Published in: Helling, Patrick/Speer, Andreas/Eide, Øyvind (Ed.): FORGE 2021: Forschungsdaten in den Geisteswissenschaften - Mapping the Landscape - Geisteswissenschaftliches Forschungsdatenmanagement zwischen lokalen und globalen, generischen und spezifischen Lösungen (FORGE2021). Meeting: Cologne, 08.09.2021 - 10.09.2021.
DOI: *https://doi.org/10.5281/zenodo.5379639*

Ein Standard in der Praxis: ISO 24624:2016 Transcription of spoken language

Schmidt, Thomas
thomas.schmidt[at]ids-mannheim.de
Leibniz-Institut für Deutsche Sprache (IDS), Mannheim, Deutschland

Hedeland, Hanna
hedeland[at]ids-mannheim.de
Leibniz-Institut für Deutsche Sprache (IDS), Mannheim, Deutschland

Frick, Elena
frick[at]ids-mannheim.de
Leibniz-Institut für Deutsche Sprache (IDS), Mannheim, Deutschland

Zusammenfassung. Der Beitrag beschreibt die Entwicklung und Anwendung des TEI-basierten ISO-Standards ISO 24624:2016 Transcription of spoken language, der seit einigen Jahren für gesprochensprachliche Forschungsdaten aus unterschiedlichen Kontexten eingesetzt wird. Ein standardisiertes Dateiformat ermöglicht Interoperabilität zwischen verschiedenen Werkzeugen und weiteren Angeboten von Datenzentren und Infrastrukturen. Durch die methodologisch fundierte Abwägung zwischen Standardisierung und Flexibilität kann der ISO/TEI-Standard zudem Forschungsdaten aus verschiedenen Forschungskontexten abbilden, und so interdisziplinäre Vorhaben erleichtern. Der Beitrag stellt einige Anwendungsbereiche aus dem Lebenszyklus gesprochensprachlicher Forschungsdaten vor, in denen auf dem ISO/TEI-Standard basierenden Erweiterungen existierender Softwarelösungen erfolgreich umgesetzt werden konnten, und zeigt weitere Beispiele für die zunehmende Verbreitung des Formats.

1 Einführung

Der Standardisierung von Dateiformaten – über einzelne Werkzeuge, Datenzentren und möglichst auch Disziplinen hinweg – kann eine Schlüsselrolle im Aufbau digitaler Infrastrukturen zukommen. Bei ausreichender Verbreitung und hinreichender technischer Unterstützung haben Standards das Potential, Arbeitsabläufe zu effektivieren, die Nachnutzbarkeit und Interoperabilität von Tools zu steigern und die nachhaltige Archivierung und Bereitstellung von Forschungsdaten zu vereinfachen. Nicht-wissenschaftsspezifische Standards wie XML (als

W3C-Recommendation für strukturierten Text) oder MPEG (für audiovisuelle Daten) belegen schon seit langem, dass und wie sich dieses Potential realisieren lässt. Für wissenschaftsspezifische Standards ist teilweise noch offen, welche Form und Art der Verbreitung am besten geeignet sind.
Im Jahr 2016 hat die ISO unter dem Titel „ISO 24624:2016 Language resource management — Transcription of spoken language“ (ab hier: ISO/TEI-Standard) einen Standard für die digitale Repräsentation von Transkripten gesprochener Sprache veröffentlicht. Dieser basiert auf Kapitel 8 der Richtlinien der Text Encoding Initiative und orientiert sich an den in Schmidt (2011) formulierten Ideen, die TEI-Richtlinien mit der Praxis, in der die Formate verschiedener Transkriptionstools wie ELAN, EXMARaLDA oder Praat als De-Facto-Standards fungieren, in Einklang zu bringen.

2 Anwendungskontexte des ISO/TEI-Standards

In diesem Beitrag werden exemplarisch verschiedene Kontexte vorgestellt, in denen der ISO/TEI-Standard zum Einsatz kommt. Dabei wird jeweils kurz der Kontext selbst beschrieben und dann auf spezifische Herausforderungen eingegangen, die sich in Bezug auf die praktische Nutzung, die Akzeptanz und die Verbreitung des Standards stellen.

2.1 Datenerzeugung: EXMARaLDA

Aus Sicht der Forschenden ist die Nutzung eines Standards nur dann eine Option, wenn sie sich mit etablierten Arbeitsabläufen vereinbaren lässt und möglichst wenig zusätzlichen Aufwand verursacht. Für Korpora gesprochener Sprache, wo die manuelle Transkription und Annotation des audiovisuellen Quellmaterials einen wesentlichen Anteil der Arbeit ausmacht, wird kein Standard Aussicht auf Akzeptanz haben, der sich nicht von mindestens einem einer kleinen Menge etablierter Tools erzeugen lässt. EXMARaLDA[1] – neben ELAN[2], Praat[3] und Transcriber[4] eines der weiter verbreiteten solchen Tools – kann in seiner neuesten Fassung nun ISO/TEI-Daten in verschiedenen Varianten importieren und exportieren. Etablierte Arbeitsabläufe müssen dafür nicht verändert werden.

