Georg Petz*

Linked Open Data. Zukunftsweisende Strategien

https://doi.org/10.1515/bfp-2023-0006

Zusammenfassung: Bibliotheken stellen bereits seit einigen Jahren ihre bibliografischen Metadaten als Linked Open Data zur Verfügung. Die Idee dahinter ist, Daten aus verschiedenen Quellen und Formaten (Datensilos), die derzeit nicht oder nur schwer zugänglich sind, in möglichst einheitlicher Form miteinander zu verknüpfen. Dieser Artikel fasst die aktuellen Entwicklungen zusammen und es werden die vier grundlegenden Möglichkeiten, Linked Open Data zu veröffentlichen, miteinander verglichen. Abschließend werden die Initiativen der Österreichischen Nationalbibliothek auf diesem Gebiet vorgestellt.

Schlüsselwörter: Linked Open Data, Web 3.0, Digitale Bibliothek, Library Labs, Semantic Web

Linked Open Data. Trend-Setting Strategies

Abstract: Libraries are opening their bibliographic metadata as linked open data. The idea behind is to harmonise heterogeneous data sources (data silos) of different origin to improve their accessibility and interoperability. This article summarises relevant developments in this area and compares the four options to provide linked open data. Finally, the initiatives of the Austrian National Library in this area are presented.

Keywords: Linked open data, web 3.0, digital library, library labs, semantic web

1 Warum Linked Open Data?

„At their core, libraries in the information age provide a public means of accessing **knowledge**“, sagt David Pescovitz, Forschungsdirektor am Institute for the Future, einer gemeinnützigen Denkfabrik mit Sitz in Palo Alto, Kalifornien, USA.[1] Wissen kann aber nicht einfach über Datenschnittstellen, wie sie oft im Rahmen von Linked Open Data (LOD) Strategien entstehen, abgerufen werden. Es ist besonders wichtig, zwischen den Begriffen Daten, Information und Wissen zu unterscheiden. Hier kann die Wissenspyramide (s. Abb. 1) helfen, dort befindet sich der Begriff Wissen an dessen Spitze und die Pyramide an sich fußt wiederum auf „Zeichen“. Erst durch Syntax können aus Zeichen Daten werden. Diese werden wieder durch Semantik zu Information und durch Pragmatik, der Anwendung von Information, kann Wissen entstehen.

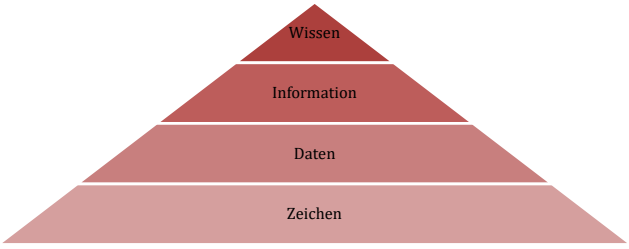

Abb. 1: Wissenspyramide

Umsetzbar ist die Vision von David Pescovitz nur mit einer konsequenten Anwendung von LOD. Was versteht man aber genau darunter? Linked Open Data, in der öffentlichen Debatte auch oft als Semantic Web bezeichnet, umfasst mehrere Aspekte. Zum einen, die eindeutige, zitierfähige und stabile Referenzierbarkeit von digitalen Objekten, zum anderen eine Gruppe von Vokabularien, die Eigenschaften z. B. von Personen, Orten und Beziehungen, wie auch physische Aspekte – beispielsweise Materialeigenschaften – beinhalten. Ziel ist es, ursprünglich voneinander getrennte Datenbestände miteinander zu verknüpfen. Auf diese Weise soll, durch die gegenseitige Kontextualisierung von Datensätzen, ein Mehrwert an Informationen entstehen. Darüber hinaus wird durch die damit einhergehende, notwendige Homogenisierung der zugrundeliegenden Datenformate die maschinelle Verarbeitung dieser Daten ermöglicht. Weiters steht eine Reihe an Abfrage- und Speichermechanismen zur Verfügung, um LOD zuverlässig speichern und abfragen zu können.

Betrachtet man die Geschichte des Semantic Webs bzw. des Internets, muss man 1960 bei Ted Nelson beginnen. Dieser gründete das Projekt Xanadu, welches bereits zahlreiche für LOD relevante Prinzipien wie z. B. eindeutige, vom Speicherort unabhängige, Adressen beinhaltete. Durchgesetzt hat sich aber erst das World Wide Web (WWW), welches 1989 von Tim Berners-Lee und Robert Cailliau am europäischen Forschungszentrum CERN in Genf entwickelt wurde.

1 https://www.businessinsider.com/libraries-of-the-future-2016-8?r=US&IR=T (alle Links abgefragt am 12.06.2023).

