MAGAZIN FÜR DIGITALE EDITIONSWISSENSCHAFTEN

Herausgegeben vom Interdisziplinären Zentrum
für Editionswissenschaften
der Friedrich-Alexander-Universität Erlangen-Nürnberg

5 / 2019

FAU University Press
Magazin für digitale Editionswissenschaften
ISSN 2364-0855

Herausgeber:
Interdisziplinäres Zentrum für Editionswissenschaften
Prof. Dr. Boris Dreyer (Sprecher)
Universität Erlangen-Nürnberg
Department Geschichte
Alte Geschichte
Kochstr. 4, Postfach 8
D-91054 Erlangen

EDITORIAL

Das »Magazin für digitale Editionswissenschaften« versteht sich als ein offenes Forum zur Vorstellung von »best practises« für Online-Editionen. In den fünf- bis zehnseitigen Darstellungen sollen anhand konkreter Beispiele aus aktuell bearbeiteten Projekten insbesondere die »technische Seite« von Online-Editionen dargelegt werden: von der digitalen Codierung bis hin zu Visualisierungsstrategien, von theoretischen Erwägungen bis hin zu pragmatischen Überlegungen. Im Zentrum stehen fachspezifische Ansprüche, Standards und Methoden bei der Editionsarbeit sowie die verwendeten digitalen Werkzeuge und Präsentationsformen.

Die Mitglieder des Interdisziplinären Zentrums für Editionswissenschaften der Friedrich-Alexander-Universität Erlangen-Nürnberg, die das Magazin tragen, wollen mit diesem Forum einen Beitrag dazu leisten, dass Kriterien für Online-Editionen in der Öffentlichkeit zur Diskussion gestellt und die Anwendung derselben Editionswerkzeuge in verschiedenen Projekten vergleichend gegenübergestellt werden können. Auf diese Weise erfolgt eine Systematisierung der genutzten Editionsmittel und es wird möglich, gemeinsame Standards auf dem immer breiter werdenden Schnittfeld zwischen Philologie und Informationstechnik zu entwickeln.

Die Herausgeber

INHALT

Bausteine auf dem Weg zu einer virtuellen Zeitmaschine – ein Editionsprojekt über lange Dauer

Andreas Maier, Daniel Stromer, Vincent Christlein, Peter Bell

Zeitreisen sind ein alter Menschheitstraum, der von Faszination und Neugier genährt wird. Die gewünschten Reiseziele sind individuell und können in der Zukunft oder in der Vergangenheit liegen, z.B. ins alte Rom zu reisen, auf den Spuren der eigenen Vorfahren zu wandeln oder an einer konkreten historischen Begebenheit wie dem Ausbruch der französischen Revolution teilzuhaben. In beiden Richtungen ist unser Wissen fragmentarisch, doch aus der Vergangenheit haben wir zumindest eine lückenhafte Überlieferung vielfältiger Quellen und Objekte. Die Intentionen zu einer Zeitreise in die Vergangenheit liegen also im Überprüfen des Geschichtsbildes und im unmittelbaren Erleben von Geschichte.

Natürlich geht eine solche Zeitreise weit über unsere heutigen physischen Möglichkeiten hinaus. Die Geschichtswissenschaft rekapituliert die Vergangenheit maßgeblich über Text, während die Kreativindustrie sie als mehr oder weniger gut recherchierte Fiktion reproduziert.

Die Time Machine Initiative hat sich nun die Aufgabe gestellt, das kulturelle Erbe in großem Umfang zu digitalisieren und aufzubereiten, um neue virtuelle Zugänge zur Vergangenheit zu schaffen, die - bei Berücksichtigung der fragmentarischen Überlieferungslage - einer Zeitreise nahekommen.

In einem groß angelegten interdisziplinären und transeuropäischen Forschungsprojekt soll eine Art Edition europäischer Geschichte entstehen, die als datengesättigte Rekonstruktion eine neue Form der Begreifbarkeit und Erfahrbarkeit schaffen kann. Die Zeitmaschine wäre also eine virtuelle Forschungsumgebung, deren Ergebnisse zusätzlich auch unmittelbar und immersiv in eine breitere Öffentlichkeit vermittelt würden.

Die Konzeption der Time Machine

Trotz der Unmöglichkeit von Zeitreisen plant ein Konsortium aus mehr als 400 europäischen Einrichtungen aus Forschung und Industrie verteilt über 34 Länder den Bau

einer virtuellen Time Machine. Inspiriert und vorbereitet ist diese Initiative von dem für sich schon ambitionierten Projekt der Venice Time Machine.[1,2]

Diese Zeitmaschine basiert auf einer großen Datenbank, in der unterschiedlichste historische Quellen gespeichert, interpretiert und verknüpft werden können, von Zahlen, Texten und Bildern über Karten und 3D-Modelle bis hin zu Musik und anderen sensorischen Informationen. Die Rolle der Zeitmaschine besteht dann darin, all diese Informationen zu verknüpfen und durch diese die Vergangenheit zu rekonstruieren. Schließlich soll es möglich sein, durch diese Daten zu navigieren, um sich in Raum und Zeit so leicht und selbstverständlich bewegen zu können, wie wir es heute im Internet tun.

Um dieses ehrgeizige Ziel zu erreichen, sind zahlreiche wissenschaftliche Durchbrüche notwendig, die zu erzielen ein interdisziplinäres Konsortium erfordert. Entsprechend kommen die Mitglieder aus den klassischen Geisteswissenschaften, den Digital Humanities und der Informatik sowie aus verschiedenen anderen flankierenden Bereichen wie Gedächtnisinstitutionen, Kulturbetrieb und Kreativindustrie. Auch der Umfang des Vorhabens muss weit über der Größe gängiger Forschungsverbünden liegen und international aufgestellt sein. Solche Projekte wurden in der Vergangenheit von der Europäischen Kommission finanziert und mit erheblichen Mitteln unterstützt. Ein Beispiel für eine in dieser Größenordnung finanzierte Initiative ist das Human Brain Project, in dem eine elektronische Nachbildung des menschlichen Gehirns aufgebaut werden soll.[3]

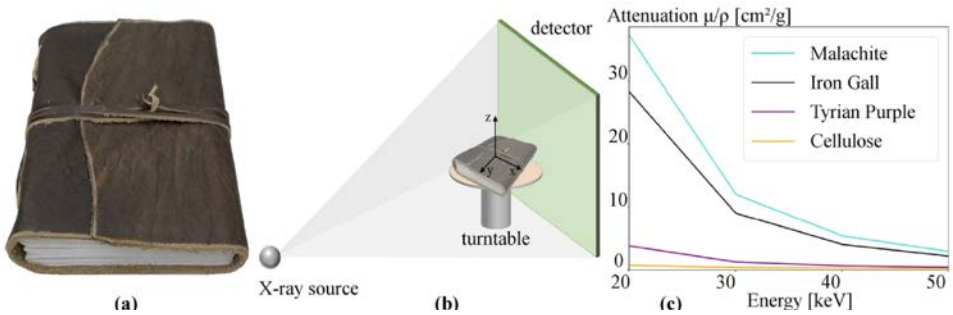

Abbildung 1: Schematische Darstellung einer Buch-Röntgen-CT Aufnahme, in: Stromer, D., Christlein, V., Martindale, C., Zippert, P., Haltenberger, E., Hausotte, T., & Maier, A. (2018). Browsing through sealed historical manuscripts by using 3-D computed tomography with low-brilliance X-ray sources. Scientific reports, 8(1), 15335.

1 Frédéric Kaplan: The Venice Time Machine, in: Proceedings of the 2015 ACM Symposium on Document Engineering. ACM (2015).

2 Beispiele unter anderem: Swiss Federal Institute of Technology, FAU Erlangen-Nürnberg, TU Wien, National Archive of Norway, Amsterdam City Archive, Le Louvre, Venice State Archives, Isreal Museum, Royal Institute for Cultural Heritage, Ministry of Interior of the Slovak Republic, ICARUS, FlixBus, IBM Italia/Switzerland, Ubisoft.

3 Henry Markram: The human brain project, in: Scientific American 306.6 (2012), S. 50–55.

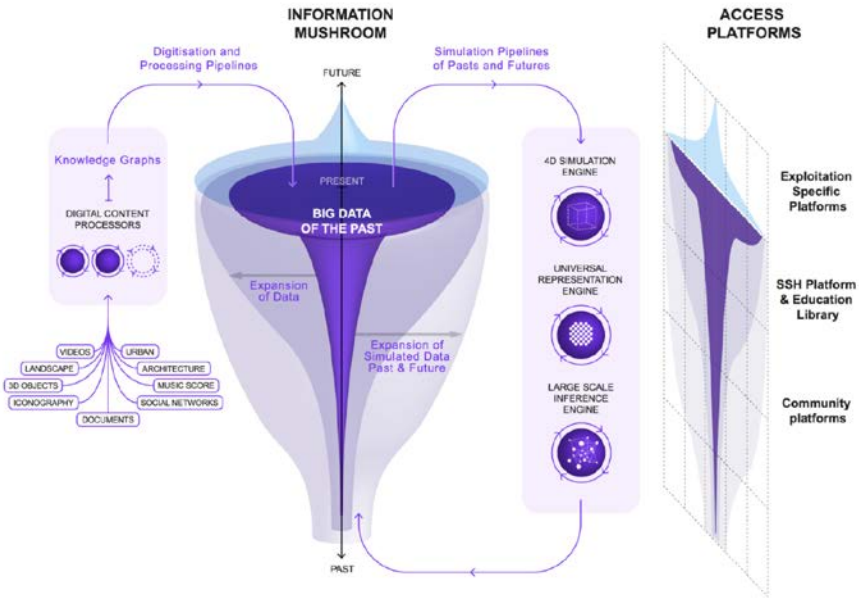

Abbildung 2: Schematische Visualisierung des Time Machine Workflows und der Infrastruktur, in: TECHNICAL ANNEX - FET-Flagship 2nd Stage proposal, Time Machine: Big Data of the Past for the Future of Europe (2018).

Die wichtigsten Herausforderungen, denen sich die Forscher gegenübersehen, lassen sich in drei Kategorien einteilen: Daten und Digitalisierung, Wissensgewinnung und -modellierung, sowie Einschränkungen und Chancen einer solchen digitalen Epistemologie (Methodenreflexion). Außerdem gibt es auf dem Weg zur Time Machine viele weitere Herausforderungen, wie Projekt- und Community Management über 34 Staaten sowie Lizenzen und rechtliche Fragen, die zu besprechen weit über den Rahmen dieses Artikels hinausgehen würden. Daher werden wir uns an dieser Stelle nur mit den Hauptherausforderungen befassen.

Für die Gegenwart und die zweite Hälfte des 20. Jahrhunderts gibt es eine extrem hohe Informationsdichte über Ereignisse, Personen und Gegebenheiten.

Grundsätzlich kann man sagen, dass, je weiter wir in die Vergangenheit gehen, desto weniger Informationen uns zur Verfügung stehen und noch weniger davon in einem elektronischen Format verfügbar sind, das als Input für die Verarbeitung mit der Time Machine geeignet wäre. Selbst für das kulturelle Erbe, d.h. Informationen, die wir für unsere kulturelle Identität als sehr wichtig erachten, sind derzeit nur zu 10% in digitaler Form verfügbar.[4] Bei Archiven und Bibliotheken liegt der Prozentsatz sogar noch nied-

4 Lorna Hughes: Infrastructures for digital research: new opportunities and challenges (2017), S. 37–53. (URL:

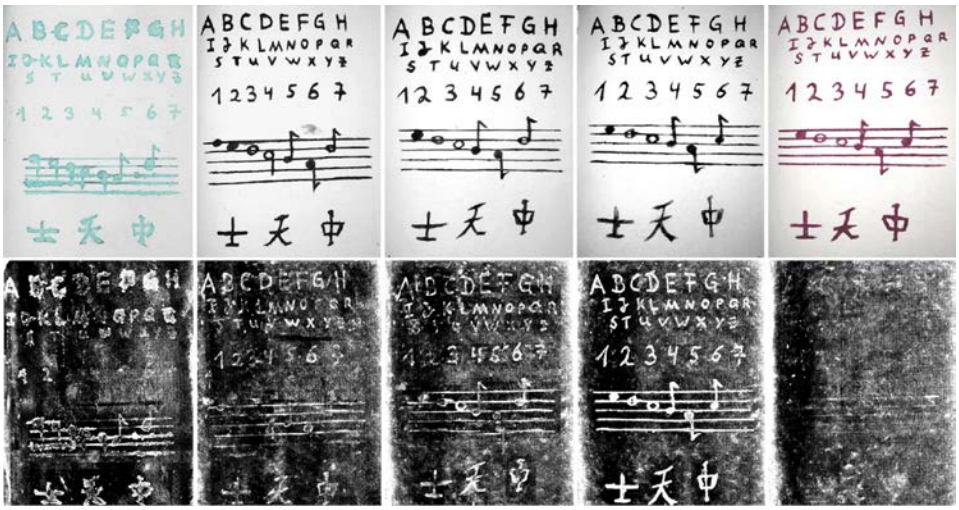

Abbildung 3: Fotografien und CT Rekonstruktionen der gleichen Seiten von Malachittinte (Kupfer), Eisengallustinte in drei verschiedenen Mischungen, und Purpurtinte. Erste Ergebnisse zeigen, dass für alle drei Farbstoffe Bildgebung prinzipiell möglich ist. Je höher der Metallanteil, desto besser der Kontrast, in: Stromer, D., Christlein, V., Martindale, C., Zippert, P., Haltenberger, E., Hausotte, T., & Maier, A. (2018). Browsing through sealed historical manuscripts by using 3-D computed tomography with low-brilliance X-ray sources. Scientific reports, 8(1), 15335.

riger. Ein erstes Ziel ist daher die massive Digitalisierung. Im Gegensatz zu herkömmlichen Aufnahmegeräten, die Seiten umblättern, könnte dieser Prozess mit volumetrischen Erfassungstechnologien, wie der Computertomografie (CT) erheblich beschleunigt werden.[5] Mobile Aufnahmegeräte wie das Scan-Zelt spielen auch eine wichtige Rolle für die hochwertige Digitalisierung im Feld.[6] Darüber hinaus ist das massive Aufnehmen von 3D-Objekten an einem Fließband bereits heute in unserer technologischen Reichweite.[7]

Diese riesigen Datenmengen erfordern jedoch auch langfristige Speichermethoden, die diese Informationen dauerhaft sichern können. Forschende von Twist Bioscience entwickeln Technologien zur Speicherung digitaler Informationen in DNA-Strängen.[8] Dies ist die kompakteste Darstellung von Informationen, die der Menschheit bekannt ist, da die Moleküle selbst die Informationen tragen. Dies ermöglicht Speicherungen, die um Größenordnungen kompakter sind als die heutigen digitalen Speicher. Zu unterstreichen

http://eprints.gla.ac.uk/158070/1/158070.pdf). Letzter Zugriff: September 2019.

