Christof Schöch

OPEN ACCESS FÜR DIE MASCHINEN

ZUSAMMENFASSUNG Die Debatten um Open Access kreisen derzeit stark um geeignete Finanzierungsmodelle, wobei aus den Augen verloren wird, dass auch die Publikationsformate grundsätzlich offener, und das heißt, jenseits des PDF-Formats als digitale Reinkarnation des gedruckten Buchs, gedacht werden müssen, wenn das Potential digitaler Technologien für die Wissensproduktion bestmöglich genutzt werden soll. Daher geht es im vorliegenden Beitrag um die Frage von wissenschaftlichen Publikationen als Daten (nicht aber um die Publikation von Forschungsdaten). Zentrale Forderung des Beitrags ist, auf eine Ablösung des PDF-Formats und die Entwicklung und Nutzung offener, standardisierter und die FAIR-Prinzipien erfüllender Datenformate für wissenschaftliche Publikationen hinzuarbeiten. Dabei, so argumentiert der Beitrag, geht es neben den Metadaten, Schlagworten, der Textstruktur und den bibliografischen Verweisen insbesondere auch um die Textinhalte: Es sind Modelle und technische Lösungen dafür nötig, wie zentrale Aussagen einer wissenschaftlichen Publikation in maschinenlesbarer Form in die Publikation selbst eingebettet werden können. Für all diese Aspekte kommt weit verbreiteten, semi-strukturierten Formaten wie XML (beispielsweise TEI, JATS oder RDF) und JSON (wie in BibJSON, ähnlich auch BibTeX) sowie dem Prinzip des Semantic Web mit der Verwendung von Linked Open Data eine wichtige Rolle zu.

SCHLAGWORTE BibTeX, Linked Open Data (LOD), Open Access, Wissenschaftliches Publizieren, XML

ABSTRACT The debates about open access currently revolve around the question of suitable financing models, losing sight of the fact that publication

Christof Schöch: Open Access für die Maschinen, in: Effinger/Kohle (Hrsg.): Die Zukunft des kunsthistorischen Publizierens. Heidelberg: arthistoricum.net, 2021, S. 79–94. https://doi.org/10.11588/arthistoricum.663.c9210 [CC BY 4.0]

formats must also be thought of as fundamentally more open. That means looking beyond the PDF format as the digital reincarnation of the printed book, if the full potential of digital technology to produce knowledge is to be exploited in the best possible way. This paper therefore deals with the question of scientific publications as data (but not with the publication of research data). The central demand of this paper is to work towards a replacement of the PDF format and the development and use of open, standardised data formats for scientific publications that meet the FAIR principles. The paper argues that, in addition to metadata, keywords, text structure and bibliographic references, text content is of particular importance. Models and technical solutions are needed for how central statements of a scientific publication can be embedded in the publication itself in a machine-readable form. For all these aspects an important role can be played by the use of semi-structured formats such as XML (for example TEI, JATS or RDF) and JSON (as in BibJSON, similar to BibTex), as well as the semantic web with the use of Linked Open Data.

KEYWORDS Academic publishing, BibTeX, Linked Open Data (LOD), open access, XML

AKTUELLE DEBATTEN UM OPEN ACCESS

Die Diskussion um das Thema Open Access in den Wissenschaften hat sich in den letzten Jahren deutlich gewandelt. Auch wenn die Praxis vielfach den Überzeugungen etwas hinterherhinkt und das Ausmaß der Nutzung von Preprint-Servern, der Anteil der Open-Access-Zeitschriften oder die freie Verfügbarkeit von Konferenz-Proceedings variieren, ist doch in weiten Teilen des Wissenschaftssystems zumindest für Zeitschriftenartikel und Konferenzpapers nicht mehr strittig, dass das Publizieren im Open Access sinnvoll und wissenschaftsadäquat ist.[1] Zu den gängigsten Argumenten gehört, dass die Ergebnisse von mit öffentlichen Mitteln geförderter Forschung auch der Öffentlichkeit in einem weiten Sinne zugänglich sein sollten und dass wissenschaftlicher Fortschritt am besten durch die weltweite, freie Verfügbarkeit wissenschaftlicher Publikationen gefördert werden kann.

Derzeit wird demnach weniger intensiv über das Warum diskutiert, dafür aber umso mehr über das Wie, wobei der Fokus der Debatten deutlich auf der Frage der angemessenen Finanzierungsmodelle liegt. Die zentrale Frage

1 Für einen fachbezogenen Überblick, siehe den Bereich *Informationen für verschiedene Fächer* auf der Plattform open-access.net (abgerufen am 8.12.2019).

lautet, wie die für Publikation und Dissemination entstehenden Kosten einerseits, die Kosten für langfristige Weiterentwicklung und Verfügbarkeit von Publikationsinfrastrukturen andererseits finanziert werden können, wenn dies nicht mehr wie bisher über Subskriptionen erfolgt. Unter den derzeit diskutierten und praktizierten Modellen sind unter anderem die sogenannten Article Processing Charges (die von den Autor*innen beziehungsweise ihren Institutionen oder Projektfördernden zu tragenden Publikationsgebühren), die groß angelegten Read-and-Publish-Abkommen (wie die vom DEAL-Konsortium angestrebten Verträge mit Großverlagen auf nationaler Ebene) oder neue kollektive Finanzierungsmodelle (wie beispielsweise das Mitgliedschaftsmodell der *Open Library of Humanities*) zu nennen.[2] Wie kann eine Umschichtung gelingen, die sich weg von Subskriptionsbudgets in Bibliotheken und hin zur Förderung von Verlagen und Initiativen bewegt, die im Open Access publizieren? Wie kann vermieden werden, dass die bisherige Zugangsungerechtigkeit (nur wer bezahlen kann, darf wissenschaftliche Ergebnisse anderer lesen) lediglich durch eine Publikationsungerechtigkeit (nur wer bezahlen kann, darf wissenschaftlichen Ergebnisse publizieren) ersetzt wird?[3] Gerade diese letzte Frage hat auch eine stark internationale Dimension und ist insofern auch von politischer Bedeutung.

