

Automatische Extraktion und semantische Modellierung der Einträge einer Bibliographie französischsprachiger Romane

Lüschow, Andreas

andreas.lueschow@gmx.de
Universität Trier, Deutschland

Einleitung

Unter dem Begriff des *Semantic Web* (Berners-Lee, Hendler, Lassila 2001) werden Techniken, Standards und Methoden zusammengefasst, mit deren Hilfe im Internet verfügbare Daten der semantischen Verarbeitung durch Maschinen zugänglich gemacht werden können. Durch die Einführung und Nutzung von offenen Standards wie z. B. RDF (Schreiber & Raimond 2014) soll hierbei die Interoperabilität unterschiedlicher Datenquellen sichergestellt werden. Diese Standards beziehen sich auf die Art, wie Informationen repräsentiert werden und wie Verknüpfungen mit anderen Informationen hergestellt werden können. Daher wird oftmals auch der Begriff der *Linked Data* verwendet (Bizer, Heath, Berners-Lee 2009). In einer Visualisierung der Linked-Data-Cloud von 2017 (Freyberg 2017: 29) sind die Geisteswissenschaften als eigener Bereich nicht explizit aufgeführt, was die geringe Veröffentlichung geisteswissenschaftlicher semantischer Daten widerspiegelt bzw. vermuten lässt, wenngleich z. B. im Bereich der Graphentechnologien durchaus einige Projekte existieren (Kuczera 2017).

Metadaten als Basis literaturwissenschaftlicher Forschung

Dabei sind solche Daten Basis vieler (literatur-)wissenschaftlicher Fragestellungen: Soll bspw. eine quantitative Textanalyse einer großen Anzahl von Romanen durchgeführt werden, müssen zunächst einmal die in Frage kommenden Werke ermittelt und ausgewählt werden. Die Erstellung solcher möglichst repräsentativen Samples ist allerdings ohne eine Kenntnis

der gesamten Romanproduktion einer Epoche, der dort behandelten Themen und Motive und weiterer Angaben über die inhaltliche Ausgestaltung der zu betrachtenden Textproduktion nicht ohne Weiteres möglich.

Hierbei helfen können Nachschlagewerke wie z. B. Fachbibliographien, in denen bibliographische Metadaten verzeichnet sind. Teilweise liegen solche Metadaten bereits als Linked Data vor, da Bibliothekskataloge (retro-)digitalisiert wurden. Diese Metadaten sind als Basis literaturhistorischer Arbeit jedoch häufig nicht ausreichend, da für eine zielgerichtete Auswahl relevanter Literatur oftmals mehr als die üblicherweise erschlossenen bibliographischen Angaben notwendig sind.

Einen weiteren, großen Anteil an der prinzipiell verfügbaren Literatur haben jedoch auch Werke, die nicht digitalisiert, sondern nur in gedruckter Form vorliegen. Die *Bibliographie du genre romanesque français 1751-1800* (Martin, Mylne, Frautschi 1977) fasst alle von den Autoren auffindbaren französischsprachigen Romane aus der zweiten Hälfte des 18. Jahrhunderts zusammen. Neben bibliographischen Daten zu Autoren, Werktiteln, Verlegern u. a. sind, soweit möglich, auch Angaben zu weiteren Auflagen (Reeditionen) und zum Inhalt der Werke zusammengetragen worden. Die Bibliographie enthält somit inhaltliche Informationen zu den einzelnen Romanen, die weit über eine Auflistung bibliographischer Metadaten hinausgehen. Solche Informationen sind wie o. g. notwendige Voraussetzung für die Erstellung repräsentativer Samples, u. a. zur weiteren literaturhistorischen Untersuchung der Textproduktion einer Sprache bzw. Epoche.

