NACHHALTIGKEIT UND LANGZEITVERFÜGBARKEIT VON DIGITALEN EDITIONEN IM SEMANTIC WEB

Jörg Wettlaufer, Göttingen

What makes a cool URI?
A cool URI is one which does not change.
What sorts of URI change?
URIs don't change: people change them.[1]
(c)1998 Tim Berners–Lee

Einführung

Das Thema der Nachhaltigkeit und langfristige Verfügbarkeit von digitalen Ressourcen im Internet ist ein Dauerbrenner in der kritischen Auseinandersetzung mit der Digitalen Transformation, in der sich Gesellschaft und Wissenschaft momentan befinden. Digitale Editionen sind in besonderer Weise mit dieser Problematik konfrontiert, da sie idealiter über Jahrzehnte oder sogar Jahrhunderte zuverlässig der Forschung zur Verfügung stehen sollen. Während die Standardisierung im Bereich der textbasierten Primärdaten durch Extensible Markup Language (XML) und den darauf aufbauenden Standard der Text Encoding Initiative (TEI)[2] eine gewisse Zukunftssicherheit hinsichtlich Lesbarkeit der Daten ermöglicht, gilt dies nicht in demselben Maße für das Layout und die Präsentation von Editionen. Doch im Semantik Web potenziert sich diese Problematik noch. Während eine lokale Instanz einer Edition vor allem mit Problemen der Nachhaltigkeit von (sofern verwendet) Content-Management-Systemen und der Softwarearchitektur der Präsentationsschicht überhaupt zu kämpfen hat, stellt sich für Digitale Editionen im Semantic Web die Verlinkung der maschinenlesbaren Ressourcen untereinander, die geradezu konstituierend für das Konzept des Semantic Web sind, als eigentliches Nachhaltigkeitsproblem dar, sofern das System nicht auf rein lokale Ressourcen beschränkt ist. Redundanz, ein Grundprinzip der Netzwerkkommunikation des Internet, existiert praktisch nicht auf der Ebene von Linked Open Data (LOD)[3]. Sie widerspricht vielmehr den grundlegenden Prinzipien des Semantic Web. Fällt eine Ressource aufgrund von Hard-

1 https://www.w3.org/Provider/Style/URI (Zugriff 2.8.2019). Ich danke Patrick Sahle für Hinweise zum Thema und Klaus Meyer-Wegener und Florian Kragl für Kommentare zum Manuskript dieses Artikels.
2 https://tei-c.org/ (Zugriff 2.8.2019).
3 http://linkeddata.org/ (Zugriff 2.8.2019).

ware- oder Softwareproblemen aus, kann also nicht über die maschinenlesbare Schnittstelle (i.d.R. SPARQL[4]) abgerufen werden, steht sie zumeist auch in der Präsentationsschicht der Digitalen Edition nicht zur Verfügung, die diese Ressource konsumiert. Ein probates Gegenmittel ist daher seit längerem die Verwendung von sog. »Datendumps«, die lokale Kopien aller notwendigen Ressourcen vorhalten. Der überaus große Nachteil dieser Lösung allerdings ist, dass sie, wie oben schon angedeutet, den Prinzipien der dezentralen Datenhaltung des Semantic Web, wie sie ursprüngliche von Tim Berners Lee definiert wurden,[5] diametral entgegenstehen und zudem Probleme, die doppelte Datenhaltung mit sich bringt, hervorruft, d.h. Veränderungen in den Ursprungsdaten können nur durch einen erneuten »Datendump« übernommen werden. An dieser Stelle sehen wir das Grundproblem der Nachhaltigkeit digitaler Ressourcen überhaupt am Werk – je stärker der Datenaustausch automatisiert wird, desto anfälliger wird er hinsichtlich fehlender Standardkonformität bzw. Weiterentwicklungen von Standards und Routinen. Ein Teufelskreis, aus dem es bislang keinen wirklich überzeugenden Weg gibt. Denn beide Aspekte, Automatisierung und Standardentwicklung, sind an sich wünschenswert und Konstituenten der Digitalen Transformation überhaupt.

Das Problem von Nachhaltigkeit und Langzeitverfügbarkeit begleitet digitale Publikationen daher seit den ersten Versuchen in den 70er und 80er Jahren und hat sich seit der Etablierung des Internets in den 90er Jahren des letzten Jahrhunderts verschärft. Neben den schnellen Entwicklungszyklen in der angewandten Informatik ist dies vor allem auch dem Netzwerkprinzip selber geschuldet. Auf der Ebene von lokalen Datenzentren wird seit einigen Jahren an Konzepten gearbeitet, die diese Probleme adressieren. In der Schweiz wurde in den letzten Jahren eine »Nationale Infrastruktur für Editionen« (NIE-INE)[6] etabliert, die auf die Homogenisierung der technologischen Basis von Editionsprojekten setzt. In Österreich wurde mit dem «Kompetenznetzwerk Digitale Editionen» (KONDE)[7] ein umfangreiches Verbundprojekt geschaffen, in dem Digitale Editionen gesichert und bereitgestellt werden sollen. Auch in Deutschland gibt es vergleichbare, wenn auch stärker föderal ausgerichtete Initiativen, die auf die Schaffung einer gemeinsamen Infrastruktur für Editionsprojekte und damit längerfristig auf eine technologische Standardisierung von Digitalen Editionen zielen.[8] Diese Entwicklung ist natürlich auch

4 https://www.w3.org/TR/rdf-sparql-query/ (Zugriff 2.8.2019).
5 Tim Berners Lee, James Hendler & Ora Lassila: The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities, Scientific American: Feature Article: May 2001.
6 https://www.nie-ine.ch/ (Zugriff 2.8.2019).
7 http://www.digitale-edition.at/ (Zugriff 2.8.2019).
8 Hier ist z.B. die virtuelle Forschungsumgebung »textgrid« (https://textgrid.de/ Zugriff 2.8.2019) zu nennen, die heute ein Projekt in DARIAH.DE ist, das ebenfalls über ein Repositorium versucht, Forschungsdaten langfristig zu bewahren. Siehe auch Mirjam Blümm, Stefan Schmunk, Peter Gietz, Wolfram Horstmann &

vor dem Hintergrund der laufenden Debatte über die Etablierung einer „Nationalen Forschungsdateninfrastruktur" (NFDI)[9] zu einzuordnen, deren Konsortien sich aktuell formieren.

Im Folgenden soll zunächst der Stand der Diskussion zum Thema der Nachhaltigkeit Digitaler Editionen referiert werden. In einem nächsten Schritt werden die Grundlagen und Prinzipien des Semantic Web vorgestellt und die Stabilität von URIs thematisiert. Anschließend sollen Beispiele für Digitale Editionen im Semantik Web vorgestellt und vorhandene Lösungsansätze und Aktivitäten, die der nachhaltigen Bereitstellung von Digitalen Editionen im Semantic Web dienen können, diskutiert werden.

