Maximilian Eibl, Martin Gaedke. (Hrsg.): INFORMATIK 2017,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2017 1717

Konzept und Prototyp einer dezentralen Wissensinfrastruktur zu Hochschuldaten für Mensch und Maschine

Vera Meister[1], Jonas Jetschni[2] und Sebastian Kreideweiß[3]

Abstract: Der Beitrag beschreibt den Stand der Entwicklungen für eine dezentrale Wissensinfrastruktur zu Hochschuldaten, welche Mehrwertdienste für Mensch und Maschine unterstützt. Im Fokus stehen zunächst wenig volatile Daten zu Studiengängen, die aktuell mit hohem Aufwand in den verschiedensten technischen und organisationalen Strukturen vorgehalten werden. Das zeigt eine aktuelle Analyse der Ausgangslage. Das Architekturkonzept kann als Knowledge Graph beschrieben werden, der Webseiten von Hochschul-Content-Management-Systemen (CMS) als primäre Wissensquellen nutzt. Dies wird zunächst durch CMS-Extensions erreicht, die auf Semantic-Web-Technologien, insbesondere auf schema.org, JSON-LD und SPARQL setzen. Die Anbindung weiterer strukturierter und semi-strukturierter Wissensquellen erfolgt in transparenten Datenintegrationsprozessen, welche individuelle Orientierung ebenso wie anforderungsspezifische Datenaktualisierung unterstützen. Neben der systematischen Darstellung des Architekturkonzeptes werden Meilensteine der prototypischen Implementierung erläutert. Der Beitrag schließt mit einem Ausblick auf Anforderungen und Rahmenbedingungen einer produktiven Implementierung.

Keywords: Knowledge Graph, dezentrale Wissensinfrastruktur, strukturierte Hochschuldaten, Maschinenlesbarkeit, Semantic Web

1 Einführung und Motivation

Die Bereitstellung aktueller, umfassender und exakter Daten zu Hochschulen und insbesondere zu ihren Studienangeboten kann als wesentliche Voraussetzung für eine adäquate Allokation von individuellen, betriebswirtschaftlichen und sogar volkswirtschaftlichen Ressourcen angesehen werden. Relevante Entscheidungen in diesem Kontext sind beispielsweise die Auswahl eines Studiengangs und Studienortes durch einen Studieninteressierten oder der Abschluss einer Kooperationsvereinbarung über duale Studiengänge bzw. über gemeinsame Forschungsprojekte zwischen einem Unternehmen und einer Hochschule. Auf volkswirtschaftlicher Ebene stehen Fragen der öffentlichen Finanzierung und der Sicherung des Fachkräftebedarfs im Mittelpunkt.

Als primäre Quelle der betreffenden Daten sind Hochschulen in eine Vielzahl von Publikationsaktivitäten eingebunden, die sich technisch je nach Adressat deutlich unterscheiden

[1] Technische Hochschule Brandenburg, Fachbereich Wirtschaft, Magdeburger Str. 50, 14770 Brandenburg a.d.H., vera.meister@th-brandenburg.de
[2] ebenda, jonas.jetschni@th-brandenburg.de
[3] CPS-IT GmbH, Gustav-Meyer-Allee 25, Gebäude 13/5, 13355 Berlin, sebastian.kreideweiss@cps-it.de

doi:10.18420/in2017_172

können. So werden mit großem Aufwand Webseiten im Rahmen von Content Management Systemen (CMS) gepflegt, um die breite Öffentlichkeit zu informieren. Zentrale Portale werden regelmäßig mit den aktuellen Daten befüllt – meist in manuellen Redaktionsprozessen. Darüber hinaus werden Daten für die öffentliche Hand bereitgestellt, zumeist in Form semistrukturierter Dateien. Tatsächlich können sich Informationsbedürfnisse und Rezeptionsformen der beteiligten Parteien sehr stark unterscheiden, sodass die Vielzahl an Publikationsprozessen auf den ersten Blick gerechtfertigt erscheint. Webseiten und Portale zielen primär auf menschliche Nutzer, Datenbanken erlauben maschinelle Nutzung, sind jedoch zumeist nicht öffentlich zugänglich bzw. in den zugrundeliegenden Schemata proprietär und untereinander inkompatibel.

Das hier vorgestellte Konzept einer dezentralen Wissensinfrastruktur zielt zunächst auf eine Verringerung des Pflegeaufwandes durch die automatische Bereitstellung konsolidierter Daten, die sowohl von Menschen als auch von Maschinen rezipiert werden können. Es nimmt ferner die Möglichkeiten für eine umfassende Integration von Daten aus den verschiedensten Quellen auf Basis eines abgestimmten Vokabulars in den Blick und zeigt, wie in der Folge ganz neue Wissensdienste aus der Nachnutzung dieser Datenbasis kreiert werden können.

Im folgenden Abschnitt soll zunächst die Ausgangssituation sowohl aus organisationaler als auch aus technischer Sicht näher analysiert werden. Dabei wird auf den Stand des Wissens auch außerhalb der betrachteten Domäne Bezug genommen. Die bereits kurz skizzierten Probleme und Bedarfe werden systematisch dargestellt und exemplarisch quantifiziert. Darauf aufbauend wird in Abschnitt 3 ein Lösungsvorschlag in Form eines integrierten Konzeptes für eine dezentrale Wissensinfrastruktur – einen sogenannten Knowledge Graphen – entwickelt. Abschnitt 4 widmet sich detailliert dem aktuellen Stand der prototypischen Implementierung dieser dezentralen Wissensinfrastruktur und nimmt dabei alle konzeptionellen und technischen Komponenten in den Blick. Abschnitt 5 thematisiert Maßnahmen zur Entwicklung der Community sowie erste Schritte zur produktiven Implementierung.

2 Ausgangssituation

Die Analyse der Ausgangssituation stützt sich auf direkt und indirekt erhobene Daten. In Voruntersuchungen [JM16] wurden an zehn Hochschulen in Deutschland Experteninterviews mit relevanten Akteuren durchgeführt. Darüber hinaus wurde auf Untersuchungen zum Einsatz von CMS an deutschen Hochschulen zurückgegriffen [Re17]. Nicht zuletzt wurde das Leistungsportfolio eines der größten deutschen Hochschulinformationsportale untersucht [Ze17]. Der Ermittlung des aktuellen Entwicklungsstandes bei der Einbindung von organisationalen CMS in Knowledge Graphen liegt eine umfassende Literaturanalyse zugrunde. Die folgende Darstellung der Ergebnisse der Analyse soll sich auf die wichtigsten Akteure, Medien und Technologien beschränken. Bereits hier wird eine große Bandbreite sichtbar, sodass die Problemlage klar erkennbar wird.