Zielsetzung

Im Rahmen des hier präsentierten Vorhabens – einer Masterarbeit im Studiengang Digital Humanities an der Universität Trier – wurde die o. g. Bibliographie eingescannt und mittels *Optical Character Recognition* (OCR) in maschinenlesbaren Text umgewandelt. Auf dieser Grundlage wurden mithilfe eines Verfahrens des überwachten maschinellen Lernens die einzelnen Einträge extrahiert, in ein selbst entwickeltes semantisches Modell überführt und mit externen Daten verknüpft, sodass die Bibliographie nunmehr als RDF-Datensatz vorliegt und weiterverwendet werden kann.¹ Zielsetzung der Arbeit war es, die in der Bibliographie enthaltenen Informationen unter Nutzung bibliographischer Standards und aktueller, verbreiteter Datenmodelle auf eine Art und Weise zu repräsentieren, die zukünftig weitere Verarbeitungen und Anreicherungen ermöglicht. Die so entstandene digitale Bibliographie kann darüber hinaus als Basis für buchwissenschaftliche, literaturhistorische und verwandte Forschungen dienen, da in ihr sowohl formale als auch inhaltliche Metadaten zur französischsprachigen Romanproduktion eines definierten Zeitraums enthalten sind.

Metadatenextraktion

Ablauf

Der Ablauf der Metadatenextraktion ist in Abbildung 1 dargestellt.

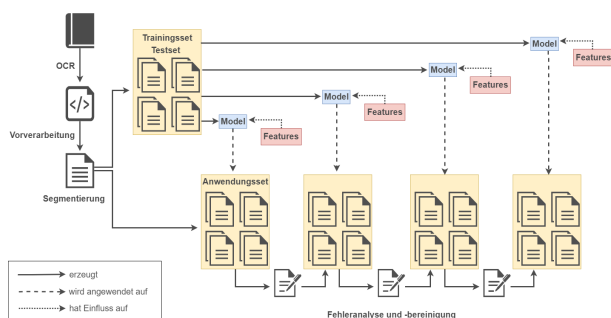


Abbildung 1: Ablauf der Metadatenextraktion

Nach dem Einscannen der gedruckten Vorlage, der OCR, der Vorverarbeitung (Korrektur von Fehlern, Entfernen von Vorwort und Abbildungen, einheitliche Zeichenkodierung etc.) wurden die einzelnen Jahreslisten der Bibliographie und innerhalb dieser die einzelnen Einträge/Romane durch XML-Markup voneinander getrennt (Segmentierung).

Anschließend wurde ein Trainingsset erstellt, mit welchem der verwendete Algorithmus trainiert werden konnte. Für die Trainingsdaten wurde aus jedem Jahrzehnt ein Jahr ausgewählt und die Metadaten der in diesem Jahr erschienenen Romane wurden manuell mit XML-Markup ausgezeichnet. Zur Evaluation der Modelle wurde ein Teil der Daten als Testset zurückgehalten.

Das maschinelle Lernen verlief iterativ, sodass jeweils Modelle für unterschiedlich „tiefe“ Metadatenebenen gelernt wurden, da eine mehrstufige Anwendung mehrerer Modelle oftmals bessere Ergebnisse als die Verwendung eines einzigen Modells für die gesamten Daten erzielt (Kovacevic et al. 2011: 388) und simpler strukturierte Modelle weniger Trainingsdaten benötigen (Candeias 2011: 28). Ein erstes Modell wurde bspw. zur Bestimmung der Makrostruktur der Metadaten verwendet (Titel, Autor, Publikationsdetails etc.), weitere Modelle verfeinerten jeweils die Auszeichnung innerhalb einer dieser Gruppen (z. B. Differenzierung der Publikationsdetails: Ort, Verleger, Jahr, Format, Seitenangabe). Insgesamt wurden sechs Modelle trainiert, die durch stichprobenartige Analyse der erzeugten Daten sukzessive angepasst wurden, bis keine Verbesserungen mehr möglich waren. Das jeweils beste Modell einer Iteration wurde dann auf die restlichen, noch nicht im Trainings- bzw. Testset enthaltenen Jahreslisten angewendet.

Algorithmus und Features

Zur Modellbildung wurden *Conditional Random Fields* (CRF), ein Verfahren des überwachten maschinellen Lernens, verwendet (Lafferty, McCallum, Pereira 2001), das sich in den letzten Jahren zu einem wesentlichen Verfahren im Rahmen der Informationsextraktion entwickelt hat (vgl. z. B. Groza, Grimnes, Handschuh 2012). CRF kombinieren die Vorteile von *Hidden-Markov-Modellen* (HMM) und *Support Vector Machines* (SVM), zwei weiteren gut untersuchten Verfahren (Peng, McCallum 2004: 329).

