hosted by Hamburg University Press

apropos [Perspektiven auf die Romania]

Sprache∫Literatur∫Kultur∫Geschichte∫Ideen∫Politik∫Gesellschaft

Linguistische Online-Ressourcen auf Basis traditioneller Werke *Anforderungen und digitale Möglichkeiten am Beispiel des* Romanischen Etymologischen Wörterbuchs

Florian Zacherl

apropos [Perspektiven auf die Romania]
hosted by Hamburg University Press
2022, 9
pp. 254-275
ISSN: 2627-3446

Online

https://journals.sub.uni-hamburg.de/apropos/article/view/1895

Zitierweise

Zacherl, Florian. 2022. „Linguistische Online-Ressourcen auf Basis traditioneller Werke. Anforderungen und digitale Möglichkeiten am Beispiel des *Romanischen Etymologischen Wörterbuchs.*“ *apropos [Perspektiven auf die Romania]* 9/2022, 254-275.
doi: https://doi.org/10.15460/apropos.9.1895

2022, n°9
pp. 254-275
doi: https://doi.org/10.15460/apropos.9.1895

Florian Zacherl

Linguistische Online-Ressourcen auf Basis traditioneller Werke

Anforderungen und digitale Möglichkeiten am Beispiel des *Romanischen Etymologischen Wörterbuchs*

Florian Zacherl
ist wissenschaftlicher Mitarbeiter in der IT-Gruppe Geisteswissenschaften an der Ludwig-Maximilians-Universität München.
Florian.Zacherl@itg.uni-muenchen.de

Keywords

Digitale Lexikographie – Wörterbücher – Webportale

1. Einleitung

Die manuelle Aggregation von Informationen aus gedruckten linguistischen Quellen stellt oftmals einen aufwendigen Prozess dar. Eine digitale Online-Präsentation hat das Potential diesen deutlich zu beschleunigen. Dieser Artikel analysiert am Beispiel des Romanischen Etymologischen Wörterbuchs (REW, entspricht Meyer-Lübke 1935) die Konzeption eines Webangebots auf Basis einer solchen traditionellen Quelle, das die digitalen Möglichkeiten umfassend nutzt. Der Fokus dieses Artikels liegt dabei auf den Anforderungen, die eine solche Online-Ressource erfüllen sollte, und der Frage, welche zusätzlichen Möglichkeiten sie im Gegensatz zur gedruckten Vorlage bieten kann. Primär wird die Funktionalität betrachtet, die menschlichen Nutzenden den direkten Zugriff ermöglicht (im Gegensatz zur maschinellen Nutzung).

Das REW erschien erstmals 1911, wobei als Grundlage hier die dritte und finale (neubearbeitete) Auflage dient. Diese enthält anhand von 10701[1] (vorwiegenden lateinischen) Lemmata im wesentlichen Listen der zugehörigen „romanischen Vertreter“ (Meyer-Lübcke 1935, XI) sowie vor allem in umstrittenen oder unklaren Fällen weiterführende Informationen und vom Autor abgelehnte Etymologien. Aufgrund seiner Anlage als Sammlung und kritische Einordnung der „wichtigeren der ungemein zahlreichen und vielfach weit zerstreuten etymologischen Untersuchungen auf dem Gebiete der romanischen Sprachen“ (Meyer-Lübcke 1935, VIII)

[1] Die Lemmata sind von 1 bis 9721 nummeriert. Die Anzahl ergibt sich unter Berücksichtigung von Einfügungen und Entfernungen im Vergleich zu den beiden früheren Auflagen.

ist das REW trotz seines Alters weiterhin ein unverzichtbares Hilfsmittel in der romanischen Etymologie. Das einzige vergleichbare neuere Werk, das denn gesamtromanischen Sprachraum abdeckt, ist das „Dictionnaire Étymologique Roman“, das aber bei weiten nicht den Umfang des REW hat und methodisch durchaus umstritten ist (vgl. z.B. Vàrvaro 2011).

Dieser Artikel ist im Kontext eines Dissertationsprojekts entstanden, welches sich primär mit der feingranulierten Erschließung von strukturierten Daten aus Wörterbuchtexten befasst und zu diesem Zweck das REW als Beispiel systematisch auswertet und digital publiziert (vgl. Zacherl in Vorb. a). Das in diesem Rahmen entstandene Webangebot ist als Entwurf bereits zugänglich (vgl. REWOnline), zum jetzigen Stand (14.10.2022) aber noch nicht vollständig funktional.

Das folgende Kapitel untersucht zunächst, welche Anforderungen eine Online-Ressource in Gegenüberstellung mit einer gedruckten Ressource sinnvollerweise erfüllen sollte, während im Weiteren daraus abgeleitete Anforderungen an die Datenmodellierung (Kapitel 3) und den Aufbau der Oberfläche eines Webangebots (Kapitel 4) betrachtet werden. Schließlich wird ein kurzer Ausblick auf die zusätzlichen Möglichkeiten einer maschinellen Nutzung gegeben (Kapitel 5).

2. Anforderungen an linguistische Online-Ressourcen

Die behandelten Funktionen werden hier grob in drei Kategorien eingeteilt, die allerdings nicht immer trennscharf unterschieden werden können: *Grundsätzliche Anforderungen* stehen für elementare Eigenschaften, die eine Online-Ressource erfüllen muss, um die Aufgabe der gedruckten Vorlage weiterhin zu erfüllen. Dies schließt auch ein, dass spezifische Nachteile von Online-Publikationen im Vergleich zu traditionellen Werken soweit wie möglich ausgeglichen werden. Mit *Arbeitserleichterungen* sind spezifisch digitale Werkzeuge gemeint, die klassische linguistische oder allgemein wissenschaftliche Arbeitsweisen vereinfachen und Aufgaben beschleunigen. Unter *erweiterte Möglichkeiten* werden Funktionalitäten eingeordnet, die sich nur durch die Umwandlung in die digitale Form ergeben und die nicht (oder nicht mit realistischem Aufwand) mit dem gedruckten Werk erreicht werden können.

2.1. Grundsätzliche Anforderungen

Zwei Basisbedingungen, die eine wissenschaftliche Publikation und insbesondere ein Wörterbuch erfüllen muss, sind der Zugriff auf die vollständige relevante Information einschließlich Verweisen auf andere Literatur und die Möglichkeit der kleinteiligen Zitation. Der Zugriff selbst erscheint trivial, aber gerade bei Wörterbüchern (oder ähnlichen stärker strukturierten Texten) bietet eine Online-Ressource die Möglichkeit die Inhalte aufzubereiten und auch in anderer Struktur (oder auch in verschiedenen wählbaren Formaten) darzustellen. Je größer allerdings der Unterschied zwischen der digitalen Darstellung und dem ursprünglichen Quellenmaterial ist, desto hilfreicher ist eine zusätzliche Zugriffsmöglichkeit auf den Originaltext, um das Ausgangsmaterial transparent zu machen und auch Fehler, die

im Prozess der technischen Transformation eventuell aufgetreten sind, auf einfache Weise erkennbar zu machen. Bei einem Wörterbuch bedeutet das weiterhin, dass nicht nur die eigentlichen Einträge, sondern auch andere, weniger strukturierte Abschnitte wie Vorwörter oder ähnliches online zugreifbar sind.

Für die Zitation werden traditionell Seitenzahlen und eventuell zusätzlich Zeilennummern verwendet, die in bei Print-Publikationen eine formatabhängige Notwendigkeit sind, bei digitalen Publikationen aber weder erforderlich noch sinnvoll sind (cf. z.B. Prätor 2011). Dies erzeugt eine eigene Problematik, die allerdings in Bezug auf Wörterbücher weniger relevant ist, da die dort übliche Referenzierung auf einzelne Lemmata weiterhin möglich ist. Relevant bleibt allerdings die mangelnde Stabilität (im inhaltlichen Sinne) von Online-Publikationen, die bei einem gedruckten Werk innerhalb einer Auflage zwingend gegeben ist. Eine Form von Versionierung, die eine statische Ausgabe von Texten oder Textbestandteilen erzeugt, ist also notwendig. Diese sollte bei einer vollständigen Nutzung der digitalen Möglichkeiten unmittelbar und kleinteilig sein. Das bedeutet, dass Korrekturen nach einer Änderung sofort zitierbar sind und die Version sich auf kleinere Sinnabschnitte bezieht, die im Falle eines Wörterbuchs die einzelnen Einträge sein könnten. Gerade die Unmittelbarkeit stellt einen großen Vorteil gegenüber traditionellen Publikationen dar, bei denen Änderungen nur in sehr großen Zeitintervallen eingepflegt werden können. Somit wäre die Möglichkeit wünschenswert solche Änderungen möglichst direkt und einfach vornehmen zu können (cf. Kapitel 2.3).

2.2. Arbeitserleichterungen

Um die Arbeit mit wissenschaftlichem Material zu erleichtern, bestehen vor allem die Möglichkeiten, erstens die Textbasis an sich aufzubereiten und/oder anzureichern und zweitens die Schaffung verbesserter Such- und Zugriffsfunktionalitäten. Das primäre Ziel einer Aufbereitung sollte eine verbesserte Lesbarkeit sein. Zudem soll der Wechsel zu anderen Teilen des Werks vereinfacht bzw. durch möglichst vollständige Einträge ggf. unnötig gemacht werden. In gedruckten Publikationen besteht grundsätzlich ein gewisser Zwang, Platz einzusparen, was gerade bei älteren Werken oftmals zu einer Vielzahl von abkürzenden Schreibweisen, Auslassungen unter bestimmten Bedingungen und sehr engen blockartigen Texten führt. Ein digitales Format hat diese Einschränkungen nicht[2] und kann somit die Darstellung an der optimalen Lesbarkeit und der Präferenz der Nutzenden ausrichten. Bestehende Abkürzungen im Quellenmaterial können entweder komplett ersetzt oder auf geeignete Weise aufgelöst werden. Dies erspart den Wechsel zu Abkürzungsverzeichnissen, Bibliographien oder Ähnlichem. Interne Referenzen (z.B. auf ein bestimmtes Lemma) sollten mit Links hinterlegt werden. Auch für externe Referenzen (z.B. Literaturangaben) bietet sich dies an, falls die entsprechenden Ressourcen digital zugreifbar sind und die notwendigen

[2] Grundsätzlich besteht natürlich ein begrenztes Datenvolumen. Diese Einschränkung ist allerdings meistens (und insbesondere im gegebenen Fall) vernachlässigbar. Die Darstellung (also insbesondere Abstände, Schriftgrößen etc.) wird nur durch die jeweilige Bildschirmgröße beschränkt, wobei auch dies durch Zoom und „unendliches“ Scrollen kaum eine realistische Begrenzung ist.

technischen Möglichkeiten (wie beispielsweise eine seitengenaue Verlinkung) aufweisen. Auch bei Entitäten, die in der Quelle keine Verknüpfung im eigentlichen Sinne aufweisen, kann eine solche aufgebaut werden. So können beispielsweise Bedeutungen oder sprachliche Formen, die an verschiedenen Stellen vorkommen, passend verknüpft werden, sodass eine zusätzliche Vernetzung der Artikel untereinander erzeugt wird (cf. Kapitel 4.2).