5 Daniel Stromer et al.: Browsing through sealed historical manuscripts by using 3-D computed tomography with low-brilliance X-ray sources, in: Scientific reports 8 (1), 15335 (2018).

6 Florian Kleber, Markus Diem, Fabian Hollaus, Stefan Fiel: Mass Digitization of Archival Documents using Mobile Phones, in: Proceedings of the 4th International Workshop on Historical Document Imaging and Processing, ACM (2017).

7 Pedro Santos et al.: CultLab3D: On the verge of 3D mass digitization, in: Proceedings of the Eurographics Workshop on Graphics and Cultural Heritage, Eurographics Association (2014), S. 65–73.

8 Andy Extance: How DNA could store all the world's data, in: Nature News, 537 (7618) (2016), S. 22.

ist, dass diese Art der Lagerung auch für die Langzeitkonservierung geeignet ist, da wir Beispiele für DNA-Befunde kennen, die 10.000 und mehr Jahre ohne Verlust ihrer Informationen überlebt haben.

Selbst wenn wir es schaffen, alle Daten zu digitalisieren und zu speichern, die wir aus mehr als 2000 Jahren europäischer Geschichte zusammentragen können, stoßen wir sofort auf zwei weitere Probleme: Der Zahn der Zeit, der an den Dokumenten nagt, sowie die daraus entstehende Unschärfe der Informationen. Ein Großteil der Daten ist der Zeit zum Opfer gefallen, wir müssen also mit Unschärfen arbeiten, und die dennoch immensen Datenmengen müssen weiterverarbeitet werden. Denn es geht bei der Digitalisierung nicht nur um die Transformation in ein anderes Medium, sondern auch um eine tiefe Erschließung des Materials, wie sie mit klassischen Methoden und vorhergehenden Aufschreibesystemen nicht möglich war.

Für die Datenverarbeitung müssen wir die Verarbeitung von Text, Bild, Audio, Karten, 3D-Objekten und deren Interpretationen verschränken. Heutzutage verwenden die meisten Systeme, die für diesen Zweck eingesetzt werden, Graphen und symbolische Darstellungen, jedoch haben wir bereits gesehen, dass die Fähigkeit des tiefen Lernens in der Lage ist, jedes symbolische System in vielen Anwendungen zu übertreffen, wie z.B. in der automatischen Übersetzung.[9] Daher ist es ein Ziel des Projekts, einen universellen Repräsentationsraum zu schaffen, der es uns ermöglicht, alle oben genannten Elemente ineinander umzuwandeln. Ein solches System hat jedoch den großen Nachteil, dass es keine Verknüpfung von Beobachtungen mit einer Inferenzkette zulässt, wie dies bei symbolischer Deduktion möglich wäre. Ein weiteres wichtiges Ziel ist es, die Verschmelzung von symbolischen, graphenbasierten und unscharfen neuronalen Netzwerken zu ermöglichen. Aufgrund dieser Fortschritte müssen wir noch historische Rekonstruktionen erstellen können. Traditionelle Methoden verwenden Computergrafiken für solche Zwecke, jedoch nimmt auch maschinelles und tiefes Lernen in dieser Disziplin zu. Daher benötigen wir Methoden, mit denen komplexe Szenen aus einfachen Beschreibungen erstellt werden können. Die nachfolgende Interpretation und Analyse von Informationen wird weiterhin von Wissenschaftler*innen durchgeführt, die in der Time Machine als virtueller Forschungsumgebung arbeiten. Die Daten werden aber auch durch ein Interface benutzerfreundlich aufbereitet, so dass diese durch ganz verschiedene Nutzergruppen verwendet werden können. Von Historiker/innen über Hobbywissenschaftler/innen, Stichwort citizien science, bis hin zu Laien und Touristen können alle Interessierten die Time Machine für ihre Zwecke bedienen. Es handelt sich um einen partizipativen Ansatz, in dem eine Teilhabe je nach Kompetenz und Interesse in verschiedenen Stadien des Prozesses möglich ist.

9 Don Monroe: Deep learning takes on translation, in: Communications of the ACM, 60(6) (2017), S. 12–14.

Ein dritter wichtiger Aspekt der Zeitmaschine ist die Gewinnung neuer Erkenntnisse. Wie in allen Beobachtungen muss der Informationsgehalt und der Grad des Vertrauens festgelegt werden. Dies erfordert erweiterte Ansätze der Erkenntnistheorie als digitale Epistemologie, die in der Lage ist, verschiedene Deutungen einer historischen Situation gleichzeitig zu handhaben. Da Geschichtsbilder auch als Werkzeuge politischer Arbeit und Meinungsbildung eingesetzt werden, ist es wichtig, dass das bereitgestellte Wissen quellenkritisch hinterfragt wird, kontroverse Ansichten parallel gezeigt und vermeintliche 'historische Wahrheiten' dekonstruiert werden. Politisch gefärbte Geschichtsbilder sind nicht nur ein Manipulationsversuch gegenüber den jeweiligen Gesellschaften, sondern erschweren auch die wissenschaftliche Auseinandersetzung mit Geschichte, die sich zumindest Objektivität anzunähern versucht.

Entsprechend muss der Zweck der verwendeten Darstellungen transparent gemacht werden. Forschende aus der Archäologie würden beispielsweise mit Grauwerten arbeiten, wenn sie anzeigen möchten, dass die Farbigkeit eines Tempels unbekannt ist. Im Gegensatz dazu würde ein Tourismusbüro eine plausible und detailreiche Rekonstruktion desselben Tempels vorziehen, um ein intensiveres Erlebnis für das Publikum zu schaffen. Eine neue historische Erkenntnis ist auch eine Interpretation von Daten und muss daher mit den ursprünglichen Informationen und der Beobachtungskette verknüpft werden, die zu dieser Einsicht führte. In den Geisteswissenschaften wird dies schon seit Jahrhunderten mittels Text und Sprache getan. Um dies digital und multimodal zu erreichen, müssen wir jedoch einen digitalen Arbeitsprozess etablieren, der ein höheres Maß an Zusammenarbeit und einen schnelleren wissenschaftlichen Fortschritt ermöglicht. Allgemeine künstliche Intelligenz, kurz KI, wird daher auch ein wichtiger Faktor für den Erfolg der Time Machine sein. So können virtuelle Agenten erzeugt werden, die Teil der Rekonstruktion werden und mit ihren Handlungen historische Situationen simulieren können. Darüber hinaus werden arbeitsintensive Aufgaben wie Datenbankabfragen durch moderne KI-Methoden gelöst, die auf automatische Frage-Antwort-Mechanismen und Sprachinterpretation abzielen.

Die Time Machine zielt darauf ab, Rekonstruktionen der Vergangenheit mit einem bis heute unbekannten Detailgrad zu erzeugen. Insofern ist es kein Zufall, dass sich wichtige Industrieunternehmen wie Ubisoft, die für ihre Assassin's Creed-Serie bekannt sind, diesem Konsortium angeschlossen haben.[10] In ihrer Vision einer Zeitmaschine – dem Animus[11] – kommen sie dem Ziel des Time Machine Projects bereits sehr nahe, denn

10 Lisa Gilbert: "The Past is Your Playground": The Challenges and Possibilities of Assassin's Creed: Syndicate for Social Education, in: Theory and Research in Social Education 45 (1) (2017), S. 1–11.

11 Animus ist in der Spielwelt der Name der Technologie, die Zeitreisen bzw. Inkarnation in eine historische Per-

Animus ermöglicht ein tiefes Erleben der Vergangenheit, indem er Personen ermöglicht zurück in das Leben der eigenen Vorfahren zu reisen. Obwohl dieses Ziel heute noch lange nicht erreicht ist, glauben die Mitglieder des Time Machine Konsortiums, dass eine solche Zeitmaschine die Forschung revolutioniert und auch im Bildungsbereich, sowie im Unterhaltungs- und Kulturbetrieb einschließlich des Tourismuses neue Perspektiven schafft.

Time Machine und Editionswissenschaft

Die Editionswissenschaft spielt in der Time Machine Initiative in zweierlei Hinsicht eine Rolle. Zum einen wird sie im wortwörtlichen und praktischen Sinne angewandt, da Millionen von Schriftstücken nach editionswissenschaftlichen Kriterien erschlossen, verglichen und geordnet werden müssen. Darüber hinaus findet sie auch in einem methodologischen Sinne, der das ganze Projekt betrifft, Anwendung. In Editionswissenschaften und Time Machine Initiative geht es jeweils um die Rekonstruktion ursprünglicher Zustände und Überlieferungsgeschichte. Gemeinsam ist ihnen der Umgang mit Unschärfe und eine möglichst große Nähe zu den Originalen. Edition und Time Machine erzeugen nicht nur Repräsentationen ihrer Gegenstände, sondern kompilieren auch die Sekundärliteratur in all ihren Kontroversen und Wendungen. Dies kann in beiden Fällen über die Belegpflicht hinausgehen und einen Ort schaffen, in dem das Wissen über einen Gegenstand zusammengefasst ist. Ungeachtet dessen, dass Editionen auch implizit oder explizit Forschungsfragen oder Erkenntnisinteressen folgen können, stellen sie eine breitere Arbeitsgrundlage dar, als stärker einem Narrativ und Forschungsfragen folgende Studien. Ein solches Plateau von dem aus weitere Forschung entsteht und verhandelt wird, möchte auch die Time Machine sein.

Unabhängig vom Grad der Digitalisierung erscheinen Quellenkritik und -analyse als entscheidende Arbeitsschritte innerhalb dieses Unternehmens, zumal das kumulierte Wissen in nachgeordneten Schritten unmittelbar in Simulationen, Rekonstruktionen und Trainingsdaten (also Lernmaterial für die künstliche Intelligenz) weiterverwendet wird. Falsch verstandene und falsch zugeschriebene Quellen würden entsprechend unmittelbar zu erheblichen Verzerrungen in den folgenden Schritten führen. Entsprechend braucht es Geisteswissenschaftler/innen, um die digitale Erschließung immer wieder kritisch gegenzulesen.

son ermöglicht. Vgl. Andy Kelly: Feel what it is like to live in Ancient Greece in Assasin's Creed Odyssey's new mode (2019). (URL: https://www.pcgamer.com/feel-what-it-was-like-to-live-in-ancient-greece-in-assassins-creed-odysseys-new-mode/). Letzter Zugriff: September 2019.

Big Data versus Monografie

Dennoch unterscheidet sich ein wichtiger Forschungsansatz der Time Machine grundlegend von der herkömmlichen Arbeitsweise in den Geisteswissenschaften. Während dort meist qualitativ, monographisch an kleinen Korpora und überschaubaren Zeitabschnitten gearbeitet wird, stellt sich das Time Machine Project der langen Dauer und einem im Grunde holistischen Ansatz. Statt anhand einer Forschungsfrage einen Wissensbestand stückweise zu akkumulieren, wird aus der Masse von Daten geschöpft. Dafür ist der Begriff Big Data berechtigt, denn obwohl die einzelnen Objektgruppen in weiter zurückliegenden Epochen überschaubar sein mögen, sind oft deren vielfältige Vernetzungen mit anderen Objekten und Akteuren sehr umfangreich und komplex.