Die Lösung der Finanzierungsfrage ist zweifelslos von großer Bedeutung, nicht nur, aber gerade auch in den Geisteswissenschaften. Der vorliegende Beitrag möchte den Fokus allerdings auf einen anderen Aspekt der Wie-Frage lenken, der in den intensiven Debatten um die Finanzierungsaspekte derzeit zu kurz kommt: nämlich auf die Frage der wissenschaftsadäquaten Publikationsformate. In der Praxis dominiert klar die PDF-Datei, die als digitale Entsprechung des gedruckten Buches oder Zeitschriftenartikels fungiert. Einige ihrer Eigenschaften erklären die Akzeptanz und den Erfolg dieses Formats: die direkte Entsprechung zwischen Druckfassung und digitaler Fassung bis ins Layout hinein; der Erhalt der Seitenzählung und

2 Vgl. Open-Access.net: *Open Access. Der freie Zugang zu wissenschaftlichen Informationen*, *Geschäftsmodelle*. n. d., unter: https://open-access.net/informationen-zu-open-access/geschaeftsmodelle (abgerufen am 3.6.2020); Speicher, Lara / Armando, Lorenzo / Bargheer, Margo / Eve, Martin Paul / Fund, Sven / Leão, Delfim / Mosterd, Max / Pinter, Frances / Souyioultzoglou, Irakleitos: *OPERAS Open Access Business Models White Paper*. OPERAS, 2018, unter: https://doi.org/10.5281/zenodo.1323707.

3 Piron, Florence: *Qui sait ? Le libre accès en Afrique et en Haïti*, *unter:* https://laviedesidees.fr/Qui-sait.html (abgerufen am 3.6.2020); Pooley, Jeff: *The Library Solution: How Academic Libraries Could End the APC Scourge,* unter: https://items.ssrc.org/parameters/the-library-solution-how-academic-libraries-could-end-the-apc-scourge/ (abgerufen am 4.5.2020).

damit der Möglichkeit, gewohnte Zitierpraktiken weiterzuführen; oder die scheinbare Unveränderlichkeit und damit Verlässlichkeit einer PDF-Datei. Außer für die Distribution und individuelle Lektüre der Publikationen ist dieses Format allerdings (trotz einiger Erweiterungen wie PDF/A für die Archivierung und Tagged PDF für bessere *accessibility*) nur eingeschränkt geeignet. Die Arbeitsgruppe Digitales Publizieren des DHd-Verbands empfiehlt in diesem Kontext beispielsweise: „Nutzen Sie PDF nicht als primäres Publikationsformat (Kodierungsschicht), sondern, wenn überhaupt, als derivatives Leseformat."[4] Sowohl für die Langzeitarchivierung als auch für die computergestützte Auswertung größerer Bestände an Publikationen sind andere Formate klar im Vorteil.

WISSENSCHAFTLICHE PUBLIKATIONEN ALS DATEN

Sobald man sich mit weniger als nur einer Handvoll von Publikationen aus der stetig wachsenden Forschungsliteratur befassen möchte, das heißt, sobald es nicht mehr nur um die Lektüre, sondern um die Verwendung der Publikationen als Datengrundlage für eine quantitative Analyse geht, zeigen sich die zahlreichen Schwächen des PDF-Formats. Wendet man die FAIR-Prinzipien (FAIR: *findable, accessible, interoperable, re-useable*) statt auf die Publikation von Forschungsdaten, auf wissenschaftliche Publikationen als Daten an, wird schnell klar, wie desaströs die aktuell dominierende Praktik der Publikation ausschließlich als PDF-Datei ist.[5] Solche Beiträge sind zwar *findable* (über persistente Identifier und Metadaten, die heutzutage häufig vorliegen) und, wenn sie im Open Access erscheinen, auch ohne größere finanzielle oder technische Hürden *accessible*. Sie sind aber eben nur äußerst eingeschränkt *interoperable* und *re-useable*: So ist der Text in einer PDF-Datei zwar extrahierbar, allerdings ohne wesentliche Strukturinformationen. Die Trennung zwischen Lauftitel, Haupttext und Anmerkungen ist bestenfalls indirekt, über typografische Hinweise oder andere Muster erschließbar. Innerhalb des Haupttextes kann nicht zuverlässig zwischen verschiedenen Textabschnitten (beispielsweise Abstract, Einleitung, analytischem oder interpretierendem Teil, Ergebnissen oder auch zwischen Haupttext und

4 DHd-AG Digitales Publizieren: *Working Paper Digitales Publizieren*, unter: http://dhd-wp.hab.de/?q=content/working-paper-digitales-publizieren (abgerufen am 4.6.2020).

5 Wilkinson, Mark D. / Dumontier, Michel / IJsbrand Jan Aalbersberg u. a.: The FAIR Guiding Principles for Scientific Data Management and Stewardship, in: *Scientific Data* 3, no. 160018, 2016, unter: https://www.nature.com/articles/sdata201618 (abgerufen am 5.6.2020).

Belegzitaten im Blocksatz) unterschieden werden. Auch semantische Information ist nicht explizit vorhanden, denn innerhalb des Textes können Entitäten (Personen, Werke, Organisationen) oder Konzepte (Fachbegriffe, Abstrakta) nicht gezielt adressiert werden. Ebenso wenig kann innerhalb der bibliografischen Angaben gezielt nach Autor*innen, Herausgeber*innen, Titeln, Publikationsdaten oder Verlagen gesucht werden.