Nachhaltigkeit Digitaler Editionen

Die Nachhaltigkeit Digitaler Editionen ist seit längerem ein Diskussionsgegenstand im Diskurs der digitalen EditorInnen.[10] Zuletzt wurde auf mehrere Tagungen das Thema behandelt. Ergebnisse aus zwei Veranstaltungen möchte ich beispielhaft hier darstellen, um die aktuelle Diskussion um die Nachhaltigkeit digitaler Editionen näher zu beleuchten.

Das Thema Nachhaltigkeit war namengebend für die DHd Tagung in Bern 2017[11]. Anna Busch hat in einem Blogbeitrag die bezüglich Digitaler Editionen relevanten Sektionen zusammengestellt und besprochen.[12] Sie resümiert: »Alle genannten Beiträge legen nahe, dass Digitale Editionen schon zu Projektbeginn Nachhaltigkeitsstrategien entwickeln müssen. Das geht einher mit einem Dringen auf Transparenz in der Vorgehensweise, den Strukturen und der Datenhaltung vor allem durch eine detaillierte und zugängliche Dokumentation. Sichergestellt werden soll zudem die Wieder- und Weiterverwendbarkeit der Daten mit einem Minimum an technischen und rechtlichen Einschränkungen (Open-Source, Open-Data, Lowtech-Lösungen, etablierte Standards). Speziell für Digitale Editionen schließt das die Verwendung von XML/TEI, konsistenten projekteigenen Transkriptionsrichtlinien, stabilen ULRs/Links und die Bereitstellung von Schnittstellen Schnittstellenbereitstellungen mit ein. Nicht zuletzt ist die Verbindung einer passgenauen technischen Lösung mit einer institutionellen Anbindung (Bibliothek, Archiv, Univer-

Heiko Hütter: Vom Projekt zum Betrieb: Die Organisation einer nachhaltigen Infrastruktur für die Geisteswissenschaften DARIAH-DE, in: ABI Technik 2016; 36(1): 10–23. Außerdem sind das Projekt »SustainLife« des DCH Köln (https://dch.phil-fak.uni-koeln.de/sustainlife.html Zugriff 2.8.2019), das Humanities Data Centre in Göttingen (http://humanities-data-centre.de/ Zugriff 2.8.2019) sowie andere regionale Datenzentren und Initiativen in diesem Zusammenhang relevant.

9 http://www.rfii.de/de/nationale-forschungsdateninfrastruktur-nfdi/ (Zugriff 2.8.2019).

10 Vgl. Elena Pierazzo: Digital Scholarly Editing. Theories, Models and Methods, Ashgate 2015, (chaper 8) http://www.jstor.org/stable/j.ctt1fzhh6v.9 (Zugriff 2.8.2019).

11 http://www.dhd2017.ch/ (Zugriff 2.8.2019).

12 https://dhd-blog.org/?p=7772 (Zugriff 27.2.2017).

sität) der Schlüssel zur Sicherstellung der langfristigen Aufbewahrung und der Erhaltung der dauerhaften Verfügbarkeit einer digitalen Ressource.« Dieses Fazit zeigt deutlich die Bereiche auf, in denen digitale Editionsprojekte achtsam sein müssen, um für eine lange Zeit verfügbar zu sein. Insbesondere die institutionelle Anbindung erscheint momentan als die einzig gangbare Lösung für die dauerhafte Bereitstellung von Editionsprojekten mit eigener Präsentationsschicht. Aber auch auf anderen Ebenen wird nach nachhaltigen Lösungen gesucht. Stichworte sind hier Automatisierte Kuratierung, Versionierung und Infrastruktur als Code. Diese Bemühungen sind Teil einer stärker reflektierten Entwicklungspraxis von Software im Forschungskontext. Research Software Engineering formiert sich innerhalb der Informatik und auch der Digital Humanities zu einem Feld, dem in letzter Zeit immer mehr Aufmerksamkeit gewidmet wird.[13] Konkrete Projekte haben diese Herausforderungen aufgegriffen und versuchen Antworten zu geben.

Das von der DFG geförderte Kooperationsprojekt „SustainLife – Erhalt lebender, digitaler Systeme für die Geisteswissenschaften", dass in einer Zusammenarbeit zwischen dem Data Center for the Humanities der Universität zu Köln (DCH, siehe http://dch.phil-fak.uni-koeln.de) und dem Institut für Architektur von Anwendungssystemen der Universität Stuttgart (IAAS, siehe http://www.iaas.uni-stuttgart.de) durchgeführt wird, arbeitet z.B. an Lösungsvorschlägen für diese Probleme. Gegenstand des Projekts ist die Adaption und Weiterentwicklung von Verfahren und Technologien aus dem Cloud-Deployment für die Digital Humanities (DH) mit dem Ziel, Management und Provisionierung von DH-Anwendungen zu optimieren und deren Sicherung und nachhaltigen Betrieb zu realisieren.[14]

In dem Workshop »Nachhaltigkeit Digitaler Editionen« am 17.09.2018 an der Nordrhein-Westfälischen Akademie der Wissenschaften und der Künste wurde das Thema zuletzt ebenfalls umfassend diskutiert und auch das Projekt SustainLife vorgestellt. Einige der dort gehaltenen Vorträge erlaubten auch einen Blick auf die Realität der Langzeitverfügbarkeit digitaler Editionen. Im Folgenden möchte ich einige der dort aufgestellten Thesen aufgreifen und anschließend in den Kontext der Überlegungen zur Nachhaltigkeit Digitaler Editionen im Semantik Web stellen. [15]

13 Vgl. den Workshop in Kassel zu diesem Thema auf der INFORMATIK 2019 Tagung: https://fg-infdh.gi.de/infdh-workshop-2019/ (Zugriff 2.8.2019).

14 Claes Neuefeind, Philip Schildkamp, Brigitte Mathiak: Technologienutzung im Kontext Digitaler Editionen – eine Landschaftsvermessung, in: DHd Abstractbook 2019, S. 219-222. Siehe dazu auch J. Barzen, J. Blumtritt, U. Breitenbücher, S. Kronenwett, F. Leymann, B. Mathiak, C. Neuefeind: „SustainLife - Erhalt lebender, digitaler Systeme für die Geisteswissenschaften." In: Book of Abstracts der 5. Jahrestagung der Digital Humanities im deutschsprachigen Raum (DHd 2018), Köln 26.2.– 2.3.2018, S. 471–474. https://kups.ub.uni-koeln.de/8085/1/boa-DHd2018-.pdf sowie C. Neuefeind, L., Harzenetter, P., Schildkamp, U., Breitenbücher, B., Mathiak, J., Barzen, and F. Leymann, F.: „The SustainLife Project – Living Systems in Digital Humanities". In: Proceedings of the 12th Advanced Summer School on Service-Oriented Computing, 2018 (IBM Research Report RC25681), S. 101–112.

