Schriften des Instituts für Dokumentologie und Editorik — Band 16

Digitale Edition in Österreich
Digital Scholarly Edition in Austria

---

herausgegeben von | edited by

Roman Bleier, Helmut W. Klug

2023

BoD, Norderstedt

Bibliografische Information der Deutschen Nationalbibliothek:
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de/ abrufbar.

Digitale Parallelfassung der gedruckten Publikation zur Archivierung im Kölner Universitäts-Publikations-Server (KUPS). Stand 29. April 2023.

2023

Herstellung und Verlag: Books on Demand GmbH, Norderstedt
ISBN: 978-3-743-102-842
Einbandgestaltung: Stefan Dumont; Coverbild: wurde von Roman Bleier und Helmut Klug für ein KONDE-Poster (DHa 2017) erstellt
Satz: Roman Bleier und LuaTEX

Editionen und Graphentechnologie: Vorteile und Hürden digitaler Editionstechniken abseits von TEI-XML

Peter Hinkelmanns

Abstract

The article offers an overview of the use of graphs for the representation of texts. The focus is on variant graphs for genetic editions and the embedding of RDF data in edition projects based on TEI-XML. A possible technology stack for such an edition project with connection to the Semantic Web is also presented. Finally, a guideline offers a decision-making aid for the conception of an edition as a variant graph.

Zusammenfassung

Der Beitrag bietet einen Überblick über den Einsatz von Graphen zur Repräsentation von Texten. Dabei stehen besonders Variantengraphen für genetische Edition und die Einbindung von RDF-Daten in Editionsprojekte auf TEI-XML-Basis im Fokus. Vorgestellt wird auch ein möglicher Technologie-Stack für ein solches Editionsprojekt mit Anbindung an das Semantic Web. Abschließend bietet ein Leitfaden eine Entscheidungshilfe für die Konzeption einer eigenen Edition als Variantengraph.

Für die Abbildung von Bezügen zwischen Texten aber auch zu anderen Datensätzen eignen sich Graphen in besonderer Weise. Ein ‚Graph' ist ein Datenmodell, das aus ‚Knoten' und diese verbindenden ‚Kanten' besteht. Ein Knoten ‚Buch' kann etwa über die gerichtete Kante ‚hat Autor' mit einer ‚Person' verbunden sein. Den Einsatz von Graphen in den digitalen Geisteswissenschaften hat Andreas Kuczera (Kuczera 2017) vorgestellt. Er hält fest, dass „Graphentechnologien für die Weiterentwicklung der digitalen Geisteswissenschaften sehr interessante Perspektiven bieten." (Kuczera 2017, 196) Auch Texteditionen, allen voran an textgenetischen Fragestellungen orientierte, können von der Realisierung mit Graphen profitieren.

Der vorliegende Beitrag bietet einen Überblick über den Einsatz von (Varianten-) Graphen für Editionsprojekte und stellt dies gängigeren Methoden wie der Realisierung mit TEI-XML gegenüber. Im Mittelpunkt stehen dabei folgende Fragen: Welche Vorteile kann eine graphbasierte Edition besonders in Hinblick auf genetische Editionen bieten und welche Risiken sind mit einem zu TEI-XML alternativen Verfahren

Digitale Edition in Österreich. Hrsg. Roman Bleier und Helmut W. Klug. Schriften des Instituts für Dokumentologie und Editorik 16. Books on Demand, 2023, 19–36.

verbunden? Wie kann ein entsprechender Technologie-Stack aussehen? Welche Möglichkeiten bietet darüber hinaus eine Kombination von TEI-XML mit RDF-Daten? Als Zusammenfassung soll ein kleiner Leitfaden Unterstützung für die Konzeption eines eigenen Editionsprojektes mit Einbezug von Graphdaten bieten.

1 Modelle der Textrepräsentation

Grob lassen sich Modelle zur Repräsentation und Annotation von Text in die zwei Ansätze ‚hierarchischer Graph' und ‚gerichteter Graph' einteilen. *Markup*, also das Annotieren von Text, kann dabei interlinear (*inline*) oder separat mit Textbezug (*standoff*) erfolgen. Das bei den meisten Editionen eingesetzte Modell der *Text Encoding Initiative* (TEI) ist ein *Markup*-Modell, bei dem ein hinterlegter Text *inline* mit Annotationen versehen wird. Das Codebeispiel unten und in Abbildung 1 zeigen einen kurzen mit linguistischer Annotation der Wortart versehenen Ausschnitt aus Lewis Carrolls *Alices Abenteuer im Wunderland*:

```
<p>
    <lb/>
    <w pos="NE">Alice</w>
    <w pos="VVFIN">fing</w>
    <w pos="APPR">an</w>
    <w pos="PRF">sich</w>
    <w pos="PTKZU">zu</w>
    <w pos="VVINF">
        lang<pc>-</pc>
        <lb break="no"/>weilen
    </w>
    <pc>;</pc>
    <w pos="PPER">sie</w>
    <w pos="VVFIN">saß</w>
    <w pos="ADV">schon</w>
    <w pos="ADV">lange</w>
    <w pos="APPR">bei<w>
    <w pos="PPOSAT">ihrer</w>
    <w pos="NN">Schwester</w>
    <w pos="APPRART">am</w>
    <w pos="NN">Ufer</w>
    [...]
</p>
```

Realisiert wird das Datenmodell der TEI mit XML, einer Markupsprache, bei der ein Knoten des beschriebenen Graphen jeweils genau einen Elternknoten hat und es einen einzigen Wurzelknoten für das gesamte Dokument geben muss. Wenn also etwa ein Wort, wie im obigen Beispiel ‚langweilen', über das Zeilenende hinausläuft und sowohl die einzelnen Zeilen als auch das Wort in einem Knoten repräsentiert werden sollen, führt dies zu einer Überlappung, die das Prinzip der geforderten strengen Hierarchie des XML-Baums verletzt. Mehrere Strategien sind in TEI eingeführt worden, um Probleme der Hierarchie zu lösen. So können etwa leere Elemente und *Standoff-*

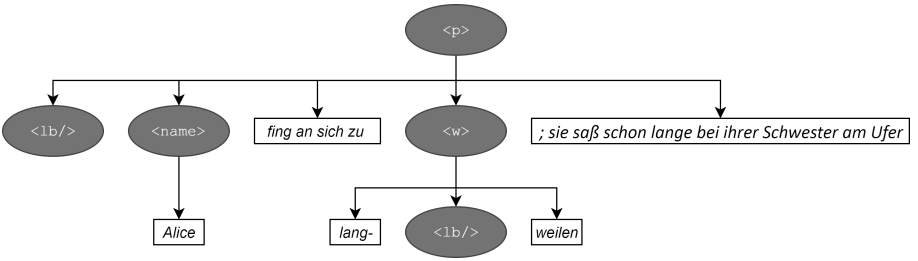

Abbildung 1: Darstellung des Code-Beispiels als hierarchischer Graph.