Tab. 1 strukturiert das Handlungsfeld Bereitstellung und Nutzung von Daten zu Studienangeboten an Hochschulen nach Akteuren, Medien und Technologien. Diese Übersicht zeigt, dass eine Reihe von Akteuren an Hochschulen damit befasst ist, Daten zu Studiengängen in eine Vielzahl von Systemen bzw. in verschiedenen Formaten einzupflegen. Die Zyklen der Pflege und Aktualisierung unterscheiden sich naturgemäß, sodass neben dem hohen Aufwand auch mit Dateninkonsistenzen zu rechnen ist. Die Leiterin der Marketingabteilung einer kleinen Hochschule (eine der in [JM16] interviewten Experten) berichtete darüber, dass jedes Semester mehr als 50 Portale zu pflegen seien.

Datenanbieter	Medien für die Datenbereitstellung	Technologien für die Datennutzung	Datennutzer
Web-Redakteure der Hochschule	Backend des CMS der Hochschule	Hochschul-Webseiten	Menschen (breite Öffentlichkeit)
Hochschul-Verwaltung	Semistrukturierte Daten (z. B. Excel)	Proprietäre Datenbanken	Systeme der öffentlichen Verwaltung
Marketing der Hochschule	Portalformulare, E-Mail	Proprietäre Datenbanken	Systeme der Portalanbieter
		Proprietäre REST APIs	Webbasierte Dienste
		Portal-Frontend mit Such- und Filterfunktionen, semantisch annotiert	Menschen mit gezieltem Informationsbedarf
			Suchmaschinen

Tab. 1: Datenbereitstellung und -nutzung zu Studiengängen an Hochschulen

Neuere Portalangebote, wie z. B. ZEIT Campus [Ze17], beginnen seit kurzem die Daten auf ihren Portalseiten sowohl per REST API bereitzustellen, als auch semantisch zu annotieren, was als ein sehr sinnvoller Schritt angesehen werden kann. Eine Analyse dieser Angebote zeigt jedoch, dass die Annotationen nicht das Ziel verfolgen, wiedernutzbare Daten zu erzeugen. Die auf den Webseiten beschriebenen Entitäten werden typisiert und mit Metadaten versehen, aber sie haben keine URI zur eindeutigen Identifikation, sondern erscheinen als Blank Nodes. Damit sind sie nicht eindeutig weiter verknüpfbar, sondern dienen allein der besseren Auffindbarkeit durch Suchmaschinen. Das bestätigt auch eine weitere Analyse mit dem Google Test-Tool für strukturierte Daten [Go17].

Tab. 1 zeigt, dass ein bedeutender Teil der Daten in proprietären Datenbanken vorgehalten wird. Diese Daten sind damit nur über eine Vielzahl verschiedener und einzeln zu pflegender Schnittstellen oder auch überhaupt nicht zugänglich. Da Hochschulbildung in Deutschland Ländersache ist, wäre zu erwarten, dass die jeweils verantwortlichen Ministerien dafür sorgen, dass öffentlich zugängliche Daten zu Studiengängen vorgehalten werden. Tatsächlich erfolgt das in den einzelnen Bundesländern auf sehr unterschiedlichem Niveau. Oft wird auf gedruckte Broschüren verwiesen (Berlin und Brandenburg). Einige Länderministerien binden zentrale Portaldienste ein, wie z. B. Bayern den Hochschulkompass der Hochschulrektorenkonferenz.

Bezugnehmend auf Tab. 1 bleibt zu analysieren, inwiefern die CMS der Hochschulen als dezentrale Quellen für strukturierte Daten in Frage kommen. Für die Problematik der semantischen Annotation in CMS finden sich in der angewandten Forschung der letzten 10 Jahre eine Reihe von Lösungsansätzen: (i) Nutzung speziell erweiterbarer Rich-Text-Editoren [He17], (ii) PlugIns für Thesaurus-Manager, wie z. B. PoolParty [Se17], (iii) komplexe Architekturen für die semantische Anreicherung von CMS-Daten auf Basis von Textanalyse und Regelauswertung mit Apache Stanbol [BD13], (iv) semantisches Mapping der in den CMS-internen relationalen Datenbanken gespeicherten Daten, z. B. mittels D2RQ [Cy17].

Den Lösungsansätzen (i)-(iii) ist gemeinsam, dass sie einen zusätzlichen, z. T. spezielles Know-how erfordernden, Aufwand für die Web-Redakteure darstellen, was ihre praktische Implementierung maßgeblich erschwert, wenn nicht sogar verhindert. Ansatz (iv) ist in einer Branche bzw. Community nur dann umsetzbar, wenn die Streubreite bei den eingesetzten CMS nicht zu groß ist. Hochschulen in Deutschland nutzen überwiegend TYPO3, wie Abb. 1 zeigt.

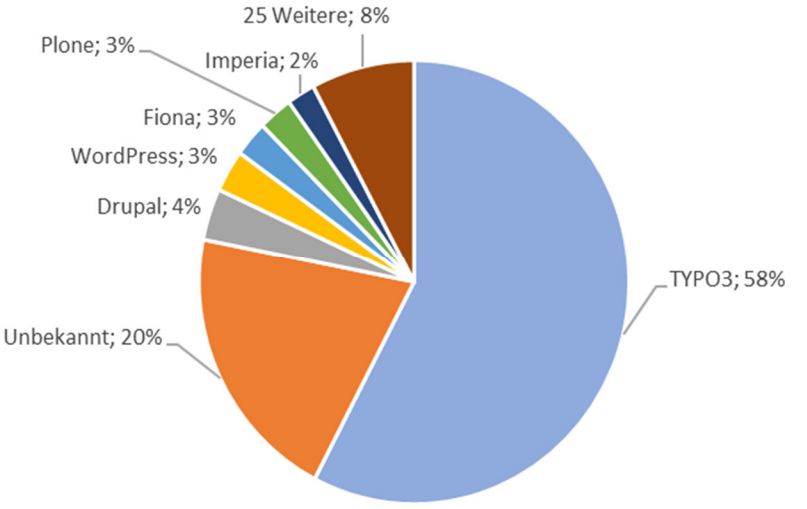

Abb. 1: CMS-Verteilung von Hochschulwebseiten in Deutschland (2017)

