ZfdG

Zeitschrift für digitale Geisteswissenschaften

Beitrag aus:
Sonderband 2 der ZfdG: Digitale Metamorphose: Digital Humanities und Editionswissenschaft. Hg. von Roland S. Kamzelak und Timo Steyer. 2018. DOI: 10.17175/sb002

Titel:
Der nächste Schritt? Semantic Web und digitale Editionen

Autor/in:
Jörg Wettlaufer

Kontakt:
jwettla@gwdg.de

Institution:
Akademie der Wissenschaften zu Göttingen (ADWG)

GND:
121084280

ORCID:
0000-0003-1957-8059

DOI des Artikels:
10.17175/sb002_007

Nachweis im OPAC der Herzog August Bibliothek:
889896755

Erstveröffentlichung:
15.03.2018

Letzte Überprüfung aller Verweise:
07.03.2018

GND-Verschlagwortung:
Edition | Elektronische Publikation | Semantic Web |

Zitierweise:
Jörg Wettlaufer: Der nächste Schritt? Semantic Web und digitale Editionen. In: Digitale Metamorphose: Digital Humanities und Editionswissenschaft. Hg. von Roland S. Kamzelak /Timo Steyer. 2018 (= Sonderband der Zeitschrift für digitale Geisteswissenschaften, 2). PDF Format ohne Paginierung. Als text/html abrufbar unter DOI: 10.17175/sb002_007.

Jörg Wettlaufer

Der nächste Schritt? Semantic Web und digitale Editionen

Abstracts

Digitale Editionen im Semantic Web? Dieser Beitrag stellt die Frage nach der Relevanz und weiteren Entwicklung von semantischen Technologien bei digitalen Editionen in den Geisteswissenschaften. Die heute als Standard verbreiteten XML-TEI basierten Editionen von Texten können auf etablierte Publikationsstrukturen aufsetzen. Doch der TEI Standard hat, entgegen den Intentionen seiner Schöpfer, nicht zu einer Interoperabilität von Editionen, sondern vielmehr zu einer immer stärkeren Auffächerung des Markups geführt. Daraus resultiert der Bedarf an Lösungen zur Aggregation, Nachnutzung und Vernetzung von Editionen sowie auch zur Erschließung über eine maschinenlesbare Semantik, die über Linked Open Data (LOD), Normdaten und andere Metadaten jedweder Form Verknüpfungen herzustellen in der Lage ist.

Digital editions in the semantic web? This paper explores the question of the relevance and further development of semantic technologies in digital editions in the arts and humanities. The XML TEI-based editions of texts can be set on top of established publication structures. However, the widespread use of the TEI standard has led not to the interoperability of editions – despite the intentions of its creator—but rather to the increasingly strong specialization of the markup by discipline. This situation has led to the need for solutions for aggregation, re-usability, and the networking of editions as well as for encoding using a machine-readable semantic – one that is capable of generating other forms of connections using Linked Open Data (LOD), authority files, and other metadata.

1. Einführung

> »I will never touch that crap of RDF or the semantic web; this is a pipe dream of reality ignoring academics and I will not have it. I will only use JSON-LD.«

Das Semantic Web hat nicht nur Freunde und Befürworter, wie dieses Zitat von Phil Archer aus einer Keynote von 2014 auf launige Weise veranschaulicht. Es drückt – etwas drastisch, aber doch zutreffend – die Distanz vieler IT Entwickler zu Standards des Semantic Web aus, indem es zwei sehr kompatible Standards, die aufeinander aufbauen und doch zugleich aus unterschiedlichen Traditionen entstammen, in einen Gegensatz zueinander setzt. Ist der nächste Schritt für digitale Editionen also wirklich das von einigen so wenig geliebte Semantic Web auf der Basis des Ressource Description Framework (RDF)?[1] Kann über eine Verknüpfung und Kontextualisierung von semantischen Auszeichnungen mit Hilfe von Ontologien ein Mehrwert für die wissenschaftliche Beschäftigung mit Texten geschaffen werden? In den aktuellen Überblickswerken zu Digitalen Editionen kommt eine solche Perspektive bislang gar nicht vor.[2] Die semantische Ebene ist hier die der korrekten Auszeichnung, des annotierenden

[1] http://www.w3.org/RDF/.
[2] Vgl. Pierazzo 2015. Sahle 2013, S. 391f. erwähnt an einer Stelle die Möglichkeiten der Verknüpfung von Personen, Orten, Institutionen über Normdaten mit anderen Ressourcen und beschreibt eine solche

Markups nach etablierten Standards wie z. B. den Richtlinien der Text Encoding Initiative (TEI).[3] Zwar erheben sich inzwischen kritische Stimmen, z. B. die von Sebastian Rahtz, dass der TEI Standard aufgrund seiner starken Auffächerung nur noch bedingt für eine Verknüpfung unterschiedlicher digitaler Editionsprojekte geeignet sei.[4] Aber nur vergleichsweise wenige Arbeitsgruppen beschäftigen sich bisher mit der Frage, welche Rolle das Semantic Web in der geisteswissenschaftlichen Forschung allgemein und für die digitale Edition von Texten im Speziellen in Zukunft haben könnte.[5]

Ich möchte in diesem Beitrag versuchen, eine Antwort auf diese Frage aus einer spezifischen Perspektive und auf der Grundlage der Beschreibung von zwei miteinander verknüpften Projekten zu geben. Bei dem ersten Beispiel handelt es sich um ein Forschungsprojekt, das von 2012 bis 2015 für die Akademie der Wissenschaften zu Göttingen am dortigen Göttingen Centre for Digital Humanities (GCDH) durchgeführt wurde. In diesem Verbundprojekt, an dem neben der Universität Göttingen, der dortigen Staats- und Universitätsbibliothek (SUB) auch die Gesellschaft für Wissenschaftliche Datenverarbeitung (gwdg) sowie die Akademie der Wissenschaften und die Herzog August Bibliothek in Wolfenbüttel beteiligt waren und in dem gemeinsam versucht wurde, dieser Verheißung namens Semantic Web bzw. den ›semantischen Technologien‹ und Linked Open Data (LOD) in den Geisteswissenschaften etwas näher zu kommen, konnten erste Erfahrungen hinsichtlich semantisch angereicherten Editionen gesammelt werden. Das zweite Beispiel beschäftigt sich mit einem aktuell laufenden Projekt namens PANDORA, das von Christopher H. Johnson im Rahmen eines Langfristprojekts der Akademie der Wissenschaften zu Göttingen entworfen wurde und in gewisser Weise eine Fortführung der am GCDH betriebenen Forschungen darstellt.Es arbeitet mit demselben Material, beschreitet jedoch auf der Ebene der Frameworks und Tools andere Wege. Beide Projekte verbindet aber nicht nur die Materialbasis, sondern auch das Bestreben, digitale Editionen innerhalb des Semantic Web zu verlinken und so Teil einer größeren, maschinenlesbaren Wissensbasis zu machen, die in Zukunft einmal ganz neue Formen der Kontextualisierung von Wissen ermöglichen könnte.