15 Die folgenden Informationen sind entnommen aus: Peter Dängli: Die Nachhaltigkeitsproblematik digitaler

Patrick Sahle zählte in seiner Einführung zur Tagung eine Reihe von Beispielen auf, die nachdenklich stimmen. Nach seiner Beobachtung kämpfen bereits viele ältere Digitale Editionen mit einer eingeschränkten Verfügbarkeit oder sind nicht länger zugreifbar. Als Beispiele nannte er das Thomas Raddall Electronic Archive Project (2001-2004)[16] oder die im Jahr 2000 als CD-ROM veröffentlichte Stjin Streuvels-Edition[17], die nur noch eingeschränkt und unter Verlust des ursprünglichen User Interfaces zugänglich sind. Das Alcalá Account Book Project sei bis auf einen zweiteiligen Artikel sowie einige Metadaten gänzlich verschwunden. Überreste der einstigen Online Edition sind nur noch über das Internet Archive (wayback machine) sichtbar, das natürlich keine funktionale Benutzung erlaubt.[18]

Wer Online-Projekte über Jahre und Jahrzehnte betreut weiß, welche Herausforderungen dies mit sich bringt. Die zunehmende personelle Mobilität, fehlende einheitliche Standards (TEI bringt hier aus verschiedenen Gründen nicht wirklich einen Fortschritt[19]) und der Projektcharakter der meisten Vorhaben tun ihr Übriges und können häufig auch erfolgreiche Projekte nicht vor dem digitalen Vergessen bewahren.

Hinsichtlich Aussehen, Funktionalitäten und technischer Architekturen kann eine große Heterogenität der Digitalen Editionen festgestellt werden, die eine nachhaltige Bereitstellung einer benutzerfreundlichen Oberfläche erschwert. In einem Vortrag auf derselben Tagung fasste Thomas Stäcker die Herausforderungen in folgenden Kernthesen zusammen: (1) Nachhaltigkeit von Editionen kann durch Bewahrung von zweidimensionalen Repräsentationen, die Resultat von Prozessschritten sind, nicht erreicht werden. Eine digitale Edition besteht vielmehr in der Summe der Verarbeitungsmöglichkeiten, die in ihrem Modell vollständig beschrieben werden können. (2) Jeder Versuch, eine konkrete technische Realisierung einer Edition zu bewahren, ist eo ipso zum Scheitern verurteilt. (3) Die vollständige Beschreibung aller Komponenten der Edition in maschinenlesbarer und standardisierter Form ist wichtige Voraussetzung für ihre Nachhaltigkeit.[20] Johannes Stigler stellte in diesem Zusammenhang ebenfalls fünf Thesen zum Thema Nachhaltig-

Editionen – Workshopbericht (2019) https://dhd-blog.org/?p=11033 (Zugriff 2.8.2019). Siehe dort auch für weitere Verweise.

16 Die URL des Projekts ist seit längerem nicht mehr erreichbar: http://www2.library.dal.ca/archives/trela/trela.htm (Zugriff 2.8.2019).

17 Edward Vanhoutte: A Linkemic Approach to Textual Variation: Theory and Practice of the Electronic-Critical Edition of Stijn Streuvels' De teleurgang van den Waterhoek, HUMAN IT, Vol. 4 (2000), No. 1, https://humanit.hb.se/article/download/197/235.

18 Siehe https://web.archive.org/web/20141101045836/http://archives.forasfeasa.ie/ (Zugriff 2.8.2019).

19 Aufgrund der vielfältigen und teilweise sich überschneidenden Möglichkeiten der Auszeichnung von Personen, Orten und anderen Informationen sind mit TEI Markup versehene Texte leider nicht ohne Berücksichtigung des zugrundeliegenden Schemas darstellbar. Dies hat in Deutschland zum Quasi-Standard des DTA Basisformats geführt. Vgl. http://www.deutschestextarchiv.de/doku/basisformat/ (Zugriff 23.8.2019).

20 Thomas Stäcker: Vortragsfolien: XML oder nicht XML – das ist hier die Frage, Düsseldorf 2018, Folie 22. https://web.archive.org/web/20181230112643/http://dch.phil-fak.uni-koeln.de/sites/dch/NDE-Workshop/Staecker.pdf (Zugriff 2.8.2019).

keit auf und lenkt damit den Blick auf weitere Aspekte. Seine erste These lautet: „Der Status Quo zum Thema Nachhaltigkeit resultiert aus den Prämissen der Forschungsförderung". 2. „Digitale Langzeitarchive sind Publikationsinstanzen für Digitale Editionen". 3. & 4. & 5. (zusammengefasst): Kuratierbarkeit, Objektorientierung und Modellierung sind notwendige Strukturmerkmale nachhaltiger Repräsentationsform Digitaler Editionen.[21]

Diesen Thesen ist zuzustimmen. Allein, sie beschreiben nur das Problem und zeigen noch nicht den Weg auf, mit der Herausforderung umzugehen. Auch wenn der Analyse von Samuel Müller »Langzeitsicherung ist immer eine Frage von Institutionen und nicht von Technologien.«[22] uneingeschränkt zuzustimmen ist, bleibt doch weiter die Frage nach der technologischen Umsetzung, die es Institutionen am Ende erlaubt, Digitale Editionen möglichst geschmeidig und harmonisch in die Sammlungen digitaler Publikationen aufzunehmen und langfristig zu pflegen bzw. vorzuhalten. Durch die Explizierung von Semantiken und die Verlinkung von Daten aus Editionen im Semantic Web wird diese Aufgabe nicht einfacher. Bevor wir uns also den Digitalen Editionen im Semantic Web (SW) konkret zuwenden, möchte ich kurz die oben schon mehrfach erwähnten Grundprinzipien des SW nach Tim Berners-Lee in Erinnerung rufen, da sich hieraus die eigentlichen Herausforderungen ergeben. Die folgenden Ausführungen stützen sich dabei auf die Darstellung des Konzepts in der deutschen Wikipedia[23].