Zur Ermittlung dieser Daten wurde ein Dienst der Universität Erlangen-Nürnberg zur Bestimmung von Webseiten-Generatoren genutzt [Re17]. Auf diese Weise lässt sich ermitteln, dass knapp zwei Drittel der 425 untersuchten Hauptseiten deutscher Hochschulen mit dem CMS TYPO3 erzeugt werden. Differenziert man in 273 staatliche und 112 private und 40 konfessionelle Hochschulen, so hält sich TYPO3 in jeder dieser Teilgruppen auf Platz 1. Alle anderen CMS liegen im Bereich unter 5 %, allerdings nimmt mit etwa 20% der Sektor Unbekannt (i.S.v. nicht bestimmbar oder kein CMS vorhanden) den nächst größeren Anteil ein. 25 CMS mit einem Anteil von unter 1 % wurden zur besseren Übersichtlichkeit in der Rubrik „25 Weitere“ aggregiert.

Daten zu Studiengängen an Hochschulen werden also durch verschiedene Akteure über eine Vielzahl von Medien und Systemen bereitgestellt. Diese Daten aktuell zu halten und zu pflegen stellt einen hohen Aufwand dar. Wegen der vielen parallelen Prozesse ist mit Inkonsistenzen zu rechnen. Aktuell werden die so gesammelten Daten nicht in geeigneter Form offen zur Nachnutzung bereitgestellt. Da eine deutliche Mehrheit von Hochschulen in Deutschland das CMS TYPO3 nutzt, erscheint ein CMS-basierter Lösungsansatz erfolgsversprechend. Eine solche technische Implementierung sollte auf offene Standards setzen und zugleich den Web-Redakteuren keine zusätzliche Arbeit abfordern.

3 Lösungskonzept: Knowledge Graph

Ausgangspunkt des Konzepts für eine Wissensinfrastruktur zu Hochschuldaten war das abstrakte Modell von Datenbereitstellung und -nutzung, wie es bereits in Tab. 1 angewandt wurde. Die aus der Ausgangs-/Problemlage identifizierten Anforderungen lassen sich wie folgt aggregieren:

1. Die Dateneingabe soll keinen dauerhaft zusätzlichen Aufwand erzeugen; ein Großteil der redundanten Datenpflegeaufgaben soll eliminiert werden.
2. Die CMS der Hochschulen sollen dabei als initiale Datenquelle eine prominente Rolle spielen.
3. Für die Datenauszeichnung, -speicherung, -weiterverarbeitung und -bereitstellung sollen offene Standards verwendet werden, die die Semantik der Daten zweifelsfrei transportieren und eine breite Nachnutzung durch verschiedene Akteure erlauben. Akteure umfassen dabei sowohl Menschen als auch Maschinen.
4. Das Verlinkungspotenzial zwischen den dezentral eingesammelten Daten soll bestmöglich ausgeschöpft werden.

Eine fünfte Anforderung resultiert nicht unmittelbar aus der Analyse der Problemlage, ergibt sich aber aus den Anforderungen 3 und 4 sowie aus der dynamisch wachsenden und sich etablierenden Webinfrastruktur verlinkter offener Daten [Ab17]:

5. Die eingesammelten Daten sollen durch geeignete Daten aus offenen Quellen erweitert und damit auch qualitativ aufgewertet werden.

Eine so konzipierte Wissensinfrastruktur wird in der Fachliteratur als Knowledge Graph bezeichnet. Dieser Arbeit liegt die Definition von Paulheim zugrunde, nach der ein Knowledge Graph (i) Entitäten der echten Welt und deren Beziehung untereinander in einem Graphen organisiert; (ii) mögliche Klassen und Beziehungen von Entitäten in einem Schema definiert; (iii) es ermöglicht, beliebige Entitäten miteinander in Verbindung zu setzten; (iv) verschiedene fachliche Domänen umfasst [Pa16]. Abb. 2 zeigt das Architekturmodell dieser technischen Infrastruktur ergänzt um Wissensquellen und abgeleitete Wissensdienste. Es folgt damit dem bereits bekannten Paradigma auf einer höheren Ebene. Aufgrund der durchgängig gesicherten Semantik der Daten ist es gerechtfertigt, hier von

Wissensquellen und Wissensdiensten zu sprechen [PS15].

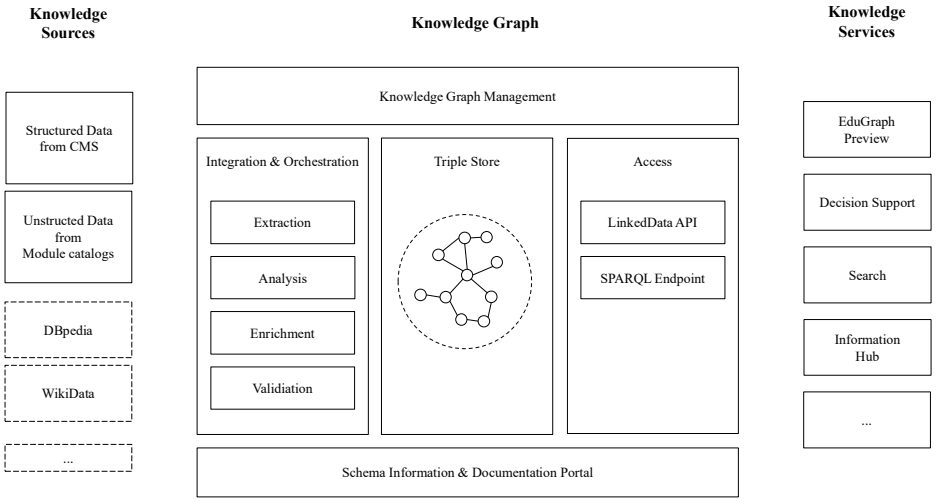

Abb. 2: Zielarchitektur einer dezentralen Wissensinfrastruktur zu Hochschuldaten

Alle grundlegenden Elemente dieser Zielarchitektur wurden bereits prototypisch implementiert [JM17]. Eine detaillierte Beschreibung folgt im nächsten Abschnitt.