Im Mittelpunkt dieser Bemühungen stand und steht die Problematik der Weiterentwicklung und Verknüpfung von digitalen Editionen über Ontologien und Normdaten in Richtung eines Wissensnetzwerks, das aufgrund seiner an die natürliche Sprache angelehnten semantischen Kompetenz eben als ›Semantic Web‹ bezeichnet wird. Was aber ist das Semantic Web und was bedeuten ›buzz words‹ wie LOD konkret für diejenigen, die sich mit der wissenschaftlichen Edition von Texten zum Beispiel im Kontext der Geschichtswissenschaft oder Wissenschaftsgeschichte beschäftigen? Am Anfang dieses kurzen Ausblicks auf die Entwicklung und die Potentiale des Semantic Web soll ein Zitat von Tim Berners Lee stehen, dem geistigen Vater des WWW und zugleich auch Begründer und Propagandist des Semantic Web als dessen natürliche Weiterentwicklung.

Verknüpfung als spezielle Sicht auf elektronische Texte im Allgemeinen und den angewandten TEI-Standard im Besonderen. Eine Ausnahme von dieser Regel stellt die Arbeit von Eva Christina Glaser dar, vgl. Glaser 2013.

[3] http://www.tei-c.org/.

[4] Rahtz / Burnard 2013, S. 193–196.

[5] Vgl. Meroño-Peñuela et al. 2015. Tomasi et al. 2013, S. 145 –158. Eide 2013, S. 26 –30. Rahtz 2010. Siehe auch: Wettlaufer 2015.

The Semantic Web is not a seperate Web but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation.[6]

Es geht also darum, das WWW, so wie wir es kennen, durch Anreicherung mit wohl definierten semantischen Informationen auch für Computer verstehbar zu machen, so dass die Ambiguitäten und Kontext-Abhängigkeiten der natürlichen Sprache auch für Maschinen ›verstehbar‹ werden. Ob Computer und Menschen dann in Kooperation besser miteinander arbeiten werden, oder – wie zuletzt Stephen Hawkings prophezeite – demnächst intelligente Maschinen die Macht übernehmen und das Web 4.0 vielleicht Skynet heißt, soll hier nicht weiter diskutiert werden. Wichtig ist aber, dass das Semantic Web nur als eine Erweiterung des WWW gedacht ist, nicht als eigenes und autarkes Gebilde im Sinne eines endzeitlichen Skynet.

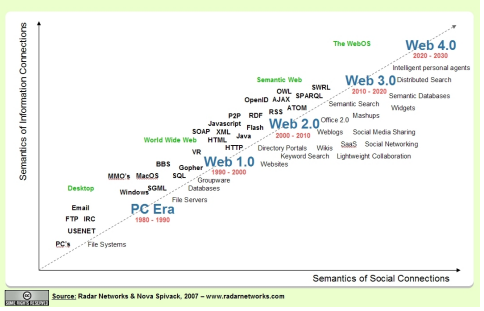


Abb. 1: www Timeline. Quelle: [online]

Nova Spivack, ein weiterer Pionier und Visionär des Internets hat schon 2007 die Entwicklung des Internets erstaunlich korrekt visualisiert und vorhergesagt. Nach dem Web 3.0, also dem von Tim Berners Lee propagierten Semantischen Web, sieht er ein Web 4.0 mit intelligenten persönlichen Agenten und einer ›augmented reality‹ entstehen – eine Entwicklung die jetzt, zehn Jahre später, tatsächlich Realität ist. Was ist aber mit dem Web 3.0., der Phase, in der wir uns nach Aussage der Grafik gerade befinden, tatsächlich gemeint? Das Semantic Web wird in Abgrenzung zum Web 2.0., dem dynamischen Mitmachweb von Facebook, Twitter und Co., auch als Web 3.0 bezeichnet und wäre somit die kommende Stufe der Entwicklung in der maschinellen Informationsverarbeitung. Seit den Anfängen vor über 15 Jahren ist das Semantic Web inzwischen nicht mehr Vision einzelner, sondern wird offiziell vom W3C Konsortium unterstützt und ist über Standardisierung in einer stetigen Weiterentwicklung begriffen. Insbesondere die folgenden Standards und Technologien spielen hierbei eine Rolle:

[6] Berners-Lee et al. 2001.

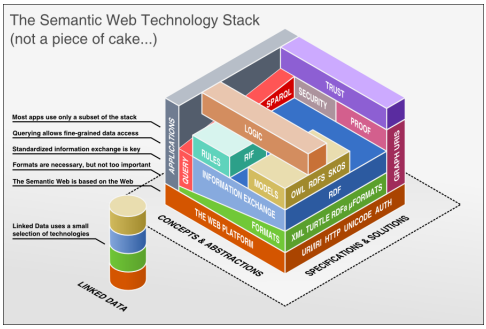


Abb. 2: Semantic Web Technology Stack. Quelle: [online]

Die Grundlage des sogenannten ›semantic web stack‹ bilden die Bezeichner URI/IRI, die Webprotokolle und die Unicode Spezifikation der Zeichenkodierung. Der ›international resource idendifier‹ oder deren Generalisierung als ›uniform resourcelocator‹ bildet die Grundlage des auf XML Serialisierung beruhenden Datenmodells von RDF, dem ›ressource description framework‹. Darauf aufbauend spielen Datenformate wie XML oder Serialisierungen des RDF wie Turtle eine Rolle. Das RDF Konzept mit seinem an die natürliche Sprache aufgebauten Formalismus von Subjekt, Prädikat und Objekt (SPO) bildet dabei den Kern des Semantic Web, da hier ermöglicht wird, einer Ressource (URI) bestimmte Eigenschaften in einer Form zuzuschreiben, die von Maschinen verstanden werden kann und beliebig erweiterbare Aussagen über Ressourcen zulässt. Subjekt und Objekt sind dabei die sog. Knoten, die durch Kanten (Prädikate) miteinander verbunden werden. Diese Kanten nun bestimmen die Art der Beziehung zwischen den Knoten genauer. Das Konzept steht und fällt mit der Eindeutigkeit der Identifizierung von Ressourcen. Das WWW bietet hier mit seiner auf eindeutigen Adressen und Standardisierung beruhenden Adressierung einen soliden Ausgangspunkt, um zumindest dort Ressourcen eindeutig zu verorten. Das grundlegende Problem des WWW, nämlich die Stabilität der Ressourcen und damit die nachhaltige Adressierbarkeit sind damit aber leider noch nicht behoben. Noch ein weiterer grundlegender Designfehler des WWW kann auch durch das Semantic Web nicht eliminiert, aber vielleicht doch abgemildert werden. Nicht nur Adressen, auch Aussagen haben einen temporalen Aspekt. Sie werden zu einer bestimmten Zeit gemacht und sind in der Regel auch nur eine Zeit lang gültig, zumindest was Wissen aus der Domäne der Geisteswissenschaften betrifft. Diese temporale ›Blindheit‹ des WWW hat auch das Semantic Web geerbt, kann aber über sein SPO-Modell zumindest theoretisch auch beliebige temporale Aussagen über eine Ressource machen. In der Praxis wird dies aber leider allzu häufig übersehen.