Grundlagen und Prinzipien des Semantic Web

Das Semantic Web erweitert das World Wide Web (WWW), um Daten für Rechner einfacher austauschbar und verwertbar zu machen. Aufgrund der semantischen Disambiguierung von Worten in natürlicher Sprache können Mehrdeutigkeiten, die sich für Menschen normalerweise nur aus dem Verwendungskontext erschließen, auch für Maschinen eindeutig aufgelöst werden. Beispielsweise kann so für den Begriff (die Zeichenkette) »Bank« in einem Webdokument eindeutig entschieden werden, ob ein Sitzmöbel oder ein Geldinstitut gemeint ist. Zur maschinenlesbaren Kodierung dient vor allem der

21 Johannes Stigler: Fünf Thesen zum Thema Nachhaltigkeit: Die Sicherstellung der Verfügbarkeit von (Text-) Daten als Aufgabe von Langzeitarchivierung. Erfahrungsbericht aus einem nationalen Forschungsdateninfrastrukturprojekt. http://dch.phil-fak.uni-koeln.de/sites/dch/NDE-Workshop/Stigler.pdf .

22 Samuel Müller: Vortragsmanuskript: „Die Nationale Infrastruktur für Editionen (NIE-INE): Aufgaben und Lösungswege zur langfristigen Präsentation digitaler Editionen", 2018, S. 2. https://web.archive.org/web/20181230112712/http://dch.phil-fak.uni-koeln.de/sites/dch/NDE-Workshop/Mueller.pdf (Zugriff 2.8.2019).

23 https://de.wikipedia.org/wiki/Semantic_Web (Zugriff 22.6.2019). Siehe auch den Beitrag von Klaus Meyer-Wegener in diesem Band.

RDF-Standard (Resource Description Framework)[24], der in einfachen Tripeln von Subjekt, Prädikat, Objekt Informationen expliziert.

Aus den in RDF hinterlegten Informationen kann durch Verknüpfung ein sehr großer (Wissens-)Graph entstehen, der potentiell alle Dinge von Interesse identifiziert und – mit einer eindeutigen Adresse versehen – als Knoten anlegt, die wiederum durch Kanten (ebenfalls jeweils eindeutig benannt) miteinander verbunden sind. Einzelne Dokumente im WWW beschreiben dann eine Reihe von Tripeln, und die Gesamtheit all dieser Tripel entspricht dem globalen Graphen (von Tim Berners-Lee auch Giant Global Graph[25] genannt). Zur Realisierung des Semantic Web dient neben RDF auch das Konzept von URIs (Unified Resource Identifiers) in der doppelten Rolle zur Identifizierung von Entitäten und zum Verweisen auf weitergehende, semantisch verknüpfte Ressourcen. Grundsätzlich gibt es URLs (Uniform Resource Locators), URNs (Uniform Resource Names) und URIs (Uniform Resource Identifiers). Inzwischen haben sich auch noch IRIs (International Resource Identifiers) hinzugesellt, die URIs mit internationalen Zeichensätzen ermöglichen.

Jeder URN und jede URL ist ein URI. Aber nicht jeder URI ist ein URN oder ein URL. Ein URN hat ein bestimmtes Schema (der vordere Teil eines URI vor dem Doppelpunkt), enthält jedoch keine Anweisungen zum Zugriff auf die identifizierte Ressource. Wir Menschen ordnen dies möglicherweise automatisch einer Zugriffsmethode in unserem Kopf zu (z.B. URNs mit digitaler Objektkennung wie doi: 10.1093/llc/fqv047, für die wir DOIs verwenden, die sich auf https://doi.org/10.1093/llc/fqv047 abbilden lassen), aber die Anweisung ist nicht in der URN erhalten. Eine URL ist nicht nur ein Bezeichner, sondern auch eine Anweisung zum Auffinden und Zugreifen auf die identifizierte Ressource.[26] Im Zusammenhang mit der Nachhaltigkeit von Ressourcen im Semantic Web interessieren, wie schon gesagt, vor allem URIs und IRIs, denn sie sind die Ressourcen, aus denen RDF-Tripel gebildet werden.

In der deutschen Wikipedia gab es 2015 eine intensive Diskussion über das das Prinzip »Cool URIs don't change«. Dabei wurde die Meinung vertreten, dies sei in der Wikipedia nicht vollständig umzusetzen. Es sei vielmehr geradezu das Prinzip der Wikipedia, einen ständig sich verändernden Inhalt anzubieten. Auch wenn sich die URI nicht ändere, könne sich der Inhalt doch stark verändern. Dies lasse sich nur durch die Verlinkung auf konkrete Versionen vermeiden. Daher würden auch Weiterleitungen keine Lösung für das zur Diskussion stehende Problem darstellen.[27] Während eine solche „Flexibili-

24 https://www.w3.org/RDF/ (Zugriff 2.8.2019).

25 Tim Berners-Lee (2007-11-21). "Giant Global Graph". https://web.archive.org/web/20160713021037/http://dig.csail.mit.edu/breadcrumbs/node/215 (Zugriff 2.8.2019).

26 Weitere Informationen unter https://tools.ietf.org/html/rfc2392 (Zugriff 2.8.2019).

27 https://de.wikipedia.org/wiki/Wikipedia:Meinungsbilder/cool_URIs_don't_change (Zugriff 2.8.2019). Ebenfalls nicht umsetzbar sei das Prinzip dort, wo Artikel ihre Lemmata aus fachlichen Gründen ändern. War

tät" bei der Wissensorganisation der Benutzbarkeit der Wikipedia als menschlesbare und auch an als Menschen als Rezipienten adressierte gegenwartsbasierte Wissenssammlung kaum Abbruch tut, sieht es bei der maschinenlesbaren DBpedia[28] oder auch bei Wikidata[29] ganz anders aus. Sobald hier eine Entität oder ein Konzept wegfällt oder sich verändert, würden eine Reihe von Verlinkungen nicht mehr funktionieren. Daher werden Entitäten dort abstrakt bezeichnet (Q+Zahl) und ihre semantische Beschreibung ist nur eine Eigenschaft dieses Objekts.

Das Prinzip »Cool URIs don't change« gewinnt zudem an Bedeutung, je länger eine Ressource besteht und je mehr Verlinkungen von außen darauf zeigen. Als Beispiel sei das Projekt Freebase genannt, das 2014 beendet und in das »geschlossene« System Wikidata überführt wurde. Diese Veränderung zog zugleich einen massiven Wechsel von URIs nach sich.[30] Das Fatale an einer solchen Veränderung ist nun, dass ein Wechsel von URIs oder IRIs im Konzept des Semantic Web nicht vorgesehen ist. Sie sind und bleiben per Definition stabil, was aber nicht der Realität des WWW und des SW entspricht.