***Kontaktperson: Mag. DI Georg Petz**, georg.petz@onb.ac.at

2001 entstand der Begriff *Semantic Web* bzw. *Web 3.0*, einer Erweiterung des bestehenden WWW, wieder maßgeblich durch Tim Berners-Lee.[2] Der semantischen Weiterentwicklung des WWW folgte 2007 die Einführung des Begriffs LOD. Web 3.0 wird dabei oft fälschlicherweise für Web3 gehalten, darauf wies Tim Berners-Lee auf dem Web Summit 2022 in Lissabon, laut dem US-Sender CNBC[3], hin. Während Web3 eine dezentralisierte, blockchainbasierte Version des Webs ist, unterstreicht Web 3.0, mit dem Schwerpunkt auf die Wiederverwendung und Verknüpfung von Daten, vielmehr Tim Berners-Lees Konzept eines semantischen Webs.

In den letzten Jahren haben vor allem die Datenbestände größerer staatlicher und nichtstaatlicher Organisationen massiv zugenommen. Transparenzbestrebungen sowie die Erwartung der Schaffung von neuen wirtschaftlichen Verwertungsmöglichkeiten haben dazu geführt, dass vor allem in „öffentlichem Besitz“ stehende bzw. von nichtkommerziell orientierten Organisationen bereitgestellte Daten immer öfter frei über das Internet verfügbar gemacht werden. Die „EU-Richtlinie über die Weiterverwendung von Informationen des öffentlichen Sektors“ (PSI 2003)[4] hatte bereits das Ziel, vorhandene Informationen aus dem öffentlichen Sektor möglichst unbürokratisch zugänglich zu machen, und es wurde in Form des Informationsweiterverwendungsgesetzes (IWG)[5] 2005 in Österreich eingeführt. Dabei wurden kulturelle Institutionen jedoch explizit ausgenommen. Mit der 2015 in österreichisches Recht übernommenen Novelle zur PSI-Richtlinie (PSI 2013)[6] wurde ein klares Bekenntnis zu Open Data gegeben und auch Bibliotheken, Museen und Archive (§ 3 Abs. 1 Z 8, S. 2) miteinbezogen. Ein Vorschlag dieser Art findet sich auch in Punkt 14 der Empfehlungen für die Umsetzung von Open Access in Österreich des Open Access Network Austria (OANA):[7] „Bereits digitalisierte Bestände in öffentlich finanzierten Archiven, Museen, Bibliotheken und Statistikämtern, sollten bis 2020 der Wissenschaft und der Öffentlichkeit zur freien und kostenlosen Weiterverwendung zur Verfügung gestellt werden.“

Traditionell sind die Datenbestände der jeweiligen Institutionen zueinander sehr heterogen, da in verschiedenen Organisationen in der Regel unterschiedliche Softwareprodukte (Datenbanken etc.) eingesetzt werden sowie teilweise sehr verschiedene „Traditionen“ der Erstellung und Verwaltung von digitalen Informationen etabliert sind. Aus diesem Grund sind diese Daten unterschiedlicher Provenienz untereinander oft nicht kompatibel, auch wenn es häufig inhaltliche Überschneidungen gibt. Genau diese potenziellen Überschneidungen zwischen den unterschiedlichen Datensammlungen sind aber ein sehr wertvolles Element, da sie zur gegenseitigen Anreicherung der jeweiligen Daten genutzt werden können. Dies kann beispielsweise anhand von eindeutig identifizierten Personen, Orten oder Begriffen geschehen. Um dies zu erreichen, müssen die einzelnen Datenbestände zuvor jedoch in eine möglichst einheitliche Form gebracht werden.

Genau hier setzt die Idee von LOD an und bietet ein durchgängiges Rahmenkonzept, um verfügbare Datenbestände auf eine Art und Weise bereitzustellen, die deren Zusammenführung bestmöglich unterstützt. Dazu gehört die Nutzung eines einheitlichen Datenmodells sowie gemeinsamer Richtlinien, Vokabulare und Normdateien, um eindeutig identifizierbare Personen, Gegenstände etc. (Entitäten) einheitlich zu bezeichnen und dadurch Verknüpfungen über sie zu ermöglichen. Darüber hinaus wird auch die technische Infrastruktur beschrieben, die für die Veröffentlichung der Daten genutzt werden kann. Dabei wird vor allem darauf Wert gelegt, die bereits etablierte Infrastruktur des Webs zu nutzen.

LOD stellt somit das Fundament der Vision des Semantischen Web (Web of Data) dar, welches als Resultat einer weitläufigen Vernetzung bestehender Datenbestände entstehen kann. Die Prinzipien von LOD sorgen hier für die Interoperabilität dieser verschiedenen Quellen und stellen darüber hinaus auch Mechanismen zur Verfügung, die kombinierten Daten abzufragen.

2 Faktoren für die Qualität von LOD

2.1 Fünf-Sterne-Modell

Ein Maß für die Qualität von LOD ist das Fünf-Sterne-Modell zur Kennzeichnung offener Daten von Tim Berners Lee aus 2010.[8] Es können maximal fünf Sterne vergeben werden, wobei das Modell kaskadierend ist:

2 Berners-Lee et al. (2001).

3 https://edition.cnn.com/2022/12/16/tech/tim-berners-lee-inrupt-spc-intl/index.html.