Jenseits des RDF Modells erlauben die Web Ontology Language (OWL) und RDF-Schema weitergehende Modellierungen von Aussagen und Wissen über Ontologien oder standardisierte Schemata, die schon einfache Schlüsse über die logische Konsistenz und Validität von Aussagen erlauben. Wenn man seine Daten nach diesem Muster abgelegt hat, kann man die einzelnen Triple – wie man die Umsetzung des SPO-Modells auch nennt - miteinander in Beziehungen setzen bzw. verlinken, z.B. durch gemeinsame Identifikatoren wie eine GND Nummer für Personen oder eine ISBN Nummer für Bücher. So entstehen verlinkte Datenbestände, die in sog. ›triple stores‹ gespeichert und über eine spezielle Abfragesprache

(SPARQL Protocol And RDF Query Language, kurz SPARQL) durchsucht und ausgelesen werden können.[7] SPARQL ist an die bekannte Abfragesprache SQL für relationale Datenbanken angelehnt, funktioniert aber etwas anders, da die Datenstruktur bei Graphen eine andere ist als bei relationalen Tabellen. SPARQL vergleicht Triple-Aussagen mit den gespeicherten Knoten und Kanten in einer Wissensbasis. So ist es z. B. immer vorteilhaft vorab zu wissen, welche Daten und Konzepte in einem ›triple-store‹ abgelegt sind, um sinnvolle und effiziente Abfragen zu formulieren.

Aufbauend auf den Standards des Semantic Web ist LOD eine praktische Umsetzung der Technologien zur Bereitstellung der maschinenlesbaren Daten im WWW. Für die gute Praxis von LOD gibt es inzwischen fünf einfache Regeln:[8] # stelle deine Daten im WWW unter einer offenen Lizenz bereit; ## stelle Daten in einem strukturierten und maschinenlesbaren Format bereit; ### verwende offene, nicht proprietäre Formate; #### verwende URIs um Dinge zu bezeichnen und den RDF Standard, damit deine Daten verlinkt werden können; ##### verlinke deine Daten mit anderen Daten, um Kontexte herzustellen.

Das fünf Sterne Modell von LOD stößt allerdings in den Geisteswissenschaften manchmal auf Vorbehalte,[9] da es sich um eine offene Vision des Daten- und Wissensmanagements handelt, die in der Regel eine Creative-Commons-Lizenz voraussetzt.[10] Daher auch Linked ›Open‹ Data!

In welchen Beziehungen steht diese Technologie nun zum Themenbereich der digitalen Editionen? Offensichtlich befinden sich digitale Editionen im WWW, viele davon unter einer freien Lizenz und erfüllen somit schon einmal den ersten Stern der LOD. Wenn sie in XML TEI oder auch einfach nur als Unicode Text vorliegen, sind der zweite und dritte Stern auch erreicht. Es geht also ›nur‹ noch um Ebene 4 und 5, RDF Standards und Verlinkung der Daten in die LOD Cloud. Sind digitale Editionen damit schon Teil des Semantic Web? Leider steckt die eigentliche Herausforderung in diesen zwei letzten Ebenen.

2. ›Semantic Blumenbach‹

Um diese Herausforderungen etwas anschaulicher darzustellen, greife ich auf ein Projekt im Rahmen des Digital Humanities Forschungsverbund Niedersachsen am Göttinger GCDH zurück, in dem Semantic Web Technologien für die Verknüpfung zwischen XML/TEI P5 Texten und Sammlungsobjekten aus den Göttinger Universitätssammlungen erprobt wurden.[11] Die dabei verwendeten Technologien und Standards sind gerade auch im Hinblick auf digitale Editionen von Texten interessant. Nebenbei bringt das gemeinsame Datenformat RDF die Möglichkeit mit sich, Informationen in neuer Weise miteinander automatisiert verknüpfen zu

[7] https://www.w3.org/TR/sparql11-query/.
[8] http://5stardata.info/de.
[9] Vgl. Wettlaufer/Wuttke 2015.
[10] Vgl. http://opendefinition.org/.
[11] Wettlaufer et al. 2015, S. i187 –i198. Siehe auch http://dhfv-ent2.gcdh.de/blumenbach/wisski/ (verlinkte Ressource wird derzeit überarbeitet und ist daher vorübergehend nicht erreichbar).

können und so neue Zusammenhänge aufzudecken. Konkret ging es um die Modellierung der Beziehung von Texten bzw. digitalen Editionen der Schriften des Göttinger Gelehrten Johann Friedrich Blumenbach (1852-1840) mit den von ihm gesammelten naturhistorischen Objekten.[12]

Die Herausforderung des Projekts bestand darin, Texte und Objekte, die innerhalb der Geisteswissenschaften meist von ganz unterschiedlichen Fachbereichen und ›scientific communities‹ untersucht werden, auf Basis des RDF-Frameworks miteinander in semantische Beziehungen zu setzen. Dazu bedurfte es einer Ontologie, die umfassend genug ist, die Strukturmerkmale sowohl von Texten als auch Objekten abbilden und so die semantischen Beziehungen zwischen diesen beiden Klassen auf verschiedenen Ebenen etablieren zu können. Nachdem zunächst EDM, das Europeana Data Model,[13] als Referenzontologie präferiert worden war, stellte sich nach Aufnahme der Arbeit schnell heraus, dass dieses Modell in seiner aktuellen Version keinen Zugriff auf Einheiten wie einzelne Worte oder Markup ermöglichte, die aber Grundlage der semantischen Verknüpfung zwischen Texten und Objekten sein sollten. Wesentlich flexibler präsentierte sich das Conceptual Reference Model (CIDOC CRM), das im Bereich der Museen weit verbreitet ist.[14] Im Rahmen des Semantik Web Frameworks ›wissenschaftliche Kommunikationsinfrastruktur‹, kurz WissKI,[15] wurde das auf CIDOC CRM aufbauende Erlangen CRM (ECRM) entwickelt, eine OWL Description Logic Version der CIDOC CRM Ontologie, die wie CIDOC selbst insbesondere für Museen entwickelt wurde.[16] Diese Ontologie ist ›event-orientiert‹, versucht also Ereignisse wie die Produktion eines Manuskripts oder seine Rezeption prozesshaft zu modellieren.

