in der Artikelstruktur eingibt (also an der Stelle, an der eigentlich nur ein Kommentar zu allen Relationspartnern stehen darf), weil er nur überprüft, ob der Kommentar online korrekt erscheint. Zwar wird auch das XML-Tagging der Artikel in *elexiko* Korrektur gelesen, aber gerade diese Skopusbeziehungen sind sehr schlecht zu überblicken. Dies ist nur ein subtiles Beispiel für dieses Phänomen, aber gerade das Denken von der Präsentation her kann zu einem Problem bei der Inhaltsstrukturmodellierung werden, ist aber bei einer langjährigen lexikographischen Routine kaum zu vermeiden.

6. Schlussbemerkung

Für das Wörterbuchnetz OWID hat sich der hier dargestellte Modellierungsansatz bewährt. Die Modellierung ist ausgerichtet auf unterschiedliche lexikographische Ressourcen, die neu erarbeitet werden, die bestimmte Angabebereiche teilen und so zum Teil gemeinsame Inhaltsstrukturen haben, zum Teil individuell verschieden strukturiert sein müssen. Der vorliegende Beitrag sollte einen Einblick in die praktische Arbeit von OWID bieten und so möglichst die Diskussion zwischen lexikographischen Projekten mit unterschiedlichen Modellierungsansätzen befruchten.

7. Literatur

Benutzungsforschung. Internet: www.benutzungsforschung.de. (Stand: Oktober 2011).

CLARIN – Common Language Resources and Technology Infrastructure. Internet: http://www.clarin.eu. (Stand: Oktober 2011).

elexiko (2003ff.), in: OWID – Online Wortschatz-Informationssystem Deutsch, hg. v. Institut für Deutsche Sprache, Mannheim. Internet: www.owid.de/elexiko_/index.html. (Stand: Oktober 2011).

Engelberg, Stefan/Müller-Spitzer, Carolin (im Erscheinen): Dictionary portal. In: Dictionaries. An international encyclopedia of lexicography. Supplementary volume: Recent developments with special focus on computational lexicography, hg. von Gouws, Rufus H./Heid, Ulrich/Schweickard, Wolfgang/Wiegand, Herbert Ernst. Berlin/New York.

E-ValBU – Das elektronische Valenzwörterbuch deutscher Verben. Internet: http://hypermedia2.ids-mannheim.de/evalbu/index.html. (Stand: Oktober 2011).

Feste Wortverbindungen (2007ff.), in: OWID – Online Wortschatz-Informationssystem Deutsch, hg. v. Institut für Deutsche Sprache, Mannheim. Internet: www.owid.de/Wortverbindungen/index.html. (Stand: Oktober 2011).

Harras, Gisela/Winkler, Edeltraud/Erb, Sabine/Proost, Kristel (2004): Handbuch deutscher Kommunikationsverben. Teil 1: Wörterbuch. (= Schriften des Instituts für Deutsche Sprache 10.1). Berlin/New York.

Herberg, Dieter/Steffens, Doris/Tellenbach, Elke (1997): Schlüsselwörter der Wendezeit. Wörter-Buch zum öffentlichen Sprachgebrauch 1989/90. (= Schriften des Instituts für deutsche Sprache 6). Berlin/New York.

IDS – Institut für Deutsche Sprache. Internet: http://www.ids-mannheim.de. (Stand: Oktober 2011).

Klosa, Annette (Hg.) (2008): Lexikografische Portale im Internet. (= OPAL – Online publizierte Arbeiten zur Linguistik 1/2008). Mannheim.

Klosa, Annette (2011): Korpusgestützte Angaben zu Grammatik und Wortbildung. In: Klosa, Annette (Hg.): *elexiko*. Erfahrungsberichte aus der lexikographischen Praxis eines Internetwörterbuchs. (= Studien zur deutschen Sprache 55). Tübingen, S. 145-156.