4 Prototypische Implementierung

Da die dezentrale Wissensinfrastruktur ein komplexes System verschiedenster konzeptioneller und technischer Komponenten darstellt, gibt es nicht „den einen“ Prototypen, sondern eine Landschaft mehr oder weniger reifer, verknüpfter, vertikaler Prototypen. Sie umfassen das Schema des Knowledge Graphen, eine TYPO3 Extension zur automatischen Auslieferung annotierter Daten über CMS-Seiten, eine Analyse-Anwendung zur Auswertung unstrukturierter Daten, insbesondere von Modulkatalogen für die Spezifikation der Schwerpunkte von Studiengängen, ein Konzept zur Anreicherung der Daten durch Linked Open Data, einen Prozess zur Datenintegration und -persistierung, standardkonforme Schnittstellen für die Datenbereitstellung sowie exemplarische Wissensdienste mit Mehrwert für ausgewählte Nutzergruppen.

4.1 Spezifikation eines Schemagraphen

Nach der oben zitierten Definition eines Knowledge Graphen definiert er mögliche Klassen und Beziehungen von Entitäten der betrachteten Wissensdomänen in einem Schema. Im Hinblick auf eine automatische und reliable Integration von Daten aus unterschiedlichen Quellen sowie eine offene und breite Nachnutzbarkeit der Daten, sollte das Schema

so weit als möglich auf bereits spezifizierte und breit anerkannte, strukturierte Vokabulare setzen.

Zunächst fiel die Aufmerksamkeit auf LRMI (Learning Resource Metadata Initiative) der Dublin Core Education Community [BS15]. Da dieses Vokabular jedoch nur einen kleinen Teil der Domäne abdecken würde und zudem ein zentraler Entwickler dieser Spezifikation (Phil Barker) 2015 die Aufgabe übernahm, eine Education Extension für schema.org [Br17] zu entwickeln, fiel die Wahl auf diesen Ansatz, der sich in der Community der Webentwickler zu einem Quasi-Standard entwickelt. Tatsächlich ist das primäre Anwendungsfeld von schema.org die semantische Auszeichnung von Webseiteninhalten für Suchmaschinen. In Abschnitt 2 wurde diese Art der semantischen Auszeichnung bereits im Kontext des Portals ZEIT Campus diskutiert. Ungeachtet dessen erweist sich schema.org jedoch als ausreichend formal und streng, um auch Dienste zur Datenintegration, -anreicherung und -bereitstellung zu unterstützen. Ein weiteres Argument für die Nutzung dieses Vokabulars ist darin zu sehen, dass es geboten erscheint, den Graben zwischen konventionellen Webentwicklern und Vertretern des Semantic Web zu überbrücken. Das könnte mit schema.org gelingen. Dass es sich hier nicht um einen formal spezifizierten Standard des W3C handelt, sondern um eine Initiative von Big Playern im Suchmaschinengeschäft, wird aus Sicht der Autoren dadurch geheilt, dass die Weiterentwicklung und Pflege des Vokabulars von einer anerkannten Community Group des W3C getragen wird.

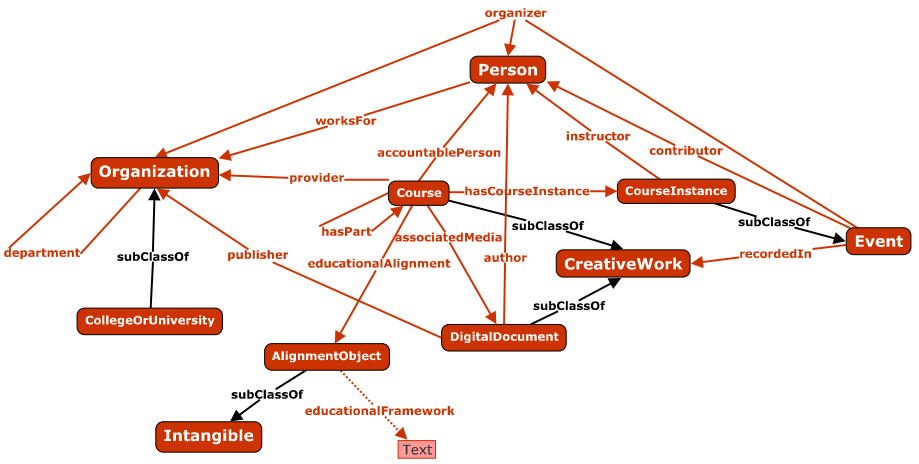

Abb. 3: Zentrale Klassen und Relationen des Knowledge Graph Schemas [MJ16]

Abb. 3 zeigt eine Projektion des Knowledge Graph Schemas auf essentielle Entitäten und deren Relationen, die alle dem Vokabular schema.org angehören. Metadaten sowie proprietäre Elemente für sehr spezifische Klassen und Relationen im Kontext von Studiengängen sind hier nicht abgebildet. Für die Zukunft ist denkbar, dass eine Community-getriebene Entwicklung, ggf. auch in Richtung einer assoziierten Extension zu schema.org, auf

den Weg gebracht wird. Aktuell wird daran gearbeitet, den Prozess der Pflege und Erweiterung des Knowledge Graph Schemas unter Berücksichtigung der Restriktionen der Domäne und der Implikationen auf die Wissensinfrastruktur zu modellieren und zu formalisieren.

4.2 Automatische Bereitstellung von Hochschuldaten

Wie in Abschnitt 3 herausgearbeitet, sollen Hochschul-CMS als initiale Datenquelle eine prominente Rolle spielen, dabei aber kein zusätzlicher Aufwand für Web-Redakteure bei der Dateneingabe und -pflege entstehen. Die Grundidee setzt darauf, dass ein Großteil der relevanten Daten in CMS ohnehin in strukturierter Form in den internen relationalen Datenbanken hinterlegt und somit grundsätzlich formal abrufbar ist. Von einem externen Mapping unter Einsatz z. B. der D2RQ-Technologie wurde abgesehen, da dies den Betrieb einer weiteren zentralen Serverinfrastruktur erforderte, die auf verteilte, interne Datenbanken zugreifen müsste. Das würde eine Reihe von Fragen z. B. zum Datenschutz nach sich ziehen. Aussichtsreicher schien vielmehr, das CMS selbst zu befähigen, strukturierte, maschinenlesbare Daten automatisch und en-passant zu menschenlesbaren Webseiteninhalten auszuliefern.