Die in den Algorithmus eingespeisten Daten (hier: Wörter bzw. Token) werden als Sequenzen von Zuständen modelliert und auf Grundlage dieser beobachteten Zustände werden Label für die einzelnen Elemente vergeben. Im Gegensatz zu HMM berücksichtigen CRF jedoch mögliche Beziehungen der Elemente untereinander – im vorliegenden Fall also der Metadatenfelder bzw. der berücksichtigten Features. Da die Einträge der Bibliographie einem definierten Schema folgen (z. B. steht immer zuerst die Autorenangabe, dann folgt der Titel), ist dieser Algorithmus zur Modellierung der vorliegenden Daten besonders geeignet.

Tabelle 1: In den Modellen berücksichtigte Features

Feature	Erklärung
word	Einzelnes Wort wie es im Text vorkommt
word.lower	Wort in Kleinbuchstaben
word[-3:], word[-2:], word[-1:]	Die letzten Zeichen des Wortes
word[-1:].isalpha	Endet das Wort mit einem Buchstaben?
word[:3], word[:2], word[:1]	Die ersten Zeichen des Wortes
word[:1].isalpha	Ist das erste Zeichen ein Buchstabe?
word.isupper	Besteht das Wort nur aus Großbuchstaben?
word.istitle	Beginnt das Wort mit einem Großbuchstaben?
word.isdigit	Besteht das Wort nur aus Zahlen?
word.isalpha	Besteht das Wort nur aus Buchstaben?

Damit ein CRF-Modell trainiert werden kann, müssen Features erhoben werden, die den Inhalt der einzelnen Metadatenfelder repräsentieren. Tabelle 1 gibt die genutzten Features wieder. Diese Features wurden nicht nur für das jeweilige Wort, sondern auch für das vorherige und das nachfolgende Wort erhoben. So kann im Modell bspw. gelernt werden, dass auf ein bestimmtes Wort stets eine Zahl folgt.

Die genutzten Features wurden ausgehend von einer manuellen Analyse der Einträge in der Bibliographie und basierend auf den ausführlichen Erläuterungen der Autoren zur Sammlung und Strukturierung der Daten im Vorwort der Bibliographie ausgewählt. In der gedruckten Vorlage wurde Großschreibung bspw. zur Hervorhebung von Familiennamen verwendet und Angaben zum Inhalt eines Romans folgten fest definierten einleitenden Begriffen.

Eine ausführliche Evaluation unterschiedlicher Feature-Kombinationen fand im Rahmen der Arbeit nicht statt, da bereits die o. g. simplen Features zu ausreichend hoher Genauigkeit der Metadatenextraktion führten. Weitere Optimierungen hätten überdies vom eigentlichen Ziel der Arbeit weggeführt. Die zur Unterscheidung der

einzelnen Metadatenfelder günstigsten Features wurden jedoch erhoben, um die Wirksamkeit und innere Struktur der gelernten Modelle zu überprüfen. Hierbei zeigte sich z. B., dass die einleitenden Wendungen zur inhaltlichen Beschreibung der Romane auch vom Algorithmus als solche gelernt und zur Auszeichnung neuer Daten verwendet wurden.

Um auch weniger strukturierte Datengrundlagen als Bibliographien mit dem entwickelten Workflow verarbeiten zu können, bestünde hier ein möglicher, näher zu untersuchender Ansatzpunkt für eine genauere Analyse hilfreicher Features und die eventuelle Einführung weiterer Features.

Evaluation

Das maschinelle Lernen wurde mithilfe der Programmiersprache *Python* und der dort verfügbaren Bibliothek *sklearn-crfsuite*² implementiert. Die Evaluation der Modelle geschah mit der zu *sklearn-crfsuite* kompatiblen Bibliothek für wissenschaftliche Programmierung *scikit-learn*³. In der folgenden Tabelle sind die gängigen Maße Precision, Recall und der F1-Score für die sechs gelernten Modelle angegeben.

Tabelle 2: Evaluation der einzelnen Modelle

Modell	Precision	Recall	F1
entry (Makrostruktur des Eintrags)	0,954	0,953	0,951
det (Publikationsdetails)	0,987	0,986	0,986
res (Schlagwörter)	0,919	0,907	0,908
au (Autorennamen)	0,975	0,986	0,980
ae (Makrostruktur weiterer Editionen)	0,997	0,997	0,997
ae_se (einzelne Einträge weiterer Editionen)	0,961	0,960	0,960

Für alle Metadatenfelder konnte eine sehr hohe Genauigkeit erreicht werden. Der so erzeugte Datensatz mit allen Einträgen aus der Bibliographie ist somit nahezu vollständig korrekt ausgezeichnet.