4 https://eur-lex.europa.eu/legal-content/DE/TXT/PDF/?uri=CELEX:32003L0098.

5 https://www.jusline.at/gesetz/iwg/gesamt.

6 https://eur-lex.europa.eu/legal-content/DE/TXT/PDF/?uri=CELEX:32013L0037.

7 Bruno Bauer et al. (2015).

8 https://dvcs.w3.org/hg/gld/raw-file/default/glossary/index.html.

Tab. 1: Fünf-Sterne-Modell

Anzahl Sterne	Voraussetzungen				
*	**OL**				
**	OL	**RE**			
***	OL	RE	**OF**		
****	OL	RE	OF	**URI**	
*****	OL	RE	OF	URI	**LD**

Das bedeutet, die Vergabe von Sternen ist aufbauend. Der erste Stern wird für das Anbieten der eigenen Daten unter einer offenen Lizenz (OL) vergeben. Stern Nummer zwei setzt die Zurverfügungstellung in einem strukturierten Format (RE, readable) voraus. Wenn es sich bei dem strukturierten Format auch noch um ein nicht proprietäres Format (OF, open format) handelt, wird dies mit dem dritten Stern prämiert. Damit Daten persistent auffindbar sind, müssen für die vierte und vorletzte Stufe sogenannte URIs (Unique Resource Identifiers, siehe Abschnitt 3.1) vergeben werden. Alle Sterne werden zuerkannt, wenn die eigenen Daten nun auch mit anderen Daten verknüpft (LD, Linked Data) werden, um die sprichwörtlichen Datensilos aufzubrechen.

2.2 FAIR

Ein weiteres Maß für die Qualität von LOD bieten die FAIR Guiding Principles for scientific data management and stewardship.[9] Sie sind 2016 als Artikel in der Fachzeitschrift Nature veröffentlich und bis dato bereits 4 900-mal zitiert worden. Absicht der Autor*innen war die Erstellung eines Leitfadens, um die Auffindbarkeit (Findability), Zugänglichkeit (Accessibility), Interoperabilität (Interoperability) und Wiederverwendbarkeit (Reuse) wissenschaftlicher Daten zu erleichtern.

2.2.1 Findable

Für die Auffindbarkeit von Daten ist es wesentlich, dass diese einen global eindeutigen und dauerhaften Identifikator erhalten. Beispiele hierfür sind der Digital Object Identifier (DOI) bzw. das Handle-System, der Archival Resource Key oder die National Bibliography Number (urn:nbn). Im Unterschied zu gewöhnlichen Hyperlinks, welche veränderlich sind und somit keine langfristige Auffindbarkeit und Zitierbarkeit gewährleisten, können diese *Persistent Identifier* (PI) dauerhaft zitiert werden. Ebenfalls essenziell ist die Anreicherung der Daten mit menschen- und maschinenlesbaren Metadaten. Daten wie auch Metadaten müssen in ein Repositorium geladen, indexiert und schlussendlich auffindbar sein.

2.2.2 Accessible

Daten und Metadaten sollten verfügbar gemacht und langzeitarchiviert werden. Über Standard-Kommunikationsprotokolle wie zum Beispiel https sollen sie möglichst niederschwellig abgerufen werden können. Metadaten sollen verfügbar bleiben, selbst wenn die Daten nicht mehr angeboten werden. Authentifizierung und Autorisierung sind von dem Repositorium zu unterstützen. Es müssen aber nicht alle Daten automatisch unter einer freien Lizenz angeboten werden. Hier unterscheidet sich FAIR stark vom Fünf-Sterne-Modell von Tim Berners Lee. Die FAIR-Prinzipien schreiben nämlich keine offene Lizenz vor. Somit gilt: „FAIR ≠ fair“ und Daten können mittels Authentifizierung und Autorisierung vor Zugriff „geschützt“ werden.

2.2.3 Interoperable

Die Daten sollten derart vorliegen, dass sie ausgetauscht, interpretiert und in einer (semi)automatisierten Weise mit anderen Datensätzen kombiniert werden können. Metadaten müssen auf kontrollierten Vokabularen, Klassifikationen, Ontologien oder Thesauren basieren, die wiederum den FAIR-Prinzipien folgen. Maschinenlesbare Formate für Metadaten wie XML[10] oder JSON[11] sind unumgänglich. Durch Angaben wie „ist Teil von“ oder „ist eine Version von“ wird eine Verknüpfung zwischen Datensätze geschaffen. Verknüpfungen von Metadaten zu anderen Daten sollten über persistente Identifikatoren erfolgen.

2.2.4 Reusable

Metadaten dienen bekanntlich der Beschreibung von Daten und sind erforderlich, um eine Wiederverwendbarkeit der Daten zu ermöglichen. Ebenfalls erleichtert es den Vergleich mit anderen Daten sowie die Nachnutzung in Nachfolgeprojekten. Die Provenienz der Daten erleichtert zudem deren Wiederverwendbarkeit. Eine eindeutige Lizenz für die Bedingungen zur Nachnutzung soll für Mensch und Maschine auffindbar und verständlich sein.