Ihr wahres Potenzial können wissenschaftliche Publikationen unter diesen Umständen nicht ausspielen. Dies können sie erst, wenn sie nicht nur digital und frei zugänglich veröffentlicht werden, sondern auch in strukturierten und semantisch angereicherten Formaten verfügbar sind. Entsprechende Publikationsstrategien, bei denen soweit wie möglich nicht nur menschen-, sondern auch maschinenlesbare Publikationen entstehen, werden seit gut zehn Jahren (angelehnt an die Idee des Semantic Web) unter dem Stichwort Semantic Publishing diskutiert.[6] Ähnlich wie im Falle des Aufbaus und der Publikation geisteswissenschaftlicher Datensätze, sind große Mengen wissenschaftlicher Publikationen zwar nützlich, noch besser aber sind semantisch und strukturbezogen angereicherte Publikationen, die so selbst zu Datensätzen werden.

Einige in diesem Zusammenhang einschlägige Anwendungsszenarien seien hier kurz skizziert. Die linguistische Analyse von Wissenschaftssprache interessiert sich so beispielsweise für die sprachlichen Eigenschaften der Texte verschiedener Disziplinen oder unterschiedlicher Typen von wissenschaftlicher Literatur; sie könnte sich mit strukturbezogen annotierten Publikationsdaten aber auch für Vokabular, Stilistik und Argumentationsmustern funktional verschiedener Abschnitte wissenschaftlicher Texte (wie Einleitung, Hauptteil oder Fazit) befassen. Die quantitative Forschung zur Fachgeschichte beispielsweise könnte durch detaillierte und groß angelegte Analysen von Zitationsnetzwerken auf der Grundlage strukturierter Bibliografien auf wesentlich breitere, empirische Grundlagen gestellt werden. Und die korpusbasierte Erarbeitung von Forschungsständen zu einem bestimmten Autor, Werk oder Problem profitiert schon heute von Abstracts und Schlagworten, könnte aber wesentlich präziser und reichhaltiger arbeiten, wenn die wesentlichen Entitäten im Text annotiert und die Kernaussagen aller Publikationen maschinenlesbar abrufbar und automatisch zu einem Netzwerk von Aussagen verknüpfbar wären.

6 Shotton, David: Semantic Publishing: The Coming Revolution in Scientific Journal Publishing, in: *Learned Publishing* 22, 2009, S. 85–94, unter: https://doi.org/10.1087/2009202.

Die wesentlichen Anforderungen an die Möglichkeiten, die solche maschinenlesbaren wissenschaftlichen Publikationen bieten sollten, lassen sich wie folgt zusammenfassen:

1. strukturierte und standardisierte Kodierung von dokumentbezogenen Metadaten (unter anderem bibliografische Angaben; Stichworte; Lizenz; persistente Identifikatoren wie DOIs)

2. explizite Kodierung von Textstrukturen (unter anderem Haupttext versus Anmerkungen; Einleitung, Hauptteil, Fazit; gegebenenfalls Daten, Hypothesen, Methoden, Ergebnisse; Autor*innentext versus Zitate)

3. strukturierte Kodierung von bibliografischen Verweisen (unter anderem bibliografische Angaben einschließlich persistenter Identifier wie DOIs für Forschungsliteratur und gegebenenfalls Primärquellen)

4. maschinenlesbare Auszeichnung der Entitäten (Akteure, Organisationen, Orte, Zeiten) und Konzepte in einem Beitrag (Abstrakta, Fachbegriffe)

5. maschinenlesbare Repräsentation der Kernaussagen eines Beitrags

Der Nutzen der ersten vier hier genannten Anforderungen ist weitgehend unstrittig. Überwiegend liegen auch technische Lösungen vor, die lediglich genutzt und, um diese Nutzung zu fördern, von den bestehenden Publikationsinfrastrukturen (besser) unterstützt oder stärker für die Nachnutzung durch Dritte geöffnet werden müssten.

Die strukturierte Kodierung von dokumentbezogenen Metadaten (Anforderung 1) wird derzeit überwiegend separat von den Artikeltexten selbst in den Datenbanken der Anbieter gehandhabt, wo sie selbstverständlich für Discovery-Zwecke intensiv genutzt werden. Eine stärkere Integration könnte dadurch umgesetzt werden, dass entsprechende Metadaten in den „Properties"-Bereich einer PDF-Datei eingebettet werden. Andere Verfahren sind, den DOI als Verweis auf den Beitrag und die entsprechenden Metadaten vorzuhalten oder, wenn mit einem semi-strukturierten Format gearbeitet wird, diese Metadaten im entsprechenden Bereich der XML-Datei (beispielsweise in JATS oder TEI) zu kodieren.[7]

Für die explizite Kodierung von Textstrukturen beispielsweise durch Zuordnung von Textabschnitten zu strukturellen oder semantischen Klassen (Anforderung 2) gilt, dass dies im Kontext von PDF-Dateien (trotz

7 Siehe zu JATS: https://jats.nlm.nih.gov/; und zu TEI: https://tei-c.org; (abgerufen am 8.12.2019).

eigentlich vorhandener Möglichkeiten) in der Regel nicht umgesetzt wird. Zu sehr dominiert das layoutbezogene Verständnis der PDF-Datei, zu wenig entwickelt sind auch die Infrastrukturen, die solche Informationen nutzen könnten. Hier ist man auf die Möglichkeiten von XML-basierten Formaten wie JATS oder TEI angewiesen, die allerdings im Zeitschriftenbereich (noch) eine marginale Rolle spielen. Bisher akzeptieren nur die wenigsten Zeitschriften oder Verlage Manuskripte in solchen Formaten. Ausnahmen von dieser Regel sind das *Digital Humanities Quarterly* (DHQ) und das *Journal of the Text Encoding Initiative* (jTEI), die XML-TEI verwenden.[8] Die Journal-Anbieter Public Library of Science (PLOS) und Open Library of Humanities generieren für ihre Zeitschriftenartikel eine XML-Fassung in JATS und bieten diese zum Download an. Im naturwissenschaftlichen und informatischen Bereich wird häufig LaTeX akzeptiert. Elsevier beispielsweise konvertiert dieses intern zu XML, publiziert dieses aber nicht.