Zum besseren Verständnis von herkömmlicher und hier vorgestellter Arbeitsweise bietet sich ein Vergleich aus der bildenden Kunst an: Während ein Maler mit der weißen Leinwand beginnt und Strich für Strich zu einem Bild und damit zur Bedeutungsstiftung gelangt, schafft ein Bildhauer Bedeutung durch das Wegnehmen und Ausdifferenzieren von Material. Entsprechend wird in der Geisteswissenschaft eher additiv gearbeitet, während die Time Machine subtraktiv vorgeht. Im Gegensatz zum Bildhauer haben die Forschenden noch keine Vorstellung, welche Informationen genau der Datenblock enthält. Doch die Erschließung des Materials durch die darin ausgebildeten Muster führt zum Verständnis.

Natürlich existiert diese idealtypische Trennung in der Praxis nicht, denn mit jedem Archiv, das hinzutritt und in dem eine erste Quelle aufgenommen und annotiert wird, ergeben sich additive Schritte. Dennoch kommen alle Informationen zusammen und bilden eine Datenbasis, anhand welcher kulturelles Erbe erschlossen und rekonstruiert werden soll. Diese Verknüpfung der Daten ist nicht nur eine schon weithin bekannte Vision des Semantic Web, sondern auch geschichtswissenschaftlich unbedingt notwendig. Dies liegt besonders an der großen Zerstreuung historischer Objekte. Durch beispielsweise weitläufige Handelsverbindungen, Krieg, Migration oder einfach nur falsche Zuschreibung befinden sich viele Objekte nicht am Ort ihrer Entstehung. Automatische Verfahren z.B. der Schreibererkennung, der automatischen Bild- und Textanalyse ermöglichen es Objekte etwa aus Nürnberg wieder virtuell mit ihrem Entstehungsort zu verknüpfen bzw. jede Referenz über den Ort etwa in der Korrespondenz zu Dritten mit dem Ort zu verbinden.

Auch die Simulation in Form von Data Augmentation kann zu dieser Verknüpfung beitragen. Ein einfaches Beispiel dafür ist der Erhaltungszustand. Unter normalen Bedingungen werden zwei Objekte mit stark unterschiedlichem Erhaltungszustand nicht automatisch aufgefunden. Der Computer könnte aber das Digitalisat eines Ausgangsdokuments so manipulieren, dass verschiedene Stadien eines besseren bzw. schlechteren

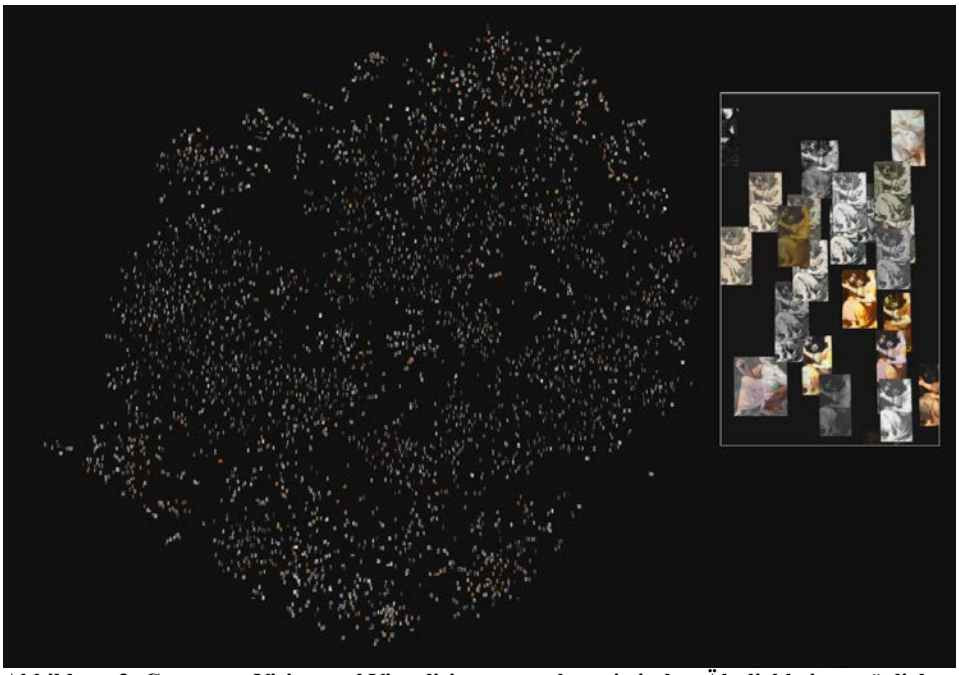

Abbildung 3: Computer Vision und Visualisierung nach statistischer Ähnlichkeit ermöglichen eine Kombination aus Distant- und Close Viewing. Letztlich das vergleichende Sehen an einer überschaubaren Anzahl von Einzelwerken eingebettet in eine Karte geordneter Bilder (Plot von Leonardo Impett und Peter Bell, 2019).

Zustands simuliert werden und mit diesen Replikaten verschiedener Qualität eine bessere Ausgangslage für die Suche herbeiführen.

Ein Beispiel für den subtraktiven Forschungsansatz mit Big Data lässt sich am Umgang mit Kunstwerken zeigen. Der Computer beginnt an einem Datensatz von 5 Millionen Digitalisaten von Kunstwerken und anderen historischen Bildern zunächst Gruppen zu bilden: Porträts, Landschaften ohne, oder mit vernachlässigbarem, Personal (Staffage), Stillleben und schließlich Szenen mit menschlicher Interaktion. Diese Bilder können nach Figurenanzahl oder Haltungen der Figuren geordnet werden. Auch die Szene lässt sich hinsichtlich ihres Inhalts grob einordnen und Objekte erkennen, dadurch werden Ikonographien und ihre Veränderungen über die Jahrhunderte identifizierbar. Quer dazu liegen die Achsen, der Technik, des Stil und des Modus. Das Arbeiten aus der Masse heraus oder – mit der geläufigeren Metapher – aus der Distanz (Distant Reading/Viewing) ergibt eine andere Übersicht über das Material, wie sie die klare Fokussierung auf eine Forschungsfrage nicht bewerkstelligen kann.

Hieraus ergibt sich viel Potential zur Dekanonisierung (und damit auch Dekolonialisierung) von kulturellem Erbe und die Betrachtung paralleler Entwicklungen. Verschiedene Gruppen innerhalb des Konsortiums arbeiten an einer solchen Ordnung der Bilder. Um die vielen Dimensionen der visuellen Repräsentationen verarbeitbar zu machen, be-

steht die Aufgabe darin, verschiedene Sichten auf das Material zu erzeugen, etwa in dem nur die Posen des Bildpersonals untersucht werden.[12]

Ausblick

Vieles, was hier beschrieben wurde, hat noch den Charakter des Vorläufigen und Skizzenhaften, da die Initiative erst im März 2019 eine Anschubfinanzierung von der Europäischen Union erhalten hat, um eine Roadmap zu erarbeiten. Die bisherige klar benennbare Leistung des Time Machine Konsortium liegt in diesem selbst. Es ist eine einzigartige Gruppe aus europäischen Forschenden im Umfeld der Digital Humanities, der Informatik und den Gedächtnisinstitutionen entstanden, die mit Industriepartnern an einem sehr konkreten Ziel arbeiten möchte.

Dieses Ziel ist gleichsam Methode: Geschichte in ihrer Mannigfaltigkeit zu erfassen, in einer sehr hohen Skalierung. Ebenso weit gefasst ist der Prozess der Digitalisierung und inhaltlischen Aufbereitung eines historischen Objektes bis hin zur gesellschaftlichen Vermittlung. Die Time Machine stellt eine Pipeline dar, die viele wissenschaftliche und kulturelle Bereiche in einer Reihe von Schritten integriert und so interdisziplinär verknüpft. Die Time Machine wird vermutlich nicht in einer Dekade eine stufenlose und tiefe Repräsentation der europäischen Vergangenheit umsetzen können, doch entscheidend ist, dass sie den Prozess dieser Digitalisierung strukturell etablieren kann und Werkzeuge schafft, die eine tiefe Erschließung in der Fläche, z.B. von vielen kleinen Archiven und Gemeinden, ermöglicht.

Der genaue Rahmen der Projektorganisation, das Methodenspektrum und die Finanzierung werden in diesem Jahr eruiert. Es wird außerdem eine Organisationsform für die Beteiligten gegründet. Schon jetzt zeigen aber einige lokale Time Machines, z.B. Venedig, Amsterdam und Nürnberg, exemplarisch wie eine europäische Time Machine Geschichte erschließt und aufbereitet[13].

12 Peter Bell, Leo Impett: Ikonographie und Interaktion. Computergestützte Analyse von Szenen der Evangelien, in: Das Mittelalter, Perspektiven mediävistischer Forschung. Themenheft Digitale Mediävistik (erscheint 2019).

13 Time Machines zu finden unter: https://www.timemachine.eu/time-machines/. Letzter Zugriff: 14. Sepetember 2019.

Literatur

Peter Bell, Leo Impett: Ikonographie und Interaktion. Computergestützte Analyse von Szenen der Evangelien, in: Das Mittelalter, Perspektiven mediävistischer Forschung. Themenheft Digitale Mediävistik (erscheint 2019).

Andy Extance: How DNA could store all the world's data, in: Nature News, 537 (7618) (2016).

Lorna Hughes: Infrastructures for digital research: new opportunities and challenges (2017), S. 37–53. (URL: http://eprints.gla.ac.uk/158070/1/158070.pdf)

Frédéric Kaplan: The Venice Time Machine, in: Proceedings of the 2015 ACM Symposium on Document Engineering. ACM (2015).

Florian Kleber et al.: Mass Digitization of Archival Documents using Mobile Phones, in: Proceedings of the 4th International Workshop on Historical Document Imaging and Processing, ACM (2017).

Henry Markram: The human brain project, in: Scientific American 306.6 (2012), S. 50–55.

Don Monroe: Deep learning takes on translation, in: Communications of the ACM, 60(6) (2017), S. 12–14.

Pedro Santos et al.: CultLab3D: On the verge of 3D mass digitization, in: Proceedings of the Eurographics Workshop on Graphics and Cultural Heritage, Eurographics Association (2014), S. 65–73.

Daniel Stromer et al.: Browsing through sealed historical manuscripts by using 3-D computed tomography with low-brilliance X-ray sources, in: Scientific reports 8 (1), 15335 (2018).

Lisa Gilbert: "The Past is Your Playground": The Challenges and Possibilities of Assassin's Creed: Syndicate for Social Education, in: Theory and Research in Social Education 45 (1) (2017), S. 1–11.

SEMANTIC WEB – EINE KURZE EINFÜHRUNG

Klaus Meyer-Wegener

Auf Wunsch der Mitglieder des „Interdisziplinären Zentrums für Editionswissenschaft“[1] an der Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) habe ich am 22. Oktober 2018 in der Mitgliederversammlung einen Vortrag über das „Semantic Web“ gehalten. Dieser Beitrag zum Magazin ist nun die schriftliche Ausarbeitung dieses Vortrags.

Ausgangspunkt für das Semantic Web[2] war das ubiquitäre World-wide Web (WWW), oft auch einfach als „Internet“ bezeichnet, obwohl das eigentlich nur die darunterliegende Rechnernetz-Infrastruktur ist. Nachdem man dort mit dem Klicken auf Links als navigierende Suche nach Information begonnen hatte, zeichnete sich sehr bald ein Bedarf nach deklarativer Suche ab, der dann von den verschiedenen Suchmaschinen befriedigt wurde. Das ist allerdings immer eine Volltextsuche, d.h. sie findet Texte, wenn ein vorgegebenes Wort oder mehrere vorgegebene Wörter in ihnen vorkommen. Sie findet sie aber nicht, wenn diese Wörter nicht vorkommen, auch wenn der Text inhaltlich viel mit dem Begriff zu tun hat. So ist in älteren Texten noch das Wort „Negerkuss“ zu finden, aktuell wird dafür aber „Schaumkuss“ verwendet – auch bei der Suche. Deshalb stellte sich die Frage, ob man statt mit Wörtern nicht auch nach „Bedeutung“ suchen könnte. Das kann man, aber dazu muss sie den WWW-Texten hinzugefügt werden. Ob das dann von Hand geschieht oder durch einen Text-Analyse-Algorithmus, ist dann noch eine andere Frage. Zuerst muss man klären, wie diese „Bedeutung“ (Semantik) überhaupt aussehen kann.

Und es gab noch einen weiteren Anlass, das WWW weiterzuentwickeln: Es enthält zunächst einmal nur Texte, inzwischen auch Bilder, Tonaufnahmen und Videos, aber das sind allesamt unstrukturierte oder bestenfalls schwach strukturierte Daten. Daneben gibt es aber auch noch in großem Umfang strukturierte Daten, die man sich als Formulare mit Feldern und Inhalten, als Tabellen, als Karteikartensammlungen und dergleichen vorstellen kann. Alle Datenbanken enthalten solche strukturierten Daten – und können sie sehr flexibel auswerten, was mit Texten leider nicht geht. Auch Tabellenkalkulation (in Microsoft Excel und ähnlichen Programmen) verwendet solche strukturierten Daten. Man möchte sie in vielen Fällen in den eigenen Datenbestand integrieren und mit ihm zusammen auswerten, man denke nur an Wirtschaftsentwicklung oder Raumpla-

1 https://www.ized.fau.de/ (letzter Zugriff am 14.08.2019).