9 Wilkinson et al. (2016).

10 https://www.w3.org/TR/1998/REC-xml-19980210.

11 https://www.json.org/json-en.html.

2.3 FAIR-Prinzipien für Bibliotheken, Archive und Museen

Koster und Woutersen-Windhouwer[12] haben konkrete Empfehlungen für die Anwendung der FAIR-Prinzipien in Bibliotheken, Archiven und Museen aufgestellt. Zur Gewährleistung der Auffindbarkeit von Daten wird auf eindeutige Identifikatoren auf Basis von URIs (s. Abschnitt 3.1) mit aussagekräftigen Metadaten verwiesen. Zugänglichkeit auf Applikationsebene soll über Programmierschnittstellen (API[13]s), dem *OAI-Protocol for Metadata Harvesting* (OAI-PMH[14]) zum Einsammeln und Weiterverarbeiten von Metadaten sowie Search/Retrieve via URL (SRU[15]), einem technischen Standard für Suchanfragen und der Abfragesprache SPARQL, erfolgen. Um Interoperabilität zu gewährleisten, wird das *Resource Description Framework* (RDF), ein System zur strukturierten Beschreibung von Ressourcen, empfohlen. Möglichst große Wiederverwendbarkeit wird durch die *Creative Commons*-Lizenzvariante CC0[16] erreicht.

2.4 LOUD

Das Fünf-Sterne-Modell zur Kennzeichnung offener Daten von Tim Berners Lee vernachlässigt die Perspektive von potenziellen Datennutzern. Robert Sanderson hat auf der Europeana Tech 2018[17] die fünf Sterne von *Linked Open Usable Data* (LOUD) vorgestellt. Der Begriff LOD wurde um den Buchstaben U erweitert, der für Usability steht, das Ausmaß, in dem ein Produkt, System oder Dienst durch bestimmte Benutzer*innen in einem bestimmten Anwendungskontext genutzt werden kann, um bestimmte Ziele effektiv, effizient und zufriedenstellend zu erreichen.[18]

Die Buchstabenfolge ABCDE kann als Merkhilfe für die 5 Eigenschaften für LOUD verwendet werden.

1. **A**bstraction (right **A**bstraction for der Audience)
   Eine der Zielgruppe angemessene Abstraktion ist von großer Bedeutung. Softwareentwickler*innen brauchen einen anderen Zugang zu den Daten als Fachexpert*innen.
2. **B**arriers (few **B**arriers to enter)
   Niedrige Einstiegshürden sind wichtig, um schnell zu einem Ergebnis mit den zur Verfügung gestellten Daten zu kommen.
3. **C**omprehensible (**C**omprehensible by introspection)
   Durch unmittelbar verständliche Daten ist es durch bloßes Betrachten möglich zu verstehen, worum es sich handelt. Ein für Entwickler*innen vertrautes Format wie z. B. JSON-LD (s. Abschnitt 4.4) eignet sich besonders gut dafür.
4. **D**ocumentation (**D**ocumentation with working example)
   Regeln lassen sich nicht intuitiv erahnen, daher braucht es eine entsprechende Dokumentation mit funktionierenden Beispielen.
5. **E**xceptions (few **E**xceptions, many consistent patterns)
   Entwickler*innen, die mit einer Programmierschnittstelle arbeiten, müssen alle deren Ausnahmen kennen, deshalb sind wenige Ausnahmen und eine möglichst einheitliche Struktur erstrebenswert.

3 Einheitliche Beschreibung und Identifikation von Daten

Die Idee von LOD basiert auf der Bereitstellung verschiedener heterogener Datenbestände in Form eines einheitlichen Datenmodelles. Eindeutige Identifikatoren sind die Grundlage, um auf die verschiedenen Ressourcen über einheitliche Datenschnittstellen zugreifen zu können. Die Abkürzungen URI, URL, URN, IRI und RDF spielen hierfür eine wichtige Rolle und werden im Folgenden kurz erläutert.

3.1 URI

Uniform Resource Identifier (URI) werden zur eindeutigen Bezeichnung digitaler Ressourcen (wie bspw. Webseiten, Dateien aber auch E-Mail-Empfänger) genutzt.

Ein Beispiel für eine URI wäre mailto:georg.petz@onb.ac.at.

3.2 URL

Der *Uniform Resource Locator* (URL) ist eine Teilmenge der URIs und beschreibt die Zugriffsadresse einer digitalen Ressource. Dadurch wird nicht nur eine Identifizierung, sondern auch eine Lokalisierung der dahinterliegenden Ressource ermöglicht. https://onb.ac.at/ wäre ein Beispiel für eine URL.