Für die strukturierte Kodierung von bibliografischen Verweisen (Anforderung 3) wiederum liegen eine ganze Reihe gut etablierter Datenformate vor, unter denen sich BibTeX sicherlich als besonders zentral herausgestellt hat. Zahlreiche Tools, wie das kostenpflichtige Citavi oder das kostenfreie Zotero, ermöglichen die komfortable Verwaltung solcher Daten sowie ihre Nutzung beim Schreiben wissenschaftlicher Texte.[9] Ganz überwiegend werden diese Formate und Programme heute allerdings lediglich dazu genutzt, um eine einheitlich nach einem arbiträren Zitationsstil formatierte Bibliografie zu generieren, die dann dem Text beigefügt wird, allerdings unter Verlust der expliziten Strukturiertheit der Daten. Auch existierende Konzepte für die Erweiterung solcher Daten beispielsweise um Angaben zum Verwendungszweck einer Referenz in einer Publikation unter Verwendung einer Ontologie wie der *Citation Type Ontology* (CiTO) werden kaum genutzt.[10] Diese Daten besser zu nutzen, erfordert durchaus weitreichende Infrastrukturanpassungen, beispielsweise die Möglichkeit, einer gegebenen Publikation gewissermaßen als Supplement eine BibTeX-Datei mit den bibliografischen Angaben beizufügen.[11]

8 Siehe auch den Erfahrungsbericht von eLife: Harrison, Melissa: Collecting XML at Article Submission at eLife: Two Steps Forward, One Step Back?, in: *NCBI*, 2016 unter: https://www.ncbi.nlm.nih.gov/books/NBK350147/ (abgerufen am 5.6.2020).

9 Für Citavi, siehe https://www.citavi.com/; für Zotero, siehe https://www.zotero.org/; (abgerufen am 8.12.2019).

10 Zu CiTO, siehe: http://purl.org/spar/cito (abgerufen am 8.12.2019).

11 Auch Zeitschriften, die eine Einreichung in LaTex + BibTeX akzeptieren, publizieren am Ende in der Regel eine PDF-Datei.

LINKED OPEN DATA FÜR DIE KODIERUNG VON INHALTEN

Kommen wir nun aber zu den beiden letzten Anforderungen, die sich unmittelbar auf den Einsatz von Linked Open Data beziehen. Seit David Shottons Artikel von 2009 hat sich das Publikationswesen stark verändert. Dennoch gilt wohl weiterhin, was er damals formulierte: "With a few shining exceptions, online journals currently provide no semantic mark-up of text that would facilitate increased understanding of the underlying meaning."[12] Die von Shotton genannten Beispiele haben sich nicht durchgesetzt, die semantische Wende des wissenschaftlichen Publikationswesens steht noch aus. Dies hat sicherlich vielfältige Gründe, unter denen wohl auch mangelndes Bewusstsein für den Nutzen und die Möglichkeiten zur Umsetzung einer solchen semantischen Kodierung von Bedeutung ist. An diesem Punkt möchte der vorliegende Beitrag ansetzen.

Die maschinenlesbare Auszeichnung der Entitäten und Konzepte in einem Beitrag (Anforderung 4) ist im Grunde nichts Neues: Es handelt sich hier letztlich um die Grundlage für das Erstellen eines Stichwortverzeichnisses oder Registers, wie sie bei Sachbüchern üblich sind. Diese beziehen sich in der Regel auf Entitäten (wie Personen, Organisationen, Orte und Werktitel) einerseits, auf Konzepte (Abstrakta, Konzepte, Fachbegriffe) andererseits. Neu im Kontext digitaler Publikationen ist allerdings, dass die Indizierung nicht nur innerhalb einer Publikation das Register mit den Fundstellen im Text verbindet und so die Publikation erschließt, sondern dass die Entitäten und Konzepte durch die Verknüpfung mit Normdaten[13] eindeutig identifiziert, in eine domänenspezifische Ontologie integriert und so als Linked (Open) Data Teil des Semantic Web werden können.[14] Zudem sollten solche Auszeichnungen selbstverständlich nicht nur manuell durch die Autoren eingefügt, sondern verfügbare Werkzeuge zur automatischen Annotation (zum Beispiel durch Named Entity Recognition) und zur Eingliederung in jeweils relevante Ontologien genutzt werden.

Durch eine solche Integration der Artikelinhalte in das Semantic Web wird nicht nur eine Indizierung über zahlreiche Publikationen hinweg möglich, sondern die indizierten Entitäten und Konzepte können zugleich dynamisch durch weitere Informationen angereichert werden: erwähnte Personen beispielsweise durch Lebensdaten und Wirkungsort(e) oder

12 Shotton 2009 (wie Anm. 6), S. 87.

13 https://de.wikipedia.org/wiki/Normdatei (abgerufen am 8.12.2019).

14 Einen Überblick bietet Dengel, Andreas (Hg.): *Semantische Technologien. Grundlagen, Konzepte, Anwendungen*, Heidelberg 2012.