Markup eingesetzt werden, um Überlappungen zu ermöglichen. Im Beispiel sind die Zeilenanfänge mit dem leeren Element `<lb/>` markiert und nicht die gesamte Zeile von einem Element umschlossen, da die Dokumentstruktur in TEI als nachgeordnet eingestuft und durch leere Elemente realisiert wird. Somit kann die Worttrennung beim Beispielwort ‚langweilen' mit einem umschließenden Element `<w>` zeilenübergreifend markiert werden, ohne die Hierarchie formal zu verletzen.

Mit der Erweiterung von TEI um für genetische Editionen erforderliche Elemente hat sich eine Arbeitsgruppe befasst, die ihre Ergebnisse in einem Entwurf vorgestellt hat (Burnard et al. 2010). Teile dieses Entwurfes sind in die TEI-Guidelines übernommen worden (Text Encoding Initiative Consortium 2019; vgl. Wout Dillen 2016). Mit dem Modell der TEI lassen sich somit auch komplexe genetische Editionen realisieren. Ein kurzes Beispiel aus einer Greifswalder Handschrift (Abbildung 2) demonstriert eine Möglichkeit zur Transkription von intertextueller Varianz in Codebeispiel 2:

```
<lb/>[...] das man
<lb/>an dem ort der ſtadt ſo alzeit wuſte iſt, noch
<lb/>ferner lieber wolt eine <del>Einode vnd</del> Wuſtenei
<lb/>am Cloſter
<subst>
    <del seq="1">ſehen</del>
    <add seq="2">wiſſen</add>
</subst> oder daſſelbige einen Pau
<lb break="no"/>ren liebere gonnen [...]
```

Ob die auf dem XML-Format gründenden strengen Hierarchien ein Problem für die Codierung von Texten darstellen, ist seit Veröffentlichung des TEI-Modells auf XML-Basis intensiv diskutiert worden.[1] Mittels der drei Elemente `<subst>`, `<add>`

[1] Ein umfassender Beitrag zur Problematik der Überlappungen liegt mit DeRose (2004) vor. Weitere Auseinandersetzungen mit dem Thema finden sich, chronologisch geordnet, etwa bei Huitfeldt 1994, 237; Buzzetti 2002, 76; Ide und Romary 2006, 227; Burghardt und Wolff 2009, 54; Schmidt und Colomb 2009, 498–99; Di Iorio, Peroni und Vitali 2011; Stührenberg 2012, §42; Schmidt 2012, 129–30; Sahle 2013, 103; Kuczera 2016; Efer 2016, 36–38; Haentjens Dekker und Birnbaum 2017; Bleeker 2017, 85; Turska und Spandini 2018; Bruder und Teufel 2018, 161.

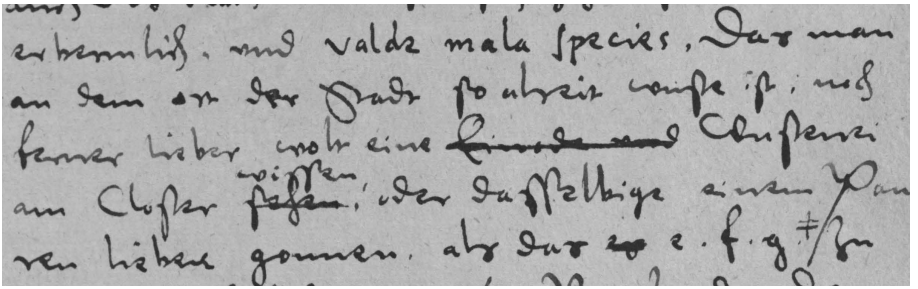

Abbildung 2: Textpassage (Universitätsarchiv Greifswald, Altes Rektorat Hbg. 134, 16r).

und <del> können Textvarianten innerhalb eines Textzeugen nach Regeln der TEI ausgezeichnet werden. Intertextuelle Varianz kann in TEI etwa über einen Apparat über das Element <app>, dokumentiert werden, um etwa separate Transkriptionen mehrerer Textzeugen aufeinander zu beziehen.[2] Mit GODDAG (Sperberg-McQueen und Huitfeldt 2004), GrAF (Ide und Suderman 2007) oder TAGML (Haentjens Dekker und Birnbaum 2017; Haentjens Dekker et al. 2018) sind Markupmodelle als Alternative zu XML vorgestellt worden, die überlappende Hierarchien erlauben. Dass TEI trotz der Basis XML überlappende Hierarchien bewältigen kann, hat James Cummings herausgestellt (Cummings 2018, 13–14). Gleichwohl bedarf es dazu einer Erweiterung des *Inline-Markups* durch *Standoff*-Annotationen.

2 Genetische Editionen als Variantengraph

In der Editions- und Literaturwissenschaft erfreuen sich textgenetische Editionen anhaltender Popularität. Vertreter der jüngeren Forschungsgeschichte sind etwa die *Faust-Edition* (Bohnenkamp, Henke und Jannidis 2016), die *Edition der Manuskripte Samuel Becketts* (van Hulle und Nixon 2011–), die *Edition der Werke Arthur Schnitzlers* (Burch et al. 2016) oder auch die *Edition des Mann ohne Eigenschaften* Robert Musils (Bosse, Boelderl und Fanta 2016–), die eine vollständige Wiedergabe der Textgenese anstreben. ‚Textgenese' bezeichnet das Nachvollziehen der Entstehung von Texten. Dabei kann ein ‚Text' Bearbeitungen mehrerer Schreiberinnen und/oder Schreiber umfassen, sich über unterschiedliche Textträger erstrecken und sich in parallele Textfassungen verzweigen. Das Wesen einer genetischen Edition beschreibt Paolo D'Iorio treffend: „Das Ziel einer genetischen Edition ist es, die existierenden Textdokumente

[2] Vgl. z. B. das Datenmodell der Software Versioning Machine (Vetter et al. 2016) oder jenes, das Projekt Welscher Gast Digital (Šimek 2014, 4) benutzt.

zueinander in Beziehung zu setzen und so zu kommentieren, dass die Textgenese nachvollziehbar wird." (D'Iorio 2017, 196)

Alternativen zum TEI-Modell, auch für genetische Editionen, sind von der Forschung wiederholt diskutiert worden. Die Arbeitsgemeinschaft *Graphentechnologien in den Digitalen Geistes- und Sozialwissenschaften* hat das Thema auf Tagungen und in Workshops diskutiert (Kuczera et al. 2017–). Ausführlich beschäftigen sich auch die Dissertationsschriften Thomas Efers (Efer 2016) und Elli Bleekers (Bleeker 2017) mit der Abbildung von Texten als Graphen bzw. mit genetischen Texteditionen. Gemein ist den Ansätzen eine Abkehr vom *Markup*-Ansatz, bei dem ein Text gleich dem analogen Markieren mit dem Textmarker annotiert wird.