Der Einstieg in die Ressource kann ebenfalls variabler gestaltet werden als bei einem gedruckten Werk. In einem solchen gibt es zum Teil ein Inhalts- bzw. Lemmaverzeichnis, in anderen Fällen hilft nur die (beispielsweise alphabetische) Ordnung beim Auffinden der einzelnen Einträge. Zusätzlich sind in den meisten Fällen Wortverzeichnisse vorhanden, die aber nicht zwangsläufig vollständig sind. Das folgende Zitat illustriert dieses Problem, das wiederum primär mit den Platzbeschränkungen von gedruckten Werken zusammenhängt:

> Die Wortverzeichnisse der anderen Sprachen sind möglichst vollständig, das deutsch-romanische bietet naturgemäß nur eine Auswahl, ist gegen die erste Ausgabe in den Stichwörtern kaum erweitert worden, erschöpft auch nicht den im Texte enthaltenen Stoff, da eine noch weitere Ausdehnung des Raumes ausgeschlossen war […] (Meyer-Lübke 1935, 815)

Eine digitale Ressource kann diese Möglichkeiten ungemein erweitern, sei es durch Volltextsuche bzw. spezialisierte Suchmöglichkeiten (z.B. nach bestimmten Formen, Bedeutungen, Sprachen, Literaturangaben etc.) oder auch durch vollständige automatisiert aus den tatsächlichen Vorkommen generierte Verzeichnisse der entsprechenden Entitäten. So ist insbesondere bei passender Datenstrukturierung grundsätzlich sowohl ein semasiologischer als auch ein onomasiologischer Zugang möglich, unabhängig davon, wie dies im Quellenmaterial der Fall war.

2.3. Erweiterte Möglichkeiten

Eine originär digitale Option ist die direkte Einbindung von Nutzenden, wenn ihnen die Möglichkeit gegebenen wird, auf bestimmte Weise einen eigenen Beitrag zu liefern. In der Sprachwissenschaft werden zum Teil Formen von Crowdsourcing für Transkriptionsaufgaben oder die Erhebung neuer Daten eingesetzt. Dies kann über externe Portale wie Zooniverse (cf. *Zooniverse* 2009) geschehen, wie etwa beim ISTROX-Projekt (cf. ISTROX 2020), oder über eigene Tools, wie beispielsweise beim *Atlas der deutschen Alltagssprache* (cf. z.B. Möller/ Elspaß 2014) oder dem Projekt *VerbaAlpina* (cf. Krefeld/Lücke 2021). Die entsprechende Plattform dient in diesem Fall gleichzeitig der Erhebung und der Publikation. Obwohl im Falle der Digitalisierung eines bestehenden Werks keine neuen Daten erhoben werden, ist durchaus eine Einbindung von Nutzenden denkbar. Die direkte Möglichkeit von z.B. Fehlerkorrekturen (wie es beispielsweise in den Wikimedia-Projekten üblich ist) ist im wissenschaftlichen Bereich allerdings nicht verbreitet. Wenn diese berücksichtigt wird, dann maximal über Kontaktformulare oder Ähnliches. Ein Beispiel liefert das Wörterbuchnetz des Kompetenzzentrums – Trier Center for Digital Humanities der Universität Trier:

„Falls Sie einen Erfassungsfehler entdecken, dann schreiben Sie uns bitte unter Angabe der Wörterbuchsigle und der betreffenden Kontextstelle." (FAQ Wörterbuchnetz, <https://woerterbuchnetz.de/>)

Die Nachteile eines solchen Ansatzes sind ein gewisser zusätzlicher Arbeitsaufwand für Nutzende und vor allem, dass die zeitliche Dauer bis zum Erscheinen der Änderung von Außenstehenden nicht eingeschätzt werden kann,[3] was gerade im Fall von Zitationen in Publikationen mit festen Abgabefristen problematisch ist.[4]

Um diese Probleme zu lösen, ist eine niedrigschwellige und direkte Möglichkeit zur Eingabe von Änderungen nötig, die einem bestimmten Muster entsprechen.[5] Diese können bei Bedarf bzw. in bestimmten Spezialfällen durch einen Moderationsmechanismus ergänzt werden, wobei dieser zumindest den Vorteil der unmittelbaren Einpflegung der Änderung zunichtemacht. Für eine solche Fehlerkorrektur ist auch die in Kapitel 2.3. angesprochene Einbindung des originalen Quellenmaterials entscheidend, um Nutzenden einfach und intuitiv den Abgleich zwischen der gedruckten Passage und der digitalen Repräsentation zu ermöglichen. Des Weiteren sind auch andere Anwendungsfälle möglich, wie die Verknüpfung mit externen Ressourcen, in Fällen in denen dies nicht automatisiert möglich ist.

Um eine Einbindung von externen Personen ohne übermäßigen Aufwand möglich zu machen, kann es sinnvoll sein, bereits vorhandene interne Tools für z.B. Mitarbeitende in einem Projekt auch für Außenstehende (zumindest mit eingeschränkter Funktion bzw. Komplexität) verwendbar zu machen. Voraussetzung hierfür ist die konsequente Nutzung von Webtools und eine gewisse Intuitivität der Oberfläche sowie eine ausführliche Dokumentation.

Ein weiterer spezifischer Vorteil bei der digitalen Darstellung der Informationen aus der Quelle ist die zusammenführende Nutzung der Daten, nicht nur, um Zusammenhänge wie in Kapitel 2.2. beschrieben leichter aufzufinden, sondern auch um die Daten zu aggregieren und beispielsweise statistische Auswertungen des vollständigen Werks zu erstellen. Dies ermöglicht Visualisierungen, in denen die Quelle aus den verschiedensten Gesichtspunkten beleuchtet werden kann (cf. Kapitel 4.4), wobei sich die dafür notwendigen Daten direkt aus den Anforderungen aus Kapitel 2.2 ergeben.

3. Modellierung der Wörterbuchinhalte

Aus den im vorherigen Kapitel aufgestellten Anforderungen an eine Online-Ressource lassen sich wiederum bestimmte Bedingungen für die digitale Darstellung des Quellenmaterials ableiten. Vor allem die in den Kapitel 2.2 und 2.3

[3] Intern kann der Arbeitsablauf auch dadurch verzögert werden, dass mehrere Instanzen involviert sind, beispielsweise wenn die Änderungswünsche erst von wissenschaftlichem Personal verifiziert und dann von technischem Personal eingepflegt werden müssen.

[4] Ob umgekehrt die Möglichkeit des selbständigen Einspeisens von Änderungen zu Manipulationen führt, muss in der Praxis beobachtet und analysiert werden.

[5] Hiermit sind vor allem Änderungen an der Textbasis (Einfügungen, Löschungen und Ersetzungen) sowie bestimmte regelbasierte Eingriffe in den Prozessablauf (beispielsweise die Auflösung einer abkürzenden Schreibweise) gemeint (vgl. Kapitel 3.5).

genannten Funktionen erfordern eine weitreichende Extraktion der in den Wörterbuchartikeln enthaltenen Information. Dieses Kapitel analysiert diese Bedingungen und stellt exemplarisch dar, wie die linguistischen Daten in einer relationalen Datenbank (also in tabellarischer Form) dargestellt werden können.

Folgende grundsätzlichen Bedingungen können aufgestellt werden:

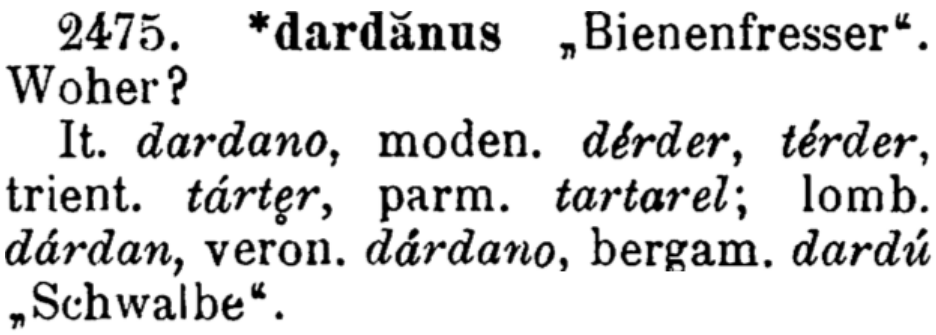
1. Für jeden Artikel (und für weitere textuelle Bestandteile der Quelle) müssen eine oder mehrere Repräsentationen erstellt werden können, die zumindest die quellentreue Transkription beinhalten.
2. Die einzelnen Grundbestandteile der Artikel, z.B. sprachliche Formen, Bedeutungen, bibliographische Angaben etc. (cf. hierzu auch Renders 2011, 118–121), müssen in geeigneter Form abgebildet werden.
3. Die Relationen zwischen diesen Bestandteilen (z.B. welche Form welche Bedeutung hat oder welche Form von welcher Form abstammt) müssen explizit dargestellt werden.
4. Änderungen bzw. Versionen müssen explizit abgespeichert werden können.
5. Datenstrukturen, die bestimmte Bestandteile mit externen Entsprechungen verknüpfen, müssen vorhanden sein.