Schließlich spielt noch die Web Ontology Language (OWL) eine gewisse Rolle für die Modellierung der unterschiedlichen Formen der Beziehung, die Ressourcen zueinander besitzen können, und für die Erstellung von Ontologien im Sinne der Informatik. Diese Ontologien ermöglichen auf einer abstrakten Ebene die Organisation des Wissens und die standardisierte Modellierung der Beziehungen von Entitäten. In den Geisteswissenschaften beliebte Ontologien sind z.B. das CIDOC Conceptual Reference Model[31] oder SKOS (Simple Knowledge Organisation System)[32].

Beispiele für nachhaltige und nicht nachhaltige Bereitstellung von Digitalen Editionen im Semantik Web

Wenn man nach Beispielen für nachhaltige Digitale Editionen im Semantic Web sucht, stellt sich schnell die Frage, ob es solche Editionen im strengen Sinne schon gibt und was eine Edition überhaupt für das Label Semantic Web qualifiziert. Einen Überblick zu Editionsprojekten überhaupt bieten die Kataloge von Patrick Sahle und Greta Franzini.[33] Allerdings sind Metadaten zur Verwendung von Semantic-Web-Technologien

das bisherige Lemma falsch und irreführend, widersprach es dem Prinzip des neutralen Standpunktes oder WP:BIO, so könne es auch nicht als Weiterleitung bestehen bleiben.

28 https://wiki.dbpedia.org/ (Zugriff 2.8.2019).

29 https://www.wikidata.org/ (Zugriff 2.8.2019).

30 https://de.wikipedia.org/wiki/Freebase. (Zugriff 2.8.2019).

31 http://www.cidoc-crm.org/ (Zugriff 2.8.2019).

32 https://www.w3.org/2004/02/skos/ (Zugriff 2.8.2019).

33 Katalog Patrick Sahle: http://www.digitale-edition.de/ Katalog Greta Franzini: https://dig-ed-cat.acdh.oeaw.ac.at/ (Zugriff 2.8.2019).

nur in dem Katalog von Greta Franzini vorhanden. Eine Suche dort bringt aber (nur) zwei Projekte zu Tage, die überhaupt RDF verwenden: Die Briefedition Vespasiano da Bisticci (Università degli Studi di Bologna) und das Petőfi Irodalmi Múzeum (Ungarn, insgesamt 5 Einzelprojekte). Und doch gibt es ja noch eine ganze Zahl weiterer Vorhaben, die zumindest einen RDF-Export der Daten bereitstellen: Dazu zählen die Baseler Jahrrechnungen und auch das Urfehdebuch von Susanna Burghartz, Sonia Calvi und Georg Vogeler, die auf der GAMS-Plattform in Graz basieren (Fedora)[34]. In der Beschreibung des Datenmodells steht »Alle textlichen und inhaltlichen Entitäten besitzen stabile Identifikatoren.«[35]. Das ist eine Grundvoraussetzung für LOD. Aber bei beiden Projekten kommt nur die eigene Institution als URI vor. Eine Verlinkung zu anderen Ressourcen hat also (noch) nicht stattgefunden. Ein anderes Kriterium wäre die Verfügbarkeit eines SPARQL-Endpoints, also einer Schnittstelle zur Abfrage der RDF-Daten mit der Anfragesprache SPARQL. Im Semantic Web erlaubt ein SPARQL-Endpoint eine automatisierte Ressourcenabfrage und die Rückgabe eines Graphen in Echtzeit, der dann in die lokalen Ergebnisse eingebunden wird oder diese erweitert.

Man kann allerdings auch einen Schritt weiter zurückgehen und die Semantik einer Digitalen Edition dort beginnen lassen, wo Entitäten oder Konzepte eindeutig identifiziert werden, wo also IDs und URIs vergeben werden.[36] Das machen inzwischen viele Editionen, z.B. auch die oben schon genannten Bisticci-Briefe aus dem Franzini Katalog.

Burckhardtsource.org ist eine digitale Bibliothek und Plattform für Editionen, die im Rahmen eines European Research Council Advanced Grant Project (EUROCORR, Juni 2010–Mai 2015) entwickelt und von Prof. Maurizio Ghelardi (Pisa, Scuola Normale Superiore) koordiniert wurde. Auf der Plattform befindet sich die kritische Ausgabe der Briefe an Jacob Burckhardt, in der eine der wichtigsten europäischen Korrespondenzen des 19. Jahrhunderts im Open Access rekonstruiert wird. Der Bearbeitungsprozess wurde mit dem auf Semantic-Web-Technologien basierenden Framework Muruca durchgeführt.[37] Die Plattform gehört zu einer Gruppe von Lösungen, die von der italienischen Firma net7 angeboten wird bzw. wurde. Die Github–Repositorien von pundit[38] und muruca zeigen, dass die Software seit 2016 nicht mehr weiterentwickelt wurde.[39] Ebenso ist das LOD-Live Portal auf burkhartsource.org seit kurzem nicht mehr erreichbar. Schon seit längerem war es zudem nicht mehr funktionsfähig.[40]

34 https://gams.uni-graz.at/ (Zugriff 2.8.2019).
35 https://gams.uni-graz.at/context:ufbas?mode=about (Zugriff 2.8.2019).
36 Persönliche Kommunikation Patrick Sahle im Juni 2019.
37 Vgl. http://www.muruca.org/ (Zugriff 2.8.2019) bzw. die Auflistung von Digitalen Editionen, die mit pundit und muruca realisiert wurden. http://www.muruca.org/portfolio/ (Zugriff 2.8.2019).
38 http://net7.github.io/pundit2/ (Zugriff 2.8.2019).
39 https://github.com/net7 am 21.7.19 (Zugriff 21.7.2019).
40 Getestet im Juni 2019.