Eingeführt wurde der Variantengraph zur Textcodierung von Desmond Schmidt und Robert Colomb (Schmidt und Colomb 2009). Sie beschreiben den Variantengraph als azyklischen, gerichteten Graphen mit einem Start- und einem Endknoten (Schmidt und Colomb 2009, 501). Dies bedeutet, dass es vom Start- zum Endpunkt zahlreiche Wege geben kann, die aber keine Rückkehr an einen zuvor besuchten Punkt erlauben. Ihr Modell bietet die Möglichkeit, Textvarianten unterschiedlicher Textzeugen darzustellen. Der Text erscheint dabei auf den Kanten des Graphen, Knotenpunkte stellen Verzweigungen in der Textüberlieferung dar: Für jeden Textzeugen existiert genau ein Pfad durch den Graphen, der dessen Text ergibt, wie zum Beispiel für *E* in Abbildung 3.

Auch Hinzufügungen, Löschungen, Ersetzungen und Umstellungen lassen sich aus dem Graphen direkt ableiten. So ersetzt *scandito* nach Knoten 1 in Version C: *certo* aus Version B während DEFHI den Text an dieser Stelle ersatzlos streichen. Umstellungen zwischen Versionen können durch weitere Kanten gekennzeichnet werden, die als gestrichelte Kanten dargestellt werden.

CollateX (Haentjens Dekker et al. 2015), welches seit 2010 als Nachfolgeprojekt der Software *Collate* entwickelt wird, ermöglicht die automatische Kollationierung unterschiedlicher Varianten eines Textes. Datenmodell ist ein Variantengraph (vgl. Abbildung 4), dessen Knoten den Text abbilden und dessen Kanten den Verlauf des Textes zeigen. Anders als im Modell von Schmidt und Colomb 2009 werden die Textzeugen tokenisiert, die einzelnen Wörter bilden also die Knoten des Graphen.

Ein Modell, das auf dem von *CollateX* basiert (Andrews und Mace 2013, 506), wird für das Projekt *Stemmaweb* genutzt (*The Stemmaweb Project* 2012–). Die verschiedenen Types der Tokens der Textzeugen bilden die Knoten des Graphen in Abbildung 5. ‚Tokens' sind die einzelnen Vorkommen der Wortformen im Text, ‚Types' hingegen das Inventar dieser Tokens. Gerichtete Kanten zeigen den Verlauf des Textes in den unterschiedlichen Textzeugen an. Zwischen den einzelnen Type-Knoten kann durch weitere Kanten die Variante näher bestimmt werden. Möglich ist etwa die Annotation orthographischer, grammatischer oder lexikalischer Varianz (Andrews und Mace 2013, 508).

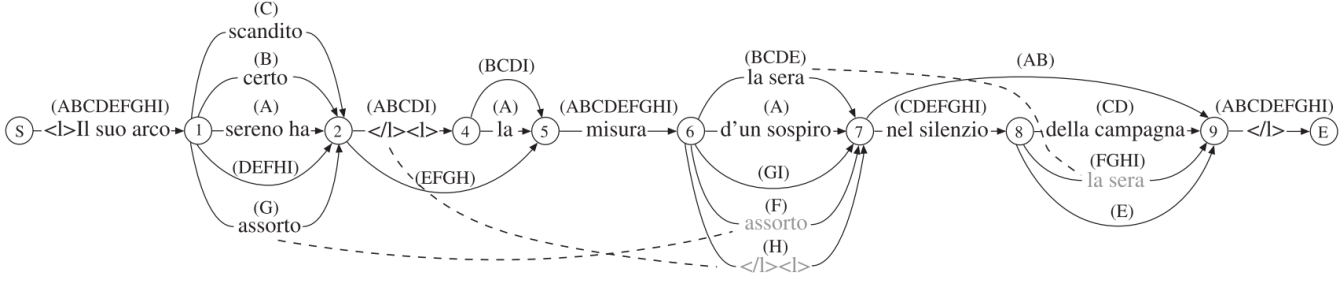

Abbildung 3: Ein Variantengraph (Abbildung aus Schmidt und Colomb 2009, 502).

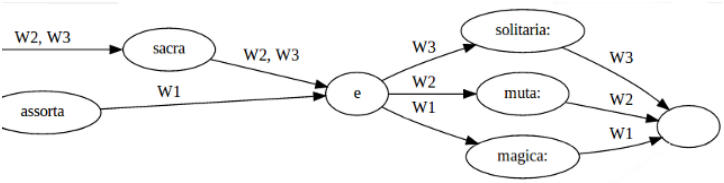

Abbildung 4: Variantengraph CollateX (Abbildung aus Haentjens Dekker und Middell 2010–).

Das graphbasierte Recherchesystem *Kadmos* ist von Thomas Efer vorgestellt worden (Efer 2016, 73–74). Das Modell unterscheidet zwischen Tokens und Types. Zwar stellt das Modell, das Abbildung 6 zeigt, keinen Variantengraph mehrerer Textzeugen dar, anders als die bereits vorgestellten Modelle besitzt es jedoch einen Knoten für jeden im Text vorkommende Type. Der Graph zeigt darüber hinaus, dass auch weitere Annotationen, wie die Zuordnung von Tokens zu Sätzen in einem Graph realisiert werden können. Konkurrierende Annotationsschichten sind dabei, anders als bei XML-*Inline-Markup*, voneinander unabhängig.

Auch der Frage der Visualisierung von Variantengraphen haben sich mehrere Untersuchungen gewidmet. Zu nennen ist hier der Ansatz des Projekts *Text Re-use Alignment Visualization* (TRAViz: Jänicke et al. 2015), weiters die Beiträge Jänicke und Wrisley 2018; Bleeker und Kelly 2018; Bleeker, Buitendijk und Haentjens Dekker 2019. Aufgrund des Fokus dieses Beitrags kann eine tiefergehende Darstellung des Themas ‚Visualisierung' leider nicht erfolgen, es bleibt jedoch festzuhalten, dass mit zunehmender Variation des Textes auch die Anforderungen an eine Visualisierung steigen. Und zwar unabhängig davon, ob TEI-XML oder ein gerichteter Graph als Basis der Edition gewählt worden ist.