In Abschnitt 2 wurde argumentiert, dass wegen der enormen Verbreitung von TYPO3 dieser Ansatz erfolgsversprechend erscheint. Diese Annahme wird dadurch bestärkt, dass sich im September 2016 ein TYPO3 Academic Committee als Interessenvertretung TYPO3-nutzender Hochschulen konstituiert hat, das eine synergetische Weiterentwicklung und Nutzung hochschulspezifischer Produktfeatures durch erfahrene Agenturen und Entwickler anstrebt.

Für die semantische Auszeichnung stehen mehrere technische Optionen zur Verfügung. Erste Wahl für die Tag-interne, integrierte Annotation sind Microdata und RDFa. Beide liegen in aktuellen Spezifikationen des W3C vor. Im vorliegenden Fall wurde jedoch auf JSON-LD, einen weiteren W3C-Standard, gesetzt, da die Annotation hier nicht in die Arbeit des Web-Redakteurs integriert, sondern als zusätzliche Datenbankabfrage implementiert werden sollte.

Ein weiteres, bereits im Abschnitt 4.1 angeführtes, Argument zielt wieder auf die Überbrückung des Grabens zwischen konventionellen Webentwicklern und Semantic-Web-Experten. Dieses Ziel wird von JSON-LD nachhaltig verfolgt, liefert es doch die strukturierten Daten in einem sehr flachen Format aus, das von flexiblen Web-Diensten bis hin zu Micro-Services präferiert wird.

Die prototypische Implementierung dieser Schlüsselkomponente der dezentralen Wissensinfrastruktur liegt als Weiterentwicklung einer hochschulspezifischen Extension von TYPO3 vor. Sie wird von zwei im erwähnten Academic Committee engagierten Agenturen vorangetrieben. Aktuell gibt es eine Reihe von Testinstallationen, die die Funktionalität des Konzeptes belegen. Abb. 4 visualisiert ein so generiertes und auslesbares JSON-

LD-Skript. Es wird sichtbar, insbesondere im Vergleich zu den auf [Ze17] implementierten Annotationen, dass die semantische Tiefe und damit die Nachnutzbarkeit der Daten bereits etwas höher ist. Weitere Verbesserungen werden dann möglich, wenn weitere Teildomänen durch entsprechende Extensions unterstützt werden. Das ist auch das Ziel bereits initiierter Entwicklungsschritte.

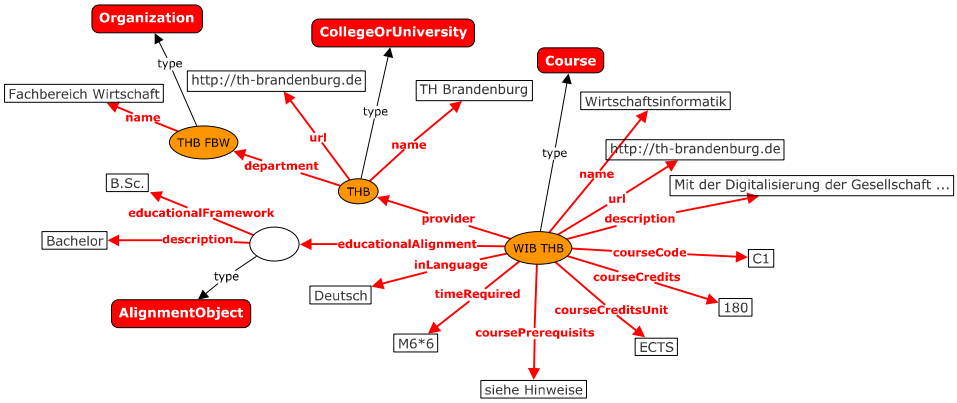

Abb. 4: In Studiengangsseite per JSON-LD-Skript eingebettete strukturierte Daten

4.3 Analyse unstrukturierter Daten

Neben den strukturierten Daten zu Studiengängen an Hochschulen, enthalten einschlägige Webseiten eine Fülle textueller Informationen, die im Sinne der Maschinenlesbarkeit als unstrukturierte Daten anzusehen sind. Aber auch für menschliche Nutzer kann die Vergleichbarkeit von textuellen Informationen aufwändig oder durch abweichende Begrifflichkeiten und Argumentationsmuster erschwert sein. Ein zentrales Informationsdokument zu jedem Studiengang ist der sogenannte Modulkatalog, der alle Module/Fächer eines Studiengangs beschreibt. Eigentlich gibt es dazu ein Spezifikationsdokument der Europäischen Kommission, was jedoch nicht streng verpflichtend ist. In der Praxis finden sich Modulkataloge in den unterschiedlichsten Ausprägungsformen und Formaten.

Die Forschungsfrage in diesem Teilprojekt bestand darin, ein Konzept zur automatischen Analyse von Modulkatalogen zu entwickeln, das es erlaubten würde, den Schwerpunkt eines Studiengangs zu ermitteln. Prototypisch wurde das für den Studiengang Wirtschaftsinformatik Bachelor umgesetzt, der nach den Empfehlungen der GI e. V. drei ungefähr gleichgewichtige Säulen aufweisen soll: Informatik, Betriebswirtschaftslehre und (genuine) Wirtschaftsinformatik. Zusätzlich wird eine vierte Säule erwähnt, die solche Fächer, wie Mathematik, Sprache und Schlüsselkompetenzen umfasst. Tatsächlich sind die Gewichtungen von Hochschule zu Hochschule sehr unterschiedlich, was z. T. mit der organisationalen Einbindung eines Studiengangs in einen Fachbereich zusammenhängt oder mit den fachlichen Präferenzen der maßgeblichen Hochschullehrer. Eine im Jahr 2014 im Rahmen einer studentischen Semesterarbeit durchgeführte Befragung an Hochschulen in

Berlin und Brandenburg ergab, dass sich sowohl Studieninteressierte als auch potenzielle Arbeitgeber für die Schwerpunkte eines Studienganges interessieren. Es wäre also hilfreich, darüber strukturierte Daten abrufen bzw. klar visualisierte Informationen ablesen zu können.