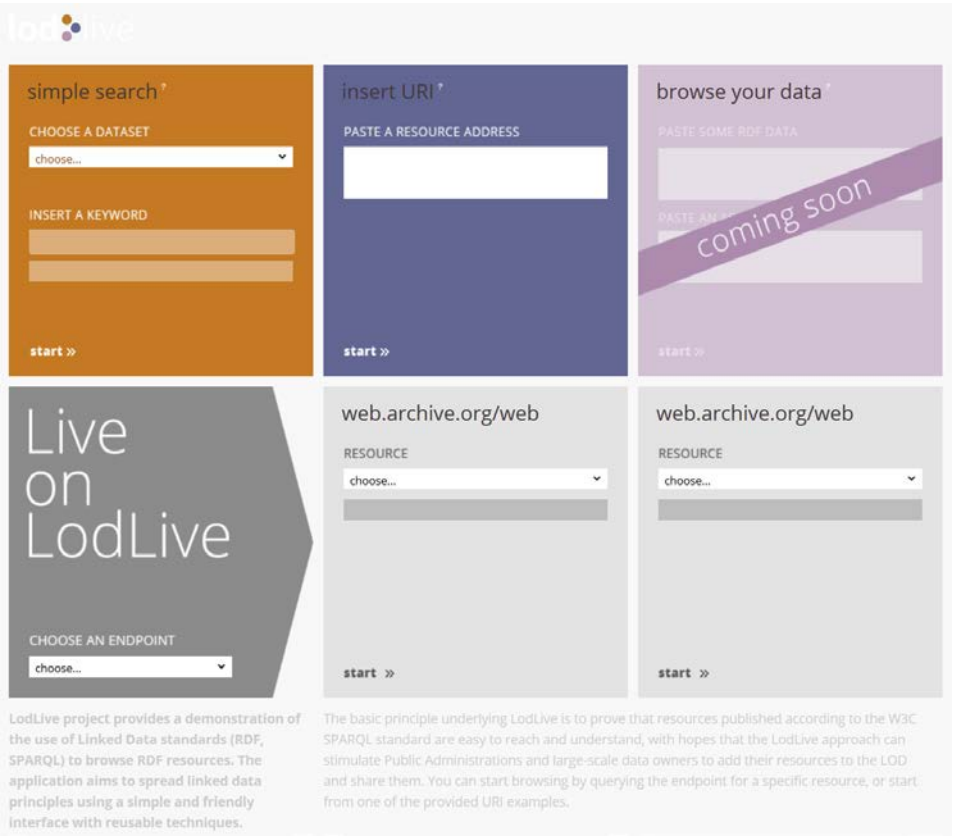

Abb. 1: LodLive at lodlive.burkhartsource.org (aktuell nicht mehr verfügbar, letzter Aufruf im Juni 2019).

Eine weitere technische Plattform, die für semantisch erweiterte Digitale Editionen verwendet werden kann, ist die Wissenschaftliche Kommunikationsumgebung (WissKI), die inzwischen in der Version 2.0 vorliegt.[41] Dieses in zwei Phasen von der DFG geförderte Projekt richtet sich im Kern an Museen, die Sammlungsverwaltung mit Semantic–Web–Unterstützung betreiben wollen[42]. Allerdings erlauben Zusatzmodule auch, semantisch angereicherte digitale Texte zu präsentieren. In einem experimentellen Vorhaben, dass die Verknüpfung von Texten und Objekten mit Hilfe von Semantic-Web-Standards erproben sollte, wurden TEI-encodierte Texte und in einer Datenbank erfasste Sammlungsobjekte über RDF-Tripel miteinander verknüpft und visualisiert.[43] Die Edition der

41 http://wiss-ki.eu/ (Zugriff 2.8.2019).

42 Die Nachhaltigkeit der Software wird durch einen Verein (Interessenvereinigung für Semantische Datenverarbeitung e.V.) und das Germanische Nationalmuseum in Nürnberg unterstützt. Weitere Informationen unter http://www.igsd-ev.de/ (Zugriff am 25.8.2019).

43 Vgl. Jörg Wettlaufer, Christopher H. Johnson, Martin Scholz, Mark Fichtner, Sree Ganesh Thotempudi: Seman-

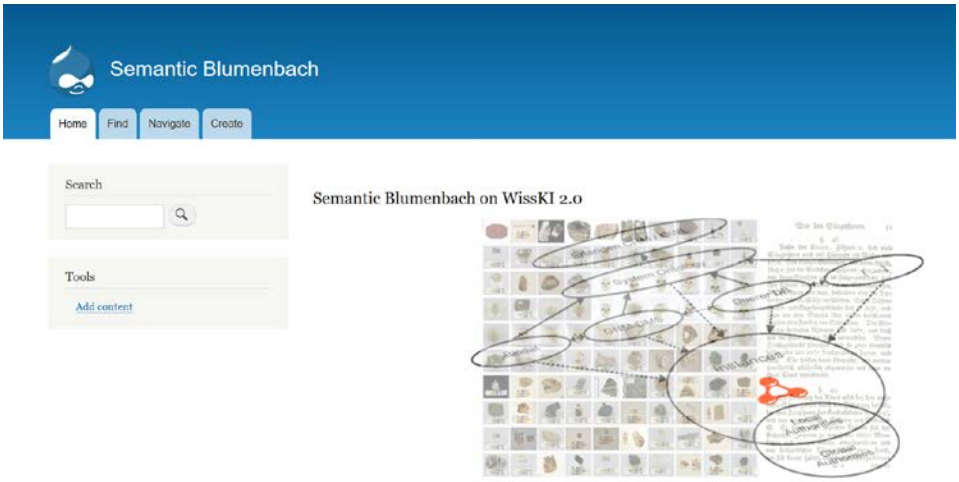

Abb. 2 Semantic Blumenbach auf WissKI 2.0 (Juli 2019)

sechsten Auflage des Handbuchs der Naturgeschichte von Johann Friedrich Blumenbach (1799), eines Göttinger Professors, der Ende des 18. und zu Beginn des 19. Jahrhunderts in Göttingen lehrte, war anschließend zunächst für etwa zweieinhalb Jahre (nach Anmeldung) online verfügbar. Nach einem fehlerhaften Serverupdate stand die Ressource dann für etwas 1,5 Jahre nicht mehr zur Verfügung. Erst ab September 2018 gelang es mit Unterstützung der Entwickler des WissKI–Systems wieder, den Server unter einer identischen URL völlig neu aufzusetzen und eine Sicherheitskopie der Daten in ein neues WissKI–2.0–System zu importieren.[44] Dabei gingen zahlreiche Erweiterungen, die nach dem eigentlichen Projektende implementiert worden waren (Live-Verknüpfung von Entitäten mit dbpedia sowie eine Geo-Visualisierung der Sammlungsobjekte), verloren. Heute läuft das System in einer Basisversion wieder, ist aber – schon aufgrund des eher experimentellen Charakters – nicht für eine nachhaltige Bereitstellung vorgesehen und wird, in absehbarer Zeit, wieder aus dem Semantic Web verschwinden.[45]

tic Blumenbach: Exploration of Text-Object Relationships with Semantic Web Technology in the History of Science, Digital Scholarship in the Humanities (DSH), Special Issue 'Digital Humanities 2014', ed. by Melissa Terras, Claire Clivaz, Deb Verhoeven and Frederic Kaplan, Vol. 30, Supplement 1, December 2015, S. i187-i198. https://doi.org/10.1093/llc/fqv047 (Zugriff 2.8.2019).

44 http://dhfv-ent2.gcdh.de/blumenbach/ (Zugriff 2.8.2019).