Besonders die dritte Bedingung hat einige nicht-triviale Konsequenzen, weswegen sie im Weiteren gesondert betrachtet wird, bevor eine Möglichkeit für ein grundlegendes Datenmodell beschrieben wird (3.2 – 3.4). Dieses erfüllt die Bedingungen 1-3. Die letzten beiden Abschnitte gehen einzeln auf Bedingung 4 (3.5) und Bedingung 5 ein (3.6).

3.1. Explizite Darstellung von Wörterbuchrelationen

Die Information innerhalb von Wörterbüchern kann explizit oder implizit kodiert sein. Mit *impliziten Informationen* sind hierbei solche gemeint, die aus allgemeinen oder quellenspezifischen Konventionen hergeleitet werden können, aber nicht direkt im Text abgebildet sind. Ein einfaches Beispiel besteht in der Zuordnung von sprachlichen Formen zu Bedeutungen. Abb. 1 zeigt einen Teil eines Wörterbuchartikels des REW, der dies illustriert. Der Eintrag enthält (das Lemma eingeschlossen) neun verschiedene Formen, eine explizite Bedeutung wird aber nur bei zwei davon angegeben. Die Zuordnung zu den anderen erfolgt über die Konventionen der Quelle, die im Vorwort angegeben sind:

> […] die romanische Bedeutung wird nur dann gegeben, wenn sie von der des Stichwortes abweicht. […] Bei den Ableitungen und Zusammensetzungen gilt eine Bedeutung für sämtliche ihr vorangehenden Formen. (Meyer-Lübke 1935, XI)



2475. *dardănus „Bienenfresser“. Woher?

It. *dardano*, moden. *dérder*, *térder*, trient. *tártęr*, parm. *tartarel*; lomb. *dárdan*, veron. *dárdano*, bergam. *dardú* „Schwalbe“.

1 | Ausschnitt REW Lemma 2475

Für eine technische Verarbeitung der Daten muss diese Information also zuerst inferiert und dann in strukturierter Form abgelegt werden. Tab. 1 illustriert dies in tabellarischer Darstellung.[6]

Sprachabkürzung	Form	Bedeutung
lat.	dardănus	Bienenfresser
it.	dardano	Bienenfresser
moden.	dérder	Bienenfresser
moden.	térder	Bienenfresser
trient.	tártęr	Bienenfresser
parm.	tartarel	Bienenfresser
lomb.	dárdan	Schwalbe
veron.	dárdano	Schwalbe
bergam.	dardú	Schwalbe

Tab. 1 | Explizite Darstellung der Bedeutungszuordnungen, die im Text von Abb. 3 enthalten sind

3.2. Grundlegende Datenmodellierung

Die entscheidende Frage ist nun, in welchem Format diese Informationen abgelegt werden. Ein weit verbreiteter Ansatz in der digitalen Sprachwissenschaft und den Digital Humanities im Allgemeinen ist die Repräsentation als annotierter Text. Dabei kommt meistens XML zum Einsatz, entweder nach einem individuellen Schema (cf. z.B. Renders 2011) oder mit Hilfe des Standardformats TEI[7] (cf. z.B. Tasovac 2020). Das zentrale Element ist hierbei der (eventuell angepasste und erweiterte) Text eines Wörterbuchartikels. Bei den genannten impliziten Daten stößt dieser Ansatz allerdings an seine Grenzen, da diese Information nicht oder nur teilweise im Ursprungstext enthalten ist. Im besonderen Maße ist dies beispielsweise bei etymologischen Relationen zwischen verschiedenen sprachlichen Formen der Fall, die überhaupt nicht explizit als solche enthalten sind, sondern nur aus der Anordnung der Formen im Artikel erschlossen werden können. Hier soll ein alternativer, datenorientierter Ansatz vorgeschlagen werden, der explizite, tabellarische Daten vorsieht, wie sie beispielsweise in Tab. 1 illustriert

[6] Die Zuordnung der Bedeutungen ist bei weitem nicht die einzige Information, die inferiert werden muss, wichtig sind im vorliegenden Beispiel auch die Herkunftsrelation zwischen der lateinischen und den romanischen Formen. Die Zuordnung der lateinischen Sprache zum Lemma wird ebenfalls aus einer Konvention hergeleitet.

[7] Das Format wird von der gleichlautenden *T*ext *E*ncoding *I*nitiative entwickelt (cf. TEI 1994).

werden.[8] Gerade für artikelübergreifende, analytische Abfragen ist dies hilfreich, da sie ohne Durchsuchen von potentiell allen Artikeltexten durchgeführt werden können.

Den Kern stellt dabei eine grundlegende Einteilung in zwei Klassen von Daten dar: *Lexikalische Daten* und *einordnende Daten*. Dies liegt in der Natur eines Wörterbuchs begründet, dass als semi-strukturierter Text aufgefasst werden kann. Weite Teile folgen einer fixen Struktur, wie die Listen mit romanischen Formen im REW, die aber immer wieder von diskursiven Abschnitten unterbrochen werden. Dieses Modell versucht diesen Voraussetzungen gerecht zu werden. Mit *lexikalischen Daten* wird die abstrakte sprachwissenschaftliche Information bezeichnet, die aus dem Wörterbuchartikel extrahiert wird. Diese Information ist vor allem für die technische Verarbeitung von Suchanfragen und Ähnlichem nötig, bleibt aber den Nutzenden weitgehend verborgen. Der Begriff *einordnende Daten* bezeichnet hier Informationen, die für die Einbettung der strukturierten lexikalischen Daten in den Kontext der Wörterbuchartikel (also beispielsweise die Anordnung und Reihenfolge der Formen im Text) nötig sind, und diskursive Elemente, die nicht oder nur teilweise ausgewertet werden können. Sie sind rein auf den Artikel bezogen und dienen zusammen mit den lexikalischen Daten dazu, den vollständigen Artikelinhalt zu rekonstruieren. Ein großer Vorteil einer solchen hybriden Modellierung ist, dass die strukturierten Daten sehr effizient für technische Anwendungen verwendet werden können, da unabhängig vom Zugriffsweg (beispielsweise semasiologisch vs. onomasiologisch) ein direkter Zugriff auf die entsprechenden Daten möglich ist, während ein annotationsbasiertes Modell weiterhin in der durch die Lemmatisierung des Quellenmaterials vorgegebenen Perspektive verbleibt, sodass Anfragen, die nicht dieser Perspektive entsprechen, zumindest deutlich aufwendiger sind. Das gilt sowohl für interne Funktionen eines Webangebots als auch für die maschinelle Datenverarbeitung von externer Seite, wenn über eine technische Schnittstelle (cf. Kapitel 5) auf die Daten zugegriffen wird. Gleichzeitig bleibt die Darstellung des zugrundeliegenden Artikels uneingeschränkt möglich.

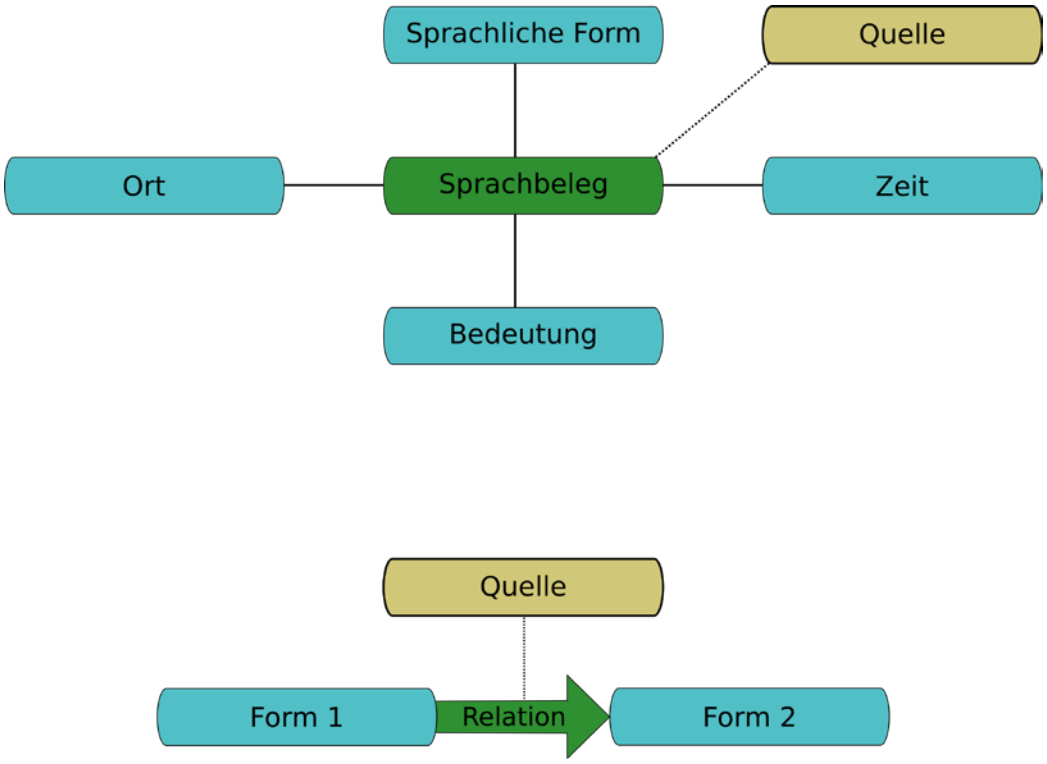
3.3. Lexikalische Daten

Die strukturierten lexikalischen Daten bestehen in diesem Modell aus sogenannten *Sprachbelegen* und Relationen zwischen sprachlichen Formen. Ein *Sprachbeleg* ordnet eine Form ein, indem er sie mit einer Bedeutung und (falls vorhanden) einem Ort sowie einem Zeitpunkt bzw. -intervall in Verbindung setzt. Relationen beschreiben verschiedene Formen von Verbindungen zwischen zwei sprachlichen Formen.[9] Im Fall eines etymologischen Wörterbuchs sind das vor allem Herkunftsrelationen wie Vorgänger oder Entlehnungen, es sind aber auch andere möglich, um beispielsweise verschiedene Flexionsformen des gleichen Lexems miteinander zu verknüpfen. Alle Datensätze werden als Teil eines globalen Informationspools verstanden. Die Beziehung zum Artikel wird als Quelle aufgefasst, aus der die

[8] Die Annotation von Text wird allerdings „im Kleinen“ verwendet, indem in Textbestandteilen, die nicht vollständig strukturiert erfasst werden können, bestimmte Entitäten annotiert werden (cf. Kapitel 3.4).