[1] Schmidt u. Wörner 2014, https://exmaralda.org.
[2] https://www.mpi.nl/corpus/html/elan/.
[3] http://www.praat.org.
[4] http://transag.sourceforge.net/.

2.2 Manuelle Annotation: WebAnno

Im Rahmen eines Projekts zur Förderung innovativer Lehrkonzepte wurde das webbasierte Annotationswerkzeug WebAnno[5] um ein Modul für die Unterstützung audiovisueller Daten sowie die Visualisierung von Transkriptionsdaten im ISO/TEI-Format erweitert[6]. Im Unterschied zu den gängigen Transkriptions- und Annotationswerkzeugen erlaubt WebAnno einen erweiterten Fokus auf größere Abschnitte des Diskurses, sowie komplexe Annotationstypen in Form von hierarchischen Bäumen oder Ketten. In der Lehre wurde die in WebAnno bereits vorhandenen Funktionen der kollaborativen Annotation und Darstellung der (Nicht-)Übereinstimmung der Annotation (Inter-Annotator-Agreement) genutzt.

2.3 Automatische Annotation: WebLicht / WebMAUS

In der europäischen Forschungsinfrastruktur CLARIN[7] wurde der Ansatz verfolgt, automatische Annotationsaufgaben über standardisierte Web-Schnittstellen zu realisieren. Waren die Schnittstellen anfangs noch auf schriftsprachliche Daten einerseits (das Format TCF und der Dienst WebLicht[8], der zahlreiche NLP-Webservices anbietet) und „phonetische" Daten andererseits (die Webservices WebMaus[9] für phonetische Segmentierung) ausgerichtet, können WebLicht und WebMaus mittlerweile auch Daten in ISO/TEI verarbeiten bzw. erzeugen. Damit erweitern sich der Einsatzbereich der Webservices und die Möglichkeiten, Aufgaben in der Erstellung und Analyse mündlicher Korpora zu automatisieren.[10]

2.4 Datenqualität: QUEST

Im BMBF-geförderten Projekt QUEST[11], das mittels Qualitäts- und Kurationskriterien für audiovisuelle annotierte Sprachdaten deren Nachnutzung fördern und vereinfachen möchte, spielt der ISO/TEI-Standard eine zentrale Rolle. Da Transkriptionsdaten aus verschiedenen Disziplinen in verschiedenen Dateiformaten erstellt werden, ermöglicht eine Abbil-

[5] Eckart de Castilho et al. 2014.
[6] Remus et al. 2019, eine Demo-Version steht unter http://ltdemos.informatik.uni-hamburg.de/webanno-mm zur Verfügung.
[7] https://www.clarin.eu/.
[8] https://weblicht.sfs.uni-tuebingen.de.
[9] https://clarin.phonetik.uni-muenchen.de/BASWebServices/interface/WebMAUSBasic
[10] vgl. Schmidt et al. 2017, Fisseni u. Schmidt 2020.
[11] https://www.slm.uni-hamburg.de/ifuu/forschung/forschungsprojekte/quest.html.

dung auf das ISO/TEI-Format einerseits überhaupt eine einheitliche Betrachtung und Verwendung der Daten, andererseits ist für diese Abbildung eine genaue Dokumentation der Originaldaten notwendig. Die erforderliche Dokumentation benutzter Formate, Schemata und Konventionen erfüllt zugleich entsprechende Anforderungen der FAIR-Prinzipien und dient daher als Operationalisierung dieser für Nachnutzende entscheidenden Aspekte der Datenqualität.[12]

2.5 Datenzugang (Query, Anzeige): ZuMult

Für die Visualisierung und Query mündlicher Korpora werden im Projekt ZuMult[13] webbasierte Zugänge entwickelt, die auf einer weiter spezifizierten Fassung des ISO/TEI-Standards basieren. Neben XSL-Stylesheets, die verschiedene Ansichten der XML-basierten Ausgangsdaten generieren, entwickelt ZuMult vor allem eine auf dem MTAS-System[14] basierende Query Engine, über die mündliche Korpora mit vielfältigen Annotationen mittels der in der Korpuslinguistik weit verbreiteten CQP Query Language[15] abgefragt werden können.[16]

2.6 Weitere Entwicklungen und Einsatzbereiche

Die genannten Kontexte sind eine Auswahl derjenigen, in denen wir selbst Anwendungen entwickeln, für die der ISO/TEI-Standard eine zentrale Rolle spielt. Fünf Jahre nach Veröffentlichung des Standards lässt sich jedoch durchaus beobachten, dass es weitere Zusammenhänge gibt, in denen der Standard sich zu etablieren beginnt. Exemplarisch genannt seien das im Umfeld des französischen CORLI-Konsortiums (CORpora, Languages and Interaction) entwickelte System TEICorpo[17], das Spoken Torlak Corpus[18] oder die Verwendung im Rahmen des Langzeitvorhabens INEL[19], das sich mit der Dokumentation indigener nordeurasischer Sprachen befasst.