CRM ist in den letzten Jahren sehr aktiv weiter entwickelt worden und eine Special Interest Group (SIG) arbeitet an der Erweiterung der Ontologie und ein mapping auf andere, domainspezifische Standards bzw. Ontologien. So gibt es z.B. die in Erlangen betriebene Erweiterung bzw. Integration des bibliothekarischen ›functional requirements for bibliographical records‹ (FRBR) Standard[17] und mit CRMArcheo auch ein Modul für die Archäologie.[18]

Die Architektur der Wissenschaftlichen Kommunikationsinfrastruktur WissKI baut auf den zu Anfang gezeigten ›semantic web stack (Abb. 2) auf. ECRM ist dabei Top Level Ontologie. Darunter liegt die Systemontologie – abgeleitet aus ECRM, in die wiederum die Anwendungsontologien (z. B. Semantic Blumenbach) eingehängt werden können. Als Frontend wird das CMS Drupal verwendet.[19]

[12] Vgl. zu dem Projekt http://www.blumenbach-online.de.
[13] http://pro.europeana.eu/page/edm-documentation.
[14] http://www.cidoc-crm.org/.
[15] Scholz / Goerz 2012, S. 1 –2.
[16] http://erlangen-crm.org/.
[17] http://erlangen-crm.org/efrbroo.
[18] Vgl. http://www.cidoc-crm.org/special_interest_meetings.html.
[19] Im Projekt wurde WissKI 1.0 verwendet, das auf Drupal 6 aufsetzt. Inzwischen wurde Version 2.0 von WissKI fertiggestellt, die auf Drupal 8 basiert. Vgl. http://wiss-ki.eu.

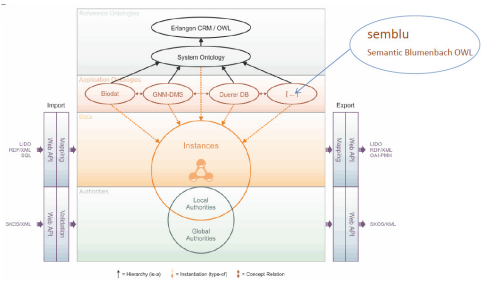


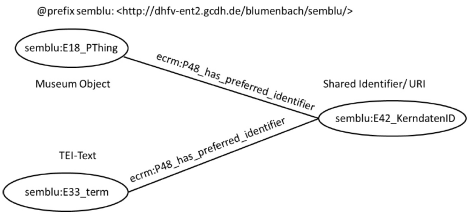
Abb. 3: Architektur WissKI 1.0. [online]

Benutzer können über sog. Pfade in WissKI die Daten in ECRM modellieren. Es handelt sich dabei meist um eine Kombination aus den Entitäten des ECRM und Erweiterungen aus der Anwendungsontologie.

SembluKapitel	Group [semblu:E33_div]
Object	semblu:E33_div -> ecrm:P106_is_composed_of -> semblu:E33_term -> ecrm:P67_refers_to -> semblu:E18_PThing -> ecrm:P48_has_preferred_identifier -> semblu:E42_KerndatenID
Person	semblu:E33_div -> ecrm:P67_refers_to -> semblu:E21_NewPerson -> ecrm:P131_is_identified_by -> semblu:E82_PersonName
Place	semblu:E33_div -> ecrm:P67_refers_to -> semblu:E44_NewPlace -> ecrm:P1_is_identified_by -> semblu:E48_PlaceName
Title	semblu:E33_div -> ecrm:P1_is_identified_by -> ecrm:E35_Title
Sammlungsobjekte	Group [semblu:E18_PThing]
Identifier	semblu:E18_PThing -> ecrm:P48_has_preferred_identifier -> semblu:E42_KerndatenID
Objektbezeichnung	semblu:E18_PThing -> ecrm:P1_is_identified_by -> ecrm:E41_Appellation
Domaene	semblu:E18_PThing -> ecrm:P2_has_type -> ecrm:E55_Type -> ecrm:P1_is_identified_by -> ecrm:E41_Appellation
InventarNr	semblu:E18_PThing -> ecrm:P48_has_preferred_identifier -> ecrm:E42_Identifier
Originalbezeichnung	semblu:E18_PThing -> ecrm:P1_is_identified_by -> ecrm:E41_Appellation

Abb. 4: Beispiel für Datenmodellierung mit Pfaden in WissKI 1.0.

Im Projekt Semantic Blumenbach wurden zudem neue WissKI Module zur Etablierung eines Workflows für den Ingest von TEI-Texten und zur Extraktion von Tripeln programmiert, die anschließend bei einem ›open database connectivity‹ (ODBC) Datenbank-Ingest mit Objekten verknüpft und disambiguiert wurden.[20] Über das Drupal Buch-Modul war schließlich eine Präsentation der Texte, in diesem Fall des etwa 450 Seiten starken Handbuchs der Naturgeschichte von Blumenbach, möglich. Aber zunächst entstehen mit Hilfe eines XSLT-Stylesheets die gewünschten Triple, in denen die semantischen Beziehungen zwischen Texten und Objekten sowie Informationen über Personen, Orte und Dinge gespeichert werden.[21] Die Modellierung der zentralen Verknüpfung in unserem Projekt zwischen Texten (hier Kapitel) und Objekten wurde über die eindeutigen Bezeichnerder Sammlungsdatensätze realisiert. Nachteil dieser Methode ist die Willkür und Individualität der entstehenden URIs, die übrigens nach den Prinzipien des Semantic Web idealerweise aus den Adressen der Sammlungen, die die Objekte aufbewahren, bestehen sollten



[20] https://github.com/mnscholz/wisski_texttei und https://github.com/WissKI/wisski_book_import.
[21] http://dhfv-ent2.gcdh.de/blumenbach/wisski/sites/all/modules/wisski_texttei/triplify.xsl (verlinkte Ressource wird derzeit überarbeitet und ist daher vorübergehend nicht erreichbar).

Abb. 5: Schematische Darstellung der Datenmodellierung in ECRM über eine gemeinsame Identifikationsnummer, die über TEI Markup vorab in die Texte eingebracht wurde.