45 Ein weiterer Anlauf, die Materialien von Blumenbach mit Hilfe von SWT zu präsentieren, misslang kurz darauf mit einer Plattform, die auf Fedora und dem iiif-Standard beruhte. Es handelt sich ein Projekt namens PANDORA, dass von 2016-2017 von der Göttinger Akademie der Wissenschaften vorangetrieben wurde, aber Mitte 2017 aufgrund des Wegfalls einer Entwicklerressource eingestellt werden musste. Siehe hierzu Christopher H. Johnson & Jörg Wettlaufer: Einführung in das PANDORA Linked Open Data Framework, in: DHd 2017. Digitale Nachhaltigkeit, Universität Bern, 13. bis 18. Februar 2017, Konferenzabstracts, Bern, S. 31–34; Jörg Wettlaufer & Christopher H. Johnson: Poster: Digitale Nachhaltigkeit bei Grundlagenforschung im Akademieprogramm: Das Beispiel „Johann Friedrich Blumenbach-online", in: DHd 2017. Digitale Nachhaltigkeit, Universität Bern, 13. bis 18. Februar 2017, Konferenzabstracts, Bern 2017, S. 234–235 sowie https://github.

Aus diesem Beispiel lässt sich exemplarisch die Problematik der Bindung von Digitalen Editionen im Semantik Web an Projektlaufzeiten sowie die Bedeutung einer institutionellen Anbindung erkennen. Wie stabil eine URI am Ende eine Ressource vorhält und ob diese Ressourcen durchgängig erreichbar sind, lässt sich wohl am ehesten an der Institution abschätzen, an der die Ressource gehostet wird. Im Fall des Semantic-Blumenbach-Projekts handelte es sich um ein universitäres Zentrum, das alle sechs Jahre hinsichtlich seiner Existenzberechtigung evaluiert wird und sich damit als langfristiger Aufbewahrungsort für die Daten kaum eignet.

Ein anderer, nachhaltigerer Ansatz wird von der NIE-INE-Infrastruktur in der Schweiz verfolgt, die ebenfalls auf Semantik-Web-Technologien setzt. Das am DaSCH[46] angesiedelte Projekt greift dazu auf generische, aber zugleich spezifisch an das Projekt angepasste Ontologien (auf der Grundlage von CIDOC-CRM und FRBR[47]) zu, über die die einzelnen Editionen untereinander und mit anderen Ressourcen verknüpft werden sollen. Ebenso ist die Präsentationsebene modular aufgebaut und kann für die einzelnen Projekte jeweils erweitert werden. Inzwischen schon online zugänglich als Prototyp ist die historisch-kritische Online-Edition von Kuno Raebers Lyrik mit dem Stand von 2017.[48] Insgesamt sind neben Raebers Lyrik aktuell noch 14 weitere Editionen auf dem Portal angekündigt.

NIE-INE unterscheidet sich von anderen Semantik-Web-Editionsportalen durch den stärker integrierten und zugleich institutionellen Ansatz. Die Editionen existieren, soweit dies schon sichtbar ist, in einer geschützten Umgebung, die institutionell über das DaSCH abgesichert ist und damit sowohl stabile URIs als auch eine kontrollierte Infrastruktur bietet. In wieweit Verlinkungen nach außen oder in das Portal hinein geplant sind, ist momentan noch nicht abzusehen.

In Österreich kümmert sich KONDE – das Kompetenznetzwerk Digitale Edition unter Führung des Zentrums für Informationsmodellierung an der Universität Graz – um die langfristige Verfügbarkeit von Digitalen Editionen. KONDE erarbeitet unter anderem ein inhaltliches und strategisches Konzept zum Aufbau einer nationalen digitalen Infrastruktur.[49] Die GAMS[50] Architektur bietet eine passende Plattform zur Bereitstellung digitaler Editionen und geisteswissenschaftlicher Daten überhaupt. GAMS basiert auf dem Open Source Projekt FEDORA (Flexible Extensible Digital Object Repository Architecture)[51] und einer selbst entwickelten Java-Applikation. GAMS setzt dabei auf eine XML basierte

com/pan-dora und https://github.com/blumenbach/ (Zugriff 2.8.2019).

46 https://dasch.swiss/ (Zugriff 2.8.2019).

47 https://www.ifla.org/publications/functional-requirements-for-bibliographic-records (Zugriff 2.8.2019).

48 http://raeber.nie-ine.ch (Zugriff 2.8.2019).

49 http://www.digitale-edition.at/ (Zugriff 2.8.2019).

50 https://gams.uni-graz.at/ (Zugriff 2.8.2019).

51 http://fedora-commons.org (Zugriff 2.8.2019).

Datenarchivierung und -präsentation, was die Einbindung von Digitalen Editionen nach diesem Standard vereinfacht. Neben XML und TEI, LIDO (Lightweight Information Describing Objects), DC (Dublin Core), METS/MODS (Metadata Encoding and Transmission Standard/Metadata Object Description Scheme) kommen mit RDF und SKOS auch Semantic Web orientierte Standards zum Einsatz.

Fazit

Digitale Editionen im Semantic Web bzw. Editionen, die auf Semantic-Web-Standards beruhen, brauchen momentan noch geschützte Umgebungen, in denen sie gedeihen können. Damit werden die Vorteile des Prinzips der verteilten Datenressourcen erst einmal aufgegeben, aber neue Modellierungsmöglichkeiten eröffnet. Außerdem benötigt langfristige Bereitstellung, darüber herrscht Einigkeit, eine Anbindung an dauerhafte Institutionen des Kulturerbes, die in der Lage sind, die neuen Aufgaben der Bereitstellung und/oder Konservierung von Digitalen Editionen auch leisten zu können.

Es ist im Kontext der Frage nach Nachhaltigkeit von Digitalen Editionen vielleicht zweitrangig, ob das Semantic Web je in der Form, wie es Tim Berners-Lee es vor über 20 Jahren vorschwebte, Realität werden wird.[52] Der Überblick zu existierenden und nicht mehr existierenden Projekten scheint jedenfalls zu bestätigen, dass die technischen Hürden einer nachhaltigen Bereitstellung von Daten in RDF hoch sind und solche Projekte häufiger als andere, die auf leichtgewichtigere Standards setzen, schon nach kurzer Zeit wieder aus dem (Semantic) Web verschwinden.

Trotzdem bieten die Versprechungen des Semantic Web nach wie vor neue Perspektiven für Digitale Editionen. Semantische Verknüpfungen im Zusammenspiel mit den Vorteilen des Hypertexts erlauben multiple Textschichten, mit deren Hilfe Editionen in ihren Kontext gesetzt werden können. Die Verknüpfung mit Normdaten ermöglicht eine tiefe und dynamische Einbettung in die dezentral akkumulierten Wissensbestände zu Personen, Orten und Organisationen. Die Modellierung von Unsicherheit oder Widersprüchlichkeit ist mit den Standards des Semantic Web maschinenlesbar möglich.[53]

52 https://twobithistory.org/2018/05/27/semantic-web.html (Zugriff 2.8.2019).