Disziplin(en). Beides erfordert Infrastrukturen in einem mehrfachen Sinne: im Sinne von Datenformaten, die eine entsprechende Anreicherung von Publikationen erlauben; von Normdatensätzen, auf die für die Disambiguierung und Anreicherung von Entitäten verwiesen werden kann; und von Publikationsinfrastrukturen, die eine entsprechende Indizierung, Verlinkung und Nutzung der Daten dann auch ermöglichen. Was die Datenformate angeht, so ist JATS hier begrenzt expressiv, während TEI alle wesentlichen Mechanismen bereitstellt. Bezüglich der Publikationsinfrastrukturen sind dem Verfasser keine Publikationsplattformen, Verlage oder Zeitschriften bekannt, die entsprechend annotierte Datenformate bei der Einreichung akzeptieren und die Daten dann auch für die Publikation nutzen würden. Allerdings gibt es hier von anderen Einsatzgebieten von Normdaten, beispielsweise in der Editionsphilologie, einiges zu lernen.[15] Nicht zuletzt erfordert die Integration ins Semantic Web aber eben auch Open Access, damit der freie Zugriff auf relevante Publikationen quer über alle Publikationsorte und nicht nur auf das Portfolio eines Anbieters beschränkt erfolgen kann. Dies steht allerdings in direktem Widerspruch zu den Interessen der Verlage, welche ihre Leser*innen auf der eigenen Plattform halten möchten.

Die letzte oben genannte Anforderung lautet, dass eine maschinenlesbare Publikation ihre Kernaussagen oder Ergebnisse in sorgfältig semantisch modellierter Form anbieten sollte. Seringhaus und Gerstein nennen dies ein „Structured Digital Abstract" und definieren dieses als ein „machine-readable XML summary of pertinent facts in the article".[16] Anders als die bis hierher diskutierten Anforderungen ist diese Anforderung weniger allgemein akzeptiert, zumindest im Kontext des Verfassens wissenschaftlicher Publikationen. Mit diesem Stand der Entwicklung hängt zusammen, dass es sowohl weniger spezifische und ausreichend weit entwickelte technische Lösungen gibt, als auch dass das Thema an sich konzeptionell noch weit weniger gut reflektiert ist. Dabei sind die technische Implementierung einerseits und die konzeptuelle Lösung andererseits zu unterscheiden. Die technische Implementierung erscheint zum aktuellen Stand der Debatte sekundär und ist vor allem eine

15 Siehe Stadler, Peter: Normdateien in der Edition, in: *Editio. Internationales Jahrbuch für Editionswissenschaft* 26, no. 1, 2012, unter: https://doi.org/10.1515/editio-2012-0013; Kamzelak, Roland S.: 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*, München 2016, unter: http://dx.doi.org/10.25366/2018.29.

16 Seringhaus, Michael R. / Gerstein, Mark B.: Publishing Perishing? Towards Tomorrow's Information Architecture, in: *BMC Bioinformatics* 8, no. 17 (2007), unter: https://doi.org/doi:10.1186/1471-2105-8-17.

Frage der Konsensbildung und der verfügbaren Tools in einer Community. Klar scheint allerdings, dass eine solche Implementierung (wie im Falle der Auszeichnung von Entitäten und Konzepten) die Mechanismen von Linked Open Data (LOD) und damit des Semantic Web nutzen sollte.

Der Fokus soll im Folgenden daher auf der konzeptionellen Seite liegen, dem vermutlich kontroversesten Aspekt des Themas. Ein Teil der Schwierigkeit ergibt sich im geisteswissenschaftlichen Kontext zudem daraus, dass hier (anders als beispielsweise in der Biologie oder Chemie, wo zahlreiche relevante Ontologien vorliegen, oder in der Linguistik, wo es mit „Linguistic Linked Open Data (LLOD)“[17] bereits umfassende Erfahrungen und einschlägige Projekte gibt) die Verwendung von Linked Open Data (zumindest jenseits der Kodierung grundlegender bibliografischer Metadaten und jenseits der digitalen Editionswissenschaften) noch kaum verankert ist. Daher sollen hier einige Überlegungen in diese Richtung anhand eines Fallbeispiels angestellt werden, das aus dem Fachgebiet des Verfassers, der Literaturgeschichte, kommt.

Die Fallstudie nimmt die Perspektive der retrospektiven Anreicherung existierender wissenschaftlicher Publikationen durch „Structured Digital Abstracts“ ein. Die hier zu sammelnden Erfahrungen werden aber auch für die Beantwortung der Frage nützlich sein, wie der Inhalt neu entstehender wissenschaftlicher Publikationen maschinenlesbar dokumentiert werden kann. Inhaltlich geht es um den französischen Roman der zweiten Hälfte des 18. Jahrhunderts. Ausgehend von einer Bibliografie aller in Frankreich zwischen 1750 und 1799 erschienen Romane (es sind rund 2000 verschiedene Titel), die bereits als LOD modelliert wurde,[18] werden die dort enthaltenen Entitäten (Romane und Romanautor*innen) nun aus einschlägiger Fachliteratur mit fachwissenschaftlich, das heißt literaturhistorisch relevanten Aussagen angereichert und damit auch erschlossen.[19]Als Beispiel soll der folgende kleine Abschnitt über den Roman *Candide* aus einer literaturgeschichtlichen Überblicksdarstellung von Erich Köhler dienen:

17 Siehe unter: https://linguistic-lod.org/ (abgerufen am 8.12.2019).

18 Vgl. Lüschow, Andreas: Automatische Extraktion und semantische Modellierung der Einträge einer Bibliographie französischsprachiger Romane, in: *Jahrestagung des DHd-Verbands 2020: Spielräume*. Paderborn, 2020, unter: https://doi.org/10.5281/zenodo.3666689; Datensatz: Lüschow, Andreas: *Bibliographie du genre romanesque français 1751–1800 – RDF Model*. Zenodo, 2019, unter: https://doi.org/10.5281/zenodo.3401428.