[12] vgl. Hedeland 2021.
[13] http://zumult.org.
[14] Brouwer et al. 2016.
[15] http://cwb.sourceforge.net/files/CQP_Tutorial/.
[16] Frick u. Schmidt 2020.
[17] Parisse et al. 2018, https://ct3.ortolang.fr/ct3/teicorpo-doc/.
[18] Vuković 2021, 2020.
[19] Ferger u. Jettka 2020, https://www.slm.uni-hamburg.de/inel/.

3 Diskussion

Als vorläufige Erkenntnis zeigt sich in der Summe, dass sich durch die Einführung des ISO/TEI-Standards neue Möglichkeiten auftun, in denen sich das versprochene Potential zu realisieren beginnt. Entscheidend für die Praxis sind dabei nach unserer Ansicht erstens eine adäquate Unterstützung durch geeignete Tools, zweitens ein für ForscherInnen greifbarer Mehrwert durch verbesserte Interoperabilität. Beides erfordert aufwändige Entwicklungsarbeit und kontinuierliche Abstimmung mit AnwenderInnen, anderen EntwicklerInnen und Infrastrukturinitiativen.

Bibliografie

Brouwer, Matthijs, Hennie Brugman und Marc Kemps-Snijders. 2017. "MTAS: A Solr/Lucene based Multi Tier Annotation Search solution." In *Selected papers from the CLARIN Annual Conference 2016, Aix-en-Provence, 26–28 October 2016. Linköping Electronic Conference Proceedings*. 136, Nr 2 (2017): 19–37.

Eckart de Castilho, Richard, Chris Biemann, Irina Gurevych und Seid Muhie Yimam. 2014. "WebAnno: a flexible, web-based annotation tool for CLARIN." In *Proceedings of the CLARIN Annual Conference 2014, October 24–25, 2014, Soesterberg, The Netherlands*.

Ferger, Anne und Daniel Jettka. 2020. "Use Cases of the ISO Standard for Transcription of Spoken Language in the Project INEL." In *Proceedings of CLARIN Annual Conference 2020. Virtual Edition*, hrsg. v. Navarretta, Constanze und Maria Eskevich, 126–130. Utrecht: CLARIN.

Fisseni, Bernhard und Thomas Schmidt. 2020. "CLARIN Web Services for TEI-annotated Transcripts of Spoken Language." In *Selected Papers from the CLARIN Annual Conference 2019. Linköping Electronic Conference Proceedings*. 172, Nr. 3 (2020): 12–22.

Frick, Elena und Thomas Schmidt. 2020. "Using Full Text Indices for Querying Spoken Language Data." *In Proceedings of*

the LREC 2020 Workshop, Language Resources and Evaluation Conference, 11–16 May 2020, 8th Workshop on Challenges in the Management of Large Corpora (CMLC-8), 40–46. Paris: European Language Resources Association, 2020.

Hedeland, Hanna. 2021. “Towards Comprehensive Definitions of Data Quality for Audiovisual Annotated Language Resources.” In *Selected papers from the CLARIN Annual Conference 2020. Linköping Electronic Conference Proceedings*. 180, Nr. 11 (2021). 93–103.

Parisse, Christophe, Céline Poudat, Ciara R. Wigham, Michel Jacobson und Loïc Liégeois. 2018. “CORLI: A linguistic consortium for corpus, language, and interaction.” In *Selected papers from the CLARIN Annual Conference 2017, Budapest, 18–20 September 2017. Linköping Electronic Conference Proceedings*. 147, Nr. 2 (2018): 15–24.

Remus, Steffen, Hanna Hedeland, Anne Ferger, Kristin Bührig und Chris Biemann. 2019. “WebAnno-MM: EXMARaLDA meets WebAnno.” In *Selected papers from the CLARIN Annual Conference 2018, Pisa, 8–10 October 2018. Linköping Electronic Conference Proceedings*. 159, Nr. 17 (2019): 166–172.

Schmidt, Thomas. 2011. “A TEI-based approach to standardising spoken language transcription.” *Journal of the Text Encoding Initiative* 1 (2011). doi:10.4000/jtei.142.

Schmidt, Thomas und Kai Wörner. 2014. “EXMARaLDA.” In *The Oxford Handbook of Corpus Phonology*, hrsg. v. Durand, Jacques, Ulrike Gut und Gjert Kristoffersen, 402–419. Oxford: OUP

Schmidt, Thomas, Hanna Hedeland und Daniel Jettka. 2017. “Conversion and annotation web services for spoken language data in CLARIN.” In *Selected papers from the CLARIN Annual Conference 2016, Aix-en-Provence, 26–28 October 2016. Linköping Electronic Conference Proceedings*. 136, Nr 9 (2017): 113–130.

Vuković, Teodora. 2020. *Spoken Torlak dialect corpus 1.0 (transcription)*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1281.

Vuković, Teodora. 2021. “Representing variation in a spoken corpus of an endangered dialect: the case of Torlak.” *Language Resources and Evaluation*. https://doi.org/10.1007/s10579-020-09522-4.