[9] Diese Art der Modellierung ist nicht direkt mit dem sogenannten *lemon*-Modell (cf. McCrae et al. 2012, 8–15) kompatibel, das für die Darstellung von lexikalischen Daten für das Semantic Web erstellt wurde und oft als Standard verwendet wird. Zacherl (in Vorb. b) führt die Gründe dafür auf.

jeweilige Information stammt. Abb. 2 veranschaulicht die beiden Informationstypen. Eine detailliertere Analyse der grundsätzlichen Darstellung von linguistischen Rohdaten und der Vorschlag einer konkreten Methodik zur Überführung von Wörterbuchtexten in relationale Daten, die nach einem solchen Datenmodell strukturiert sind, findet sich in Zacherl (in Vorb. b).



2 | Schematische Darstellung der beiden Grundelemente für das lexikalische Datenmodell

3.4. Einordnende Daten

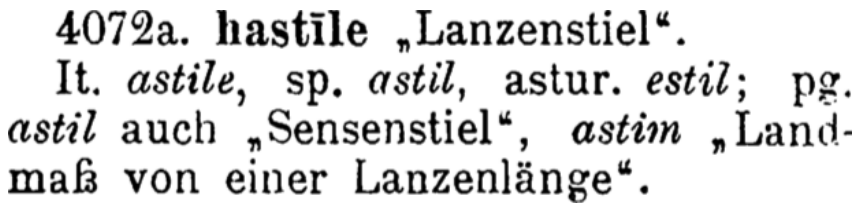
Die Grobstruktur eines Wörterbuchartikels wird über Beleglisten dargestellt. Jeder Artikel besteht aus mindestens zwei dieser Listen; eine repräsentiert die Kopfzeile mit der Angabe der Lemmata, alle weiteren beschreiben den eigentlichen Artikelinhalt. Die Listen werden grundsätzlich aufsteigend nummeriert, eine Ausnahme bilden nur Entlehnungen, die sich immer auf konkrete Sprachbelege beziehen. Diese stehen außerhalb der Nummerierung und werden den entsprechenden Belegen zugeordnet.

Die Sprachbelege selbst müssen für eine spätere Darstellung ebenfalls entsprechend ihrer Position in der Liste nummeriert werden. Grundsätzlich wäre eine einfache aufsteigende Nummerierung ausreichend, um die Belege entsprechend ihrer ursprünglichen Anordnung wiederzugeben. Um die logische Struktur besser abzubilden und auch bis zu einem gewissen Maß abweichende Darstellungen der Wörterbucheinträge zu ermöglichen (cf. Kapitel 4.1), bietet sich allerdings die Verwendung von mehreren hierarchisch funktionierenden Indizes an:

- Index 1 (Subliste): Listen von Belegen werden häufig durch Semikola oder andere Trennzeichen in unterschiedliche Gruppen aufgeteilt. Die dort

angegebenen Formen haben beispielsweise ähnliche Bedeutungen, die gleiche Wortart etc. Der erste Index nummeriert diese Einteilung.

- Index 2 (Position innerhalb der Subliste): Dieser Index wird über die Sprache[10] definiert. Wenn für eine Sprache mehrere Formen bzw. Bedeutungen gegeben werden, werden diese auf dieser Ebene noch zusammengefasst. So besteht das Beispiel aus Abb. 3 aus zwei verschiedenen Sublisten, wobei die erste aus drei Elementen und die zweite aus nur einem Element besteht. Kurzschreibweisen werden behandelt, als wären sie ausgeschrieben (d.h. Angaben wie pg., sp. *astil* werden wie pg. *astil*, sp. *astil* als zwei Elemente indiziert).
- Index 3 (Varianten innerhalb einer Sprache): Falls mehrere Formen bzw. Varianten einer Form angegeben werden, werden sie hier unterschieden (beispielsweise die portugiesischen Formen *astil* und *astim* in Abb. 3). Dabei wird wiederum nicht zwischen abkürzenden Schreibweisen und expliziter Trennung durch Kommata im Originaltext unterschieden (also pg. *(f)ata* und pg. *ata*, *fata* würden beispielsweise identisch nummeriert).
- Index 4 (Bedeutung): Falls eine Form mehrere Bedeutungen hat, werden diese durch den letzten Index nummeriert. Dabei spielt es keine Rolle, ob die Bedeutungen in der Quelle explizit angegeben sind oder inferiert werden müssen.


4072a. **hastīle** „Lanzenstiel".
It. *astile*, sp. *astil*, astur. *estil*; pg. *astil* auch „Sensenstiel", *astim* „Landmaß von einer Lanzenlänge".

3 | REW Lemma 4072a

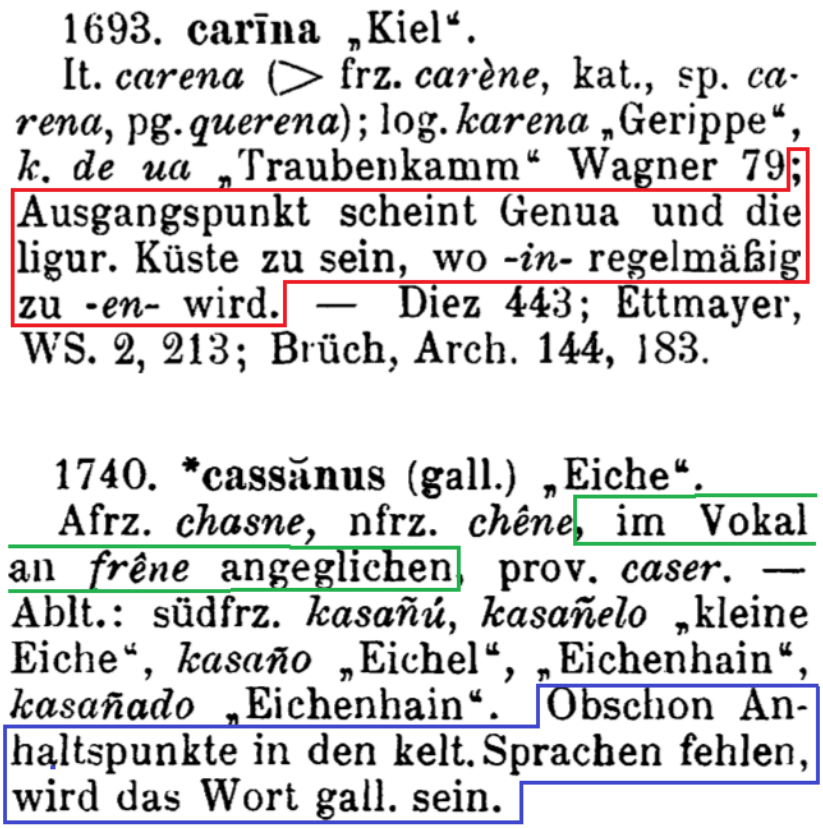
Tab. 2 illustriert am bereits erwähnten Beispiel die verschiedenen Indizes. Zusätzlich zur Indizierung werden den Sprachbelegen außerdem alle Literaturangaben zugeordnet, die sich auf diese beziehen.

10 Hier und im Folgenden wird der Begriff *Sprache* stellvertretend für eine Sprache oder einen Dialekt verwendet. Das dient nur der Vereinfachung, da beides im Normalfall durch entsprechende Abkürzungen angegeben und syntaktisch nicht unterschieden wird.

Sprachabkürzung	*Form*	*Bedeutung*	*Subliste*	*Position*	*Variante*	*Bedeutung*
it.	astile	Lanzenstiehl	0	0	0	0
sp.	astil	Lanzenstiehl	0	1	0	0
astur.	estil	Lanzenstiehl	0	2	0	0
pg.	astil	Lanzenstiehl	1	0	0	0
pg.	astil	Sensenstiehl	1	0	0	1
pg.	astim	Landmaß von einer Lanzenlänge	1	0	1	0

Tab. 2 | Indizierung der sprachlichen Formen aus dem Eintrag aus Abb. 3

Diskursive Elemente sind Texteinschübe, die keiner festen Struktur unterliegen. Die häufigsten Vertreter sind Einschübe, die sich auf einen einzelnen Beleg beziehen, Einschübe, die sich auf eine komplette Belegliste beziehen und vollständige Sätze, die strukturell in keinem Bezug zu einer Belegliste stehen (cf. Abb. 4). Die ersten beiden Typen werden als Kommentar zum jeweiligen Beleg bzw. zur jeweiligen Liste aufgefasst und diesen zugeordnet.[11] Der letzte Typ wird als „entartete" Belegliste modelliert, d.h. als eine Belegliste, die nur aus einem Kommentar besteht, aber keine Belege enthält. Somit kann jeder Artikel als eine geordnete Menge von Beleglisten aufgefasst werden.[12]



1693. **carīna** „Kiel".
It. *carena* (> frz. *carène*, kat., sp. *carena*, pg. *querena*); log. *karena* „Gerippe", *k. de ua* „Traubenkamm" Wagner 79; Ausgangspunkt scheint Genua und die ligur. Küste zu sein, wo *-in-* regelmäßig zu *-en-* wird. — Diez 443; Ettmayer, WS. 2, 213; Brüch, Arch. 144, 183.

1740. ***cassănus** (gall.) „Eiche".
Afrz. *chasne*, nfrz. *chêne*, im Vokal an *frêne* angeglichen, prov. *caser*. — Ablt.: südfrz. *kasañú*, *kasañelo* „kleine Eiche", *kasaño* „Eichel", „Eichenhain", *kasañado* „Eichenhain". Obschon Anhaltspunkte in den kelt. Sprachen fehlen, wird das Wort gall. sein.