19 Die hier folgenden Überlegungen sind Teil der Vorüberlegungen für das Projekt *Mining and Modeling Text* (MiMoText), das derzeit am Trier Center for Digital Humanities der Universität Trier anläuft und im Rahmen der Forschungsinitiative des Landes Rheinland-Pfalz gefördert wird. Siehe auch https://www.mimotext.uni-trier.de. Der Verfasser ist Sprecher des Vorhabens.

> „*Candide* ist das meistgelesene Werk Voltaires und war es wohl schon zu Lebzeiten des Autors. Als es 1759 in Genf erstmals im Druck erschien, wurde es zwar sofort verboten, aber doch nur mit dem Ergebnis, daß es im gleichen Jahr noch dreizehn Neuauflagen erlebte.“[20]

Die Grundidee ist nun, zentrale Inhalte eines solchen Textes in Form basaler Aussagen festzuhalten, die in Form von „Subjekt-Prädikat-Objekt“-Statements im Sinne der Linked Open Data formuliert und als sogenannte Tripel beispielsweise in einem Format wie RDF oder Turtle festgehalten werden.[21]

Zunächst einmal kann sich eine semantische, explizite Kodierung von Aussagen in einer wissenschaftlichen Publikation auf die annotierten Entitäten und Konzepte stützen, deren Auszeichnung im Sinne von Anforderung 4 (siehe oben) hier vorausgesetzt wird. Es liegen also bereits Mechanismen vor, mit denen man, einem Linked Open Data-Ansatz folgend, auf Entitäten und Konzepte verweisen und diese als Entitäten in Statements verwenden kann. Relevante Entitäten sind im hier verhandelten Fallbeispiel dann Personen (konkret: Autor*innen entweder von Romanen oder von Fachliteratur; hier: Voltaire), andererseits Werke (konkret: entweder Romane oder fachwissenschaftliche Einzelartikel, Einzelkapitel oder monografische Publikationen; hier: *Candide*).[22] Das Inventar der denkbaren Entitäten ist grundsätzlich als unabgeschlossene Liste zu verstehen und somit auch nicht zu kodifizieren. Darüber hinaus können auch aus grundlegenden literaturwissenschaftlichen Bereichen wie Inhalt, Stil, Genre, Epoche etc. konzeptuelle Entitäten gewonnen werden. Die Annotation (in XML-TEI und unter Verwendung von Identifiern aus Normdatensätzen wie dem VIAF oder dem Getty Thesaurus of Geographical Names) könnte dann wie folgt aussehen:

```
<p><title type="work" ref="viaf:176620251">Candide</title> ist das meistgelesene Werk <persName type="author" ref="viaf:36925746">Voltaires</persName> und war es wohl schon zu Lebzeiten des Autors. Als es <date>1759</date> in <placeName
```

20 Köhler, Erich: *Aufklärung II*, hg. von Henning Krauß und Dietmar Rieger, Stuttgart 1984, S. 8.

21 Siehe zu RDF und Turtle: Dengel 2012 (wie Anm. 14).

22 Zur bibliografischen Erschließung dieser Publikationen benötigt man einige weitere einschlägige Entitäten, wie Städte (offene Liste), Verlage (offene Liste), Jahre, wobei dieser Aspekt hier nicht vertieft werden soll.

type="city" ref="tgn:7007279">Genf</placeName> erstmals im Druck erschien, wurde es zwar sofort verboten, aber doch nur mit dem Ergebnis, daß es im gleichen Jahr noch dreizehn Neuauflagen erlebte.</p>

Weniger konzeptuelle Vorarbeiten bestehen hingegen bei den Aussagen, die nun auf der Grundlage solcher Entitäten formuliert werden können. Diese können erstens (und trivialerweise) natürlich schlicht den Text selbst mit den im Text erwähnten Entitäten verbinden:

Köhler_1984 (viaf:174648806) HAS_SUBJECT Voltaire (viaf:36925746); Candide (viaf:176620251); Genf (tgn:7007279)

Darauf aufbauend ist die zentrale Frage aber nun, wie der Textinhalt formalisiert werden kann. Diese Frage rührt bis an das Grundverständnis einer gegebenen Disziplin, denn es geht darum festzulegen, welche Art von Aussagen eine wissenschaftliche Fachcommunity nun jeweils als so grundlegend für eine bestimmte Domäne erachtet, dass sie für sie einen Aussagentyp formalisiert. Im obigen Beispiel sind eine Reihe von Aussagen enthalten, die hier in Frage kämen und die hier pseudo-formalisiert genannt werden:

- Voltaire (viaf:36925746) IS_CREATOR_OF Candide (viaf:176620251)
- Candide (viaf:176620251) HAS_PUBLICATION_DATE 1759
- Candide (viaf:176620251) HAS_PUBLICATION_LOCATION Genf (tgn:7007279)
- Candide (viaf:176620251) HAS_RECEPTION_INTENSITY high
- Candide (viaf:176620251) HAS_RECEPTION_TIME immediate;long-term
- Candide (viaf:176620251) HAS_LEGAL_STATUS censored (1759)

Dabei sind die ersten drei Aussagen nicht viel mehr als bibliografische Metadaten, wie man sie auch in einem Katalog oder einer Sachbibliografie finden könnte (und wie sie in unserem Fall bereits vorliegen). Für einen Teil dieser Aussagetypen, insbesondere wenn es sich um prosopografische und bibliografische Informationen handelt, kann man entsprechend für die Formalisierung auf vorhandene Ontologien zurückgreifen, zum Beispiel auf Dublin Core (für creator, publisher, date, title, subject) oder die SPAR Ontologies (für die

weiterführende bibliografische Modellierung).[23] Für die darauf folgenden Aussagen gilt dies aber nicht. Die zentrale Frage ist demnach, wie eine Ontologie zentraler Aussagetypen für eine bestimmte wissenschaftliche Domäne (hier: die Literaturgeschichte als Teil der Literaturwissenschaften) gestaltet sein sollte und wie ein Konsens zu diesen Themen in der Community hergestellt werden kann. Welche domänenspezifischen (und das heißt hier: genuin literaturhistorischen) Informationen sollten als basale Statements formuliert werden können?