53 Vgl. auch Andreas Kuczera & Dominik Kasper: Modellierung von Zweifel – Vorbild TEI im Graphen. In: Die Modellierung des Zweifels – Schlüsselideen und -konzepte zur graphbasierten Modellierung von Unsicherheiten. Hg. von Andreas Kuczera, Thorsten Wübbena und Thomas Kollatz. Wolfenbüttel 2019. (= Zeitschrift für digitale Geisteswissenschaften / Sonderbände, 4) text/html Format. DOI: 10.17175/sb004_003 und Michael Piotrowski: Accepting and Modeling Uncertainty. In: Die Modellierung des Zweifels – Schlüsselideen und -konzepte zur graphbasierten Modellierung von Unsicherheiten. Hg. von Andreas Kuczera, Thorsten Wübbena und Thomas Kollatz. Wolfenbüttel 2019. (= Zeitschrift für digitale Geisteswissenschaften / Sonderbände, 4) text/html Format. DOI: 10.17175/sb004_006a.

Roland Kamzalak schreibt 2016 in einem Artikel zu »Editionen im Semantic Web. Chancen und Grenzen von Normdaten, FRBR und RDF«:

> »Nun bleibt noch die Frage offen, wie denn das Semantic Web zu einem sinnvollen, Bedeutung transportierenden Informationsmedium wird. Der RelFinder kann die DBPedia verwenden, weil sie einen SPARQL-Endpoint bietet, der gezielt abgefragt wird. Damit sind aber nur die Datenquellen im Blick, die auch bekannt sind. Das ist zu wenig. Die Holschuld muss in eine Bringschuld umgewandelt werden, der Fetch- in einen Pushdienst. So wie alle mobilen Daten potenziell überall verfügbar sind, müssen auch alle semantischen Tripel überall verfügbar sein. Es muss eine Technologie entwickelt werden, die jede RDF-Quelle mit einem Sender ausstattet, statt mit einer Schnittstelle. Der Receiver fragt dann in den virtuellen Raum und empfängt relevante Daten, die dann gefiltert, übersetzt und zusammengesetzt werden müssen. Erst dann entstehen Graphen, die von Experten gespeist werden und diese dann wieder durch die Zusammenschau, die Visualisierung zu neuen Fragestellungen führen. Erst dann entsteht ein wirkliches, hochqualifiziertes semantic web.«[54]

Davon sind wir in der Tat noch weit entfernt. Die Vision eines Open Archive Initiative (OAI) Harvester für semantisch ausgezeichnete Linked-Open-Daten ist ebenso verführerisch wie die Bereitstellung semantischer Daten im WWW überhaupt. Für beide Visionen ist momentan nicht abzusehen, ob sie sich jemals verwirklichen lassen.

Heute schon Realität hingegen sind Projekte, die für spezifische Domänen Metadaten aggregieren und auf diese Weise der Forschung zur Verfügung stellen. Das mehrfach ausgezeichnete Projekt correspsearch.net[55] geht bei der Erschließung und Bewahrung der Verfügbarkeit von Briefeditionen einen besonderen Weg. Dort werden, nach dem Standard des Correspondence Metadata Interchange Format (CMIF)[56], Metadaten zu und aus Briefeditionen erfasst, die vor allem Absender, Empfänger, Schreibort und Datum umfassen. Natürlich bewahrt die Aufnahme in diese Meta-Suchmaschine für Briefeditionen eventuell nicht einzelne Projekte vor dem digitalen Untergang, aber es bleiben doch zumindest die gesammelten Metadaten verfügbar, die so auch einen Nachweis der (digitalen) Edition bieten und den letzten Speicherort benennen. Ein solcher Ansatz für Digitale Editionen überhaupt könnte, auf RDF basierend, sowohl Nachweis als auch Aggregationsportal für Metadaten aus Digitalen Editionen sein. Auf diesem Wege würde die

54 Roland Kamzelak: Digitale Editionen im semantic web. Chancen und Grenzen von Normdaten, FRBR und RDF. In: „Ei, dem alten Herrn zoll' ich Achtung gern'". Festschrift für Joachim Veit zum 60. Geburtstag, hg. von Peter Stadler und Kristina Richts, München 2016, S. 423–435, hier S. 434.

55 https://correspsearch.net/ (Zugriff 2.8.2019). Vgl. auch Stefan Dumont: »Briefe kommentieren im Semantic Web: Ein Konzept«. DARIAH-DE Working Papers Nr. 33. Göttingen: DARIAH-DE 2019. urn:nbn:de:gbv:7-dariah-2019-5-8.

56 https://correspsearch.net/index.xql?id=participate_cmi-format (Zugriff 2.8.2019).

Editionen besser erschlossen und zugleich auch nachhaltiger verfügbar gemacht. Werkzeuge für ein solches Unterfangen liegen in den Digital Humanities längst vor.[57]

Realität sind aber auch die Working Drafts der W3C Gruppe »Web Publications«, die sich seit 2017 um eine Standardisierung von Publikationen im WWW bemüht. Der letzte Entwurf zu Web Publications, Packaged Web Publications und den Web Annotation Extensions for Web Publications stammt vom 14. Juni 2019.[58] Es handelt sich dabei, verkürzt gesagt, um den Versuch, das erfolgreiche Manifest-Format der iiif–Bewegung[59] auf den Bereich der online-Publikationen zu übertragen. Basistechnologie ist hier wie dort JSON-LD[60], ein zu RDF kompatibler Standard, der den Fokus wieder zurück auf die Metadaten lenkt. Auch dies ist vielleicht eine Richtung, in die wir weiterdenken sollten, wenn wir Digitale Editionen mit und über das Semantic Web nachhaltig erschließen und zur Verfügung stellen wollen.

57 Vgl. z.B. Max Grüntgens und Torsten Schrade: Data repositories in the Humanities and the Semantic Web: modelling, linking, visualising. In: WHiSe 2016 Humanities in the Semantic Web. Proceedings of the 1st Workshop on Humanities in the Semantic Web (WHiSe), hg. von Alessandro Adamou, Enrico Daga, und Leif Isaksen, Aachen 2016, S. 53–64 (CEUR Workshop Proceedings 1608). http://ceur-ws.org/Vol-1608/#paper-07 .

58 https://www.w3.org/TR/wpub/#dfn-web-publications (Zugriff 2.8.2019).

59 https://iiif.io/ (Zugriff 2.8.2019).

60 https://json-ld.org/ (Zugriff 2.8.2019).