3 Technischer Vergleich: XML und RDF

Die Textdarstellungen als hierarchischer Graph und als gerichteter Graph setzen auf unterschiedliche Technologie-Stacks. Das Format der *Text Encoding Initiative* baut in der gegenwärtigen Umsetzung auf XML auf. Im Vorwort der *P5 Guidelines* heißt es jedoch:

> However, the TEI encoding scheme itself does not depend on this language; it was originally formulated in terms of SGML (the ISO Standard Generalized Markup Language), a predecessor of XML, and may in future years be re-expressed in other ways as the field of markup develops and matures. (Text Encoding Initiative Consortium 2019, iv. About These Guidelines)

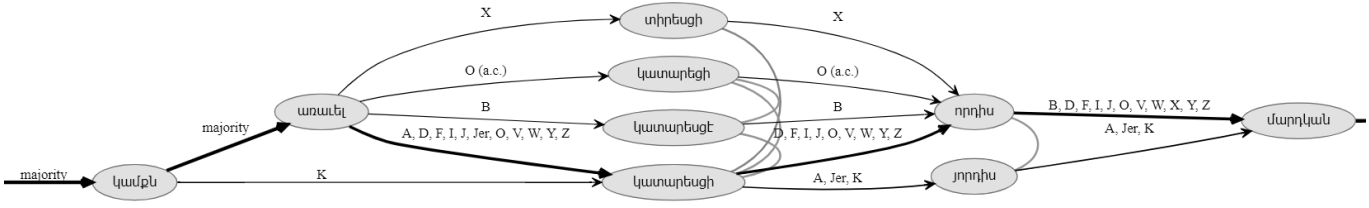

Abbildung 5: Screenshot des Text Relationship Mappers von Stemmaweb mit einem Ausschnitt aus dem ersten Abschnitt der Chronik Matthias, von Edessa (Abbildung aus *The Stemmaweb Project* 2012–).

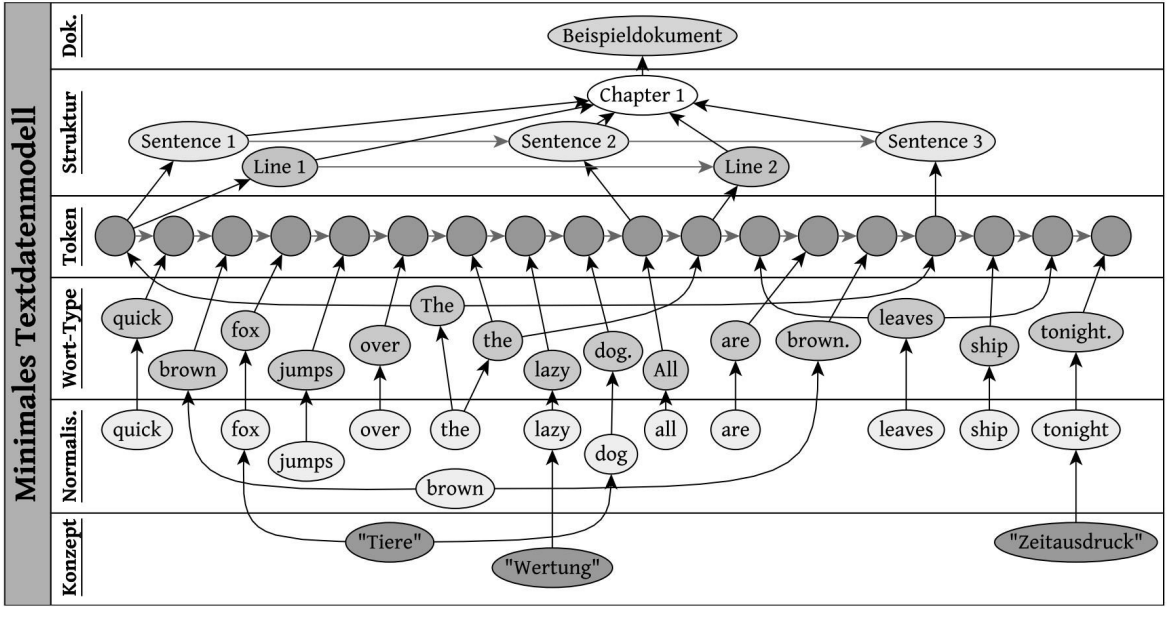

Abbildung 6: Schematische Darstellung der Instanzdatensätze und Verknüpfungen eines kurzen Beispieldokuments bei minimalistischem Textdatenmodell (Abbildung aus Efer 2016, 76).

TEI ist damit nicht an eine Serialisierung gebunden, sodass eine an die Wahl des Formats XML gerichtete Kritik zukünftig gegenstandslos werden könnte. XML bietet für wesentliche Funktionalitäten Lösungen, die es attraktiv für den Einsatz in einem Editionsprojekt machen. XML wird in Form von Textdateien gespeichert, die für die Langzeitarchivierung und -lesbarkeit besser geeignet sind als etwa Binärdateien. Mit der Sprache *XML Schema* (Gao, Sperberg-McQueen und Thompson 2012) kann die Struktur von XML-Daten beschrieben werden und XML-Daten können auf Validität gegenüber einem Schema geprüft werden. Ein Schema definiert etwa, dass ein TEI-Dokument einen Kopfbereich haben muss, in dem ein Dokumenttitel angegeben ist. Zur Abfrage von Inhalten, aber auch, um weitreichendere Operationen zu ermöglichen, wurde *XQuery* (Robie, Dyck und Spiegel 2017b) eingeführt. Mittlerweile unterstützt *XQuery* nicht nur Abfragen auf XML-Daten, sondern auch JSON. Zur Transformation eines gesamten Dokumentes gedacht ist *XSLT* (Saxonica 2007). *XSLT* ermöglicht etwa die Umwandlung eines TEI-XML-Dokuments in ein HTML-Dokument. Die für *XML Schema*, *XQuery* und *XSLT* notwendige Navigation im Baum erfolgt über *XPATH* (Robie, Dyck und Spiegel 2017a).

Die Erstellung einer auf TEI-XML basierenden Edition unterstützen zahlreiche Werkzeuge wie XML-Editoren und XML-Datenbanken. Unter den Editoren hervorzuheben sind *Atom*[3] und *Oxygen.*[4] XML-Datenbanken sind etwa *eXist-db*[5] oder *BaseX*[6] wobei sich die Zugriffsgeschwindigkeiten der Datenbanken durchaus unterscheiden.[7]

Nachdem eine Standardisierung für Editionen auf Graphbasis noch aussteht, könnten an dieser Stelle auch beliebige andere Graphdatenmodelle genannt werden. Als Beispiel eines Technologie-Stacks für eine auf einem gerichteten Graph basierende Edition soll hier RDF vorgestellt werden. Das *Semantic Web* ist die Umsetzung einer Idee von Tim Berners-Lee (Berners-Lee 2006; zusammenfassend auch Stein 2014), um semantisch ausgezeichnete Daten maschinenlesbar über das Netz verfügbar zu machen. Im Fokus steht also die Vernetzung aller in das System eingespeister Daten, ein auch für Editionsprojekte nützlicher Vorteil. Wesentliche Technologien sind das *Resource Description Framework* (Gandon und Schreiber 2014), die Web Ontology Language (OWL: W3C OWL Working Group 2012) und die *SPARQL Protocol And RDF Query Language* (W3C SPARQL Working Group 2013). RDF bildet die Basis, um Triple bestehend aus Subjekt – Objekt – Prädikat zu beschreiben, und kann in unterschiedlichen Textformaten wie *XML-RDF* oder *N3* mit dessen Subset *Turtle* serialisiert werden. Für eine Langzeitarchivierung gelten damit dieselben Voraussetzungen wie für XML-Daten. OWL beschreibt Objektklassen und deren Beziehungen unter-

[3] https://atom.io.
[4] https://www.oxygenxml.com.
[5] http://exist-db.org.
[6] http://basex.org.
[7] Für einen Überblick über die Performance und weitere Eigenschaften vgl. Chau et al. 2019.