12 Koster und Woutersen-Windhouwer (2018).
13 https://www.talend.com/de/resources/was-ist-eine-api/.
14 https://www.openarchives.org/pmh/.
15 http://loc.gov/standards/sru/.
16 https://creativecommons.org/publicdomain/zero/1.0/deed.de
17 https://youtu.be/r4afi8mGVAY, https://www.slideshare.net/azaroth42/europeanatech-keynote-shout-it-out-loud.
18 https://www.usability.de/usability-user-experience.html.

3.3 URN

Ein *Uniform Resource Name* (URN) ist eine URI mit dem Schema *urn*. URNs sind sogenannte *Persistent Identifier* (PI), mit dem Online-Ressourcen unabhängig von ihrem Speicherort eindeutig und dauerhaft identifiziert werden können.

urn:ISBN:978-3-905924-46-6 speichert bspw. die ISBN 978-3-905924-46-6.

3.4 IRI

Internationalized Resource Identifier (IRI) sind URIs mit einem (um Unicode/ISO 10646) erweiterten Zeichensatz. Während bei URIs nur die druckbaren Zeichen des ASCII-Zeichensatzes erlaubt sind, können in IRIs beinahe sämtliche Zeichen des Unicode-Standards genutzt werden. Dadurch können z. B. mehrere Sprachen umgesetzt werden.

3.5 RDF

Resource Description Framework (RDF) bildet als zugrundeliegendes Datenmodell das Herzstück für Linked Open Data. Nur mithilfe von RDF ist die Erlangung des 5. Sterns im Fünf-Sterne-Modell von Tim Bernes-Lee möglich, indem Daten aus verschiedenen Datensilos miteinander verknüpft werden. RDF basiert darauf, die in den verschiedensten Datensätzen lagernde Information in Form von „kleinsten möglichen Einheiten“ darzustellen. Diese Einheiten werden in Form von sogenannten Tripel (engl.: Triple) dargestellt und setzen sich aus den Bestandteilen Subjekt, Prädikat und Objekt zusammen. Die dadurch beschriebenen Beziehungen sind stets vom Subjekt zum Objekt gerichtet und werden durch das Prädikat benannt. Die Aussage, dass „Goethe, Johann Wolfgang von“ die bevorzugte Bezeichnung[19] laut GND für die Person Goethe ist, lässt sich in RDF als Tripel z. B. folgendermaßen abbilden:

Subjekt: <http://d-nb.info/gnd/118540238>
Prädikat: <http://www.w3.org/2004/02/skos/core#prefLabel>
Objekt: „Goethe, Johann Wolfgang von“

Darüber hinaus kann auch eine alternative Bezeichnung[20] angegeben werden:

Subjekt: <http://d-nb.info/gnd/118540238>
Prädikat: <http://www.w3.org/2004/02/skos/core#altLabel>
Objekt: „ゲーテ，ヨハン・ヴォルフガング・フォン“ (Schriftcode: Japan)

Die Werte für Prädikat und Objekt werden zumeist durch bestimmte Vokabulare bzw. Ontologien[21] bestimmt. Im Falle des oben beschriebenen Datensatzes stammen die Werte *prefLabel* (bevorzugte Name) und *altLabel* (alternativer Name), die für die Beschreibung der Prädikate herangezogen wurden, aus dem *Simple Knowledge Organisation System* (SKOS[22]). Da das Ziel des *Semantic Web* die Harmonisierung von Datenbeständen ist, ist es wünschenswert, die Vokabulare und Ontologien für vergleichbare Inhalte möglichst einheitlich zu halten. Gerade diese Vorgehensweise wird bei bibliothekarischen Datensätzen schon seit langem angewandt, wie es sich gerade durch den Einsatz einheitlicher Datenformate wie MARC[23] und Normdateien wie die GND[24] zeigt.

Mit dem Einsatz von Ontologien kann man noch einen Schritt weiter gehen, da diese, in Anlehnung an den philosophischen Begriff, dazu eingesetzt werden, Subjekte, Prädikate und Objekte in Klassen einzuteilen und jeweils zueinander in Beziehung zu setzen. Prädikate wie „hat Autor“ und „hat Übersetzer“ können einer Oberklasse „hat Bezug zu Person“ zugeordnet werden. Objekte wie beispielsweise Schlagworte können wiederum ebenfalls zueinander in Beziehung gesetzt werden, um Begriffe wie „Geschichtswissenschaften“ und „Sprachwissenschaften“ unter den Oberbegriff „Geisteswissenschaften“ einzuordnen. Durch *Inferenz* wird es im Rahmen einer semantischen Suche somit möglich, Abfragen zu stellen, die für eine gegebene Oberklasse alle Resultate, die eigentlich Unterklassen entsprechen, zurückzugeben. Resultate einer Suche nach dem Schlagwort „Geisteswissenschaften“ würden dann alle Werke für „Geschichtswissenschaften“ und „Sprachwissenschaften“ beinhalten, auch wenn diese nicht explizit mit dem Schlagwort „Geisteswissenschaften“ versehen sind.