In der Textansicht des Buch-Moduls sind dann Sammlungsobjekte mit Kapiteln der Texte Blumenbachs verknüpft. Eine feinere Modellierung mit einer Verknüpfung auf der Wort oder Seitenebene wäre im Nachhinein wünschenswert gewesen, hätte jedoch auch den Import eines Textes mit derErstellung von Triplen, der immerhin in der vorliegenden Version gut 30 Minuten dauerte, weiter verlangsamt.

Insgesamt kann man die semantische Beziehung zwischen Text und Objekten nur so präzise erfassen, wie sie zunächst via ›named entity recognition‹ ausgezeichnet werden. ›Close‹, ›narrow‹ und ›exact match‹ aus SKOS[22] können die Qualität der Beziehung zwar graduell abbilden, aber eine solche Kategorisierung kann nur bedingt automatisiert erfolgen. Es gibt am Ende nur wenige Textstellen, an denen Blumenbach sich direkt auf seine Sammlungen bezieht, der überwiegende Rest steht in der Spannung zwischen dem abstraktem Begriff einer Sache und dem konkreten Gegenstand in einer Sammlung.

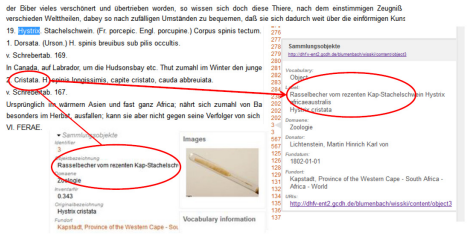


Abb. 6: Darstellung der Objekte im Kontext der TEI Edition der Werke Blumenbachs. Foto des Rasselbechers: GZG Museum / G. Hundertmark.

Diese Implementierung war am Ende nicht sehr performant, da beim Einlesen der Texte und Objekte in das WissKI System viele leere Knoten angelegt wurden, die als Ballast mitgeführt werden mussten. Auch das nachträgliche Löschen von Knoten stellte sich als problematisch heraus, da sich durch die permanente Synchronisation zwischen ›triple-store‹ (hier Arc2)[23] und der Drupal MySQL Datenbank ein Jojo-Effekt einstellte. Da WissKI nicht für unseren speziellen Anwendungsfall konzipiert worden ist, sondern von uns mit Hilfe der Erlanger Kollegen mit einigem Aufwand nur in die gewünschte Richtung erweitert wurde, erwies es sich für die Aufgabenstellung als nur bedingt tauglich. Seine eigentlichen Stärken entfaltet WissKI bei der Sammlungserfassung, wo es die Kuratoren bei der Eingabe von Objekten in die Sammlungsdatenbank unterstützt und eine komfortable Verlinkung zu Normdatensätzen von Personen und Orten anbietet. In diesem Kontext wird es heute vielfach sehr erfolgreich eingesetzt.[24]

3. ›PANDORA‹

[22] Simple Knowledge Organisation System. Siehe: https://www.w3.org/2004/02/skos/.
[23] https://github.com/semsol/arc2/wiki.
[24] Siehe auch http://www.igsd-ev.de/ und die Liste der Projekte auf http://wiss-ki.eu.

Unter anderem aufgrund der Erfahrungen mit der semantischen Edition der Blumenbach-Texte im Projekt ›Semantic Blumenbach‹ hat sich das Akademieprojekt ›Johann Friedrich Blumenbach – online‹ der Göttinger Akademie der Wissenschaften im Zusammenhang mit der geplanten digitalen Edition der gedruckten Werke und naturhistorischen Sammlungen entschlossen, bei der Bereitstellung eines Portals zur Erforschung der Werke und Sammlungen dieses Gelehrten nicht auf eine fertige Anwendung zu setzen, sondern vielmehr selber eine modular aufgebaute Umgebung zu schaffen, in der die spezifischen Bedürfnisse des Vorhabens optimal abgedeckt werden können. Bei der Konzeption dieser Portalumgebung stehen Interoperabilität, Erweiterbarkeit und Nachnutzung als zentrale Entwicklungsziele im Vordergrund. Nicht zuletzt aufgrund dieser Designziele werden eine ganze Reihe von Technologien aus dem Bereich des Semantic Web eingesetzt und die Voraussetzungen für LOD geschaffen. Ausgangspunkt des PANDORA [resentation (of) nnotations (and) otations (in a) igital bject epository rchitecture] LOD Frameworks[25] sind digitale Abbildungen von Texten und Objekten, die in einem Fedora Commons Repository[26] gespeichert werden. PANDORA ist eine Sammlung von Open Source Anwendungen, die über ein gemeinsames ›Manifest‹ Dokument die Präsentation der Daten für den Anwender n nach dem IIIF Standard organisieren.[27] Das ›Manifest‹ besteht aus einem JSON-LD[28] Dokument und wird aus einem digitalen Objektrepositorium über die dynamische Verwendung von SPARQL-Abfragen erzeugt. Es orientiert sich dabei an der Semantik und dem Konzept der ›IIIF Presentation API‹[29]. Diese Schnittstelle definiert, wie die Struktur und das Layout eines komplexen und bildbasierten Objekts in einem Standardformat dargestellt werden kann und zielt darauf ab, die Interoperabilität und Erweiterbarkeit von Präsentationen basierend auf dem vom W3C standarisierten Web Annotation Datenmodell[30] zu erleichtern. In diesem Modell ist eine Annotation jede Ressource, die aus zwei Komponenten besteht, einen ›body‹ und einen ›target‹:

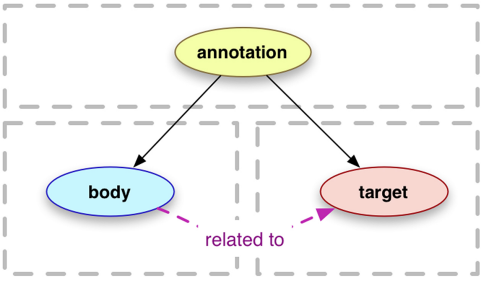


Abb. 7: Schematische Darstellung des Annotationsschemas von PANDORA.