4 | Verschiedene Arten von Texteinschüben im REW (rot: auf alle vorangegangenen Formen bezogen, grün: auf den direkten Vorgänger bezogen, blau: eigenständiger Satz)

[11] Alle anderen selteneren Formen von Texteinschüben können ebenfalls bestimmten Entitäten zugeordnet werden und werden somit analog behandelt.

[12] Ein Eintrag besteht zusätzlich zu den Beleglisten auch aus dem unbearbeiteten Originaltext, so dass dieser alternativ zu einer angereicherten Version ebenfalls dargestellt werden kann. Dies ist prinzipiell redundant, weil der Text auch rekonstruiert werden könnte, vereinfacht die Prozesse aber deutlich.

Diskursive Elemente werden nicht systematisch ausgewertet,[13] die darin enthaltenen relevanten Entitäten, die über die Struktur erkennbar sind, werden aber erkannt und mit XML annotiert, sodass sie an der Oberfläche entsprechend dargestellt werden können.

3.5. Korrekturen und Ausnahmen

Alle Änderungen werden explizit als Datensätze abgelegt. Es können zwei grundlegende Typen unterschieden werden: Fehler an der Textbasis und *Ausnahmen*, die in den Ablauf des Transformationsprozesses eingreifen, der aus dem ursprünglichen Text die entsprechenden Daten erzeugt. Beim ersten Typus können zusätzlich Korrekturen von Digitalisierungsfehlern, die beispielsweise bei einer automatischen Texterkennung oder auch bei manueller Transkription entstanden sind, unterschieden werden von Fehlern in der Quelle an sich. Letztere sollten nur in sehr offensichtlichen Fällen (fehlende schließende Klammern, offensichtliche Tipp- oder Druckfehler, etc.) korrigiert werden. Trotzdem kann eine solche Korrektur nicht immer völlig ohne Interpretation stattfinden, wenn beispielsweise unklar ist, an welcher Stelle eine schließende Klammer vergessen wurde. An der Oberfläche der Web-Anwendung können solche editorischen Eingriffe dann beispielsweise gesondert markiert werden. Unabhängig vom Typ werden sämtliche Änderungen mit einem Zeitstempel und dem entsprechenden User-Login markiert, so dass die Änderungshistorie jederzeit nachvollziehbar bleibt.[14]

Je nach Präferenz können die Korrekturen nach Erstellung sofort umgesetzt werden, wodurch eine neue Version des/der betroffenen Artikel erstellt wird oder es kann für alle oder bestimmte Änderungen (je nach Login und Art der Änderung) ein Moderationsprozess vorgeschaltet werden. Die Versionen der Artikel werden dabei explizit in der Datenbank abgelegt, d.h. nach jeder Änderung wird dort eine Kopie der Artikelrepräsentation erstellt, die diese Änderung enthält. Grundsätzlich wäre dies nicht nötig, da aufgrund der zeitlichen Markierung aus der Textbasis und den jeweiligen Korrekturen die zum damaligen Zeitpunkt gültige Version rekonstruiert werden könnte. Dahinter steckt aber die wenig realistische Annahme, dass sich der Programmcode für den Transformationsprozess zu keinem Zeitpunkt ändert. Ein weiterer Vorteil der expliziten Darstellung der Artikelversionen ist, dass deren Erstellung weniger rechenintensiv ist.

3.6. Datenanreicherung

Bei der Betrachtung der Möglichkeiten für eine externe Vernetzung ist die offensichtlichste die Verknüpfung der Literaturangaben, die im Text genannt werden, mit potentiell vorhandenen digitalen Versionen der jeweiligen Quelle, die ohne Zugangsbeschränkung verfügbar sind. Für Quellen, die über Seitenzahlen referenziert werden, ist dabei eine einfache Abbildung von Quellenabkürzung und eventuell vorhandener Bandnummer auf entsprechende Links nötig. Etwas

[13] Dies wäre nur mit sehr weit fortgeschrittenen Methoden des *Natural Language Processings* möglich.

[14] Die exakte Darstellung beider Arten von Änderungen hängt stark von der algorithmischen Umsetzung des Transformationsprozesses ab und wird deshalb an dieser Stelle nicht genauer spezifiziert. Entsprechende Beispiele finden sich in Zacherl (in Vorb. b).

komplizierter ist der Fall bei Wörterbüchern, die über Lemmanummern referenziert werden. In einem solchen Fall ist zusätzlich eine Abbildung von Seitenzahlen auf die jeweils dort gelisteten Lemmata nötig.

In Bezug auf die Bedeutungen bietet sich eine Verknüpfung mit einem kontrollierten Vokabular oder einer Ontologie an. Das hat den Vorteil, dass einerseits intern doppelt vorhandene Konzepte, die durch unterschiedliche sprachliche Ausdrücke beschrieben werden, verknüpft werden können, andererseits hat es einen Normierungseffekt, der gerade beim Zusammenführen von Daten aus verschiedenen (potentiell in unterschiedlichen Sprachen verfassten) Werken wichtig ist. Ein Beispiel für eine solches externes Portal ist die Wissensdatenbank Wikidata (cf. Wikidata 2012). Durch die Zuordnung zum entsprechenden Konzept in dieser Datenbank können z.B. die im REW vorkommenden synonymen Bedeutungsangaben „Abfall“, „Unrat“ und „Müll“ zusammengefasst werden, was einen Abgleich der entsprechend zugeordneten sprachlichen Formen deutlich erleichtert.

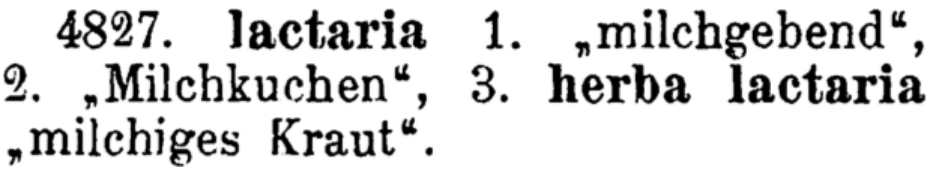
Für eine geographische Visualisierung ist außerdem eine Zuordnung von Sprachen bzw. Dialekten zu Verbreitungsgebieten notwendig. Ob eine solche Visualisierung für eine gegebene Quelle in Frage kommt, ist sehr spezifisch und ergibt tendenziell nur bei Werken Sinn, die eine große Anzahl kleinräumiger, dialektaler Formen enthalten.

4. Gestaltung der Oberfläche

Im Folgenden wird exemplarisch eine mögliche Umsetzung der Konzepte aus Kapitel 1 beschrieben und der daraus entstandene Entwurf eines Webauftritts vorgestellt.

4.1. Darstellung der Artikel

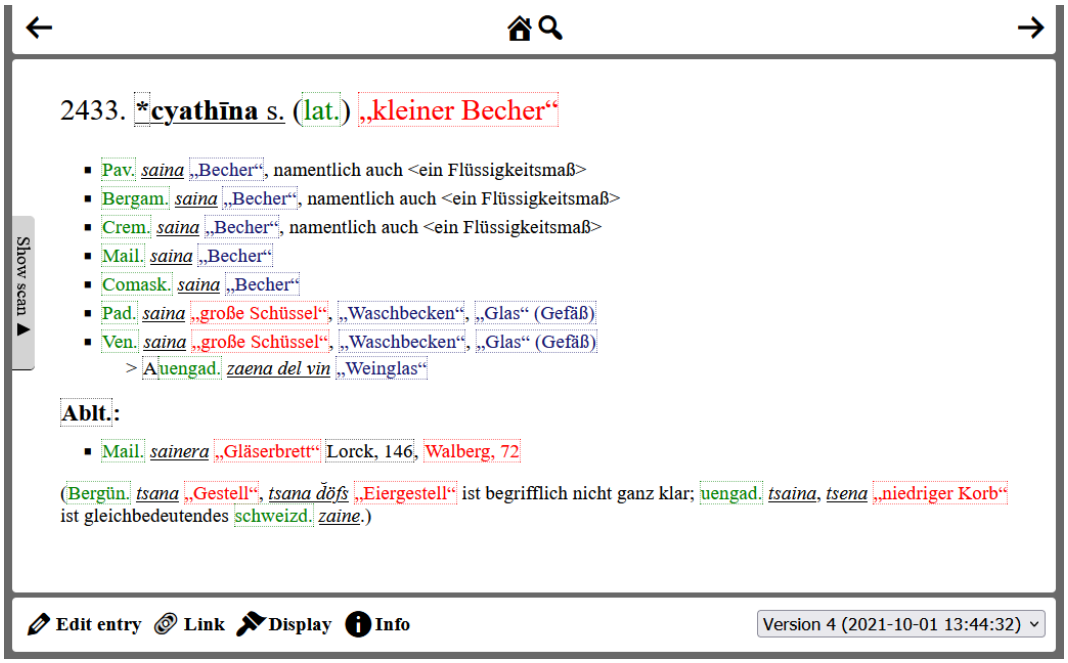
Der Kern eines jeden Online-Wörterbuchs ist die Darstellung der einzelnen Artikel. Hierbei stellt sich die grundlegende Frage, ob die Artikel einzeln (cf. z.B. TLIO) oder im Kontext der vorangegangenen und nachfolgenden Artikel dargestellt werden (cf. z.B. Wörterbuchnetz). Eine Entscheidung hängt u.a. davon ab, wie häufig einzelne Einträge Bezug auf ihre direkte Umgebung nehmen bzw. ob ähnliche Lemmata (Varianten, Komposita, etc.) in einem Artikel zusammengefasst werden oder nicht. Das REW referenziert beispielsweise grundsätzlich über Lemmanummern und fasst verhältnismäßig viele Lemmata in einem Artikel zusammen (cf. Abb. 5). Eine einzelne Darstellung der Artikel, die oftmals übersichtlicher ist, bietet sich also an.