Darüber hinaus können Ontologien dafür verwendet werden, miteinander verwandte Datensätze zu harmonisieren. Eine Ontologie, die das MARC Feld 245 (Titelangabe in einem bibliografischen Datensatz) und das RDF-Element <http://purl.org/dc/elements/1.1/title> unter dem gemeinsamen Prädikat „hat Hauptsachtitel“ vereint, könnte dadurch

19 http://terms.tdwg.org/wiki/skos:prefLabel.
20 http://terms.tdwg.org/wiki/skos:altLabel.
21 http://mario-jeckle.de/semanticWebServices/intro.html.
22 http://www.w3.org/2004/02/skos/.
23 https://www.loc.gov/marc/.
24 https://www.dnb.de/DE/Professionell/Standardisierung/GND/gnd_node.html.
25 https://lod-cloud.net/clouds/publications-lod.svg.

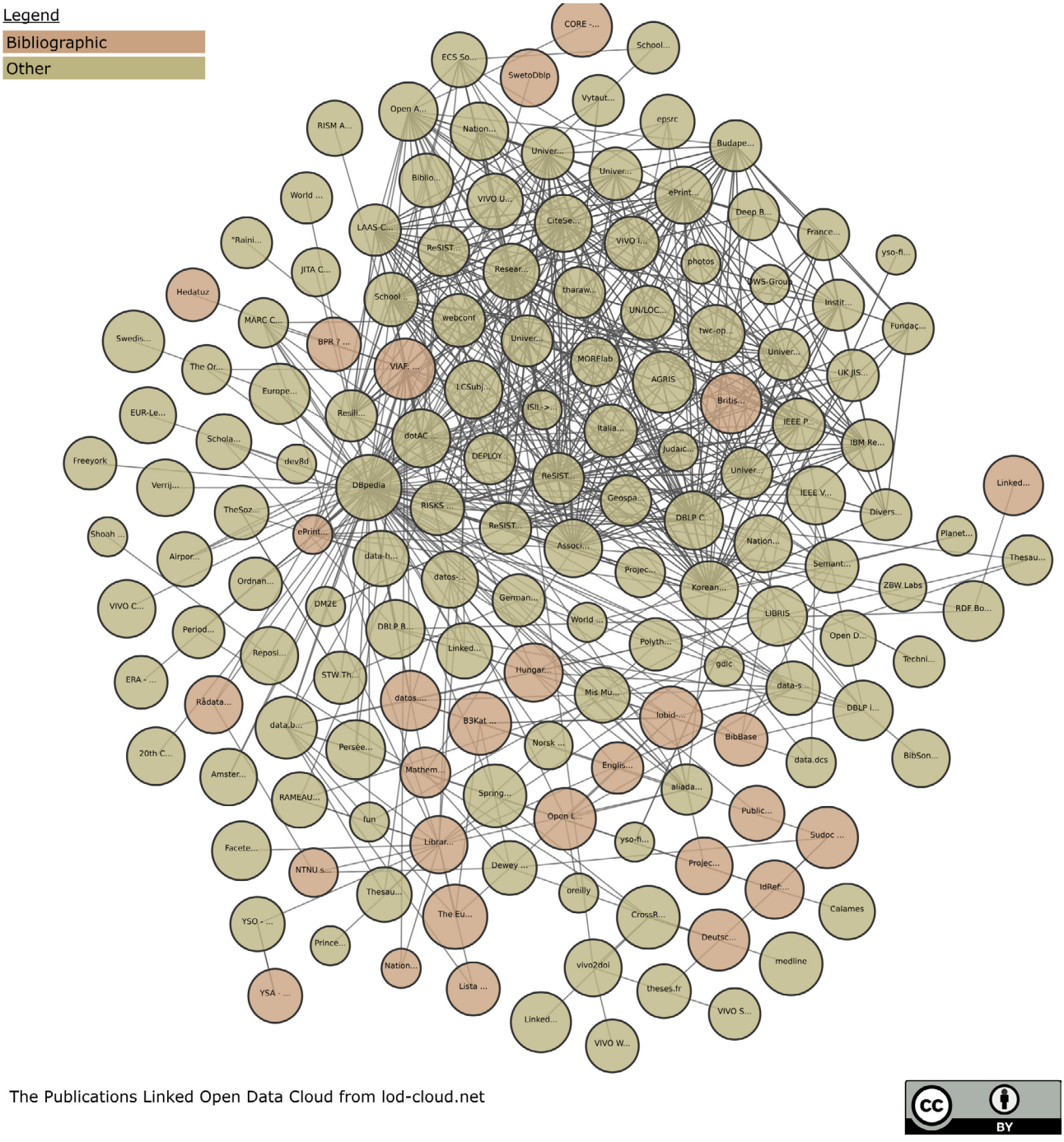

Abb. 2: Publications LOD Cloud[25]

einheitliche Abfragen über heterogene Datenformate ermöglichen.

Eine weitere wichtige Eigenschaft der LOD Repräsentationen von Datensätzen ist deren Einbettung in die Infrastruktur des Webs. Die Idee dahinter ist, die bestehenden Protokolle des WWW dazu zu nutzen, über die Inhalte eines gegebenen Tripels jederzeit so viel Kontextinformation wie möglich abrufen zu können.