2 Pascal Hitzler, Markus Krötzsch, Sebastian Rudolph und York Sure: *Semantic Web – Grundlagen*. Springer : Berlin, Heidelberg, 2008. – ISBN 978-3-540-33993-9.

nung. Während bestimmte Daten dem Datenschutz unterliegen (Personalverwaltung, Gesundheitswesen, Banken, Versicherungen), sind andere im Prinzip öffentlich (Wetter, Landkarten, amtliche Statistiken). Bevor man diese nutzen kann, muss man sie aber verstehen. Das hat zwei Aspekte: die Struktur der Daten und ihre Bedeutung. Als Struktur (oder Format) kann ein sehr breites Spektrum vorkommen: CSV, XML, JSON und viele andere mehr. Es ist hier nicht von Interesse, was diese Abkürzungen bedeuten. Wichtig ist, dass sie alle ganz verschieden sind und dass man sich auf jedes dieser Formate erst einstellen muss. Noch schwieriger ist es mit der Bedeutung: Diese strukturierten Daten weisen Bestandteile auf, die wir hier als „Felder" bezeichnen wollen, wie die Felder, die man in Formularen ausfüllen soll. Sie können auch geschachtelt sein; das Feld „Adresse" kann aus den untergeordneten Feldern „Straße", „Hausnummer", „Wohnort" usw. bestehen. Hier habe ich schon vorausgesetzt, dass die Felder eine Bezeichnung haben, was leider auch nicht immer der Fall ist. Und selbst wenn – was ist eigentlich ein „Preis"? Ist die Mehrwertsteuer schon enthalten oder noch nicht? Ist es der Preis pro Stück oder der Gesamtpreis? Solche Frage treten immer auf, wenn man es mit einem fremden Datenbestand zu tun hat, und es ist alles andere als einfach, sie zufriedenstellend zu beantworten. Andererseits muss man das aber tun, wenn man zu vernünftigen Auswertungen dieser Daten kommen möchte.

Die Idee von Tim Berners-Lee, dem „Vater" des WWW, war nun gegen Ende der neunziger Jahre, die inzwischen ubiquitäre Infrastruktur des WWW auch für die Verbreitung von strukturierten Daten zu nutzen[3]. Man soll sie herunterladen, abfragen und auswerten können wie in einer Datenbank. Im Zusammenhang mit „Big Data" und „Data Science" gewinnt das gerade jetzt erst so richtig an Bedeutung. Es bedeutet aber auch, dass die Daten dann sozusagen von überall her kommen können. Deshalb muss ihre Bedeutung erklärt werden – vom Vertrauen in die Korrektheit der Daten einmal ganz abgesehen. Da sind zunächst die Anbieter der Daten in der Pflicht, die etwas dazu sagen müssen. Vermutlich haben sie ja auch ein Interesse daran, dass ihre Daten verwendet werden[4]. Zur Beschreibung von Daten verwendet man heute sog. Ontologien – eine der vielen unreflektierten Übernahmen attraktiv klingender Begriffe in die Informatik, bei der die ursprüngliche Bedeutung in der Philosophie großzügig angepasst wurde. Hier sind es eigentlich nur Begriffssystematiken oder -systeme (Taxonomien), die eine Festlegung von Begriffen ermöglichen, um sie dann bei der Kennzeichnung von Daten zu verwenden. So wird eben auch festgelegt, was mit „Preis" bei diesen Daten gemeint sein soll. Wichtig

3 Tim Berners-Lee, James Hendler und Ora Lassila: The Semantic Web: a new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. In: *Scientific American*, 284 (5), Mai 2001, S. 34–43 (dt.: Mein Computer versteht mich. In: *Spektrum der Wissenschaft*, August 2001, S. 42–49).

4 Das sollte bei guter wissenschaftlicher Praxis immer mit einer ordentlichen Quellenangabe erfolgen, die dann auch dem Anbieter zugutekommt.

ist dabei, dass auch Beziehungen zwischen Begriffen definiert werden, die sehr zu ihrer konsistenten Verwendung beitragen können. So werden Oberbegriffe und Unterbegriffe einander zugeordnet und es werden Synonyme und Homonyme benannt, u.U. sogar Gegensätze. Ganz zentral ist auch die Zuordnung von Typ und „Instanz", wobei letzteres leider eine direkte Übernahme aus dem Englischen ist und auf Deutsch eigentlich „Exemplar" heißen müsste. Dazu gehören beispielsweise „Säugetier" und „Löwe", „Flugzeug" und „Airbus A380" sowie „Fußballspieler" und „Ronaldo". Bevor ich näher auf die Nutzung solcher Ontologien für die Kennzeichnung der Bedeutung von Daten eingehe, muss ich aber noch die Struktur der Daten etwas genauer beschreiben.

Resource Description Framework (RDF)

Dafür hat sich das „Resource Description Framework" (RDF) etabliert. Es ist sehr einfach und gerade deshalb sehr allgemein. Alles wird dargestellt in Form von Tripeln aus Subjekt, Prädikat und Objekt. Das „Prädikat" ist dabei meist ein Verb. Damit kann man elementare Aussagen formulieren wie: Ein Professor hält einen Vortrag. Nun ist das natürlich noch sehr simpel, seine Wirkung entfaltet es erst dadurch, dass das Subjekt des einen Tripels das Objekt eines anderen Tripels sein kann und umgekehrt. Damit kann man dann schon recht komplexe Zusammenhänge ausdrücken. Wenn man sich wieder ein Formular vorstellt, so kann man das nun z.B. darstellen in der Form:

```
Müller – ist – der Name,
Erlangen – ist – der Geburtsort,
17. Mai 1965 – ist – der Geburtstag
```

Damit kann man das ganze Formular mit allen seinen Daten beschreiben. Für Tabellen geht das analog. Und man kann das fortsetzen mit:

```
Erlangen – ist – eine Stadt
```

So ergibt sich ein immer größeres Netz von Aussagen, das man als Graph darstellen kann: Subjekte und Objekte sind die Knoten, Prädikate die Kanten zwischen ihnen. In Anlehnung an das WWW hat Tim Berners-Lee das „Giant Global Graph" (GGG) genannt, aber das hat sich nicht durchgesetzt.

Der Oberbegriff für Subjekt, Prädikat und Objekt ist Ressource, daher die Bezeichnung RDF. Nun ist die Frage, was alles als eine solche Ressource verwendet werden darf. Wenn man einfach jedes beliebige Wort zulässt, ist wenig gewonnen, denn dann werden die unterschiedlichen Verwendungen dieses Worts zu unterschiedlichen und wahrscheinlich sogar widersprüchlichen Aussagen führen. Jede Ressource muss also eindeu-

tig benannt werden. Dabei schadet es nicht, wenn eine Ressource mehrere Namen hat (Synonyme), denn das kann in RDF problemlos konstatiert werden:

```
Name1 – meint dasselbe wie – Name2
```

Homonyme darf es aber nicht geben, da muss ggf. umbenannt oder präziser benannt werden.

Erfreulicherweise stellt das WWW auch dafür schon einen Mechanismus bereit, denn auch Web-Seiten müssen einen weltweit eindeutigen Namen haben. Das sind die sog. „Uniform Resource Identifiers" (URIs), deren bekannteste Form der „Uniform Resource Locator" (URL) ist. Die sind hierarchisch aufgebaut, damit man sie auch dezentral vergeben kann. Eine URI kann auf eine Web-Seite führen, die das benannte Konzept genauer erläutert, muss es aber nicht.

Somit sind nun die drei Ressourcen in einem RDF-Tripel grundsätzlich URIs. Die einzige Ausnahme macht das Objekt, das auch ein einfaches Literal sein kann, also eine Zahl, ein Wort oder auch ein Satz:

```
Preis-URI – ist-URI – 15 Euro
```

Ein Beispiel aus dem Buch von Hitzler et al.[2] zeigt den Einsatz von richtigen URIs:

```
[Subjekt: „http://example.org/SemanticWeb",
Prädikat: „http://example.org/VerlegtBei",
Objekt: „http://www.springer.com/Verlag"]
```

Und nun auch noch mit einem Literal (einem Wert) als Objekt:

```
[Subjekt „http://www.springer.com/Verlag",
Prädikat: „http://example.org/Name",
Objekt: „Springer-Verlag"]
```

Wie man deutlich sieht, wird das Hinschreiben dadurch zwar präziser, aber auch etwas mühsam. Zur Vereinfachung hat man die sog. *Turtle-Syntax* entwickelt. Sie setzt die URIs in spitze Klammern und die Literale in Anführungszeichen. Am Ende steht ein Punkt.

```
<http://www.springer.com/Verlag> <http://example.org/Name> „Springer-
Verlag" .
```

Der Punkt ist keine Spielerei, sondern erlaubt die Abkürzung bei wiederholt vorkommenden Teilen. Oft will man zu ein und demselben Subjekt mehrere Aussagen machen. Dann verwendet man statt des Punkts ein Semikolon, dem dann nur noch Prädikat und Objekt folgen. Und man darf immer wieder vorkommende Teile von URIs als Präfix deklarieren und mit einem Kürzel versehen:

```
@prefix ex: <http://example.org/>
@prefix springer: <http://springer.com/>
ex:SemanticWeb ex:VerlegtBei springer:Verlag ;
        ex:Titel „SemanticWeb – Grundlagen“ .
springer:Verlag ex:Name „Springer-Verlag“ .
```

So etwas kann man mit ein wenig Übung auch als Mensch verstehen. Zugleich kann es problemlos von Computern verarbeitet werden.

Für die Speicherung großer Mengen von Tripeln bietet sich eine Tabelle mit vielen Zeilen an:

Subjekt	Prädikat	Objekt
ex:SemanticWeb	ex:VerlegtBei	springer:Verlag
ex:SemanticWeb	ex:Titel	„SemanticWeb – Grundlagen“
springer:Verlag	ex:Name	„Springer-Verlag“

Solche Strukturen kann man problemlos in Datenbanken abspeichern – und dann sehr flexibel auswerten! SPARQL[5] ist eine Sprache, in der man diese Auswertungen formulieren kann. Es würde zu weit führen, hier auch nur die wichtigsten Bestandteile von SPARQL zu beschreiben. Ich versuche es mit einem Beispiel für eine Anfrage:

```
PREFIX abc: <http://example.com/exampleOntology#>
SELECT ?capital ?country
WHERE { ?x abc:cityname ?capital ;
    abc:isCapitalOf ?y .
 ?y abc:countryname ?country ;
    abc:isInContinent abc:Africa . }
```

Die Präfix-Definition wurde oben schon eingeführt; sie hat nur eine leicht andere Syntax. In der SELECT-Zeile schreibt man die Ausgaben, die man erhalten möchte, hier also Paare von einer Hauptstadt und einem Land. Die hinter WHERE stehenden Zeilen definieren Muster für Tripel. Nur Tripel, die zu einem der Muster passen, kommen

5 SPARQL Protocol And RDF Query Language.

überhaupt in Frage. Ressourcen, die in einem Tripel-Muster mit einem Fragezeichen beginnen, sind nicht vorgegeben. Was immer ein Tripel an dieser Stelle enthält, passt. Ressourcen ohne Fragezeichen dagegen müssen genau so im Tripel enthalten sein. Hinter dem Fragezeichen steht dann noch eine Bezeichnung. Kommt die zusammen mit dem Fragezeichen noch in einem anderen Muster vor, müssen zwei Tripel an dieser Stelle in beiden Fällen dieselbe Ressource benennen. Wenn man es in normale Sprache übersetzt, sucht die oben genannte Anfrage nach einer Stadt („?x"), die die Hauptstadt von „?y" ist, und „?y" muss in Afrika liegen. Die Namen dieser Stadt und ihres Landes sollen ausgegeben werden.

RDF Schema

In der Nutzung von RDF trat sehr bald ein weiterer Bedarf zutage, nämlich die Festlegung einer gemeinsamen Struktur für eine größere Menge von Ressourcen. Dazu kann ich noch einmal das Beispiel des Formulars bemühen: Darin sind Felder vorgesehen, die man gern in jedem Fall mit Inhalten gefüllt haben würde („Pflichtfelder"). So wie RDF bisher beschrieben wurde, kann man aber nicht sicherstellen, dass bestimmte Tripel in Abhängigkeit von anderen immer vorhanden sein müssen („Jeder Angestellte muss eine Personalnummer haben!")[6]. Außerdem hat das Formular bei allen Angestellten die gleiche Struktur, es wiederholt sich. Auch das kann man nicht sicherstellen. Man muss im Moment noch bei jedem Angestellten die ganze Struktur wieder anlegen und darf dabei nichts vergessen[7]. Dass das dann für die Gesamtmenge aller Angestellten auch wirklich zusammenpasst, muss man sehr mühsam selbst kontrollieren.