Semantische Modellierung

Zurzeit existiert kein einheitlicher, akzeptierter Standard, der in der Bibliothekswelt für die semantische Repräsentation bibliographischer Daten verwendet wird. Stattdessen orientieren sich diejenigen Bibliotheken, die bereits Linked Data zur Verfügung stellen, an unterschiedlichen Datenmodellen, Schemas und Ontologien. Es existieren jedoch Versuche, die bereits entwickelten Modelle in ein möglichst generisches und von vielen Bibliotheken nachnutzbares Modell zu integrieren (Suominen, Hyvönen 2017).

Vorhandene Ontologien

Vor allem die folgenden Datenmodelle sind für die semantische Modellierung der Metadaten aus der Bibliographie relevant, da sie entweder bereits weit verbreitet sind oder spezifische Elemente enthalten, die nachgenutzt werden können.

- *FRBR: Functional Requirements for Bibliographic Records* und *RDA: Resource Description and Access* (IFLA 2009)
- *DCTerms: Dublin Core Metadata Terms* (Dublin Core Metadata Initiative 2012)
- *PRISM: Publishing Requirements for Industry Standard Metadata* (IDEAlliance 2008)
- *SPAR Ontologies* (Peroni, Shotton 2018)

Die Entwicklung der SPAR-Ontologien wird von den Autoren u. a. damit begründet, dass bisherige Systeme uneinheitlich seien und deutliche Schwächen aufwiesen. PRISM und FRBR seien bspw. „top-level vocabularies rather than something specifically developed to characterise specific aspects of scholarly publishing“ (Peroni, Shotton 2018). Gleichzeitig benutzen die SPAR-Ontologien jedoch Elemente aus den anderen o. g. Vokabularen, um Redundanzen und doppelte Element-Definitionen zu vermeiden. In der hier beschriebenen Arbeit wurde daher ebenfalls versucht, aus den o. g. Datenmodellen vorrangig diejenigen Elemente zu verwenden, die bereits im Bibliothekswesen etabliert und nicht zu spezifisch, gleichzeitig aber ausreichend detailliert sind.

Modellentwicklung

Nach einer eingehenden Analyse der in der Bibliographie vorhandenen Metadaten wurden aus den o. g. Ontologien diejenigen Elemente zur weiteren Berücksichtigung ausgewählt, die zur möglichst genauen und eindeutigen Modellierung der einzelnen Einträge der Bibliographie (siehe Abbildung 2) benötigt werden. Hierbei wurde darauf geachtet, nicht bloß die einzelnen Romane mit ihren Metadaten abzubilden, sondern auch den Aufbau und die Struktur der Bibliographie an sich. Dadurch konnte das gesamte zu erzeugende Modell an den bereits im Linked-Data-Service der *Bibliothèque nationale de France* (BnF) vorhandenen Eintrag für die *Bibliographie du genre romanesque français* angebunden werden.⁴

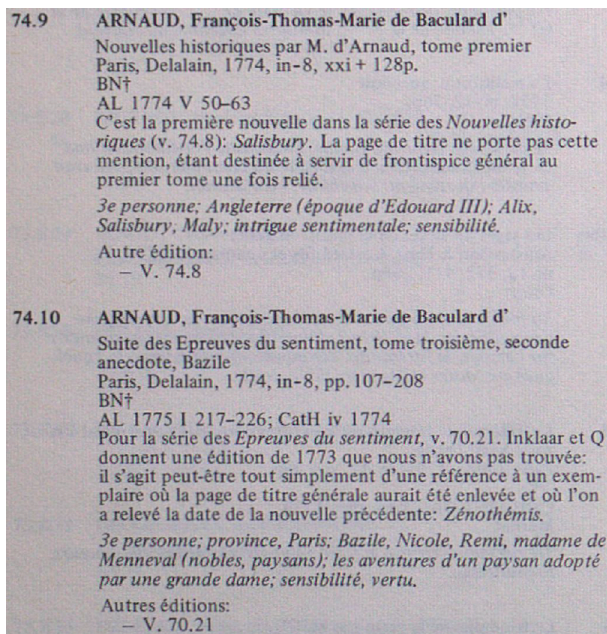


Abbildung 2: Beispieleinträge in der gedruckten Bibliographie

Durch die im Vorfeld bereits erfolgte Extraktion der einzelnen Metadatenfelder aus den OCR-Daten konnten diese schließlich direkt auf die entsprechenden Elemente in dem erstellten RDF-Modell abgebildet werden. Dies geschah überwiegend mithilfe der Programmiersprache *Java* und der dort verfügbaren Bibliothek *Apache Jena*⁵.