Das Analyse-Tool wurde auf Basis einer Apache Lucene Engine implementiert. In einer Vorverarbeitungsstufe „zerschneidet" es das Dokument in einzelne Module, ermittelt den jeweiligen ECTS-Wert und erstellt zu jedem Modul einen Indexvektor. In der eigentlichen Analyse werden die Abstände dieses Indexvektors zu Säulen-spezifischen Schlüsselwortlisten berechnet. Eine Aggregationsformel ermittelt aus diesen Daten und den ECTS-Werten die Säulengewichtung im betreffenden Studiengang. Über eine REST-Schnittstelle können die Ergebnisse als JSON-File abgerufen werden, wie das folgende Listing exemplarisch zeigt:

```
{"inf":0.242,"nn":0.177,"wi":0.404,"bwl":0.177}
```

Der hier analysierte Modulkatalog beschreibt einen Wirtschaftsinformatik-Studiengang mit einem starken Fokus auf genuine Wirtschaftsinformatik, einer normalen Ausprägung der Säule Informatik und schwächeren Anteilen aus der BWL-Säule und dem Feld der sonstigen Fächer. Die Ergebnisse des Analysetools decken sich mit exemplarischen manuellen Analysen, die im Vorfeld angestellt wurden.

4.4 Anreicherung durch Linked Open Data

Neben den formalen und fachlichen Aspekten zu Studiengängen, die aus den CMS bzw. aus der Analyse von Modulkatalogen extrahiert werden, können auch allgemeine Informationen zur geographischen Lage, zum Studienort selbst oder zu den Wetterverhältnissen vor Ort in verschiedenen Anwendungskontexten interessant sein, wie sie heute als standardisiert ausgezeichnete, offene Daten (LOD) zur Verfügung stehen.

Die dezentrale Wissensinfrastruktur zu Hochschuldaten greift aktuell auf zwei LOD-Wissensquellen zu: die DBpedia und die WikiData. Technisch erfolgt die Integration über SPARQL Queries und Statements. Einen Auszug aus einem solchen Skript zeigt Abb. 5 rechts unten. Die Query fungiert hier als Payload eines Serviceaufrufs im Rahmen eines umfassenden, vollautomatischen Datenintegrationsprozesses, der im folgenden Teilabschnitt beschrieben wird.

4.5 Integration und Bereitstellung

Während die vorhergehenden drei Teilabschnitte die prototypische Implementierung des automatischen Dateninputs aus verschiedenen Arten von Wissensquellen thematisierten (vgl. die linke Seite des Architekturmodells in Abb. 2), soll es nun um den zentralen Teil der Wissensinfrastruktur – den Knowledge Graphen selbst – gehen, der die Datenintegration vornimmt sowie für das Persistieren und die Bereitstellung der Daten sorgt. Auch

wenn für dieses Aufgabenpaket Systeme und Dienste „out of the box“ verfügbar sind, bleibt doch eine Reihe von Implementierungsfragen zu klären:

1. Durch welche Anlässe bzw. wie häufig soll der Prozess der Extraktion, Analyse und Erweiterung der Daten aus den Quellsystemen ausgelöst werden?

2. Wie soll der Prozess unter 1. technisch implementiert werden? Wie transparent soll die Implementierung sein? Welcher Automatisierungsgrad ist umsetzbar?

3. Welches Persistenz-System genügt den Anforderungen an Performanz, Datenintegrität und Robustheit im Fall von Fehlern und Störungen?

4. Welche Schnittstellen für die Bereitstellung der Daten für Zielsysteme sollen unterstützt werden?

Nicht alle diese Fragen können zum aktuellen Zeitpunkt als abschließend geklärt gelten. Zudem genügen einige Elemente der prototypischen Implementierung nicht den Anforderungen an ein Produktivsystem. So wird aktuell ein eher leichtgewichtiger Triple Store in einer ressourcenbeschränkten Serverumgebung eingesetzt. Auch die Performanz des zugehörigen SPARQL Endpoints würde schnell an ihre Grenzen kommen (Frage 3). Dagegen kann die Entscheidung für den Einsatz einer leistungsfähigen Process Engine für die Implementierung des komplexen Datenintegrationsprozesses als zukunftssicher angesehen werden (Frage 2). Abb. 5 zeigt einen Ausschnitt dieses Prozessmodells im Camunda Modeler. Das aufgeklappte Property Panel visualisiert die technische Spezifikation der Datenerweiterung durch Aufruf des SPARQL Endpoints der DBpedia.

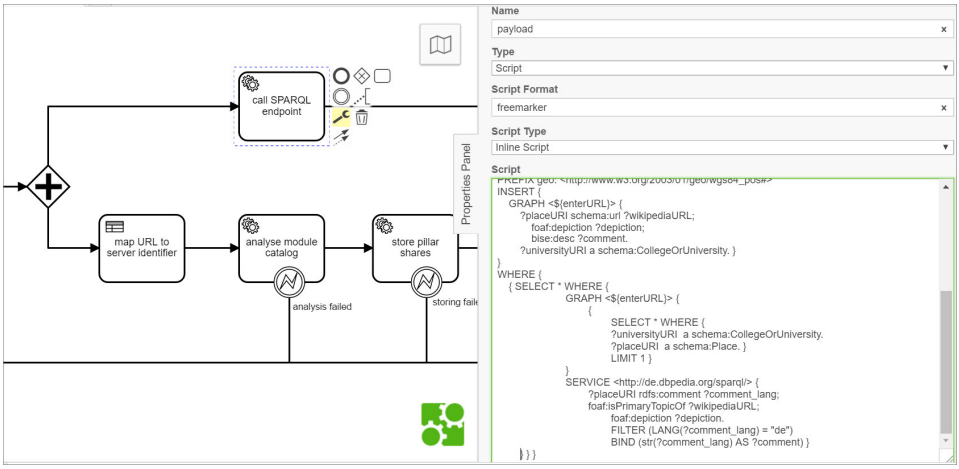

Abb. 5: Screenshot des Camunda Modelers mit Ausschnitt aus dem Integrationsprozess

Zu Frage 1: Daten zu Studiengängen zeichnen sich durch eine geringe Volatilität aus. Eine Anbindung der Datenquellen in Echtzeit bzw. eine Aktualisierung in engen Intervallen wären nicht gerechtfertigt. Prototypisch ist eine wöchentliche Aktualisierung aller Daten

implementiert. Der Datenintegrationsprozess selbst verarbeitet grundsätzlich nur Daten eines Studiengangs. Die vollständige Datenaktualisierung wird durch einen periodisch startenden übergreifenden Prozess gesteuert. In anderen Anwendungsfällen kann es interessant sein, nur die Daten einer Hochschule bzw. eines Studiengangs zu verarbeiten. In diesem Fall kann ein externes Anwendungssystem, wie z. B. das Preview Tool, einen einzelnen, studiengangspezifischen Datenintegrationsprozess auslösen.