Die RDF Grundregeln lassen sich in 4 Punkten einfach zusammenfassen:

1. Tripel: Jede Aussage besteht in RDF aus drei Einheiten, die zusammen ein Tripel bilden: Subjekt, Prädikat und Objekt.
2. *Internationalized Resource Identifiers* (IRI): Subjekt und Prädikat sind sogenannte Ressourcen, die durch eindeutige Bezeichner identifiziert werden.

3. Literal: Objekt kann entweder eine Ressource oder ein sogenanntes Literal sein.
4. Objekt einer Aussage kann das Subjekt einer anderen Aussage sein.

4 Formate

Neben einer einheitlichen Beschreibung und eindeutigen Identifikation von Daten, sind standardisierte Formate und Regelwerke für die Normalisierung und Verknüpfung heterogener Datenbestände erforderlich. Die Formate BIBFRAME, RDA/RDF, EDM und JSON-LD spielen bei LOD im Bibliotheksbereich eine besonders wichtige Rolle.

4.1 BIBFRAME

Aktuelle bibliothekarische Formate für den Austausch von Daten stoßen mit den steigenden Anforderungen an die Vernetzung von Informationen zunehmend an ihre Grenzen. BIBFRAME[26] (Bibliographic Framework), ein in RDF verfasstes Modell, soll diese Formate und ihre Schnittstellen künftig ersetzen und die Einbindung von Bibliotheksdaten in das WWW ermöglichen.

4.2 RDA/RDF

Das Regelwerk RDA[27] (Resource Description and Access) ist ein internationaler Katalogisierungsstandard und ermöglicht die einheitliche Erfassung von Veröffentlichungen sowie Normdatensätzen für Personen, Familien, Körperschaften, Konferenzen und Gebietskörperschaften. RDA-Elemente[28] können direkt in RDF abgebildet werden.

4.3 EDM

Beim Europeana Data Model (EDM)[29] handelt es sich um ein RDF Vokabular, das die Nutzung von LOD in Europeana ermöglichen soll. Es kann aber natürlich auch außerhalb von Europeana verwendet werden. Im Gegensatz zu BIBFRAME ermöglicht EDM eine maximale Nachnutzung existierender Standards wie z. B. SKOS oder DC (*Dublin Core*[30]).

4.4 JSON-LD

JSON (*JavaScript Object Notation*) ist ein von Programmiersprachen unabhängiges Datenformat und dient der Speicherung und Übertragung strukturierter Daten. JSON-LD (*JavaScript Object Notation for Linked Data*)[31] erweitert dieses und erlaubt eine Annotation der Daten. Dadurch werden sie für Webanwendungen und -services besser verstehbar und unterstützen die Veröffentlichung und Nutzung verknüpfter Daten im Web.

5 Einsatzbereiche für LOD in Bibliotheken

In Bibliotheken hat der Austausch von Datenbeständen bereits eine lange Tradition. Wurden die Daten lange Zeit im Rahmen von Verbünden oder im internationalen Austausch, z. B. innerhalb des deutschsprachigen Raumes, in der Regel nur zwischen Bibliotheken geteilt, werden sie nun in Form von LOD auch jenseits des unmittelbaren Bibliothekskontextes verfügbar gemacht.

5.1 Normdateien

Vor allem in Bezug auf die Verknüpfung von Datenbeständen anhand gleicher Entitäten, einer der Grundideen von LOD, spielen bibliothekarische Normdateien potenziell eine besonders wichtige Rolle. Nur über Normdateien kann eine eindeutige und konsistente Identifikation von Personen, Orten oder Begriffen in verschiedenen Datenbeständen gewährleistet werden.

Bei Betrachtung der *Publications LOD Cloud* (s. Abb. 2 bzw. https://lod-cloud.net/) findet sich neben den zwei großen Normdateien für den englisch- und deutschsprachigen Raum (LCSH, GND) und RAMEAU[32] für den französischsprachigen Raum auch eine dritte wichtige Quelle, das Projekt VIAF (Virtual International Authority File[33]). VIAF versucht, personenbezogene Einträge verschiedener Normdateien (inkl. der drei genannten) unter jeweils einheitlichen Bezeichnern zu vereinen, und stellt einen wichtigen Schritt zur Verknüpfung internationaler (bibliothekarischer) Datenbestände dar.

Die freie Verfügbarkeit von Normdateien als LOD ermöglicht die technisch einfache Verknüpfung von Titelda-

26 https://www.loc.gov/bibframe/.
27 http://www.rda-rsc.org/.
28 https://www.rdaregistry.info/.
29 https://pro.europeana.eu/page/edm-documentation.
30 https://www.dublincore.org/specifications/dublin-core/.
31 https://www.json.org/json-de.html.
32 https://www.cs.vu.nl/STITCH/rameau/.
33 https://viaf.org/.

tensätzen mit anderen, inhaltlich in Beziehung stehenden Informationen, die zur Anreicherung bzw. Kontextualisierung der eigenen Daten genutzt werden können. Voraussetzung ist, dass die zu verknüpfenden Datenquellen jeweils die gleichen Normdateien zur Bezeichnung ihrer Entitäten nutzen bzw. verschiedene Normdateien verwenden, die nachträglich über VIAF zusammengeführt wurden.