	XML	RDF / Semantic Web
VORANNAME	Closed-World	Open-World
ARCHIVIERUNG	Textdatei	Textdatei
MODELLIERUNG	XML Schema	OWL
VALIDIERUNG	XML Schema	SHACL
ABFRAGE	XQuery	SPARQL
TRANSFORMATION	XSLT	-(SPARQL Construct)
SPEZIFISCHES VOKABULAR FÜR EDITIONEN	TEI	-(domänenspezifische Ontologien)

Tabelle 1: Technologie-Stacks im Vergleich.

einander und SPARQL ermöglicht schließlich das Abfragen der erzeugten Graphen. Anders als XML geht RDF von einem *Open-World*-Ansatz aus. Wesentliches Prinzip ist, dass nicht vorausgesetzt werden kann, dass im *Semantic Web* ein Teilnehmer das gesamte Netz kennt. Daher kann eine Aussage mit einer OWL-Ontologie nicht auf deren Wahrheitsgehalt geprüft werden, wie dies durch den *Closed-World*-Ansatz von XML mit einem Schema möglich ist. Für die Datensätze eines Editionsprojektes können mittels der *Shapes Constraint Language* (SHACL: Knublauch und Kontokostas 2017) allerdings Regeln definiert werden, die von einem *Closed-World*-Ansatz im Projektrahmen ausgehen und daher die Validierung der Daten ermöglichen.

Implementationen von RDF-*Triple-Stores*, die zum Technologie-Stack des *Semantic Webs* gehören, sind etwa *Apache Jena*[8] oder *GraphDB*.[9] Die Entscheidung für eine Graphdatenbank unterliegt zahlreichen Kriterien, etwa den Kosten, der Performance oder der verfügbaren Bibliotheken für unterschiedliche Programmiersprachen. Eine Übersicht über die Eigenschaften verschiedener *Triple-Stores* bieten Banane und Belangour (2019).

Eine mit TEI vergleichbare Ontologie zur Beschreibung von Texten und zahlreichen weiteren Daten, wie Personen oder Orten, existiert im Kontext des *Semantic Webs* zwar nicht, jedoch können einzelne, domänenspezifische Bereiche, wie Personendaten, nach normierten Vokabularen bzw. Ontologien beschrieben werden. Ein Beispiel ist die Ontologie der *Gemeinsamen Normdatei* (GND), die unter anderem ein Vokabular zur Beschreibung von Personen, Orten, Werken und Körperschaften bereitstellt (Haffner 2012–). Abschließend sollen die skizzierten Möglichkeiten der beiden Technologie-Stacks tabellarisch gegenübergestellt werden (Tabelle 1).

Sowohl XML als auch RDF bieten eine Vielzahl von Werkzeugen, die wichtige Funktionen der Datenerhebung, des Datenabfragens und des Datenarchivierens er-

[8] https://jena.apache.org.
[9] https://graphdb.ontotext.com.

möglichen. Im Bereich der Normierung ist TEI-XML für Editionen und zahlreiche damit in Verbindung stehende Daten etabliert, während für das *Semantic Web* eher kleinteilige, domänenspezifischere Ontologien von Relevanz sind. Eine Ontologie zur Abbildung von Texten als Graph existiert bislang nicht bzw. nur in Ansätzen.[10] Obwohl TEI-XML auf anderen Technologien basiert, sind bereits Vorschläge gemacht worden, TEI auch als Ontologie verfügbar zu machen (Ciotti 2018; Ciotti und Tomasi 2016).

4 Graphen und (TEI-)Editionen

Neben einem Einsatz von gerichteten Graphen als Datenmodell für die Textwiedergabe können auch auf TEI-XML basierende Editionen mittels Graphen erweitert werden. Die TEI Guidelines enthalten für die direkte Einbettung von RDF-Daten etwa das Element `<xenoData>`. Für die Vernetzung archivalischer bzw. geisteswissenschaftlicher Ressourcen bieten die Vokabulare und Ontologien des *Semantic Webs* vielfältige Möglichkeiten, wie etwa Christopher Pollin und Lina M. Zangerl (2020) zeigen. Ontologien zur Auszeichnung geisteswissenschaftlicher Ressourcen sind unter anderem das *Datenmodel der Europeana* (Isaac 2013), die bereits zuvor genannte *Ontologie der Gemeinsamen Normdatei* oder das *CIDOC Conceptual Reference Model* (Le Boeuf et al. 2017) – letzteres ist als *Top-Level*-Ontologie ein Bezugsmodell für den Entwurf eigener Ontologien und Schnittstelle zwischen unterschiedlichen Datenmodellen.

Die Verbindung von TEI-Editionen mit Graphdaten wird auch in der *Mittelhochdeutschen Begriffsdatenbank* (MHDBDB: Universität Salzburg 1992–) umgesetzt, an deren Entwicklung und Konzeption der Autor beteiligt ist. Aufgrund der vielfältigen Annotationsschichten, die auf den Texten dieses onomasiologischen Projekts zum mittelhochdeutschen Wortschatz liegen, konnte nicht ausgeschlossen werden, dass eine einfache Inline-Annotation zu Verletzungen der Hierarchie von XML führt. Weiters stellt das Begriffssystem einen Thesaurus dar, der idealerweise mit dem RDF-Vokabular *Simple Knowledge Organization System* (SKOS: Miles und Bechhofer 2009) beschrieben wird. Aus diesen Gründen werden in der MHDBDB Texte mit TEI-XML codiert, aber auf Tokenebene mit RDF-Graphdaten vernetzt.

Die Vernetzung von TEI-basierten Briefeditionen ist im Projekt *correspSearch* umgesetzt worden (Dumont 2016). Zwar entspricht das Verfügbarmachen über ein *Application Programming Interface* (API) nicht den *Semantic-Web*-Technologien, es zeigt jedoch sehr gut, welches Potential in der Verbindung unterschiedlicher Editionsprojekte steckt. Die Transformation unter anderem von TEI-Daten in ein RDF-Modell ist ein Ziel der Ontologie der *Digital Edition Publishing Cooperative for Historical*

[10] An dieser Stelle sei auf die Überlegungen des Autors zu einer *Text Graph Ontology* hingewiesen (Hinkelmanns 2019).

Accounts (DEPCHA: Pollin und Vogeler 2019; Vogeler 2019). Die Ontologie ermöglicht es, Rechnungsbucheinträge semantisch auszuzeichnen und somit aufeinander zu beziehen. Wie in *correspSearch* können so die Daten unterschiedlicher Projekte zusammengeführt werden.