Frage 4 ist Gegenstand weiterführender Forschungen und Entwicklungen. Die generische Bereitstellung der Daten über den SPARQL Endpoint des Triple Stores ist ohne technischen Zusatzaufwand oder logische Einschränkungen in der Auswertbarkeit der Daten gegeben. Beim Übergang zum Produktivbetrieb wäre jedoch das Problem der Überlastung des Endpoints bei massiven Anfragen zu adressieren. Das Problem des Technologie-Gaps zwischen konventionellen Web-Technologien und Semantic-Web-Technologien ist prototypisch mit Hilfe eines Dienstes gelöst, der den Zugriff auf die Daten über standardisierte REST APIs erlaubt, die auf vorkonfigurierten SPARQL-Anfragen basieren [MH16].

4.6 Optionen für Mehrwertdienste

Zum Abschluss der Darstellungen zur prototypischen Implementierung einer dezentralen Wissensinfrastruktur zu Hochschuldaten soll nun die Nutzerseite in den Blick genommen werden. Tab. 1 machte bereits deutlich, dass die Hochschuldaten von verschiedenen Nutzerkategorien – sowohl Menschen als auch Maschinen – verwendet werden; allerdings sind dafür in fast allen Kontexten redundante Prozesse der Pflege in unterschiedlichen Systemen und technischen Kontexten notwendig. Das vorgestellte Konzept und die Elemente seiner prototypischen Implementierung haben gezeigt, dass diese Schwierigkeiten durch ein Bündel von Technologien, Prozessen und Entscheidungen überwunden werden können. Diese Infrastruktur unterstützt gleichberechtigt die „Informationsbedürfnisse“ von Menschen und Maschinen. Zum Beleg sollen exemplarisch zwei Use Cases und dazu passende Mehrwertdienste dargestellt werden.

Use Case 1. Web-Administratoren von Hochschulen wollen sich im Vorfeld über Möglichkeiten und Konsequenzen einer Anbindung an die dezentrale Wissensinfrastruktur informieren und dafür Qualität und Informationsgehalt der über ihr CMS gestreuten strukturierten Daten visuell prüfen.

Mehrwertdienst Preview Tool. Eine einfache Nutzeroberfläche erlaubt die Eingabe der URL einer CMS-Seite zum Studiengang. Über den Submit-Button wird der Datenintegrationsprozess angestoßen. Geeignete grafische Elemente visualisieren die integrierten Daten, die sowohl von der CMS-Seite selbst, aus der Analyse des Modulkatalogs sowie aus der Anreicherung aus externen Quellen stammen. Sofern Daten(elemente) fehlen oder nicht analysiert bzw. angereichert werden können, werden entsprechende Fehlermeldungen dargestellt. Abb. 6 zeigt eine beispielhafte Ergebnisseite des Preview-Tools – hier mit einem vollständigen Datensatz.

Use Case 2. Marketing-Verantwortliche an Hochschulen sind daran interessiert, die Studiengangdaten auf zentralen Informationsportalen aktuell zu halten. Im Moment erfordert das die manuelle Pflege einer Vielzahl von Online-Formularen.

Information Hub. Die dezentrale Wissensinfrastruktur stellt zum einen REST APIs bereit, über welche die Informationsportale aktuelle Daten abrufen können. Zum anderen kann auch ein Web-Dienst implementiert werden, der die Portale proaktiv mit Daten im Falle einer Aktualisierung versorgt. Hier liegt eine klare Maschine-zu-Maschine-Kommunikation vor.

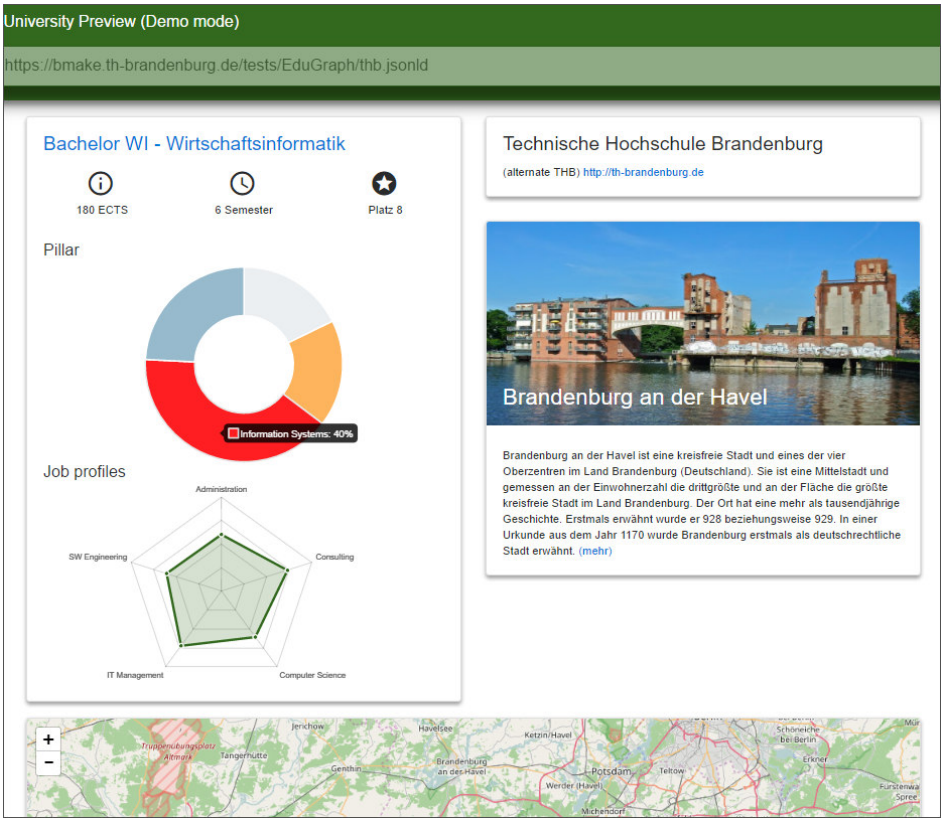

Abb. 6: Exemplarische Ergebnisvisualisierung mit dem Preview Tool

5 Ausblick Weiterentwicklung und produktive Implementierung

In die konzeptionelle Weiterentwicklung, die prototypischen sowie die geplanten produk-

tiven Implementierungen der im Beitrag beschriebenen Wissensinfrastruktur zu Hochschuldaten sind eine Reihe unterschiedlicher Akteure aus Wissenschaft, Hochschulverwaltung, Open-Source-Community, Unternehmen der freien Wirtschaft sowie eine Interessengruppe von Hochschullehrern der Wirtschaftsinformatik involviert. Insbesondere seien hier die TYPO3 Academic Association sowie der Arbeitskreis Wirtschaftsinformatik an Fachhochschulen der Gesellschaft für Informatik genannt.