Meyer, Peter/Müller-Spitzer, Carolin (2010): Consistency of Sense Relations in a Lexicographic Context. In: Barbu Mititelu, Verginica/Pekar, Viktor/Barbu, Eduard (Hg.): Proceedings of the Workshop „Semantic Relations. Theory and Applications“, 18 May 2010, at the International Conference on Language Resources and

© 2011 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.

Evaluation (LREC) 2010, Malta. Internet: http://www.lrec-conf.org/proceedings/lrec2010/workshops/W9.pdf. (Stand: Oktober 2011).

Müller-Spitzer, Carolin (2007a): Der lexikografische Prozess. Konzeption für die Modellierung der Datenbasis. (= Studien zur deutschen Sprache 42). Tübingen.

Müller-Spitzer, Carolin (2007b): Vernetzungsstrukturen lexikografischer Daten und ihre XML-basierte Modellierung. In: Hermes 38, S. 137-171.

Müller-Spitzer, Carolin (2008a): Der texttechnologische Aufbau von OWID. In: Klosa, Annette (Hg.): Lexikografische Portale im Internet. (= OPAL – Online publizierte Arbeiten zur Linguistik 1/2008). Mannheim, S. 45-55.

Müller-Spitzer, Carolin (2008b): The Lexicographic Portal of the IDS. Connecting Heterogeneous Lexicographic Resources by a Consistent Concept of Data Modelling. In: Proceedings of the 13th EURALEX International Congress. Euralex 2008. Barcelona, Spain (CD-ROM).

Müller-Spitzer, Carolin (2011): Der Einsatz einer maßgeschneiderten, feingranularen XML-Modellierung im lexikografischen Prozess. In: Klosa, Annette (Hg.): *elexiko*. Erfahrungsberichte aus der lexikographischen Praxis eines Internetwörterbuchs. (= Studien zur deutschen Sprache 55). Tübingen, S. 173-191.

OWID – Wortschatzinformationssystem Deutsch (2008ff.), hg. v. Institut für Deutsche Sprache, Mannheim. Internet: http://www.owid.de. (Stand: Oktober 2011).

Diskurswörterbuch 1945-55 (2007), in: OWID – Online Wortschatz-Informationssystem Deutsch, hg. v. Institut für Deutsche Sprache, Mannheim, www.owid.de/Diskurs1945-55/index.html. (Stand: Oktober 2011).

SprichWort-Plattform (2008-2010). Internet: http://www.sprichwort-plattform.org/sp/Sprichwort. (Stand: Oktober 2011).

Storjohann, Petra (2006): Sinnrelationen in Wörterbüchern – Neue Ansätze und Perspektiven. In: EliSe 2/2005, S. 35-61. Internet: http://www.uni-due.de/imperia/md/content/elise/ausgabe_2_2005_storjohann.pdf). (Stand: Oktober 2011).

Storjohann, Petra (2011): Paradigmatische Konstruktionen in Theorie, lexikografischer Praxis und im Korpus. In: Klosa, Annette (Hg.): *elexiko*. Erfahrungsberichte aus der lexikographischen Praxis eines Internetwörterbuchs. (= Studien zur deutschen Sprache 55). Tübingen, S. 99-129.

TEI – Text Encoding Initiative. Internet: http://www.tei-c.org/index.xml. (Stand: Oktober 2011).

TextGrid – Vernetzte Forschungsumgebung in den eHumanities. Internet: http://www.textgrid.de. (Stand: Oktober 2011).

WebLicht – Deutsche Sprachressourcen-Infrastruktur D-SPIN. Internet: http://weblicht.sfs.uni-tuebingen.de/. (Stand: Oktober 2011).

Wortverbindungen online. Plattform des Projekts Usuelle Wortverbindungen. Internet: http://wvonline.ids-mannheim.de/. (Stand: Oktober 2011).

© 2011 Institut für Deutsche Sprache, Mannheim. Alle Rechte vorbehalten.