4827. **lactaria** 1. „milchgebend“, 2. „Milchkuchen“, 3. **herba lactaria** „milchiges Kraut“.

5 | Kopfzeile REW Lemma 4827

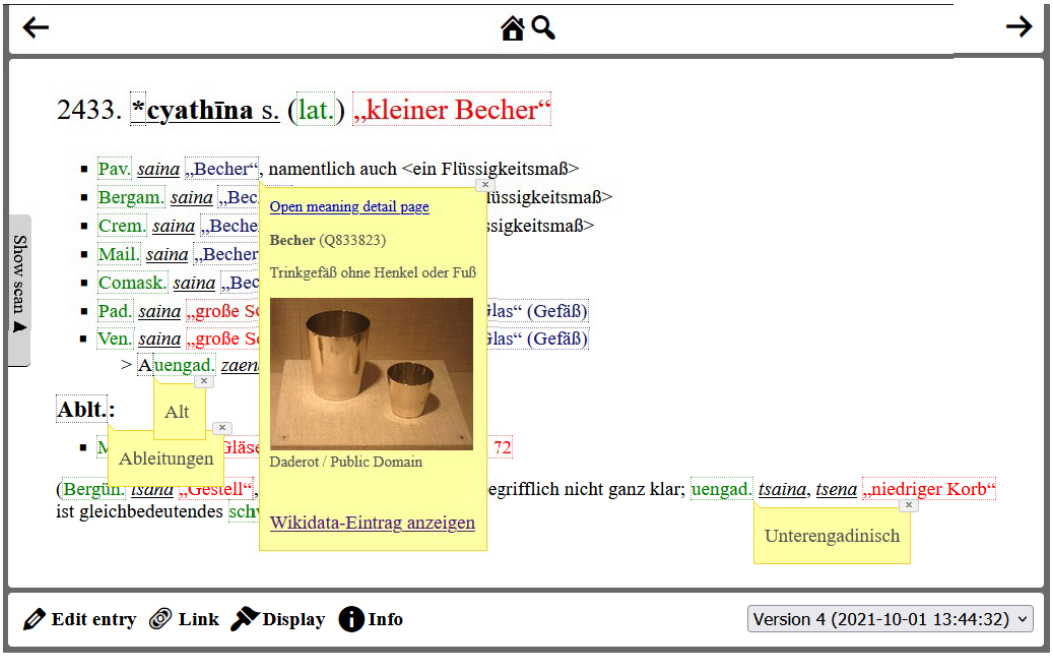
Ein großer Unterschied der standardmäßigen Darstellung (cf. Abb. 6) eines Artikels ist, dass die Beleglisten als Aufzählungen (also vertikal) angezeigt werden. Entlehnungen, die im Originaltext in Klammern hinter dem jeweiligen Beleg aufgeführt

werden, stellen eingerückte Unterlisten dar. Außerdem werden verschiedene Bestandteile eines Artikels durch verschiedene Absätze stärker visuell voneinander abgegrenzt. Einleitende Angaben, die eine Belegliste genauer spezifizieren (Ableitungen, Zusammensetzungen, etc.) werden prominenter dargestellt. All dies ist Konsequenz des bereits erwähnten Wegfalls der Platzbeschränkungen eines gedruckten Werks und hat die Absicht die Übersichtlichkeit und Lesbarkeit zu erhöhen. Diskursive Elemente behalten ihre grundsätzliche Darstellung.



6 | Standardansicht eines Artikels

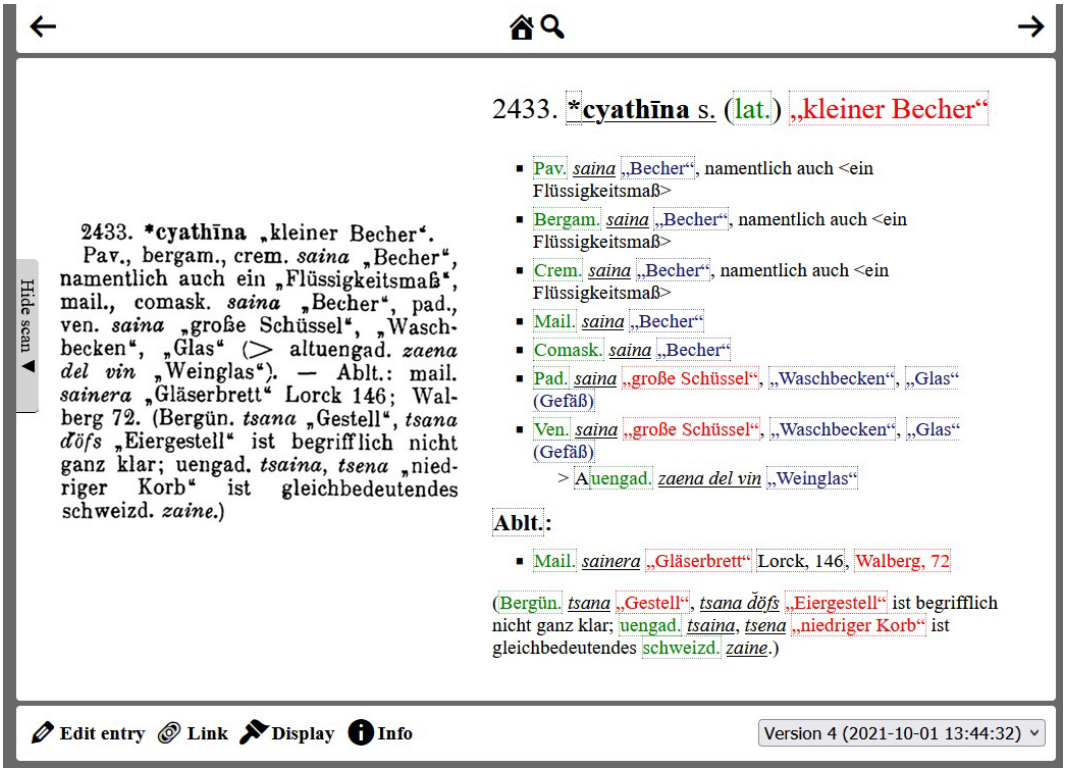
Die einzelnen Bestandteile der Artikel werden unterschiedlich hervorgehoben. Die grundsätzlichen Konventionen aus der Quelle (Lemmata sind fett markiert, andere sprachliche Formen kursiv bzw. in Kapitälchen und Bedeutungen in Anführungszeichen) werden so übernommen. Alle interaktiven Elemente, die beim Überfahren mit der Maus und Klick (bzw. Antippen auf mobilen Endgeräten) zusätzliche Informationen anzeigen, werden durch einen gestrichelten Rahmen markiert. Vor allem trifft das auf alle Formen von Abkürzungen zu, die so aufgelöst werden können. Bedeutungen und bibliographische Referenzen werden außerdem unterschiedlich formatiert, je nachdem ob sie mit externen Daten verknüpft sind oder nicht. In beiden Fällen wird (falls vorhanden) eine Verlinkung auf die externe Entsprechung angegeben, im Fall der Bedeutung werden zusätzlich noch Basisinformationen und eventuell ein Bild zur Illustration des Wikidata-Konzepts nachgeladen. Abb. 7 zeigt einige Beispiele.



7| Auflösung von Abkürzungen und Anzeige zugeordneter Konzepte aus Wikidata

Für beliebige Artikel ist jederzeit eine Anzeige des Ausschnitts aus dem Scan des Wörterbuchs über einen Button auf der linken Seite möglich.[15] Dieser dient einerseits zum Nachweis der Quelle, bietet aber auch eine unaufwendige Möglichkeit aufgefundene (potentielle) Fehler, die im Verlauf der Verarbeitung entstanden sind, im Abgleich mit dem Quellenmaterial als solche zu verifizieren und im Anschluss entsprechend zu korrigieren (vgl. hierzu Kapitel 4.3). Abb. 8 zeigt die gemeinsame Ansicht von Scan und Artikel.

[15] Zusätzlich zu den in Kapitel 3 behandelten Daten muss dafür das Bildmaterial zur Verfügung stehen, sowie eine Zuordnung von Pixelkoordinaten zu den jeweiligen Artikeltexten. Texterkennungssysteme können zum Teil beispielsweise das HOCR-Format erzeugen, in dem eine Zuordnung von Zeilen zu Pixelkoordinaten vorhanden ist, die für diese Zwecke genutzt werden kann.



8 | Artikeldarstellung mit Ursprungstext

Die weiteren Bedienelemente sind in zwei Gruppen eingeteilt. In der oberen Hälfte finden sich Navigations- und Suchfunktionen. Es kann zum vorherigen und nächsten Artikel gewechselt werden, sowie auf die Startseite, die verschiedene Einstiegsmöglichkeiten sowie ein vollständiges Inhaltsverzeichnis umfasst, das auch das Vorwort sowie die Abkürzungsverzeichnisse enthält.[16] Das Lupensymbol erlaubt den Zugriff auf verschiedene Suchfunktionalitäten wie eine uneingeschränkte Volltextsuche und spezialisierte Suchen nach bestimmten Entitäten.