In der IIIF Presentation API ist das Ziel ein ›canvas‹ (Leinwand), der eine Abstraktion des Client-Arbeitsplatzes oder Sichtbereichs darstellt. Die Anmerkung oder Notation (body)

[25] Das Framework wurde geplant und wird umgesetzt von Christopher H. Johnson im Rahmen seiner Tätigkeit für das Akademievorhaben »Johann Friedrich Blumenbach – online« mit einer Laufzeit von 2010–2025. Die Beschreibung des PANDORA Projekts stammt von Christopher H. Johnson und wurde vom Verfasser ins Deutsche übersetzt. Vgl. auch Johnson / Wettlaufer 2017. Für eine ausführliche Dokmentation sowie Beispiele siehe https://github.com/pan-dora.
[26] http://fedorarepository.org/.
[27] http://iiif.io/.
[28] https://www.w3.org/TR/json-ld/.
[29] http://iiif.io/api/presentation/2.1/.
[30] https://www.w3.org/TR/annotation-model/.

kann mit jedem verknüpften oder eingebetteten Objekt wie einem Bild, einer Beschreibung oder einem semantischen Tag verlinkt sein. Die assoziativen Beziehungen zwischen verschiedenen ›bodies‹ auf einem ›canvas‹ sind mit der ›linked-data semantik‹ im Manifest instanziiert. Die Segmentierung ermöglicht die Auswahl eines Bildbereiches oder eines ›canvas‹ unter Verwendung rechteckiger Begrenzungsrahmen oder mit der ›IIIF Image API‹[31], einem ›stream‹ von Bildausschnitten. Hotspot Verknüpfungen ermöglichen es, die Auswahl auf ein Anmerkungsobjekt zu lenken, um eine Zustandsänderung in einem anderen Anmerkungsobjekt auszulösen.

Durch die Verwendung des PANDORA IIIF Manifest Services[32] wird die Konstruktion von Präsentationen aus SPARQL Abfragen erlaubt, die eine sehr differenzierte Darstellung der Annotationen und Notationen über JSON-LD ermöglichen. Der Entwurf einer LDP Container-Hierarchie und von Sammlungs-Definitionen im Einklang mit der Semantik der IIIF Presentation API ›Annotation-Liste‹[33] und ›Layer‹[34] für die Darstellung von Textsequenzen (Linien, Wortgruppen, Absätze, Seiten, Kapitel, etc.) ist ein integraler Bestandteil von PANDORA.

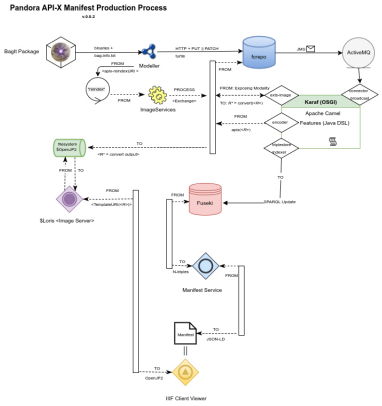


Abb. 8: PANDORA API-X Produktionsprozeß für Manifeste (Abb. Christopher H. Johnson, [online]).

Mit einer klaren Trennung der Domain- und Client-Rollen bietet das PANDORA Framework Flexibilität und Erweiterbarkeit für alle möglichen Web-Client Präsentationsmethoden. Darüber hinaus unterstützt PANDORA Node.js Instanzen, die durch socket.io und Redis Pub/Sub[35] Ereignisse verbunden sind und dadurch Redundanz und Durchsatz für dezentrale asynchrone Operationen bieten. Das Framework besteht aus aktueller Open Source Software nach Industriestandards für Linked Data. Dazu gehören das weiter oben schon erwähnte Fedora-Repository in der aktuellen Version 4, ein intern genutzter Apache Jena Fuseki Triple-Store[36], die Messaging und Integrationskomponenten Apache Camel[37] und Apache Karaf[38] sowie ein weiterer Service zur Bereitstellung der Daten als LOD namens Open Virtuoso[39]

[31] http://iiif.io/api/image/2.1/.
[32] https://github.com/pan-dora/manifest-service.
[33] http://iiif.io/api/presentation/2.1/#annotation-list.
[34] http://iiif.io/api/presentation/2.1/#layer.
[35] http://redis.io/topics/pubsub.
[36] https://jena.apache.org/.
[37] http://camel.apache.org/.

Abb. 9: Architekturskizze PANDORA LOD Framework v.0.3 (Christopher H. Johnson; [online])

PANDORA ist gekennzeichnet durch Interoperabilität, Flexibilität und Erweiterbarkeit und erlaubt, durch die Verwendung von Standard-Software, ebenfalls eine Nachnutzung der Forschungsdaten über Linked Open Data Schnittstellen. Diese Daten können über den SPARQL-Endpoint entweder lokal integriert oder extern zur Nachnutzung angeboten werden.

4. Semantische Editionen. Herausforderungen und Lösungen für die Zukunft.

Es ist durchaus gerechtfertigt, den nächsten Schritt – ›semantische‹ digitale Editionen – im Titel mit einem Fragezeichen zu versehen. Momentan nimmt das Interesse an einer Verknüpfung von Linked Open Data mit klassischen digitalen Editionen zwar rasant zu, die Durchführbarkeit und die Relevanz müssen sich jedoch in der Praxis erst bewähren. Mit dem MEDEA Projekt (Modelling Semantically Enriched Digital Editions of Accounts) läuft inzwischen ein weiteres Projekt mit semantischen Technologien, das digital edierte serielle Rechnungsquellen in die Linked Open Data Cloud bringen möchte.[40] Auch dieses Projekt ist notwendiger Weise explorativ, da es bislang kaum Erfahrungen bei der Nutzung der Semantic Web Technologie bei geisteswissenschaftlichen Editionsprojekten gibt. Die bis heute immer noch fehlende flächendeckende Umsetzung der Vision eines Semantic Web führt zu Insellösungen, die das volle Potential von Linked Data noch nicht ausschöpfen können. Es gibt allerdings eine Reihe von kommerziellen Angeboten, die semantische Technologien für geisteswissenschaftliche Editions- und Forschungsprojekte anbieten und so den Einstieg erleichtern können. Umfassende Unterstützung bei Annotation und Publikation werden von Tools wie ›Muruca‹[41] und ›Pundit‹[42] angeboten werden. Hinter diesen Akronymen verbergen sich Open Source Produkte der italienischen Firma net7, die sich seit einigen Jahren im Bereich Semantic Web für die Geisteswissenschaften engagiert. Bei Muruca handelt es sich um ein semantisches Publikationsframework, während Pundit ein Annotationswerkzeug ist. Muruca enthält eine LOD Wolke namens korbo (für die Suche, den Import und Vermehrung von LD-Ressourcen), Pundit und ein Visualisierungs-Tool (EVT).[43] Muruca kann auch mit OxGarage[44] generierte Transformationen von Texten verwenden, so dass Wissenschaftler zur Texterstellung mit ihren vertrauten Textverabeitungsprogrammen arbeiten können. Einen einfachen Weg aus Datenbanken ins Semantic Web bieten auch Tools wie D2RQ[45] oder das aktuellere italienische ontop.[46] Diese Lösungen wurden im Rahmen des DHFV-Projekts getestet. Insgesamt erwiesen sich diese Tools zur LOD Bereitstellung aus relationalen Datenbankenals tauglich, da sowohl Performanz als auch technologische Usability überzeugen