Vergleichsweise unstrittig dürften, ähnlich wie die bereits erwähnten, grundlegenden bibliografischen Aussagen, etablierte prosopografische Informationen sein, wie man sie beispielsweise in Wikidata findet:

- (Person) DATE_OF_BIRTH (Datum); DATE_OF_DEATH (Datum)
- (Person) OCCUPATION (Berufsbezeichnung)
- (Person) RELIGION (Religionsbezeichnung)
- (Person) MOVEMENT (Ideologie, Weltanschauung, Bewegung)

Etwas stärker domänenspezifische Aussagen, wie sie bislang zwar nicht offiziell im Rahmen einer Ontologie standardisiert sind, aber beispielsweise in Wikidata praktiziert werden, sind die folgenden:[24]

- (Person) INFLUENCED_BY (Person)
- (Person) AWARD_RECEIVED (Auszeichnung)
- (Werk) GENRE (Gattung)
- (Werk) CHARACTERS (Figurennamen)
- (Werk) NARRATIVE_LOCATION (Geografischer Ort)
- (Werk) SET_IN_PERIOD (Zeitspanne)
- (Werk) DERIVATIVE_WORK (Werk)

23 Wikipedia contributors: Dublin Core, in: *Wikipedia. The Free Encyclopedia*, 2019, unter: https://en.wikipedia.org/w/index.php?title=Dublin_Core&oldid=922336659, (abgerufen am 8.12.2019); Peroni, Silvio / Shotton, David: The SPAR Ontologies, in: *International Semantic Web Conference* (ISWC 2018), 2018, unter: https://doi.org/10.1007/978-3-030-00668-6.

24 Wikidata contributors: Wikidata Property Related to Works of Fiction, in: *Wikidata*, 2019. https://www.wikidata.org/wiki/Wikidata:List_of_properties/work#Wikidata_property_related_to_works_of_fiction (abgerufen am 8.12.2019).

- (Werk) INSPIRED_BY (Werk)
- (Werk) NARRATOR (Figurennamen)

Hier beginnt deutlich zu werden, dass eine systematische Auseinandersetzung mit dieser Art von Aussagen in Form einer Ontologie noch aussteht. So stehen beispielsweise „RELIGION" und „MOVEMENT" in einer unklaren Beziehung oder gibt es auf der Personenebene „INFLUENCED_BY" und auf der Werkebene „INSPIRED_BY". Für manche Aspekte könnten vorhandene Taxonomien oder Ontologien nachgenutzt werden, wie beispielsweise im Bereich der (historischen und aktuellen) Berufsbezeichnungen;[25] für andere Aspekte, wie (literarische) Gattungen, Epochen, Formen oder Themen gibt es keine vergleichbar formalisierten und konsensuellen Ressourcen. Und natürlich sind die in Wikidata bisher verwendeten Prädikate keineswegs ausreichend für eine literaturhistorisch adäquate Beschreibung literarischer Werke, Autor*innen und Epochen. Nur in Bezug auf literarische Werke selbst wären beispielsweise folgende weitere Sachinformationen relevant:

- (Werk) HAS_FORM (Prosa|Vers|Anderes)
- (Werk) HAS_NARRATIVE_PERSPECTIVE (autodiegetisch|homodiegetisch|heterodiegetisch) – Erzählform bei narrativen Werken
- (Werk) HAS_DIALOGUE_PROPORTION (Prozentsatz) – Anteil der direkten Rede in einem narrativen Werk, in Prozent der Wörter oder Sätze
- (Werk) HAS_STAGE_DIRECTIONS (Prozentsatz) – Anteil der Bühnenanweisungen in einem dramatischen Werk, in Prozent der Wörter.

Dies ist selbstverständlich nicht einmal in Ansätzen eine abschließende Auflistung; eine systematische Modellierung der Domäne steht noch aus. Zu ergänzen ist in diesem Zusammenhang allerdings noch, dass die jeweils erhobenen oder extrahierten Informationen nicht als Fakten, sondern als Aussagen aufgefasst werden: insofern jedes Statement einer Quelle zugeordnet ist, repräsentiert es die Meinung einer Fachwissenschaftlerin oder eines Fachwissenschaftlers beziehungsweise den Stand der Forschung zum

25 So beispielsweise die Historical International Standard Classification of Occupations (HISCO); siehe: Van Leeuwen, Marco H. D. / Maas, Ineke / Miles, Andrew: Creating a Historical International Standard Classification of Occupations An Exercise in Multinational Interdisciplinary Cooperation, in: *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 37, no. 4 (September 1, 2004), S. 186–197, unter: https://doi.org/10.3200/HMTS.37.4.186-197.

Zeitpunkt der Publikation. Folglich kann ein Informationssystem, das große Mengen solcher Aussagen versammelt, auch sich gegenseitig widersprechende oder anderweitig inkompatible Aussage beinhalten, ohne dass der Bestand des Systems deswegen als inkonsistent gelten müsste.