Verknüpfung mit anderen Ressourcen

Um die Möglichkeit der Anreicherung der Daten mit Informationen aus externen Ressourcen beispielhaft darzustellen, wurden die Namen der Autoren der einzelnen Romane aus dem RDF-Modell extrahiert und mithilfe von *Apache Jena* an die API der *Virtual International Authority File* (VIAF)⁶ gesendet. Von dort wurden – sofern vorhanden – die VIAF-IDs extrahiert und dem RDF-Modell hinzugefügt. Weitere externe Ressourcen könnten auf ähnliche Weise integriert werden. Voraussetzung für die erfolgreiche Nutzung der API ist, dass die Einträge im RDF-Modell keine Schreibfehler oder OCR-Fehler aufweisen. Dies kommt allerdings relativ häufig vor (Gründe sind u. a.: kleine Schrift in der Vorlage, viele Eigennamen, kurze Wörter mit wenig Kontext) und ist eines der wesentlichen Probleme des Datensatzes.

Fazit

Sowohl die Extraktion der einzelnen Metadaten aus den OCR-Texten als auch die Erstellung und anschließende Überführung in ein RDF-Modell ließen sich mit gutem Erfolg umsetzen. Die Erkennungsgenauigkeit des CRF-

Algorithmus war mit einem F1-Score von durchschnittlich 0,964 (0,908–0,997) außerordentlich hoch. Grund hierfür war sicherlich vor allem die bereits stark strukturierte Datengrundlage. Fehlende einheitliche Standards zur Repräsentation bibliographischer Metadaten und Fehler in den Textdaten sind jedoch Schwachstellen, die eine genauere Analyse und evtl. umfangreiche Bereinigung/Korrektur der zu repräsentierenden Daten nötig machen.

Das vorgestellte Projekt hat durch die Kombination von modernen Verfahren zur Informationsextraktion und die Zusammenstellung von aktuellen Ontologien zur Repräsentation bibliographischer Metadaten einen für die Datengrundlage passenden Ansatz entwickelt, der als Standard-Workflow für ähnliche Projekte verwendet werden könnte und in solchen überprüft und verfeinert werden sollte. Denkbar wären z. B. die Digitalisierung und Metadatenextraktion weiterer Bibliographien, um den erzeugten Datenbestand zu ergänzen, zu erweitern oder anzureichern. Auch die Überprüfung des hier beschriebenen Vorgehens in verwandten Kontexten (andere Nachschlagewerke, andere Sprachen, andere Epochen) unter Nutzung weiterer oder anderer Features wäre sinnvoll.

Der Workflow und die Daten werden daher am *Trier Center for Digital Humanities* im Rahmen des von der Forschungsinitiative Rheinland-Pfalz geförderten Projektes „MiMoText – Mining and Modeling Text“ weiterverwendet und erweitert. Ziel ist hier der Aufbau eines „aus unterschiedlichen Quellen gespeisten Informationsnetzwerks für die Geisteswissenschaften, das durch die Bereitstellung als Linked Open Data nicht nur frei verfügbar und mit anderen Wissensressourcen des Semantic Web verknüpfbar ist, sondern auch neuartige und effiziente Zugriffsmöglichkeiten auf fachwissenschaftliche Informationen bietet“.⁷ Die beschriebene Arbeit liefert hierfür eine geeignete Grundlage.