Im unteren Bereich finden sich Interaktionsmöglichkeiten und Dokumentation. Die ersten beiden Bedienelemente dort erlauben das Bearbeiten des Artikels (cf. Kapitel 4.3) und die Verlinkung bzw. Zitation. Bei letzterem können zwei verschiedene URLs[17] erzeugt werden, wobei eine auf die aktuelle Artikelversion verweist, also auf eine statische Darstellung, die für wissenschaftliches Zitieren genutzt werden kann, und die andere auf die jeweils neuste Version des jeweiligen Artikels. Weiterhin ist die Auswahl von unterschiedlichen Darstellungsvarianten möglich, die in verschiedenen Stufen die Ähnlichkeit zum originalen Text steigern, bis hin zur ursprünglichen Spaltendarstellung, wie sie auch der Scan zeigt. Der Info-Button enthält eine detaillierte Dokumentation der verschiedenen Notationskonventionen, was sowohl diejenigen, die aus dem ursprünglichen Wörterbuch übernommen wurden, als auch die spezifischen der digitalen Variante umfasst. Rechts unten kann schließlich zwischen den verschiedenen Artikelversionen

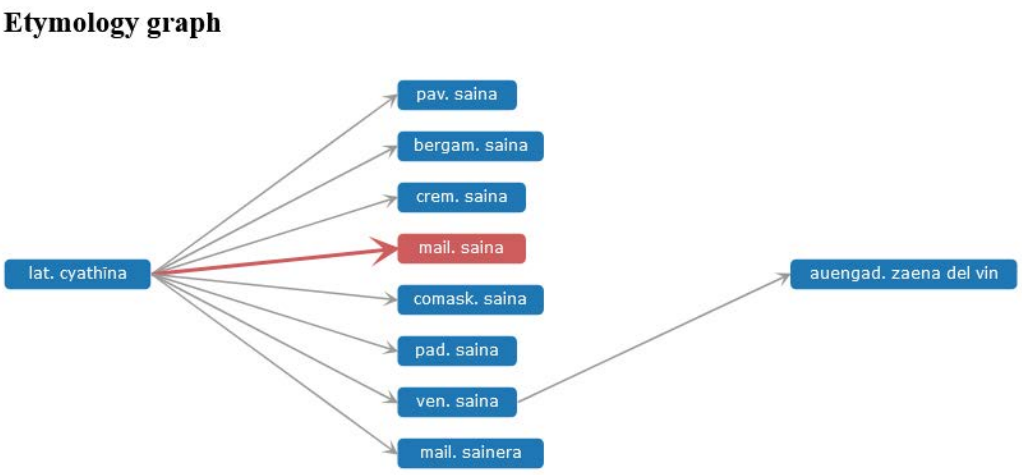
[16] Letztere sind grundsätzlich nicht nötig, da alle Abkürzungen an den Stellen aufgelöst werden, an denen sie vorkommen. Zur Übersicht und um der Struktur des Originalwerks möglichst zu entsprechen, werden sie trotzdem aufgeführt.

[17] Die URLs sollten möglichst auf persistenten Identifikatoren wie DOIs basieren.

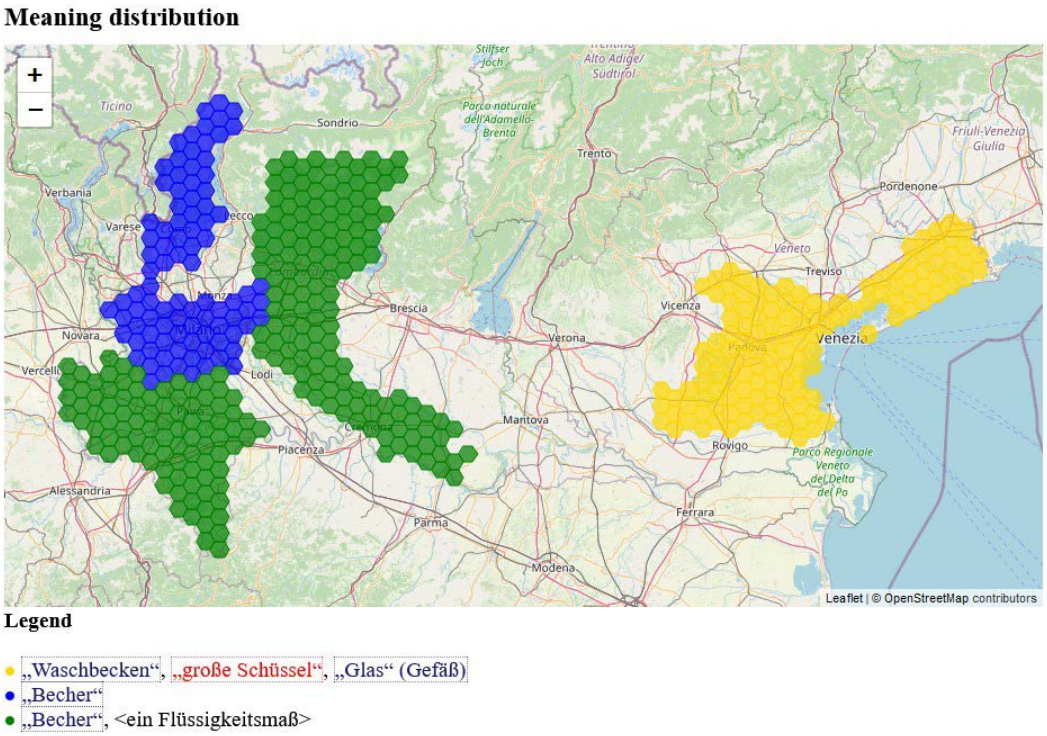
gewechselt werden (falls mehrere vorhanden sind). Wenn eine veraltete Version ausgewählt wird, erscheint im oberen Teil eine entsprechende Warnung, um die versehentliche Nutzung von Informationen, die bereits korrigierte Fehler enthalten, möglichst zu vermeiden.

4.2. Detailseiten für Wörterbuchbestandteile

Für die drei wichtigsten Entitäten im Kontext des Wörterbuches, nämlich sprachliche Formen, Bedeutungen und referenzierte Literatur, werden weiterhin sogenannte *Detailseiten* angeboten, die über die Vorkommen der jeweiligen Entitäten innerhalb der Artikel verlinkt werden und artikelübergreifend Informationen zu diesen aggregieren. So enthält beispielsweise die Detailseite einer Form alle Vorkommen derselben in den verschiedenen Artikeln mit entsprechenden Bedeutungen. Umgekehrt enthält die Detailseite einer Bedeutung eine Liste aller Vorkommen und der entsprechenden sprachlichen Formen. Weiterhin sind verschiedene Visualisierungen möglich. Abb. 9 zeigt eine Visualisierung, in der eine einzelne Form mit Hilfe einer Graphdarstellung in den Kontext der etymologischen Relationen eingebettet wird, während Abb. 10 ein Beispiel für eine geographische Visualisierung zeigt, die die Verteilung der Bedeutungen einer bestimmten Form darstellt.



9 | Visualisierung von Herkunftsrelationen

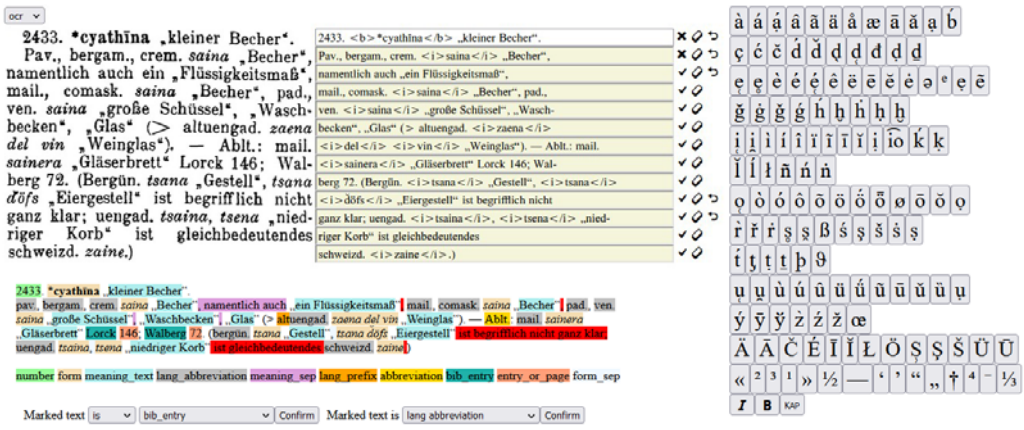


10 | Visualisierung der geographischen Verteilung von Bedeutungen einer Form

Die Detailseiten haben grundsätzlich eigene URLs, die die ID der Entität enthalten, somit können diese auch einzeln verlinkt werden.

3.3. Interaktionsmöglichkeiten

Für Nutzende der Online-Ressource sind zwei Möglichkeiten vorgesehen, wie sie selbst dazu beitragen können, die Datenbasis zu verbessern bzw. zu erweitern. Die wichtigste Möglichkeit ist wohl die zur selbstständigen Korrektur von Fehlern im digitalisierten Quellentext. Abb. 11 zeigt einen Entwurf für eine entsprechende Eingabemaske. Den Kern bildet der Ausschnitt des Scans zum jeweiligen Artikel mit zugehörigen Texteingabefeldern, die den Zeilen im Scan entsprechen und passend zu diesem angeordnet werden. Absätze im Text werden über zwei verschiedene Hintergrundfarben markiert, die sich abwechseln. Im rechten Teil findet sich eine Art virtueller Tastatur, die häufig benötigte Zeichen enthält, die mit vielen Tastatur-layouts nicht direkt eingegeben werden können. Unten findet sich schließlich eine Darstellung des Verarbeitungsergebnisses, in dem die unterschiedlichen erkannten Bestandteile farblich markiert werden. Bei jeder Änderung in einem der Textfelder wird ein Korrektur-Datensatz angelegt und die Darstellung aktualisiert, so dass vor allem die Korrektur von strukturellen Fehlern (z.B. Satzzeichen, Klammern, etc.) besonders deutlich im Ergebnis sichtbar wird.