Mit einigen zusätzlichen definitorischen Tripeln kann das aber auch das verarbeitende System übernehmen. Genau dafür wurde *RDF Schema* entwickelt. Es umfasst Ressourcen mit einer auf diese Situation zugeschnittenen Bedeutung. Auch hier ziehe ich wieder ein Beispiel heran, um das zu erläutern[8]:

```
<ex:SemanticWeb> <rdf:type> <ex:Lehrbuch> .
<ex:Lehrbuch> <rdf:type> <rdfs:Class> .
<ex:Buch> <rdf:type> <rdfs:Class> .
<ex:Lehrbuch> <rdfs:subClassOf> <ex:Buch> .
```

6 In Datenbanken geht das: Man deklariert ein Feld als NOT NULL, also nicht leer zu lassen.

7 Natürlich kann es auch Felder geben, die in einem speziellen Fall einfach nicht ausfüllen kann, z.B. „Geburtsname". Dann darf man das zugehörige Tripel durchaus weglassen. Nur bei Pflichtfeldern darf man das nicht.

8 Wenn in einem Bezeichner einer Ressource ein Doppelpunkt auftaucht, dann ist das entweder – wie oben erläutert – ein Präfix oder ein sog. *Namensraum*. Was man damit alles machen kann, ist in der einschlägigen Literatur nachzulesen, z.B. in Hitzler et al., a.a.O.

Das Prädikat „rdf:type“ sagt aus, dass das Subjekt vom dem Typ ist, der als Objekt genannt wird. Damit kann man ausdrücken, dass Herr Müller ein Angestellter ist[9]. Damit ist aber noch nicht viel gewonnen. Erst das zweite Tripel sorgt für die gewünschte Einheitlichkeit: Damit wird ausgesagt, dass es sich bei „Lehrbuch“ um eine Klasse handelt, und das bedeutet, dass dazu die Prädikate definiert sind, die ein „Lehrbuch“ haben kann bzw. muss, also z.B. „ist der Autor von“, „ist der Titel von“, „ist der Verlag von“ usw. Im dritten Tripel wird „Buch“ als eine zweite Klasse eingeführt, weil man offenbar nicht nur etwas über Lehrbücher aussagen möchte, sondern auch über Bücher im Allgemeinen. Diese Klasse verfügt ebenfalls über eine Menge von Prädikaten. Und weil „Buch“ und „Lehrbuch“ etwas miteinander zu tun haben – Lehrbücher sind ein Spezialfall von Büchern –, kann man auch das im vierten Tripel noch definieren. Was genau „rdfs:subClassOf“ bedeutet und welche Konsequenzen es hat, kann hier aus Platzgründen nicht weiter erläutert werden[10].

Weitere Elemente von RDF Schema sind die Klassen:

```
Class, Resource, Property, Literal
```

Sie können als Objekt von „rdf:type“ verwendet werden und ein Subjekt charakterisieren. Analog die Eigenschaften:

```
subClassOf, subPropertyOf, domain, range
```

Sie werden als Prädikate verwendet und verknüpfen Klassen oder Eigenschaften.

Mit diesen Mitteln kann man schon einfache Ontologien, also Begriffssysteme definieren. Man beschreibt damit, was unter einem „Lehrbuch“ oder einem „Angestellten“ zu verstehen ist. Und man kann mögliche richtige Aussagen von ganz sicher falschen trennen:

```
<ex:HansMeier> <ex:hatAlter> „-127“ .
```

Das wiederspricht einer hoffentlich an anderer Stelle mit der Eigenschaft „rdfs:range“ gemachten Aussage, dass das Alter einer Person im Bereich von 0 bis 150 liegen muss. Auch werden erste Schlussfolgerungen möglich. Aber es ist noch nicht umfassend genug, und deshalb sollen diese Dinge erst im folgenden Abschnitt ausgeführt werden.

9 Er ist dann „vom Typ“ Angestellter oder „hat den Typ“ Angestellter.

10 Wer sich mit objektorientierter Programmierung auskennt, weiß, dass es sich um eine Vererbungsbeziehung handelt.

OWL

Mit Hilfe der „Web Ontology Language“ (OWL)[11] ist dann die umfassende Definition von Taxonomien, also Begriffshierarchien möglich. Sie hat eine streng formale Semantik und erlaubt nun wirklich umfassende logische Schlussfolgerungen, speziell zur Entdekkung von Widersprüchen in einer großen Menge von Tripeln. Da die Syntax etwas umständlich ist, wird sie hier nicht gezeigt. Stattdessen werden umgangssprachliche Beispiele mit gleicher Aussage verwendet, die lesbarer sind.

Zunächst einmal gibt Klassen und Instanzen, wie es schon bei RDF Schema zu sehen war:

```
Julius „ist ein“ Mensch.
```

Und es können auch Subklassen definiert werden:

```
Alle Menschen „sind auch“ Lebewesen.
```

Den Klassen werden Eigenschaften (Properties) zugeordnet, die bei allen (!) ihren Instanzen vorhanden sein müssen:

```
Ein Mensch „hat“ einen Namen.
```

Ein Beispiel für die Verwendung von OWL ist CIDOC CRM[12], das einen Bezug zur Editionswissenschaft hat. Es definiert die Begriffe, die man benötigt, um (Museums-) Objekte eindeutig zu beschreiben. Das geschieht Ereignis-orientiert, d.h. die Beschreibung orientiert sich an Ereignissen wie der Herstellung eines Objekts (durch einen Künstler, einen Handwerker usw.), dem Fund, dem Kauf, der Aufstellung in einer Sammlung usw. Ein Beispiel:

```
<E12 Production> <P108 has-produced>
       <E24 Physical Man-Made Thing> .
```

11 Dean Allemang und Jim Hendler: *Semantic Web for the Working Ontologist: Effective Modeling in RDFS and OWL*. 2. Aufl. Morgan Kaufmann : Amsterdam a.o., 2011. – ISBN 978-0-12-385965-5. Zur Geschichte des Buchstabendrehers in der Abkürzung siehe auch http://lists.w3.org/Archives/Public/www-webont-wg/2001Dec/0169.html .

12 Das «Conceptual Reference Model» des Comité International pour la Documentation, heute: ICOM's International Committee for Documentation; siehe http://www.cidoc-crm.org/ . Das CRM definiert Klassen und Eigenschaften noch verbal, mit dem Erlangen CRM (http://erlangen-crm.org/) ist es in OWL ausformuliert worden.

Zum Ereignis „E12 Production“ ist noch eine ganze Reihe weiterer Eigenschaften definiert, ebenso wie für das Objekt „E24 Physical Man-Made Thing“. Nun ist das noch sehr abstrakt, weil es für alle möglichen (Museums-) Objekte verwendbar sein soll. Seine ganze Wirkung entfaltet es erst, wenn man die Zuordnung eigener Datenobjekte zu den Bestandteilen dieser Ontologie vornimmt:

```
“Mona Lisa” <rdf:type> <E24 Physical Man-Made Thing> .
„Mittelalter“ <rdf:type> <E2 Temporal Entity> .
```

Das Ziel ist, dass nach und nach alle Museen zu ihren Sammlungen eine Beschreibung vornehmen, die diesem vorgegebenen Muster folgt. Bisher denken sie sich eher jeweils eigene Beschreibungen aus, was zu einem geradezu babylonischen Sprach-Wirrwarr geführt hat. Mit CIDOC CRM und OWL kann man hier zu deutlich mehr Einheitlichkeit, Vergleichbarkeit, Austauschbarkeit und Zusammenführung gelangen.

Linked (Open) Data

Mit den bis hierher eingeführten Methoden lässt sich die einheitliche Beschreibung der Struktur und sogar der Bedeutung von Daten – insbesondere auch von Web-Seiten – erreichen. Und danach kann man suchen, indem man SPARQL benutzt, siehe oben. Was nun noch fehlt ist die Nutzung der Infrastruktur des WWW zur Bereitstellung von (freien) strukturierten Daten[13]. RDF kann auch das leisten. Eine Tabelle mit Einwohnerzahlen[14]:

Stadt	*Einwohnerzahl*
München	1.456.039
Hamburg	1.834.823
Berlin	3.613.495
Köln	1.080.394

wird in RDF zum Beispiel dargestellt mit:

```
<München> <hat-Einwohner> 1456039 .
<Hamburg> <hat-Einwohner> 1834823 .
<Berlin>  <hat-Einwohner> 3613495 .
<Köln>    <hat-Einwohner> 1080394 .
```

13 Christian Bizer, Tom Heath und Tim Berners-Lee: Linked Data - The Story So Far. *Int. J. Semantic Web Inf. Syst.* 5(3), 2009, S. 1–22.

14 Alle Zahlen stammen aus Wikipedia (Abruf am 18. März 2019).

Das man da nun noch viel ergänzen kann und sollte, ist nach dem oben Gesagten klar: So sollte ausgedrückt werden, dass alle vier Subjekte eine „Stadt" sind, also Instanzen der Klasse „Stadt". Weiterhin sollte gesagt werden, dass alle „Städte" eine Einwohnerzahl haben und das die über die Eigenschaft „hat-Einwohner" zugeordnet ist. Das ist mit den oben nur skizzierten Möglichkeiten von RDF Schema und OWL problemlos machbar.

Die Bereitstellung solcher Daten ist Teil des Semantic Web; nur sind es sind dann eben keine Web-Seiten – obwohl man sich die ganzen Tripel durchaus auch ansehen könnte, aber da erkennt man nur wenig. Ziel ist es ja, diese Daten insgesamt oder in (durch SPARQL definierten) Ausschnitten herunterzuladen und in eigene Auswertungen einzubeziehen, sei es nun in einer Datenbank, in Excel, in R oder in Python – und was es an Auswertungswerkzeugen sonst noch so gibt.

Anbieter solcher Daten sind typischerweise Regierungen, Zeitungen, Rundfunk- und Fernsehanstalten, Museen und manchmal auch Wirtschaftsunternehmen. Es gibt inzwischen schon eine sehr große Menge solcher Daten-Angebote; die Abb. 1 kann dazu nur einen groben Überblick geben. Die verschiedenen Farben kennzeichnen die Gruppen von Anbietern, wie sie eben genannt wurden. Als Einstieg in diese Welt ist http://linkeddata.org zu empfehlen, oder auch https://lod-cloud.net/ . Hier kann man in den verschiedensten Datenbeständen herumstöbern und auch den Zugriff über SPARQL ausprobieren.

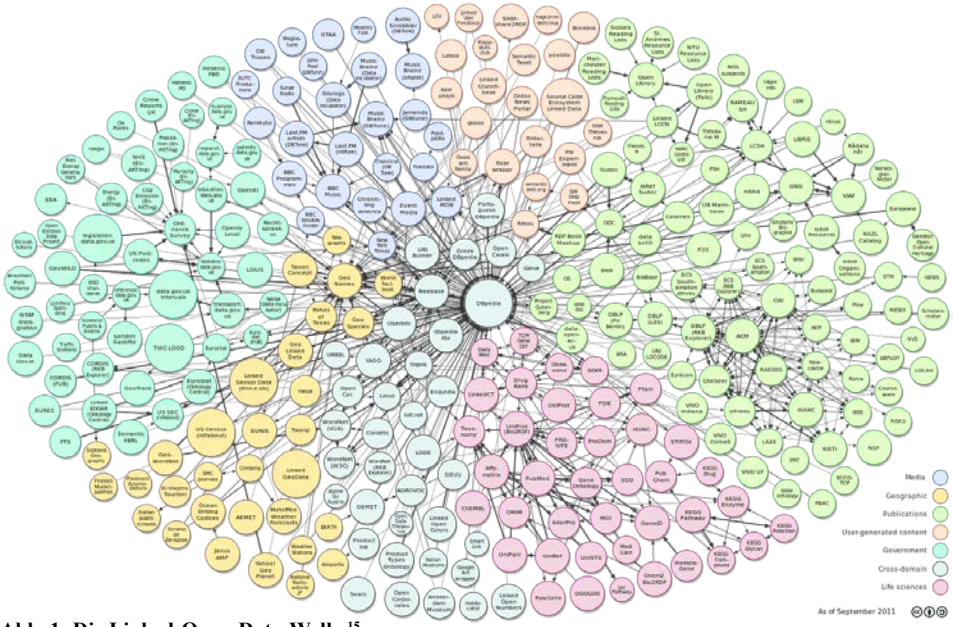

Abb. 1: Die Linked-Open-Data-Wolke[15]

15 Von Richard Cyganiak Anja Jentzsch - http://richard.cyganiak.de/2007/10/lod/, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=24597456.