[38] http://karaf.apache.org/.
[39] https://virtuoso.openlinksw.com/.
[40] http://medea.hypotheses.org/. Vgl. auch Vogeler 2014, S. 398 –400. Tomasek et al. 2016, S. 96 –98.
[41] http://www.muruca.org/.
[42] http://www.thepund.it/.
[43] http://www.netseven.it/.
[44] http://www.tei-c.org/oxgarage/.
[45] http://d2rq.org/.
[46] http://ontop.inf.unibz.it. Vgl. auch Michel et al. 2014.

und sich so bestehende Architekturen gut nachnutzen lassen. Ebenfalls interessant in diesem Zusammenhang ist ein Service der Digitalen Akademie in Mainz, die einen Xtriples genannten Dienst zur Extraktion von RDF-Statements aus XML Daten anbietet.[47] In einem einfach zu bedienenden Web-Interface können XML Daten hochgeladen und dann als Triple in verschiedenen Formaten ausgegeben werden. Der Service baut auf Apache any23 und der eXist XML Datenbank auf.[48]

Andere Akteure in der Linked Data Welt sind Ontowiki, ein semantisch erweitertesWiki – ganz ähnlich dem Konzept von WissKI - und die Firma Poolparty mit Tools wie Skossy, einem auf der SKOS Ontologie basierenden Werkzeug zur dynamischen Generierung von Thesauri.[49] Bemerkenswert in diesem Zusammenhang war auch die LOD2 Initiative der EU, die einen Semantic Web Stack für die vereinfachte Bereitstellung und Verwendung von Linked Open Data zur Verfügung stellte.[50]

Zum Schluss möchte ich noch auf einige grundsätzliche Probleme bei der Einbindung von digitalen Editionen in das Semantic Web zu sprechen kommen. Einige dieser Probleme, wie das Fehlen der zeitlichen Dimension in RDF und WWW insgesamt, wurden zu Beginn schon genannt. Doch es gibt noch andere, grundlegendere Bedenken, die zwischen geisteswissenschaftlicher Forschung und einer Formalisierung von geisteswissenschaftlichen Wissensbeständen im Sinne einer semantischen Lesbarkeit durch Maschinen stehen. Ontologien entschärfen nämlich das Problem der Subjektivität von Welterfahrung (leider) nicht gänzlich, sie können es nur etwas mindern (ebenso wie standardisierte Vokabulare). Zudem erfordert eine Triplifizierung von Aussagen eine Exaktheit und Eindeutigkeit, die die Geisteswissenschaften ab einem gewissen Grad von Abstraktion möglicherweise nicht zu leisten im Stande sind. Wenn alles in den Geisteswissenschaften in einfachen SPO Sätzen ausdrückbar wäre, dann würde die Publikationskultur in diesen Fächern sicher eine andere sein und auch die Trennung zu den MINT Fächern wäre weniger deutlich ausgeprägt.

Wenn es aber zur übergreifenden Verknüpfung von Wissensbeständen einer gemeinsamen Ontologie bedarf, erhebt sich zudem die Frage, welche dies sein soll. Projekte wie Semantic Blumenbach, Medea und auch PANDORA setzen auf CIDOC CRM. Aber können wir die Welt in den Geisteswissenschaften wirklich mit einem Konzept erfassen, das ursprünglich ein genuin museales war? Ein möglicher anderer Ansatz wäre eine Anbindung an das zurzeit größte LOD Hub im Semantic Web, die dbpedia und das wikidata Projekt.[51] Gerade das wikidata-Projekt könnte dabei in Zukunft eine besondere Rolle als Link-Hub für die Geisteswissenschaften spielen. Das dort gesammelte Wissen ist relativ stabil, die Adresse etabliert und es stehen kuratierte Daten zur Verfügung, die alle Bereiche menschlichen Erkenntnisinteresses abbilden. Wenn Semantic Web irgendwann funktionieren wird, dann werden Projekte wie dbpedia und neuerdings wikidata darin sicher weiterhin eine zentrale Rolle spielen.[52]

[47] http://xtriples.spatialhumanities.de/index.html.
[48] http://exist-db.org/ und https://any23.apache.org/.
[49] http://aksw.org/Projects/OntoWiki.html und https://www.poolparty.biz/.
[50] Vgl. http://linkeddata.org. Der Stack wird inzwischen nicht mehr angeboten [http://stack.linkeddata.org].
[51] http://de.dbpedia.org/ und https://www.wikidata.org/.
[52] Vgl. auch Sahle / Henny 2015, S. 113 –148, hier S. 118 –119.

Linked Data kann heute auch in den Geisteswissenschaften schon aktiv genutzt werden, um digitale Ressourcen miteinander zu verknüpfen und damit Zusammenhänge evident zu machen, die bislang erst nach serieller Rezeption im Kopf des oder der Forschenden entstanden. Mit sog. ›Mashups‹ können so qualitativ neue Ergebnisse aus schon in standardisierten Datenformaten digital vorliegenden Informationen gewonnen werden, die das Potential besitzen, bislang unbekannte Zusammenhänge sichtbar(er) werden zu lassen.[53] Bis zu einer solchen semantischen Verknüpfung von digitalen Editionen im Semantic Web ist es noch ein weiter Weg, aber es gibt eine Reihe von vielversprechenden Hinweisen, dass es eben dieser Weg sein wird, den wir in Zukunft beschreiten werden.

[53] Vgl. auch Endres-Niggemeyer 2013.

Bibliographie

Tim Berners-Lee / James Hendler / Ora Lassila: The Semantic Web. In: Scientific American 284 (2001), H. 5, S. 34-43.[Nachweis im GBV]

Øyvind Eide: Ontologies, data modelling, and TEI. In: The Linked TEI: Text Encoding in the Web. Abstracts of the TEI Conference and Members Meeting 2013. Hg. von Fabio Ciotti and Arianna Ciula. (TEI Conference and Members Meeting 2013, Rom, 2.-5.10.2013) Rom 2013, S. 26-30. PDF. [online] [Nachweis im GBV]

Brigitte Endres-Niggemeyer: Semantic Mashups. Intelligent Reuse of Web Resources. Heidelberg 2013. [Nachweis im GBV]

Fredo Erxleben / Michael Günther / Markus Krötzsch / Julia Mendez / Denny Vrandečić: Introducing Wikidata to the Linked Data Web. PDF. [online] In: The Semantic Web. ISWC 2014. (International Semantic Web Conference: 13, Riva del Garda, Italien, 19.–23.10.2014) 2 Bde. Cham u.a. 2014. Bd 1: S. 50-64. PDF. [online] [Nachweis im GBV]