5 Leitfaden und Ausblick

Bislang sind auf einem Variantengraph basierende Editionen die Ausnahme. Die laufenden Projekte zeigen jedoch, dass Graphen besonders für genetische Editionen und für die Vernetzung von Editionen unterschiedlicher Projekte einen interessanten Ansatz darstellen. Da eine Standardisierung bislang nur in Grundzügen existiert, ist es auch für graphbasierte Editionen ratsam, einen Export nach TEI-XML zur Archivierung der eigenen Daten einzusetzen. In der Vorbereitungsphase zu einer variantengraph-basierten Edition sollte eine Reihe von Fragen gestellt werden:

1. **Wie soll der Text segmentiert werden?** Genügt die Similarität von Zeichenketten aus unterschiedlichen Textzeugen als Segmentierungskriterium (Schmidt und Colomb 2009)? Soll eine Tokenisierung an Weißraum durchgeführt werden (Andrews und Mace 2013)? Sollen auch Phänomene unterhalb der Wortgrenze betrachtet werden?

2. **Welche Art von Varianz soll analysiert werden?** Sollen etwa Varianzen innerhalb eines Textträgers oder auch zwischen verschiedenen Fassungen eines Textes analysiert werden?

3. **Welche Datenobjekte neben der eigentlichen Textedition soll meine Edition beinhalten?** In zahlreichen Projekten gehören Metadaten zu Entitäten wie Personen, Orten oder Objekten zu den Anforderungen. Eine graphbasierte Edition sollte hier auf Ontologien wie *Europeana* (Isaac 2013) basieren und erforderliche eigene Ontologien auf das Referenzmodell CIDOC-CRM (Le Boeuf et al. 2017) beziehen.

4. **Mit welchen Normdatenrepositorien und anderen Editionsprojekten möchte ich meine Edition vernetzen?** Durch das *Semantic Web* und Projekte wie *correspSearch* (Dumont 2016) wird die Möglichkeit eröffnet, die eigene Edition mit externen Daten, etwa aus der *Gemeinsamen Normdatei* (Deutsche Nationalbibliothek) anzureichern.

5. **Welche technologische Basis soll meine digitale Edition haben?** Neben den erwähnten verfügbaren Infrastrukturen von Editionsprojekten können auch Eigenentwicklungen eingesetzt werden. Von Vorteil ist die Erstellung der Graphedition mit *Semantic-Web*-Technologien.

6. **Wie soll meine Edition visualisiert werden?** Die Visualisierung des Textes ist elementarer Bestandteil einer textgenetischen Edition, deren Komplexität einen eigenen Beitrag rechtfertigen würde, zusammenfassend dazu etwa Bleeker und Kelly (2018).

Zum gegenwärtigen Zeitpunkt gibt es noch keine einheitliche Vorgehensweise für die Umsetzung einer Edition als Variantengraph. Das in den letzten Jahren gestiegene Interesse an Graphen und der Vernetzung über das *Semantic Web* könnte gleichwohl auch neue Entwicklungen im Bereich der Editionswissenschaft nach sich ziehen. Aber auch eine Verbindung von TEI-XML mit RDF-Graphdaten kann eine lohnenswerte Vorgehensweise für das eigene Editionsprojekt sein. Auf der einen Seite profitiert das Projekt so von der erprobten Textkodierung mit TEI-XML, auf der anderen Seite werden die eigenen Daten über das *Semantic Web* erschlossen.

Einige wenige Projekte bieten bereits die Möglichkeit, eine eigene graphbasierte genetische Edition umzusetzen. Frei nutzbar ist die Infrastruktur des Projektes *Stemmaweb* (*The Stemmaweb Project* 2012–). Eingespeiste Transkriptionen in unterschiedlichen Formaten können automatisch kollationiert und deren Stemmata sowie der daraus entstandene Variantengraph manuell nachbearbeitet werden. Ebenfalls zur freien Nutzung freigegeben ist *CollateX*, das es ermöglicht, mehrere Textversionen zu vergleichen und einen Variantengraphen zu generieren (Haentjens Dekker und Middell 2010–).

Es kann festgehalten werden, dass es – zum Glück! – nicht die eine Lösung für die Umsetzung einer Edition als Graph oder unter Einbeziehung von Graphdaten gibt. Dieser Beitrag und Leitfaden soll demnach nicht abschließend sein, sondern die vielfältigen Möglichkeiten mit und jenseits TEI-XML für Editionsprojekte aufzeigen und ermutigen, auch die eigenen Projekte von diesen Möglichkeiten profitieren zu lassen; Gründe dafür gibt es viele, exemplarisch genannt können etwa die im Vergleich zu TEI-XML bessere Abbildung von Textstufen in Variantengraphen oder auch die Anbindung an das *Semantic Web* sein.

Literatur

Andrews, T. L. und C. Mace. 2013. „Beyond the tree of texts: Building an empirical model of scribal variation through graph analysis of texts and stemmata.“ *International Journal of Human-Computer Studies* 28 (4): 504–21. doi:10.1093/llc/fqt032.

Banane, Mouad und Abdessamad Belangour. 2019. „A Comparative Study of RDF Triple Stores.“ *SSRN Journal*. doi:10.2139/ssrn.3349399.

Berners-Lee, Tim. 2006. „Linked Data.“ Zugriff: 4. September 2017. https://www.w3.org/DesignIssues/LinkedData.html.

Bleeker, Elli. 2017. „Mapping invention in writing: Digital infrastructure and the role of the genetic editor.“ Dissertation, Universität Antwerpen. hdl:10067/1556760151162165141.

Bleeker, Elli, Bram Buitendijk und Ronald Haentjens Dekker. 2019. „From graveyard to graph." *International Journal of Digit Humanities* 1: 141–163. doi:10.1007/s42803-019-00012-w.

Bleeker, Elli und Aodhán Kelly. 2018. „Interfacing Literary Genesis." In *Digital Scholarly Editions as Interfaces*, hg. v. Roman Bleier, Martina Bürgermeister, Helmut W. Klug, Frederike Neuber und Gerlinde Schneider, 193–218. Norderstedt: Books on Demand.

Bohnenkamp, Anne, Silke Henke und Fotis Jannidis. 2016. *Johann Wolfgang Goethe: Faust: Historisch-kritische Edition.* 2. Beta-Version. Zugriff: 26. April 2017. http://beta.faustedition.net/.

Bosse, Anke, Artur Boelderl und Walter Fanta. 2016–. *Musil online.* Zugriff: 30. September 2019. http://musilonline.at/.

Bruder, Daniel und Simone Teufel. 2018. „Data models for Digital Editions: Complex XML versus Graph structures." In *Kritik der digitalen Vernunft. Abstracts zur Jahrestagung des Verbandes Digital Humanities im deutschsprachigen Raum, 26.02.–02.03.2018 an der Universität zu Köln, veranstaltet vom Cologne Center for eHumanities (CCeH)*, hg. v. Georg Vogeler, 158–62. Köln: Universität zu Köln.