Die initiale Entwicklung der TYPO3 Extension umfasste als Proof of Concept zunächst nur die CMS-Seiten von Studiengängen. Naheliegend und ohne konzeptionellen Mehraufwand sind im nächsten Schritt die bereits im Knowledge Graph Schema (vgl. Abb. 3) vorgesehenen Konzepte und Relationen zu implementieren. Tab. 2 zeigt in der Übersicht, auf welchen CMS-Seiten, welche Konzepte und Relationen zu implementieren sind und welche Arten von Informationen damit zugänglich gemacht werden können. Auf damit zusammenhängende Attribute (Data Properties) soll hier nicht eingegangen werden. Eine detaillierte Auflistung des geplanten Modellierungsumfangs findet sich unter [MJ16].

CMS-Seiten	Konzepte	Relationen	Informationsarten
Modul	Course, Person	hasPart, accountablePerson, associatedMedia	Modul-beschreibungen
Personen	Person, CreativeWork	author, organizer, worksFor	Aktivitäten und Arbeitsergebnisse
Lehr-veranstaltungen	Person, CourseInstance	hasCourseInstance, instructor, contributor	Veranstaltungspläne
Publikationen	Person, CreativeWork	author, recordedIn	Publikationslisten, Forschungsthemen

Tab. 2: Ausblick auf weitere Implementierungen in der TYPO3 Extension

Semantische Standards befördern in besonderer Weise iteratives Arbeiten, insofern sind auch Erweiterungen des bisherigen Schemas umsetzbar. Für eine gelingende produktive Implementierung können folgende drei Maßnahmen als essentiell identifiziert werden:

1. Entwicklung und Implementierung von Managementstrukturen,
2. Abstimmung und Konsolidierung des Knowledge Graph Schemas und der damit verbundenen Pflege- und Entwicklungsprozesse,
3. Schaffung einer relevanten Datenbasis durch produktive Implementierungen der TYPO3-Extension in Hochschul-CMS.

Bezugnehmend auf das Architekturschema in Abb. 2 handelt es sich um die horizontalen Blöcke des Knowledge Graphen (Maßnahmen 1 und 2) sowie um die initiale (namensgebende) dezentrale Wissensbasis (Maßnahme 3).

Die Maßnahmen 1 und 2 werden aktuell in Vorstudien evaluiert. Ein Projektantrag zur

Vorbereitung der produktiven Implementierung wurde auf den Weg gebracht. Träger sind hier primär die Wissenschaft sowie kooperierende Agenturen. Maßnahme 3 ist eingebunden in Initiativen des TYPO3 Academic Committee. Es ist davon auszugehen, dass der exemplarisch für TYPO3 verfolgte Entwicklungsansatz zur Publikation strukturierter Daten über Hochschul-CMS auch für andere Systeme implementierbar ist. Die dezentrale Wissensinfrastruktur könnte folglich mit geringem Aufwand eine weite Abdeckung sowie vielfältige Nutzungsoptionen für Mensch und Maschine im Hochschulumfeld erreichen.

Literaturverzeichnis

[Ab17] Abele, A., McCrae, J.: Linking Open Data cloud diagram 2017, http://lod-cloud.net/, Stand: 18.04.2017.

[BD13] Behrendt, W.; Damjanovic, V.: Developing Semantic CMS Applications. The IKS Handbook, 1. Auflage, Salzburg Research, Salzburg, 2013.

[Br17] Brickley, D.: Schema.org, https://schema.org/, Stand: 18.04.2017.

[BS15] Barker, P.; Sutton, S.: LRMI Metadata Terms in RDF, http://dublincore.org/dcx/lrmi-terms/, Stand: 09.02.2015.

[Cy17] Cyganiak, R.: D2RQ - Accessing Relational Databases as Virtual RDF Graphs, http://d2rq.org/, Stand: 18.04.2017.

[Go17] Google: Test-Tool für strukturierte Daten, https://search.google.com/structured-data/testing-tool?hl=de, Stand 24.06.2017.

[He17] Herath, S.: Schema.org configuration tool (RDF UI), https://www.drupal.org/project/rdfui, Stand: 18.04.2017.

[JM16] Jetschni, J.; Meister, V.; Prototypische Umsetzung eines dezentralen Studienführers für die Wirtschaftsinformatik an Fachhochschulen im deutschsprachigen Raum. In: Tagungsband zur 29. AKWI-Jahrestagung – Angewandte Forschung in der Wirtschaftsinformatik, Brandenburg a.d.H., S. 318-329, 2016.

[JM17] Jetschni, J.; Meister, V.: Prototypical Implementation of a Community Knowledge Graph, https://edugraph.github.io/architecture, Stand: 18.04.2017.

[MH16] Meroño-Peñuela, A.; Hoekstra, R.: grlc Makes GitHub Taste Like Linked Data APIs. In: 13th Workshop on Services and Applications over Linked APIs and Data, 20

[MJ16] Meister, V; Jetschni, J.: Strukturiertes Wissen an Hochschulen, https://edugraph.github.io/, Stand 24.06.2017.

[Pa16] Paulheim, H.: Knowledge Graph Refinement: A Survey of Approaches and Evaluation Methods. Semantic Web Journal, S. 1-23, 2016.

[PS15] Paschke, A.; Schäfermeier, R.: Einordnung und Abgrenzung des Corporate Semantic Webs. In: Corporate Semantic Web, Springer Berlin Heidelberg, S. 11-21, 2015.

[Re17] Regionales Rechenzentrum Erlangen der Universität Erlangen-Nürnberg, Übersicht der verwendeten Generatoren, Editoren oder Systeme an Hochschulen in Deutschland,

http://statistiken.rrze.fau.de/webauftritte/hochschulen/, Stand: 15.02.2017.

[Se17] Semantic Web Company: Semantic Drupal Modules based on PoolParty Semantic Suite, https://drupal.poolparty.biz/, Stand 18.04.2017.

[Ze17] ZEIT ONLINE GmbH: ZEIT Campus – Suchmaschine für Studiengänge, http://studiengaenge.zeit.de/, Stand 24.06.2017.