Im derzeit laufenden Projekt Mining and Modeling Text (MiMoText) geht es zwar nur indirekt um die Frage, wie wissenschaftliche Publikationen zukünftig erschlossen werden sollten. Zentral ist hingegen das Ziel, in einem Korpus bestehender (insbesondere auch älterer, überblickshafter) Fachliteratur dem skizzierten Ansatz folgend und weitgehend automatisch ein bestimmtes Inventar literaturhistorischer Aussagetypen zu identifizieren und semantisch zu modellieren. Auch ältere Literaturgeschichtsschreibung soll wieder sichtbar und in großem Umfang nutzbar gemacht werden, indem sie als Linked Open Data unter Verwendung von domänenspezifischen Ontologien modelliert und publiziert wird. Zudem sollen auf diese Weise auch über selten gelesene Romane Informationen ermittelt werden und in das so entstehende, literaturhistorische Informationssystem einfließen. Auf diese Weise entsteht, eine ausreichend umfangreiche Menge semantisch erschlossener Fachliteratur vorausgesetzt, ein literaturhistorisches Informationssystem, das eine Reihe von Anwendungsszenarien unterstützt. Beispielsweise wird es so möglich, die Rezeptionsgeschichte einer bestimmten Autorin nicht nur quantitativ zu ermitteln (beispielsweise über die Anzahl der relevanten Publikationen pro Jahr), sondern auch inhaltlich nachzuvollziehen, indem Entwicklungen in den jeweils angesprochenen Themen, in den Bewertungstendenzen oder den jeweils mobilisierten Vergleichsautor*innen analysiert werden. Ebenso wird es möglich sein, auf der Grundlage der im System enthaltenen (inhaltlichen, stilistischen, bewertenden, einordnenden usw.) Aussagen literarische Werke zu identifizieren, die den jeweils gewählten Kriterien zufolge Gemeinsamkeiten haben und sich so für weiterführende, vergleichende Analysen eignen könnten.

Diese Form der recht aufwendigen, retrospektiven semantischen Erschließung würde zukünftig überflüssig, wenn neu publizierte Fachliteratur ihre wesentlichen Aussagen eben bereits in Form von Linked Open Data mitveröffentlichte. Sei es, dass die Identifikation der relevanten Entitäten und das Formulieren der entsprechenden Aussagen von den Autor*innen selbst geleistet werden oder dass Verfahren entwickelt werden, die dies automatisch auf Grundlage des Volltextes bewerkstelligen.[26] Für die Zukunft wäre eine

26 Dies könnte unter Nutzung von Verfahren geschehen, wie sie seit einiger Zeit im Bereich des Argument Mining und der automatischen Semantic Annotation entwickelt werden; siehe: Lippi, Marco, / Torroni, Paolo: Argumentation Mining: State of the Art and Emerging Trends, in: *ACM Transactions on Internet Technology* 16.2 (2016), S. 1–25, unter: https://doi.org/10.1145/2850417; Uren, Victoria / Cimiano, Philipp / Iria, José / Handschuh,

größere Präzision (bei gleichzeitig geringerer Abdeckung) zu erwarten, wenn sich eine entsprechende Praktik beim Verfassen von wissenschaftlichen Publikationen etablieren würde, ähnlich wie es derzeit üblich ist, beim Einreichen von Artikeln oder Kapiteln auch eine Reihe von Schlagworten anzugeben beziehungsweise Begriffe aus einer fachwissenschaftlichen Ontologie auszuwählen. Bis dahin ist es allerdings noch ein weiter Weg und zukünftige wissenschaftliche Publikationsweisen und die retrospektive Erschließung der Fachgeschichte bleiben voneinander nicht unberührt: Einerseits entstehen durch eine zeitgemäße Praktik der semantischen Annotation wissenschaftlicher Publikationen in Kombination mit dem Volltext auch Trainingsdaten für das Entwickeln automatischer Verfahren. Andererseits kann die retrospektive Erschließung auch dazu beitragen, Anforderungen an zukünftige semantische Annotationsverfahren wissenschaftlicher Publikationen und an zugrunde liegende, domänenspezifische Ontologien zu präzisieren.

Die Vision des Verfassers ist jedenfalls, dass wir in naher Zukunft Forschungsergebnisse nicht mehr nur in natürlichsprachiger Prosa formulieren und als PDF-Dateien publizierte Artikel oder Bücher produzieren, verbreiten und rezipieren, und dass diese Prosa auch nicht unverbunden mit der dazugehörigen Publikation von Datensätzen und dem Programmiercode erfolgt. Vielmehr wird dieser Prosatext mit relevantem Code und Datensätzen verbunden, mit reichhaltigen Metadaten versehen, in seiner Textstruktur ausgezeichnet, unter Verwendung strukturierter bibliografischer Daten mit Entitäten und Konzepten annotiert und in Form von LOD-Statements zusammengefasst publiziert werden. Dass der in natürlichsprachiger Prosa ausformulierte Fließtext dadurch obsolet wird, soll nicht behaupten werden; aber der Fließtext wird in Zukunft nicht mehr allein stehen, sondern eingebettet sein in einen reichhaltigen, maschinenlesbaren Kontext von Daten, Code, Metadaten, Zitationsdaten und modellierten Aussagen.[27]

ORCID®
Christof Schöch https://orcid.org/0000-0002-4557-2753

Siegfried / Vargas-Vera, Maria / Motta, Enrico und Ciravegna, Fabio: Semantic Annotation for Knowledge Management: Requirements and a Survey of the State of the Art, in: *Web Semantics: Science, Services and Agents on the World Wide Web* 4/1 (2005), S. 14–28.

27 Den Herausgebern des vorliegenden Bandes, Maria Effinger und Hubertus Kohle, möchte ich an dieser Stelle herzlich für die Initiative danken, einige der im Beitrag angesprochenen Impulse für die Publikation des Bandes aufzugreifen. Ein Datensupplement zu dieser Publikation wurde auf Zenodo.org unter dem folgendem DOI abgelegt: https://doi.org/10.5281/zenodo.3898418.