Zum Schluss

Das Semantic Web ist nun schon etwa zwanzig Jahre alt. Es wird intensiv genutzt, wie gerade die zuletzt genannten Web-Seiten mit langen Listen von Einstiegspunkten demonstrieren. Dennoch ist das Potenzial bei weitem noch nicht ausgeschöpft. Was man noch ein bisschen vermisst ist eine bequeme Art, aus schon vorhandenen Daten die ganzen Tripel zu erzeugen, die man in RDF für ihre Darstellung benötigt. Das will niemand von Hand machen! Bei Datenbanken gibt es solche Werkzeuge schon, aber nur ein kleiner Teil der Daten ist in Datenbanken gespeichert. Viel mehr liegt in Dateien (Excel …) oder in Textform vor. Es gibt erste, noch unzureichende Versuche, Text in RDF zu transformieren[16], aber da ist noch viel zu tun. Dennoch bleibt es für mich eine lohnende Herausforderung, dieses Ziel weiter zu verfolgen. Man stelle sich nur vor, was es bedeutete, wenn man die Bücher aus Bibliotheken in eine dann sicher riesige Menge von Tripeln transformieren könnte, die sich mit Schlussfolgerungen auswerten ließen.

16 Zum Beispiel: Brahim Batouche, Claire Gardent und Anne Monceaux: Parsing Text into RDF Graphs. In: Proc. SEPLN 2015, Actas del XXXI Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural, Alicante, 16-18 Sept 2015.

MIT GRAPHBASIERTER EDITION ZUR SEMANTISCHEN MULTIDIMENSIONALITÄT

Andreas Kuczera

Einleitung

Dieser Beitrag beschreibt Umfang und Funktionalität von "The Codex", einem Projekt, in dem die Möglichkeit zur mehrdimensionalen Annotation von Texten mit dem Management von Erschließungshierarchien vereint wird.[1] In relationalen oder XML-Datenbanken ist es oft schwer, Texte in ihrem direkten Kontext wahrzunehmen, da die narrative oder argumentative Struktur leicht verloren geht. Andererseits erleichtern sie statistische Auswertungen.

Auch XML ist ein leistungsstarkes Werkzeug für die Modellierung von Text, die beispielsweise mit XPATH ausgewertet werden können. Bei XML führt jedoch das Markup selbst zu Diskontinuitäten in Text und Annotationsdaten. Auch überlappende Annotationen, wie sie häufig bei der gleichzeitigen Auszeichnung von Layout und Semantik benötigt werden, können nicht direkt in der baumartigen hierarchischen Struktur von XML abgebildet werden. Dies erfordert Workarounds wie Standoff Markup, die wiederum die Lesbarkeit des XML-Dokuments beeinträchtigen.[2]

Codex zielt darauf ab, die Kluft zwischen überlappenden Annotationshierarchien und lesbarem Text zu überbrücken um eine „mehrdimensionale Integration" zu erreichen. Auf Markup wird völlig verzichtet. Stattdessen werden Standoff Properties zur Annotation verwendet. In den digitalen Geisteswissenschaften gibt es verschiedene Ansätze für Standoff Properties.[3] Codex verfolgt hier einen Ansatz, bei dem der Nutzer einen Text in Echtzeit annotieren kann und diese Annotationen frei überlappen können. In der Grundkonfiguration bietet Codex eine Auswahl von Stil-, Layout- und semantischen Annotationstypen. Es können auch Fußnoten, Marginalien usw. innerhalb des Editors hinzugefügt

1 Der vorliegende Beitrag ist eine überarbeitete Übersetzung eines Beitrages des Sonderbandes 4 der ZfdG, Die Modellierung des Zweifels – Schlüsselideen und -konzepte zur graphbasierten Modellierung von Unsicherheiten. Herausgegeben von Andreas Kuczera, Thorsten Wübbena, Thomas Kollatz mit dem Titel The Codex.net -- An Atlas of History, von Iian Neill and Andreas Kuczera.

2 Georg Vogeler, Patrick Sahle, XML, in: Fotis Jannidis/Hubertus Koh- le/Malte Rehbein (Hg.), Digital Humanities, Eine Einführung, Stuttgart 2017, S. 128-146, hier S. 134.

3 Zu Standoff-Markup im Rahmen der TEI vgl. https://wiki.tei-c.org/index.php/Stand-off_markup zuletzt abgerufen am 6.2.2019).

Abbildung 1. Stil-, Layout- und semantische Annotationstasten im Codex-Editor.

und annotiert werden. Annotationen selbst können in der browserbasierten, graphischen Oberfläche mit dem Erschließungsgraph verknüpft und selbst editiert und ergänzt werden.

Eines der Ziele von Codex ist es, über die Annotation von Entitäten und Ereignissen und deren Verknüpfung in den verschiedensten Quellen einen „Atlas von Beziehungen" zu erstellen. Dies ist mehr als eine Metapher für die Art der Annotationen und der Visualisierung ihrer Verbindungen: Atlas bezieht sich auf ein Netzwerk von Beziehungen, die auf verschiedene Weise projizierbar sind. Ziel ist es also nicht nur Entitäten im Text zu annotieren, sondern diese Entitäten als Knoten und Kanten in eine erschließende Graphstruktur einzubinden. Das Ergebnis ist sowohl ein annotiertes Dokument als auch eine Graphdatenbank, in der die Annotation in ihrem semantischen Kontext eingebettet ist.

Annotation

Es gibt momentan drei Hauptkategorien von Annotationen: Stil, Layout und Semantik. Beim Stil werden häufig verwendete typografische Stile wie z.B. Kursivschrift, Fettdruck, Unterstreichung, Durchstreichung, Hochstellung etc. angeboten. Dies ist aber komplett durch den Nutzer konfigurier- und erweiterbar.

Die interessanteren Kategorien sind jedoch Layout und Semantik. Diese gliedern sich wie folgt.

Layout

Seite, Absatz, Zeile, Satz, Spalte, etc.

Diese Annotationen beschreiben das Layout einer Seite in einem Manuskript. Da sich Standoff Properties frei überlappen können, stellen Überschneidungen, wie z.B. Absätze, die über einen Seitenwechsel hinweg reichen, kein Problem dar.

Trenn- und Bindestriche als Zero Point Annotation (ZPO)

Trenn- und Bindestriche stellen eine Herausforderung für die Kommentierung von Manuskripten dar; während der Editor die Position des Bindestrichs festhalten möchte, gibt es oft gute Gründe, Trennstriche bei der Weiterverarbeitung von Texten nicht zu berücksichtigen. Die Bindestrich-Annotation in Codex löst dieses Problem, indem sie die

Trennstriche als Zero Point Annotationen darstellt. Eine "nulldimensionale" Annotation ist ein Sonderfall einer Standoff-Property, die einen Anfangsindex, aber keinen Endindex hat. Auf diese Weise bezieht sich eine Annotation effektiv auf eine Position im Text zwischen den Zeichen. Der Bindestrich selbst wird nicht im Text gespeichert, die Worte bleiben unberührt. Im Editor wird die Annotation aber an der Position des Trennstrichs im Original in rot angezeigt. Zero Point Annotations sind ein verallgemeinerbares Merkmal von Standoff Properties und können auch für andere Fälle verwendet werden. Festzuhalten ist, dass beim Export ausgewählt werden kann, ob Trennstriche mit exportiert oder unterdrückt werden sollen.

Semantik

In Codex gibt es für jede semantische Annotation eine entsprechende semantische Einheit in der Graphdatenbank. Die verfügbaren Typen von Annotationen und Entitäten sind dabei frei konfigurierbar und werden nicht durch einen bestehenden Standard definiert. So kann das System in jedem Projektzusammenhang nach Wunsch konfiguriert werden. Jede Entität wird als eine Kombination von Knoten und Kanten in der Graphdatenbank modelliert. In Codex sind die für die Annotationen verwendeten Entitäten eigenständige Datensätze und sind damit unabhängig vom edierten Text.

Agents

Ein Agent bezieht sich auf jede Art von Entität, die im Text erwähnt wird. Dies können z.B. Personen, Orte, Objekte, Kollektive, Organisationen, Familien und andere Gruppen sein. Im Agentknoten selbst wird weiter differenziert, ob es sich um eine Person, einen Ort oder eine Gruppe etc. handelt. Metadaten über Agents sind ebenso in den Agentknoten gespeichert. Beispiele solcher Eigenschaften sind das Geschlecht einer Person, ihre Größe, ihr Gewicht etc.

Agents können auch über dynamisch erzeugte Beziehungen miteinander verbunden werden. (In Codex werden diese Meta-Relations genannt). Das folgende Beispiel zeigt einige der genealogischen Beziehungen von Lorenzo de' Medici, seine Verbindung zu Agents, die Kollektive darstellen sowie seine Anwesenheit in einer Gruppe von sechs florentinischen Botschaftern in Rom 1471.

RELATIONSHIP ID	VALUE	PROPERTY	TIME
3a79992e-ef90-4ecf-9f0c-6097931725bd	Male	Gender	

Abbildung 2. Das Geschlecht von Lorenzo de' Medici als Property.

RELATIONSHIP ID	RELATION	AGENT 2
2e0a3d13-8c5a-4762-903f-2d83df11ca2b	Part of	Six ambassadors [1471/09/23]
46a56b46-f962-4735-ba82-b7ff7dae6922	Brother of	Giuliano de' Medici
980feab9-ed6f-4b7d-b13d-0518333e1283	Part of	The Medici family
993428e1-efb9-43b3-a299-7ba23460f730	Child of	Cosimo de' Medici
cb1ad5d7-e02e-4393-8ca9-2ffa997b607e	Child of	Madonna Lucrezia [de' Medici]
fbc97384-90df-4a2e-8fbb-d0917cca45dd	Part of	Three Florentine ambassadors [1483/11/10]
2afb8754-1705-456a-ae32-97087991fe27	Married to	Madonna Clarice
00700f9b-5369-4a5e-a426-d1dead67d710	Parent of	Giovanni de' Medici

Abbildung 3. Eine Liste der Beziehungen von Lorenzo de‘ Medici.

Claim / Statement

Ein Claim bezieht sich auf eine Aussage über einen oder mehrere Agenten, in der Regel an einem Ort und zu einem Zeitpunkt. Ein Claim ist im Wesentlichen eine Aussage, die in der Regel im Rahmen eines Events erfolgt, aber auch einen Gedanken oder eine Meinung darstellen kann. Ein Claim in Codex wird nicht als Tatsachenaussage verstanden, sondern ist eine Datenstruktur, die einem RDF-Statement ähnelt. So ist beispielsweise die Aussage, dass „Lorenzo de‘ Medici auf seinem Anwesen in Careggi gestorben ist“, die Luca Landucci in seinem Tagebucheintrag vom 8. April 1492 gemacht hat, in Codex als

```
(Subject) Lorenzo de Medici
(Event) died
(At) Careggi
```

modelliert (Siehe Abbildungen 4 und 5). In Codex werden also “Tatsachen” als Statements gespeichert.

Texte

Eine Text-Einheit in Codex besteht aus Plaintext und einer Sammlung von Standoff-Properties. Der Einfachheit halber können Texte mit einem „Typ“ versehen werden, der

	ROLE	TYPE	NAME	AGENTS	TIME
View	Event	died	Lorenzo de' Medici died on his estate at Careggi	(According To) Luca Landucci (b.1436 - d. 1516; apothecary) (At) Careggi (Subject) Lorenzo de' Medici (b. 1449/01/01 - d. 1492/05/08; active from 1469/12/02)	On 1492/April/8 -:-:-

Abbildung 4. Das Statement von Luca Landucci über den Tod von Lorenzo de' Medici in Codex.

ihre Funktion angibt (z.B. „`Body`", „`Footnote`", „`Marginnote`", etc.). Dabei gibt es keine Einschränkungen hinsichtlich der Art des gespeicherten Textes. Im Falle von Luca Landuccis Tagebuch wird jeder Tagebucheintrag (von dem es mehrere pro Seite in der Quelle gibt) in einem separaten Textknoten gespeichert, während jeder Michelangelo-Brief als Ganzes in einem Textknoten gespeichert wird. Wird ein Text kommentiert, ist der Kommentartext wieder annotierbar, so dass der Kommentar selbst auch wieder annotiert werden kann. Die Annotationskette kann beliebig weitergeführt werden.

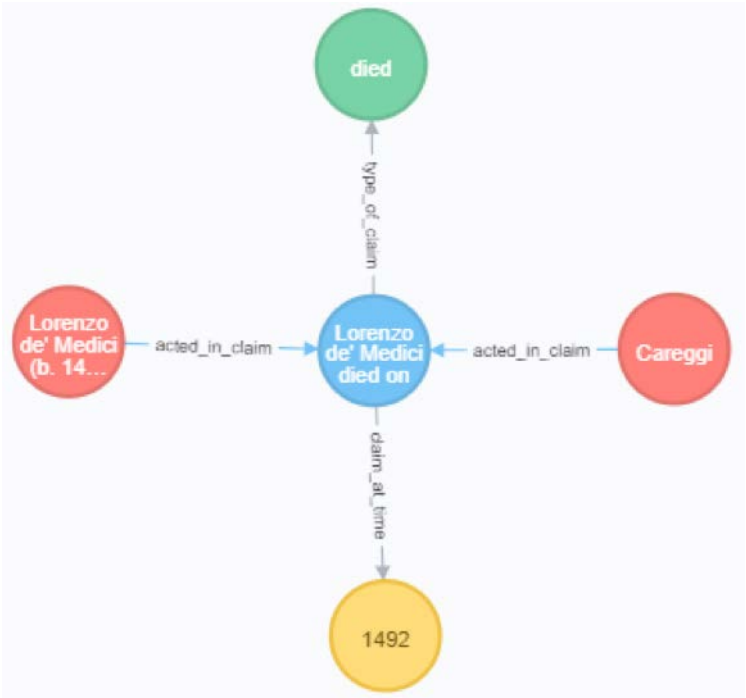

Abbildung 5. Eine Darstellung des obigen Statements als Knoten und Kanten im Neo4j-Browser. Der blaue Knoten ist ein Claim, die roten Knoten sind Agents, der grüne Knoten ist ein Concept und der gelbe Knoten ist die Zeitangabe.