Burch, Thomas, Stefan Büdenbender, Kristina Fink, Vivien Friedrich, Patrick Heck, Wolfgang Lukas, Kathrin Nühlen et al. 2016. „Text[ge]schichten: Herausforderungen textgenetischen Edierens bei Arthur Schnitzler." In *Textgenese und digitales Edieren: Wolfgang Koeppens „Jugend" im Kontext der Editionsphilologie*, hg. v. Katharina Krüger, 87–105. Berlin, Boston: de Gruyter.

Burghardt, Manuel und Christian Wolff. 2009. „Stand off-Annotation für Textdokumente: Vom Konzept zur Implementierung (zur Standardisierung?)." In *Von der Form zur Bedeutung: Texte automatisch verarbeiten*, hg. v. Christian Chiarcos, Richard Eckart de Castilho und Manfred Stede, 53–59. Tübingen: Narr.

Burnard, Lou, Fotis Jannidis, Elena Pierazzo und Malte Rehbein. 2010. „An Encoding Model for Genetic Editions." Zugriff: 25. März 2019. http://www.tei-c.org/Activities/Council/Working/tcw19.html.

Buzzetti, Dino. 2002. „Digital Representation and the Text Model." *New Literary History* 33 (1): 61–88. https://www.jstor.org/stable/20057710.

Chau, Kien Tsong, QianLing He, XuQing Hu und Rui Wu. 2019. „Comparison on Performance of Text-Based and Model-Based Architecture in Open Source Native XML Database." In *2019 IEEE 4th International Conference on Signal and Image Processing (ICSIP)*, 340–44: IEEE Xplore. doi:10.1109/SIPROCESS.2019.8868709.

Ciotti, Fabio. 2018. „A Formal Ontology for the Text Encoding Initiative." *Umanistica Digitale* 2 (3). doi:10.6092/issn.2532-8816/8174.

Ciotti, Fabio und Francesca Tomasi. 2016. „Formal Ontologies, Linked Data, and TEI Semantics." *Journal of the Text Encoding Initiative* 9. doi:10.4000/jtei.1480.

Cummings, James. 2018. „A world of difference: Myths and misconceptions about the TEI." *Digital Scholarship Humanities* 34 (Supplement 1): i58-i79. doi:10.1093/llc/fqy071.

D'Iorio, Paolo. 2017. „Die Schreib- und Gedankengänge des Wanderers. Eine digitale genetische Nietzsche-Edition." *Editio* 31 (1): 191–204. doi:10.1515/editio-2017-0011.

DeRose, Steven J. 2004. „Markup Overlap: A Review and a Horse." In *Extreme Markup Languages 2004 Proceedings.* Zugriff: 25. März 2019. http://xml.coverpages.org/DeRoseEML2004.pdf.

Deutschen Nationalbibliothek. „Gemeinsame Normdatei: (GND).“ Zugriff: 8. Juni 2022. https://www.dnb.de/DE/Professionell/Standardisierung/GND/gnd_node.html.

Di Iorio, Angelo, Silvio Peroni und Fabio Vitali. 2011. „A Semantic Web approach to everyday overlapping markup.“ *Journal of the American Society for Information Science and Technology* 62 (9): 1696–1716. doi:10.1002/asi.21591.

Dumont, Stefan. 2016. „correspSearch: Connecting Scholarly Editions of Letters.“ *Journal of the Text Encoding Initiative* 10. doi:10.4000/jtei.1742.

Efer, Thomas. 2016. „Graphdatenbanken für die textorientierten e-Humanities.“ Dissertation, Universität Leipzig. urn:https://nbn-resolving.org/nbn:de:bsz:15-qucosa-219122.

Gandon, Fabien und Guus Schreiber. 2014. „RDF 1.1 XML Syntax.“ Zugriff: 23. August 2019. http://www.w3.org/TR/2014/REC-rdf-syntax-grammar-20140225.

Gao, Shudi, C. M. Sperberg-McQueen und Henry S. Thompson. 2012. „W3C XML Schema Definition Language (XSD).“ Zugriff: 18. Februar 2020. http://www.w3.org/TR/xmlschema11-1.

Haentjens Dekker, Ronald und David J. Birnbaum. 2017. „It's more than just overlap: Text As Graph.“ In *Proceedings of Balisage: The Markup Conference 2017. Balisage Series on Markup Technologies* 19. doi:10.4242/BalisageVol19.Dekker01.

Haentjens Dekker, Roland, Elli Bleeker, Bram Buitendijk, Astrid Kulsdom und David J. Birnbaum. 2018. „TAGML: A markup language of many dimensions.“ In *Proceedings of Balisage: The Markup Conference 2018. Balisage Series on Markup Technologies* 21. doi:10.4242/BalisageVol21.HaentjensDekker01.

Haentjens Dekker, Ronald und Gregor Middell. 2010–. „CollateX: Software for Collating Textual Sources.“ Documentation. Zugriff: 23. September 2019. https://collatex.net/doc/.

Haentjens Dekker, Ronald, Dirk van Hulle, Gregor Middell, Vincent Neyt und Joris van Zundert. 2015. „Computer-supported collation of modern manuscripts: CollateX and the Beckett Digital Manuscript Project.“ *Digital Scholarship Humanities* 30 (3): 452–70. doi:10.1093/ll c/fqu007.

Haffner, Alexander. 2012–2019. „GND Ontology.“ Zugriff: 30. September 2018. http://d-nb.info/standards/elementset/gnd.

Hinkelmanns, Peter. 2019. „Text Graph Ontology.“ Zugriff: 5. Januar 2020. https://github.com/Wolkenstein/tgo-ontology.

Huitfeldt, Claus. 1994. „Multi-dimensional texts in a one-dimensional medium.“ *Computers and the Humanities* 28 (4-5): 235–41. doi:10.1007/BF01830270.

Ide, Nancy und Laurent Romary. 2006. „Representing Linguistic Corpora and Their Annotations.“ In *Proceedings of the Fifth Language Resources and Evaluation Conference (LREC'2006)*, hg. v. Nicoletta Calzolari, et al., 225–28. Genua: European Language Resources Association. Zugriff: 18. Februar 2020. http://www.lrec-conf.org/proceedings/lrec2006/pdf/562_pdf.pdf.

Ide, Nancy und Keith Suderman. 2007. „GrAF: A Graph-based Format for Linguistic Annotations.“ In *Proceedings of the Linguistic Annotation Workshop*, hg. v. Branimir Boguraev et al., 1–8. Prague: Association for Computational Linguistics. Zugriff: 18. Februar 2020. https://aclanthology.org/W07-1501.pdf.

Isaac, Antoine. 2013. „Europeana Data Model Primer.“ Zugriff: 18. Februar 2020. https://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation/EDM_Primer_130714.pdf.

Jänicke, Stefan, , Annette Geßner, Greta Franzini, Melissa Terras, Simon Mahony und Gerik Scheuermann. 2015. „TRAViz: A Visualization for Variant Graphs." *Digital Scholarship in the Humanities* 2013 (2): i83–i99. doi:10.1093/llc/fqv049.