Eva Christina Glaser: Digitale Edition als Gegenstand bibliothekarischer Arbeit: Probleme, Umsetzung und Chancen am Beispiel der Wolfenbütteler Digitalen Bibliothek (WDB). Berlin 2013 (= Berliner Handreichungen zur Bibliotheks- und Informationswissenschaft, 339). URN: urn:nbn:de:kobv:11-100206926 [Nachweis im GBV]

Christopher H. Johnson / Jörg Wettlaufer: Einführung in das PANDORA Linked Open Data Framework. PDF. [online] In: DHd 2017: Digitale Nachhaltigkeit. Konferenzabstracts. (DHd 2017, Bern, 13.-18.02.2017) Bern 2017, S. 31–34. PDF. [online]

Albert Meroño-Peñuela / Ashkan Ashkpour / Marieke van Erp / Kees Mandemakers / Leen Breure / Andrea Scharnhorst / Stefan Schlobach / Frank van Harmelen: Semantic Technologies for Historical Research: A Survey. [online] In: Semantic Web Journal 6 (2015), H. 6, S. 539-564. [online]

Franck Michel / Johan Montagnat / Catherine Faron-Zucker: A survey of RDB to RDF translation approaches and tools. 2014. [online]

Elena Pierazzo: Digital Scholarly Editing: Theories, Models and Methods. Farnham u.a. 2015. [online] [Nachweis im GBV]

Sebastian Rahtz / Lou Burnard: Reviewing the TEI ODD system. PDF. [online] In: Proceedings of the 2013 ACM Symposium on Document Engineering. (DocEng: 13, Florenz, 10.-13.09.2013) New York, NY. 2013. [Nachweis im GBV]

Sebastian Rahtz: TEI to CRM. (TEI Ontologies SIG meeting, Zadar, 2010) Oxford 2010. PDF. [online]

Patrick Sahle: Digitale Editionsformen. Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels. 3 Bde. Preprint-Fassung. Norderstedt 2013. Bd. 3: Textbegriffe und Recodierung. URN: urn:nbn:de:hbz:38-50130

Patrick Sahle / Ulrike Henny: Klios Algorithmen: Automatisierte Auswertung von Wikipedia-Inhalten als Faktenbasis und Diskursraum. In: Wikipedia und Geschichtswissenschaft. Hg. von Thomas Wozniak / Jürgen Nemitz / Uwe Rohwedder. (Deutscher Historikertag: 50, Göttingen, 23.-26.09.2014) Berlin 2015, S. 113–148. [Nachweis im GBV]

Martin Scholz / Günther Goerz: WissKI: A Virtual Research Environment for Cultural Heritage. DOI: 10.3233/978-1-61499-098-7-1017 In: 20th European Conference on Artificial Intelligence. Hg. von Luc De Raedt / Christian Bessiere / Didier Dubois / Patrick Doherty / Paolo Frasconi / Fredrik Heintz / Peter Lucas. (ECAI: 20, Montpellier, 27.-31.08.2012) Amsterdam 2012, S. 1 –2. (= Frontiers in Artificial Intelligence and Applications, 242) [online]

Kathryn Tomasek /Georg Vogeler / Kathrin Pindl / Cliff Anderson / Anna Orlowska / Øyvind Eide: MEDEA (Modeling Semantically Enriched Digital Editions of Accounts). In: Digital Humanities 2016. Conference Abstracts. Hg. von Maciej Eder / Jan Rybicki. (DH 2016, Krakau, 11.-16.07.2016) Krakau 2016, S. 96–98. PDF. [online]

Francesca Tomasi / Fabio Ciotti / Maurizio Lana / Fabio Vitali / Silvio Peroni / Diego Magro: Dialogue and linking between TEI and other semantic models. In: The linked TEI: Text Encoding in the Web. Hg. von Fabio Ciotti / Arianna Ciula. (TEI Conference and Members Meeting 2013, Rom, 2.-5.10.2013) Rom 2013, S. 145-150. PDF. [online]

Georg Vogeler: Modelling digital edition of medieval and early modern accounting documents. [online] In: Digital Humanities 2014. Conference Abstracts. Hg. von Cyril Bornet. (DH 2014, Lausanne, 08.-12.7.2014) Lausanne 2014, S. 398–400. PDF. [online]

Jörg Wettlaufer: Tagungsbericht. Historische Semantik und Semantic Web. (Workshop, Heidelberg, 14.-16.09.2015) In: H-Soz-Kult, Heidelberg 12.11.2015. [online]

Jörg Wettlaufer / Ulrike Wuttke: Tagungsbericht. Offene Lizenzen in den Digitalen Geisteswissenschaften. (Workshop, München, 27.-28. April 2015) In: H-Soz-Kult, München 09.06.2015. [online]

Jörg Wettlaufer / Christopher H. Johnson / Martin Scholz / Mark Fichtner / Sree Ganesh Thotempudi: Semantic Blumenbach. Exploration of Text-Object Relationships with Semantic Web Technology in the History of Science. In: Digital Scholarship Humanities 30 (2015), H. Suppl. 1, S. 187-197. DOI: 10.1093/llc/fqv047

Jörg Wettlaufer / Christopher H. Johnson: Digitale Nachhaltigkeit bei Grundlagenforschung im Akademieprogramm: Das Beispiel „Johann Friedrich Blumenbach-online“. In: In: DHd 2017: Digitale Nachhaltigkeit. Konferenzabstracts. (DHd 2017, Bern, 13.-18.02.2017) Bern 2017, S. 234-235. PDF. [online]

Abbildungslegenden und -nachweise

Abb. 1: www Timeline. Quelle: [online].

Abb. 2: Semantic Web Technology Stack. Quelle: [online].

Abb. 3: Architektur WissKI 1.0. [online].

Abb. 4: Beispiel für Datenmodellierung mit Pfaden in WissKI 1.0.

Abb. 5: Schematische Darstellung der Datenmodellierung in ECRM über eine gemeinsame Identifikationsnummer, die über TEI Markup vorab in die Texte eingebracht wurde.

Abb. 6: Darstellung der Objekte im Kontext der TEI Edition der Werke Blumenbachs. Foto des Rasselbechers: GZG Museum / G. Hundertmark.

Abb. 7: Schematische Darstellung des Web Annotation Schemas, das in PANDORA verwendet wird [online]).

Abb. 8: PANDORA API-X Produktionsprozeß für Manifeste (Abb. Christopher H. Johnson, [online]).

Abb. 9: Architekturskizze PANDORA LOD Framework v.0.3 (Christopher H. Johnson; [online]).