Meta-Relations

Eine Meta-Relation ist eine Beziehung zwischen Agents mit zwei besonderen Merkmalen:

Meta-Relations sind in Codex dynamisch definierbar. Nach unseren Erfahrungen ist es von Vorteil Beziehungstypen anfangs nicht fest vorzugeben. Wenn der Benutzer hier frei ist, wird die spontane Erstellung von Beziehungstypen gefördert, man ist näher am Forschungszusammenhang. Dies schließt eine später Neu- und Umstrukturierung der Erschließungshierarchien nicht aus.

Meta-Relations sind bidirektional, d.h. der Benutzer kann beide Richtungen der Beziehung unterschiedlich benennen (z.B. Elternteil von / Kind von). Dies erleichtert das Denken in generischen Beziehung, z.B. „Abstammung". Ein Vorteil bei Cypher-Abfragen besteht darin, dass die Verbindung eines Agents zu Meta-Relation gefunden werden kann, ohne seine Rolle in der Beziehung kennen zu müssen.

Meta-Relations sind innerhalb einer Hierarchie modellierbar, so dass die Beziehungen selbst ein Graph sind. Beispielsweise kann man einen übergreifenden Beziehungstyp wie „zwischenmenschliche Beziehungen" definieren und darunter untergeordnete Typen wie „soziale Beziehungen", „Familienbeziehungen", „berufliche Beziehungen" usw. schachteln, so dass man Beziehungen zwischen Personen auf einer abstrakten Ebene abfragen kann. So kann eine Suche nach den „Freunden" einer Person leicht um „Mitarbeiter", „Bekannte" oder „Vertraute" ergänzt werden.

Eine Meta-Relation ist also eine Annotation, die in der Graphdatenbank durch einen Meta-Relation-Knoten dargestellt wird. Damit wird es möglich, Beziehungen von Agents in Texten zu kommentieren. In Abbildung 6 ist die orange Linie unter „Sohn von Antonio" eine Meta-Relation, die angibt, dass Luca Landucci der Sohn von Antonio Landucci war.

Concepts

Ein Concept in Codex ist eine Klasse oder ein Typ, der als Ganzes gesehen ein Konzept in der gemeinsamen Ontologie des Systems darstellt. Zu beachten ist, dass Ontologie hier nicht im Sinne einer „universellen" oder „Welt"-Ontologie zu verstehen ist, sondern lediglich als Subgraph, der von anderen Entitäten (wie Agents, Claims, Meta-Relations

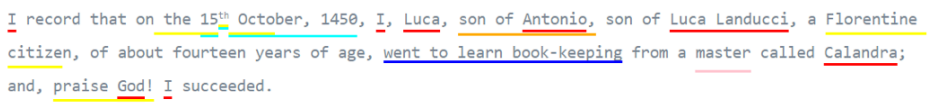

Abbildung 6. Beispiel für eine Meta-Relations-Annotation (orange unterlegt).

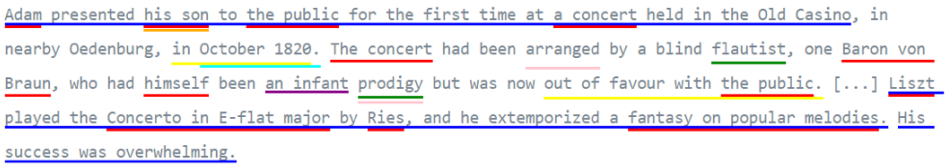

Abbildung 7. Konzeptkommentare (dunkelgrün) aus Alan Walkers Biographie von Franz Liszt.

usw.) im Sinne von gemeinsamen, wiederverwendbaren Konzepten geteilt wird. Anstatt eine universelle, top-down Ontologie zu verwenden, können die Nutzer sie bei Bedarf selbst definieren. Codex enthält bereits eine Reihe von Ontologien, wie z.B. Ontologien für Arten von Ereignissen, Orte, Beziehungen, Berufe. Das Concept ‚Event' ist der Wurzelknoten der Event-Ontologie und enthält alle Arten (und Subtypen) von Events, die in der Projektdomäne auftreten. Concepts können mehrere übergeordnete Elemente haben da ein Graph nicht an die Einschränkungen eines Baums gebunden ist. Die Änderung der Struktur der Ontologie (z.B. das Verschieben eines Kindkonzepts auf einen anderen Elternteil) ist in Codex einfach möglich, so dass ontologische Strukturen flexibel gehalten werden, um dem sich entwickelnden Verständnis der Projektdomäne gerecht werden zu können.

In Abbildung 7 stellen die dunkelgrünen Unterstriche das Konzept für „Flötist" und „Wunderkind" dar.

Datensätze und Datenpunkte

Ein Datenpunkt in Codex ist definiert als ein Maß in Raum und Zeit. Ein Datensatz ist wiederum eine Sammlung von Datenpunkten. Im folgenden Beispiel ist die Aussage, dass „drei Menschen an diesem Tag gestorben sind" als Datensatz der Datenpunkte „3", „Menschen", den Ort „Florenz" und die Zeit „23. April 1483" (Datum des Tagebucheintrags von Landucci) zusammenfasst.

Die Idee eines Datenpunktes besteht darin, es dem Editor zu ermöglichen, schnell numerische Daten aus einem Text zu extrahieren, die von statistischem Interesse sein können.

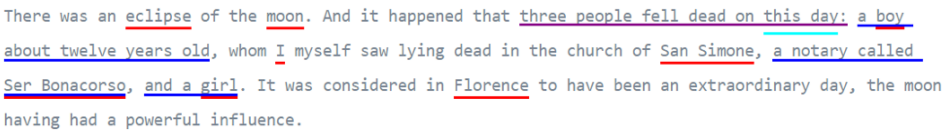

Abbildung 8. Datenpunkte (dunkelviolette Unterstreichung) in einem Tagebucheintrag von Landucci.

Abbildung 9. Das Eingabefenster für eine Zeiteinheit.

Einige praktische Beispiele für Datensätze, die aus historischen Quellen extrahiert werden können, sind epidemiologische Daten, Wetteraufzeichnungen, Volkszählungen, Kriminalitätsstatistiken usw.

Zeit

Die Modellierung von Zeitangaben ist in unterschiedlicher Genauigkeit möglich. Die Zeit-Entität besteht aus 9 Komponenten, die alle optional sind. Die möglichen Angaben umfassen `on`, `before`, `after`, `circa`, `early` und `late`. Die Jahreszeiten sind durch `Winter`, `Summer`, `Autumn`, `Spring` vertreten. Die Vielfalt der Optionen soll die Realitäten der Datumsangaben in historischen Texten widerspiegeln. Es ist geplant, die W3C Time Ontology als Grundlage hierfür zu nutzen.

In Codex ist es möglich, implizite Datumsangaben explizit zu machen. Die Zeitangabe in Abbildung 8 (Unterstreichung in türkis) mit dem Text „this day" wird mit dem angegebenen Datum des Tagebucheintrags (23. April 1483) verknüpft.

Nach dem Überblick zu den wichtigsten Annotationstypen wird nun das Standoff-Property-Modell vorgestellt. Es bildet die Grundlage um mit Codex Text-as-Graph editierbar zu machen.

Standoff-Properties

Eine Wortkette ist ein Graphmodell eines Textes, bei dem jedes Wort als Token-Knoten behandelt wird und die Serialität des Textes durch die Zuordnung jedes Knotens zu seinem Nachbarn mit einer NEXT_TOKEN-Kante modelliert wird.

Wie Standoff Properties stellt Text als Kette von Wortknoten eine markupfreie Alternative zu XML-Dokumentenformaten dar und und beherrschen die Modellierung von überlappenden Annotationshierarchien. Momentan fehlen noch einfache Nutzoberflächen für das direkte Management von Text-as-a-Graph-Modellen durch den Nutzer. Daher wird in Codex das Text-as-a-Graph-Konzept mit Hilfe eines Standoff-Property-Editors umgesetzt. Dies ist ein Kompromiss zwischen den multidimensionalen Möglichkeiten

Abbildung 10. Text als Kette von Wortknoten.

des Graphen und der technischen Robustheit und Nachhaltigkeit des Standoff-Property-Formats.

Die Entfernung von eingebettetem Markup macht den Textstrom sowohl für Menschen als auch für Maschinen leicht lesbar. Es löst auch das Problem der überlappenden Annotation, da die Eigenschaften getrennt vom Text gespeichert werden und keine hierarchischen Kodierungskonflikte entstehen. Mehrere Eigenschaften können sich über ihre Start- und Endindexe auf die gleichen Textbereiche oder überlappende Bereiche beziehen. Standoff-Properties sind von Natur aus diskrete Objekte, die in einer „flachen" Hierarchie koexistieren, d.h. ohne vorgeschriebene Hierarchie. Wird eine Annotation als Standoff-Property erstellt, kann sie unmittelbar mit dem Erschließungsgraphen verknüpft werden. Dies stellt die direkte Verbindung von Entitäten in der Graphdatenbank zu den entsprechenden Textbereichen her.

Eine Standoff-Property ist also im wesentlichen eine Datenstruktur mit folgenden Properties::

`Typ`: Eine Zeichenkette, die den Namen (d.h. den Typ) der Annotation angibt.

`StartIndex`: Eine ganze Zahl, die die Indexposition des ersten Zeichens der Annotation darstellt: 0 <= x < n, wobei x der Index ist und n die Länge des Textes ist.

`EndIndex`: Eine ganze Zahl, die die Indexposition des letzten Zeichens der Annotation innerhalb der Länge darstellt (gleiche Regel wie der StartIndex).

`Wert`: Eine Zeichenkette, die Daten darstellt, die spezifisch für die Annotation sind, wie beispielsweise die eindeutige Kennung einer referenzierten Entität, oder alternativ einen Farbwert, eine Textgröße, eine Schriftart usw.

`GUID`: Eine 32-stellige Zeichenkette, die als eindeutiger Identifikator für die Standoff-Property dient. Dies ist erforderlich, um sie in der Datenbank zu speichern.

`UserGUID`: Eine `GUID` (siehe oben), die den Benutzer repräsentiert, der die Annotation erstellt hat.

`Index`: Eine optionale ganze Zahl, die die Reihenfolge angibt, in der die Standoff-Propertys erstellt wurde.

`Text`: Eine optionale Zeichenkette, die den Quelltext darstellt, auf den sich die Annotation bezieht. Dies ist ein optionales Attribut, um die Anzeige von Standoff-Properties in der Datenbank zu erleichtern.

`Layer`: Eine optionale Zeichenkette, die beliebige Gruppen von Annotationstypen (Layer) repräsentiert (z.B. Layout, Struktur, Semantik).

`IsZeroPoint`: Ein boolescher Wert, der angibt, ob die Standoff-Property eine Zero-Point-Annotation ist.

`IsDeleted`: Ein boolescher Wert, der angibt, ob die Standoff-Eigenschaft als gelöscht markiert wurde.

Die bisher üblichen Standoff-Markup-Formate können in der Regel nicht mit der nachträglichen Änderung des Datums (des annotierten Textes) umgehen. Im Codex

Standoff-Property-Editor wird daher der Text als NodeList von HTML-SPAN-Elementen dargestellt, die über Referenzzeiger mit einem Array von JavaScript-Objekten mit den Standoff-Properties verbunden sind. Da eine verknüpfte Liste verwendet wird, um den Text darzustellen, und die Standoff-Properties mit dem Text mit Zeigern und nicht mit Indizes verknüpft sind, können Zeichen frei zu dem Text hinzugefügt oder entfernt werden, ohne dass Indexe während der Bearbeitung neu berechnet werden müssen. Erst wenn der Benutzer die Daten aus dem Editor exportiert (z.B. zum Speichern), werden die dynamischen Zeiger in statische Indexnummern umgewandelt. Der Plaintext und die Properties werden als JSON-Objekt exportiert, das in eine Textdatei gespeichert oder in ein Datenspeicherformat nach Wahl des Nutzers konvertiert werden kann. Codex übersetzt den JSON-Export in Standoff-Property-Knoten, die in einer Graphdatenbank gespeichert werden.