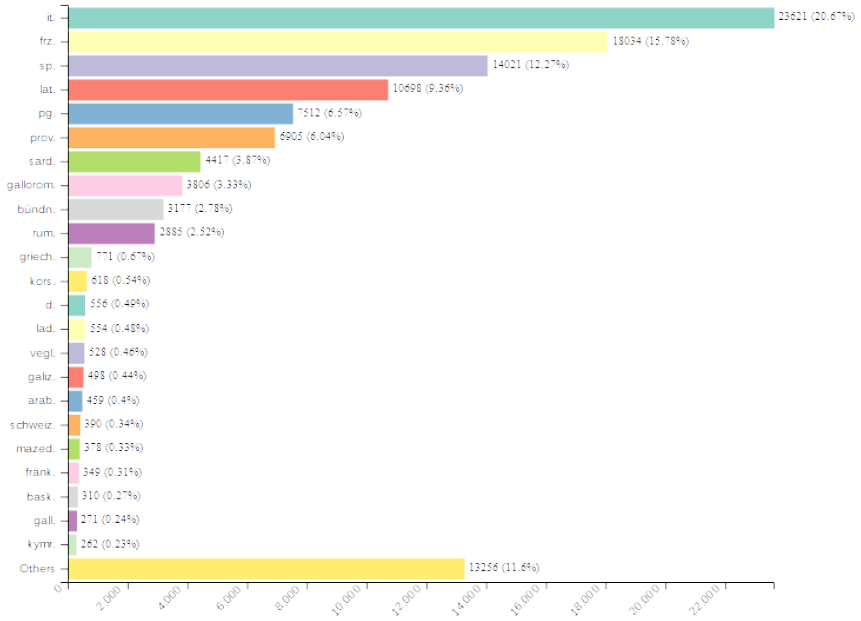
11 | Entwurf einer Oberfläche zur Korrektur

Eine weitere Interaktionsmöglichkeit bietet das Gebiet der Anreicherung und Vernetzung. Die in Kapitel 3.6 genannten Daten können nur zum Teil automatisiert erstellt werden. In den übrigen Fällen, sei es bei der Zuordnung von Bedeutungen zu Wikidata-Einträgen, von Sprachabkürzungen zu Verbreitungsgebieten oder von literarischen Quellen zu entsprechenden Online-Repräsentationen, wird bei einem fehlenden Datum stattdessen ein Oberflächenelement angezeigt, das Nutzenden das Nachtragen ermöglicht. So kann zum Beispiel beim Fehlen einer Wikidata-Verknüpfung eine Abfrage an die Wikidata-Server geschickt werden, die aus der Bedeutungsbeschreibung nach Möglichkeit verschiedene Kandidaten liefert, aus denen der richtige ausgewählt werden kann. Falls keiner der Kandidaten sinnvoll ist, kann auch manuell die ID eines Eintrags angegeben werden, wobei grundsätzlich auch immer die Möglichkeit besteht, bisher nicht vorhandene Konzepte in Wikidata neu zu erstellen.

4.4. Statistische Auswertung

Abb. 12 zeigt exemplarisch eine der statistischen Visualisierungen, die aus den aus dem REW extrahierten Daten erzeugt werden können. In diesem Fall wird der prozentuale Anteil der verschiedenen Sprachen dargestellt. Dafür werden alle Formen verwendet, die strukturiert erfasst wurden, also nicht ausschließlich innerhalb von diskursiven Elementen vorkommen (cf. Kapitel 3.4). Dialekte werden unter den übergeordneten Sprachen zusammengefasst.[18] Da die Darstellung direkt aus dem aktuellen Datenbestand generiert wird, sind die zugrundeliegenden Zahlen nicht statisch, sondern können grundsätzlich bei jeder Änderung oder Korrektur leichten Schwankungen unterliegen.

[18] Diese Einschränkungen dienen der Übersichtlichkeit und sind nicht obligatorisch.



12 | Visualisierung des Anteils der Sprachen bezogen auf alle Formen, die aus dem REW strukturiert erfasst wurden (Stand 31.01.2022)

Weitere statistische Auswertungen, beispielsweise die Häufigkeitsverteilung der literarischen Quellen oder die Aufgliederung der Lemmata nach Sprachen ist in ähnlicher Weise möglich.

5. Ausblick

Dieser Beitrag konzentriert sich auf Möglichkeiten und Methodiken, um die digitalisierten Informationen aus einem Wörterbuch für die direkte menschliche Nutzung zu verwenden. Gerade das sehr kleinteilige Datenmodell, das in Kapitel 3 beschrieben wurde, eignet sich allerdings auch gut für eine maschinelle Nutzung der Daten. So ist sowohl der Zugriff über eine technische Schnittstelle der Online-Ressource als auch die Umwandlung der sprachlichen Kerndaten, d.h. der lexikalischen Daten wie sie in Kapitel 3.3. definiert wurden, in Formate des Semantic Webs wie RDF (cf. RDF 2014) ohne weiteres möglich. Dies eröffnet weitreichende Möglichkeiten wie individuelle Abfragen auf dem Datenbestand oder auch zusätzliche Visualisierungen auf Basis externer Tools.

Auch ein vollständiger Export kann vorgesehen werden. Das bezieht sich nicht nur auf die Kerndaten (wie beispielsweise eine Liste der Lemmata oder die vollständigen Artikeltexte), sondern auch auf sekundäre Daten, wie sie in Kapitel 3.6 für die Anreicherung des Ursprungsmaterials beschrieben werden. Beides kann somit in anderen Projekten wiederverwendet werden.[19] Sinnvollerweise sollten die Exportdaten (in gewissen Zeitschnitten) ebenfalls in passenden Repositorien archiviert werden.

[19] Voraussetzung hierfür ist die Veröffentlichung unter Verwendung einer entsprechend offenen Lizenzierung. Im vorliegenden Beispiel werden alle Daten unter Creative Commons BY-SA zur Verfügung gestellt.

Bibliografie

ISTROX. 2020. „Launch of ISTROX on Zooniverse: Press Release.“ <https://istrox.ling-phil.ox.ac.uk/news/2020/08/04/launch-istrox-zooniverse-press-release> 30.01.2022.

KOMPETENZZENTRUM – „Trier Center for Digital Humanities: Wörterbuchnetz“. <https://woerterbuchnetz.de>. 05.02.2022.

KOMPETENZZENTRUM – „Trier Center for Digital Humanities: „FAQ Wörterbuchnetz.“ <https://woerterbuchnetz.de>. 31.01.2022.

KREFELD, Thomas & Stephan Lücke. 2021. s.v. „Crowdsourcing” *VerbaAlpina-de* 21/2 (Erstellt: 16/1, letzte Änderung: 21/1), Methodologie. <https://doi.org/10.5282/verba-alpina?urlappend=%3Fpage_id%3D493%26db%3D212%26letter%3DC%2312>.

MCCRAE, John et al. 2012. „Interchanging lexical resources on the Semantic Web.“ *Lang Resources & Evaluation* 46, 701–719. <https://doi.org/10.1007/s10579-012-9182-3>.

MEYER-LÜBKE, Wilhelm. 1935. *Romanisches etymologisches Wörterbuch* 3., vollst. neubearb. Aufl. Heidelberg: Winter. <https://nbn-resolving.org/urn:nbn:de:bvb:355-ubr07799-0>.

MÖLLER, Robert & Stephan Elspaß. 2014. „Zur Erhebung und kartographischen Darstellung von Daten zur deutschen Alltagssprache online: Möglichkeiten und Grenzen.“ In *20 Jahre digitale Sprachgeographie*, ed. Tosques, Fabio, 121–131, Berlin: Humboldt-Universität, Institut für Romanistik. <https://www2.hu-berlin.de/vivaldi/tagung/beitraege/pdf/20_jahre_web_version.pdf>.

PRÄTOR, Klaus. 2011. „Zur Zukunft des Zitierens. Identität, Referenz und Granularität digitaler Dokumente.“ *Editio* 25 Heft 2011, 170-183. <https://doi.org/10.1515/9783110239362.170>.

RDF. 2014. „Resource Description Framework.“ <https://www.w3.org/RDF/>. 19.02.2022.

RENDERS, Pascale. 2011. *Modélisation d'un discours étymologique. Prolégomènes à l'informatisation du Französisches Etymologisches Wörterbuch*. <https://orbi.uliege.be/handle/2268/94407>.

REWONLINE = Zacherl, Florian. 2021-. *Digitale Aufbereitung des Romanischen etymologischen Wörterbuches von Wilhelm Meyer-Lübke*. <https://www.rew-online.gwi.uni-muenchen.de>.

TASOVAC, Toma. 2020. *The Historical Dictionary as an Exploratory Tool: A Digital Edition of Vuk Stefanovic Karadzic's Lexicon Serbico-Germanico-Latinum*. <http://hdl.handle.net/2262/92750>.

TEI. 1994. „Text Encoding Initiative.“ <https://tei-c.org/>. 19.02.2022.

TLIO = *Tesoro della Lingua Italiana delle Origini*. 1997. <http://tlio.ovi.cnr.it/TLIO/>. 05.02.2022.

VÀRVARO, Alberto. 2011. „Il DÉRom: un nuovo REW?“ *Revue de linguistique romane* 75, 297-304.

WIKIDATA. 2012. <https://www.wikidata.org>. 19.02.2022.

ZACHERL, Florian. In Vorb. a. *Digitale Tiefenerschließung traditioneller Lexikographie – am Beispiel des Romanischen Etymologischen Wörterbuchs*, univ. Diss. Ludwig-Maximilians-Universität München.

ZACHERL, Florian. In Vorb. b. „Automatisierte Erschließung von strukturierten Daten aus Wörterbuchtexten.“ In *Digitale romanistische Sprachwissenschaft: Stand und Perspektiven*, ed. Becker, Lidia et al., Tübingen: Narr

Francke Attempto.
ZOONIVERSE. 2009. <https://www.zooniverse.org/>. 19.02.2022.

Zusammenfassung

Das Zusammenführen von Informationen aus verschiedenen Quellen im Rahmen der linguistischen Forschung kann einen nicht zu unterschätzenden Aufwand darstellen. Webportale, die diese in digitalisierter Form enthalten, bieten eine mögliche Lösung für dieses Problem, müssen aber dazu bestimmte Anforderungen erfüllen. Dieser Beitrag analysiert diese Anforderungen und untersucht weitere originär digitale Möglichkeiten, die sich in diesem Kontext ergeben. Darauf aufbauend ermittelt er, was dies für Struktur und Format der zugrundeliegenden Daten bedeutet und zeigt am Beispiel eine konkrete Umsetzung der aufgestellten Prinzipien.

Abstract

Combining information from different sources in the context of linguistic research can lead to an effort that should not be underestimated. Web portals that contain a digital representation of that information offer a possible solution to this problem, but need to fulfil certain criteria to achieve this goal. The article analyses these criteria and examines further possibilities, exclusive to the digital form, that arise in this context. On this basis, the contribution identifies requirements for suitably structured and formatted data and presents an example that illustrates a concrete implementation of the principles set out.