Jänicke, Stefan und David J. Wrisley. 2018. „Interactive Visual Alignment of Medieval Text Versions." In *2017 IEEE Conference on Visual Analytics Science and Technology (VAST)*, IEEE Xplore. doi:10.1109/VAST.2017.8585505.

Knublauch, Holger und Dimitris Kontokostas. 2017. „Shapes Constraint Language (SHACL)." Zugriff: 18. Februar 2020. https://www.w3.org/TR/shacl/.

Kuczera, Andreas. 2016. „Digital Editions beyond XML: Graph-based Digital Editions." In *Proceedings of the 3rd HistoInformatics Workshop, Krakow, Poland, 11 July 2016*, hg. v. Marten Düring, Adam Jatowt, Johannes Preiser-Kappeller and Antal van Den Bosch, 37–46. http://ceur-ws.org/Vol-1632/paper_5.pdf.

—, Aline Deicke, Julian Jarosch. 2017–. „Graphentechnologien: Graphentechnologien in den digitalen Geistes- und Sozialwissenschaften." Zugriff: 30. September 2019. https://graphentechnologien.hypotheses.org.

—. 2017. „Graphentechnologien in den Digitalen Geisteswissenschaften." *ABI Technik* 37, Nr. 3: 179–96. doi:10.1515/abitech-2017-0042.

Le Boeuf, Patrick, Martin Doerr, Christian Emil Ore und Stephen Stead. 2017. „Definition of the CIDOC Conceptual Reference Model: Version 6.2.2." Zugriff: 1. Februar 2022. https://www.cidoc-crm.org/sites/default/files/2017-09-30%23CIDOC%20CRM_v6.2.2_esIP.pdf.

Miles, Alistair und Sean Bechhofer. 2009. „SKOS Simple Knowledge Organization System: Reference." Zugriff: 18. Februar 2020. http://www.w3.org/TR/skos-reference.

Pollin, Christopher und Georg Vogeler. 2019. „Datenmodell für historische Rechnungsunterlagen: Version 1.1." Zugriff: 30. September 2021. https://gams.uni-graz.at/context:depcha.

Pollin, Christopher und Lina M. Zangerl. 2020. „Der Nachlass als Netzwerk: Zur Entwicklung einer Nachlass-Ontologie am Beispiel des Projekts ‚Stefan Zweig digital'." In *Digital Humanities Austria 2018: Empowering researchers*, hg. v. Marlene Ernst, Peter Hinkelmanns und Lina M. Zangerl, 123–27. doi:10.1553/dha-proceedings2018s123.

Robie, Jonathan, Michael Dyck und Josh Spiegel. 2017a. „XML Path Language (XPath) 3.1." Zugriff: 18. Februar 2020. https://www.w3.org/TR/xpath-31/.

—. 2017b. „XQuery 3.1: An XML Query Language." Zugriff: 18. Februar 2020. https://www.w3.org/TR/xquery-31/.

Sahle, Patrick. 2013. *Digitale Editionsformen 3: Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels: Textbegriffe und Recodierung*. Norderstedt: BoD.

Saxonica, Michael K. 2007. „XSL Transformations (XSLT) Version 2.0." Zugriff: 18. Februar 2020. http://www.w3.org/TR/xslt20/.

Schmidt, Desmond. 2012. „The Role of Markup in the Digital Humanities." *Historical Social Research* 37 (3): 125–46. http://www.jstor.org/stable/41636601.

Schmidt, Desmond und Robert Colomb. 2009. „A data structure for representing multi-version texts online." *International Journal of Human-Computer Studies* 67 (6): 497–514. doi:10.1016/j.ijhcs.2009.02.001.

Šimek, Jakub. 2014. Welscher Gast Digital. TEI-Handbuch. Version 0.6. Zugriff: 5. Januar 2021. http://digi.ub.uni-heidelberg.de/wgd/pdf/TEI-Handbuch_0-6.pdf.

Sperberg-McQueen, C. M. und Claus Huitfeldt. 2004. „GODDAG: A Data Structure for Overlapping Hierarchies." In *Digital Documents: Systems and Principles*, hg. v. Peter King und Ethan V. Munson, 139–60. Berlin, Heidelberg: Springer. doi:10.1007/978-3-540-39916-2_12.

Stein, Christian. 2014. „Linked Open Data: Wie das Web zur Semantik kam." *Bibliothek. Forschung und Praxis* 38 (3): 447–55.

Stührenberg, Maik. 2012. „The TEI and Current Standards for Structuring Linguistic Data." *Journal of the Text Encoding Initiative* 3. Zugriff: 17. Juni 2016. http://jtei.revues.org/523.

Text Encoding Initiative Consortium. 2019. „P5: Guidelines for Electronic Text Encoding and Interchange: Non-hierarchical Structures." Zugriff: 5. Februar 2019. http://www.tei-c.org/release/doc/tei-p5-doc/en/html/NH.html.

The Stemmaweb Project: Tools and techniques for empirical stemmatology. 2012–. Zugriff: 30. September 2019. https://stemmaweb.net.

Turska, Magdalena und E. Spandini. 2018. „XML/TEI Stand-off Markup: One step beyond." *Digital Philology: A Journal of Medieval Cultures* 8 (2): 225–39. doi:10.1353/dph.2019.0025.

Universität Salzburg. 1992–. „Mittelhochdeutsche Begriffsdatenbank." Zugriff: 18. Februar 2020. http://www.mhdbdb.sbg.ac.at/.

van Hulle, Dirk und Mark Nixon. 2011–. „Samuel Beckett: Digital Manuscript Project." Zugriff: 30. September 2019. https://www.beckettarchive.org.

Vogeler, Georg. 2019. „The ‚assertive edition'." *International Journal of Digital Humanities* 1: 309–22. doi:10.1007/s42803-019-00025-5.

Vetter, Lara, Jarom McDonald, Sean Daugherty, Tanya Clement, Susan Schreibman, Roman Bleier und Joshua D. Savage. 2016-01-21. Dokumentation: Versioning Machine 5.0. Zugriff: 05. Januar 2021. http://v-machine.org/documentation.

W3C OWL Working Group. 2012. „OWL 2 Web Ontology Language: Document Overview (Second Edition)." W3C Recommendation 11 December 2012. Zugriff: 8. Juni 2018. https://www.w3.org/TR/owl2-overview.

W3C SPARQL Working Group. 2013. „SPARQL 1.1 Overview: W3C Recommendation." Zugriff: 05. Januar 2021. http://www.w3.org/TR/2013/REC-sparql11-overview-20130321/.

Wout Dillen. 2016. „Sequentiality in Genetic Digital Scholarly Editions: Models for Encoding the Dynamics of the Writing Process." In *Digital Humanities 2016: Conference Abstracts*, hg. v. Maciej Eder und Jan Rybicki, 174–75. doi:10.17613/M6GB9B.