Beispiel

Nachfolgend ein Auszug aus dem JSON-Export des in Abb. 11 dargestellten Textes. Die gelben Hervorhebungen zeigen die typischen Teile der Standoff Properties. Die orangefarbenen Texte wiederholen den jeweils annotierten Text. Der Plaintext ist hellblau unterlegt.

Das volle Potenzial des Standoff-Property-Modells für die Integration von Text- und Graphstrukturen wird, abgesehen von der guten Lesbarkeit des Plaintextes ohne Mar-

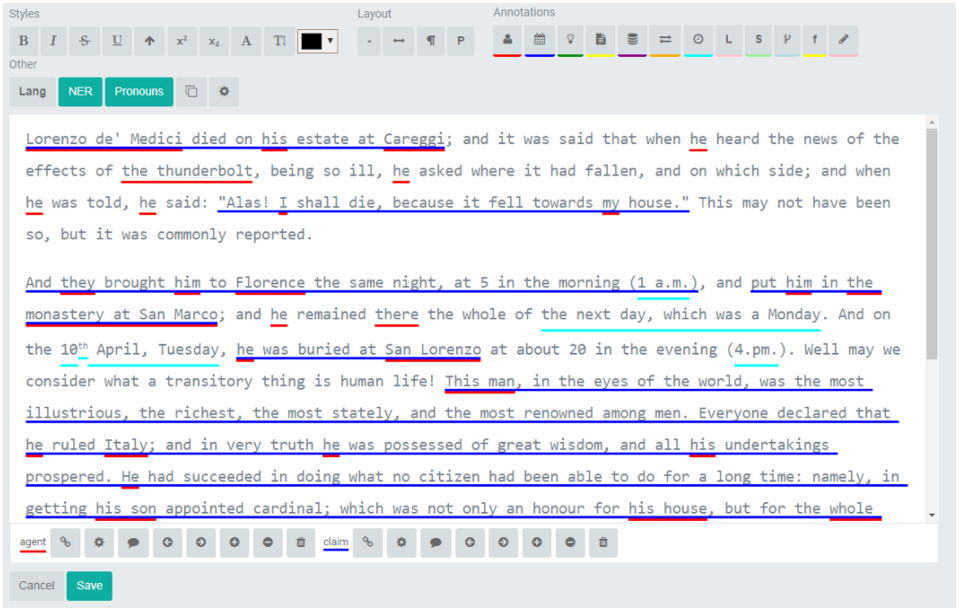

Abbildung 11. Luca Landucci's Tagebucheintrag vom 8. April 1492 über den Tod von Lorenzo de' Medici.

```
{
 “text”: “Lorenzo de’ Medici died on his estate at Careggi; and it was
said that when he heard the news of the effects of the thunderbolt,
being so ill, he asked where it had fallen, and on which side; ... “,
 “properties”: [{
       “index”: 23,
       “guid”: “cde24f38-81cb-4110-b368-b5b1f4ed4d53”,
       “type”: “agent”,
       “layer”: null,
       “text”: “Lorenzo de’ Medici”,
       “value”: “e45fed44-17a0-4c2c-9c00-858667a17904”,
       “startIndex”: 0,
       “endIndex”: 17,
       “isZeroPoint”: false,
       “isDeleted”: false,
       “userGuid”: “fb067f75-a121-47c1-8767-99271c75cfc0”
 }, {
       “index”: 25,
       “guid”: “5e33d2a8-dea0-4bbf-b4f6-8ccf40dac4d9”,
       “type”: “agent”,
       “layer”: null,
       “text”: “Careggi”,
       “value”: “d8eda97c-79d7-43ad-b43f-fd3e3a05c68e”,
       “startIndex”: 41,
       “endIndex”: 47,
       “isZeroPoint”: false,
       “isDeleted”: false,
       “userGuid”: “fb067f75-a121-47c1-8767-99271c75cfc0”
 }, {
       “index”: 46,
       “guid”: “94eebe67-0136-4f54-a4c6-6e7072dfb3de”,
       “type”: “claim”,
       “layer”: null,
       “text”: “Lorenzo de’ Medici died on his estate at Careggi”,
       “value”: “175742e2-a342-455d-a1b9-3fe3a96e3fd9”,
       “startIndex”: 0,
       “endIndex”: 47,
       “isZeroPoint”: false,
       “isDeleted”: false,
       “userGuid”: “fb067f75-a121-47c1-8767-99271c75cfc0”
 }]
}
```


Abbildung 12. Auszug aus dem JSON-Export von Landucci‘s Tagebucheintrag.

kup, den überlappenden Annotationen mit ihren direkten Verknüpfungen zu präzisen Textstellen vor allem durch zwei Merkmale klar:

Standoff-Properties können in sogenannten Layern gruppiert werden, wobei ein Layer entweder implizit durch den Annotationstyp oder explizit durch einen im Layer-Attribut gespeicherten Wert definiert wird.

Die Attribute Startindex und Endindex bieten die Möglichkeit, Annotationen zu kombinieren, die im gleichen Textbereich enthalten sind oder sich überlappen.

Diese Möglichkeiten zur Schichtung und Kombination haben bereits zu interessanten Funktionen im Codex-Editor geführt (z.B. bei der Verwaltung von Named Entities und Pronomen-Annotationen). Sie bieten aber auch weitere Erkenntnisperspektiven. So wird es möglich über eine Cypher-Abfrage alle Annotationen in einem Text (oder in allen Texten) zu finden, die sich überlappen. Kombinationen von Annotationen können den von ihnen kommentierten Text wiederum mit neuen Bedeutungen anreichern.

Schließlich ist es mit Standoff-Properties möglich, verschiedene Annotationskonzepte in einem System zu vereinen, z.B. Layout, Semantik und Syntaxanalyse.

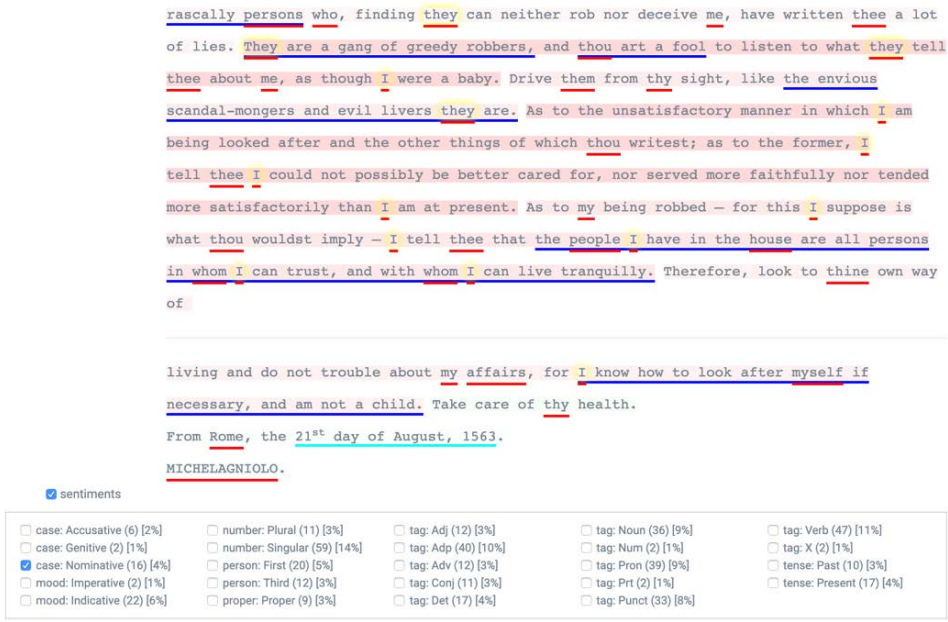

Abb. 13: Visualisierung der Syntax- und Sentimentanalyse auf Grundlage von Standoff Properties.

Weitere Annotationsebenen

In der folgenden Abbildung werden Sentiment- und Syntaxanalyse für einen Michelangelo-Brief dargestellt. Im Rahmen des hier beantragten Projekts werden auch Annotationen zu Syntax, Schreiberhänden, Sprache und Sprachgebrauch etc. denkbar.

Im folgenden werden die technischen Grundlagen von SPEEDy (Editorkomponente) und des Codex-Systems rund um den Editor beschrieben. SPEEDy dient dabei zur direkten Transkription von Text. In Codex werden die Annotationsinformationen strukturiert gespeichert und damit auch erschlossen.

Fazit

Codex verwendet Standoff Properties, um Interpretation und Text im Graph zu vereinen. Annotationen werden dem Text auf Zeichenebene zugeordnet und können uneingeschränkt überlagert werden. Mit kommentierten Annotationen ergeben sich sogar Perspektiven, den wissenschaftlichen Diskurs nachvollziehbar zu machen. Dabei kann jede Annotationen - einschließlich Events, Meta-Relationen, Agents etc. - jederzeit auf ihre textliche Grundlage zurückgeführt werden.

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.

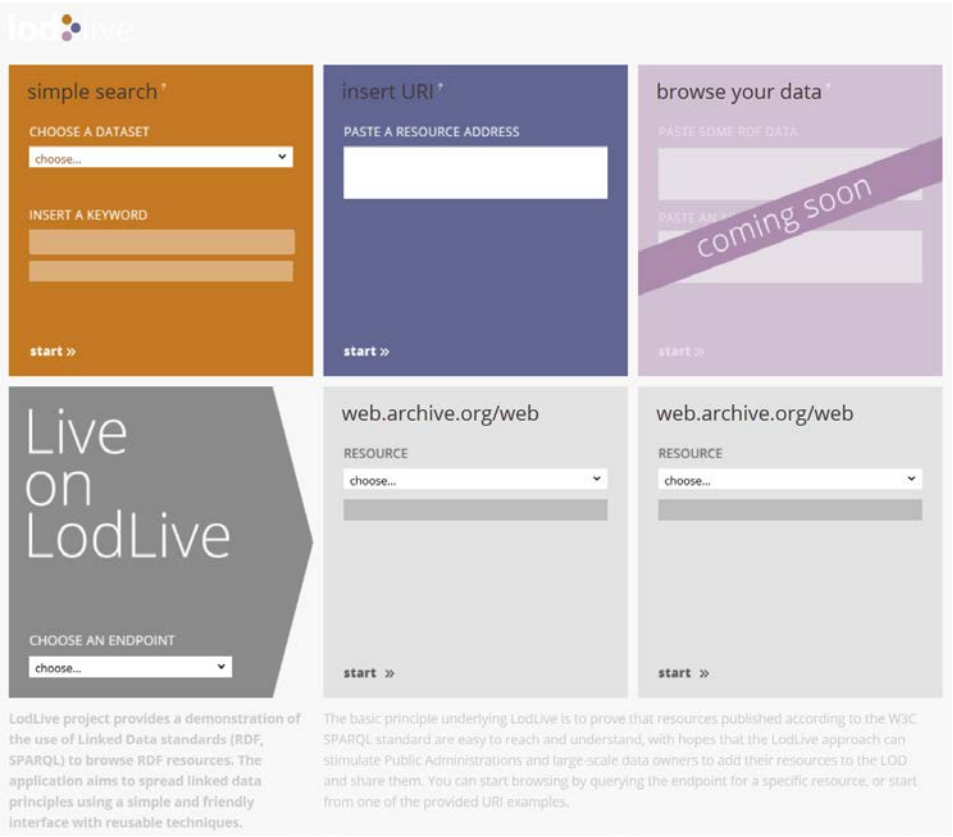

Abb. 1: LodLive at lodlive.burkhartsource.org (aktuell nicht mehr verfügbar, letzter Aufruf im Juni 2019).

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-

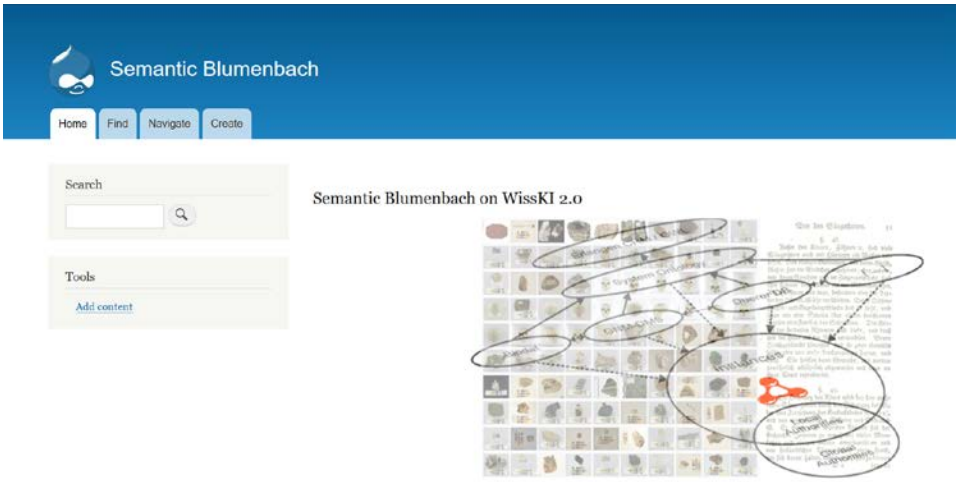

Abb. 2 Semantic Blumenbach auf WissKI 2.0 (Juli 2019)

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).