Fußnoten

1. Der Datensatz ist verfügbar unter <https://zenodo.org/record/3401429> (Lizenz: CC-BY).
2. <https://github.com/TeamHG-Memex/sklearn-crfsuite>
3. <https://scikit-learn.org/stable/index.html>
4. <http://data.bnf.fr/ark:/12148/cb34586696d>
5. <https://jena.apache.org/>
6. <https://platform.worldcat.org/api-explorer/apis/VIAF>
7. <https://kompetenzzentrum.uni-trier.de/de/projekte/projekte/m/>

Bibliographie

- Berners-Lee, Tim / Hendler, James / Lassila, Ora** (2001): „*The Semantic Web*“, in: *Scientific American* 284.5: 29–37.
- Bizer, Christian / Heath, Tom / Berners-Lee, Tim** (2009): „*Linked Data – The Story So Far*“, in: *International*

Journal on Semantic Web and Information Systems 5.3: 1–22 <http://tomheath.com/papers/bizer-heath-berners-lee-ijswis-linked-data.pdf> [letzter Zugriff 03. Januar 2020].

Candeias, Ricardo Pereira (2011): *Metadata Extraction from Scholarly Articles*. Master Thesis, Universidade Técnica de Lisboa. <https://fenix.tecnico.ulisboa.pt/downloadFile/395143160947/dissertacao.pdf> [letzter Zugriff 03. Januar 2020].

Dublin Core Metadata Initiative (2012): *DCMI Metadata Terms*. DCMI Recommendation. <http://dublincore.org/documents/2012/06/14/dcmi-terms/> [letzter Zugriff 03. Januar 2020].

Freyberg, Linda (2017): "Density of Knowledge Organization Systems", in: *Knowledge Organization for Digital Humanities. Proceedings of the 15th Conference on Knowledge Organization WissOrg '17 of the German Chapter of the International Society for Knowledge Organization (ISKO)* 25–30.

Groza, T. / Grimnes, A. / Handschuh, S. (2012): "Reference Information Extraction and Processing Using Conditional Random Fields", in: *Information Technology and Libraries* 31.2: 6–20.

IDEAlliance – International Digital Enterprise Alliance (2008): *The PRISM Namespace – Final* http://www.prismstandard.org/specifications/2.0/PRISM_prism_namespace_2.0.pdf [letzter Zugriff 03. Januar 2020].

IFLA Study Group on the Functional Requirements for Bibliographic Records (2009): *Functional Requirements for Bibliographic Records – Final Report*. (IFLA Series on Bibliographic Control, Vol. 19) <https://www.ifla.org/publications/functional-requirements-for-bibliographic-records> [letzter Zugriff 03. Januar 2020].

Kovacevic, Aleksandar / Ivanovic, Dragan / Milosavljevic, Branko / Konjovic, Zora / Surla, Dusan (2011): "Automatic extraction of metadata from scientific publications for CRIS systems", in: *Program* 45.4: 376–396.

Kuczera, A. (2017): "Graphentechnologien in den Digitalen Geisteswissenschaften", in: *ABI Technik*, 37.3: 179–196.

Lafferty, John D. / McCallum, Andrew / Pereira, Fernando C. N (2001): "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data", in: *Proceedings of the Eighteenth International Conference on Machine Learning (ICML '01)* 282–289.

Martin, Angus / Mylne, Vivienne / Frautschi, Richard (1977): *Bibliographie du genre romanesque français 1751-1800*. London, Paris: Mansell, France expansion.

Peng, Fuchun / McCallum, Andrew (2004): "Accurate Information Extraction from Research Papers using Conditional Random Fields", in: *Proceedings of Human Language Technology Conference and North American Chapter of the Association for Computational Linguistics (HLTNAACL)* 329–336 [https://](https://www.cs.umass.edu/~mccallum/papers/hlt2004.pdf)

www.cs.umass.edu/~mccallum/papers/hlt2004.pdf [letzter Zugriff 03. Januar 2020].

Peroni, Silvio / Shotton, David (2018): "The SPAR Ontologies", in: Vrandečić D. et al. (eds.): *The Semantic Web – ISWC 2018*. Lecture Notes in Computer Science. Cham: Springer 119–136.

Schreiber, Guus / Raimond, Yves (2014): *RDF 1.1 Primer*. W3C Note. <http://www.w3.org/TR/2014/NOTE-rdf11-primer-20140624> [letzter Zugriff 03. Januar 2020].

Suominen, Osmo / Hyvönen, Nina (2017): "From MARC silos to Linked Data silos?", in: *o-bib. Das offene Bibliotheksjournal* 4.2: 1–13.