Ein bekanntes Beispiel für den erfolgreichen Einsatz von Normdateien zur Verknüpfung verschiedener Datenbestände ist die seit dem Jahr 2007 sowohl automatisch als auch manuell durchgeführte Indexierung von Wikipedia-Artikeln über Personen mit GND-Bezeichnern. Dadurch wird es möglich, Datensätze, welche die GND als Normdatei nutzen, auf einfache Weise mit den entsprechenden Wikipedia-Artikeln zu verbinden. Die LOD Version von Wikipedia, *DBpedia*,[34] erleichtert den technischen Vorgang der Verknüpfung in großem Maße, da sie die entsprechenden Daten in strukturierter Form anbietet. Nicht zu verwechseln mit Wikidata,[35] einer Wissensdatenbank die Daten (z. B. Geburtsdaten) *Wikipedia* zur Verfügung stellt.

5.2 Titeldaten

Während bibliothekarische Normdateien ein wichtiges Mittel zur Verknüpfung zwischen verschiedenen Datenquellen darstellen, handelt es sich bei den Titeldaten um die unmittelbaren Metadaten-Repräsentation der Sammlungen der jeweiligen Institutionen. Die Veröffentlichung von Titeldatensätzen in strukturierter und einheitlich zugänglicher Form ermöglicht es Dritten, die außerhalb des unmittelbaren Bibliothekskontextes stehen, diese Daten in vielfältigster Weise einzusetzen.

Das Gros der bibliothekarischen Institutionen hat sich entschieden, ihre Daten entsprechend der LOD-Idee anzubieten. Dies hat dazu geführt, dass der Nachfolger für das in die Jahre gekommene MARC21-Datenformat, BIBFRAME, auf Grundlage von LOD-Prinzipen entworfen wurde. Auch namhafte Hersteller von Bibliothekssoftware wie *Ex Libris* oder OCLC haben dementsprechend ihre Produkte um LOD-Funktionalitäten erweitert.

6 Bereitstellung von LOD

Es existieren vier grundlegende Möglichkeiten, Daten als LOD zu veröffentlichen. Sie unterscheiden sich hinsichtlich des Aufwandes für ihre Umsetzung aber auch in ihrer Flexibilität für den Einsatz durch Softwareentwickler*innen.

a) Abruf von Datensätzen für einzelne Titel/Person etc.
   Diese Art des Zugriffes kann am ehesten mit dem Navigieren einer Person in einem Online-Katalog verglichen werden. Anhand eines eindeutigen Bezeichners wird zuerst ein Titel- oder Normdatensatz abgerufen und entlang der in diesem Datensatz vorkommenden Verknüpfungen dann entsprechend zu anderen Datensätzen gewechselt. Dieses „Entlanghanteln“ von einem Datensatz zum nächsten wird auch als *Follow-your-nose-Prinzip*[36] bezeichnet. Hier kommt in der Regel die sogenannte *Content Negotiation*[37] zum Einsatz, wobei ein und dieselbe URL je nach Anfrage entweder HTML für einen menschlichen Besucher oder die RDF-Repräsentation für eine maschinelle Abfrage zurückliefert.
   Die Vorteil dieser „klassischen“ Form der Bereitstellung von LOD liegt vor allem in der technisch einfachen Umsetzung. Dagegen können von einem für den Sucheinstieg erforderlichen bestimmten Einstiegspunkt ausgehend nur jene Ressourcen erreicht werden, die damit direkt oder indirekt zusammenhängen. Zudem ist diese Variante für den Download kompletter Datenbestände nicht geeignet.
b) Abruf bzw. Download des kompletten Datensatzes auf den eigenen Rechner („Dump“)
   Sämtliche angebotenen Datensätze werden dabei in Form von einer oder mehreren Dateien auf der Webseite der Institution zum Download bereitgestellt. Besonders für Forschungszwecke interessant, es wird dabei eine große Menge an Daten gleichzeitig für z. B. Daten-Analysen zur Verfügung gestellt.
   Diese ebenfalls einfach umzusetzende Variante, kommt ohne einen konkreten „Einstiegspunkt“ aus und erfordert nur geringe zusätzliche Hard- und Softwareinfrastruktur. Die Nachteile dieser Variante liegen vorrangig in der Aktualität der angebotenen Daten. Die Erzeugung der Download-Dateien geschieht i. d. R. nur in (un-) regelmäßigen Abständen, zudem müssen Anwendungsentwickler*innen den gesamten Datensatz lokal installieren, um ihren Applikationen flexiblen Zugang auf die Daten zu ermöglichen. Dies kann dazu führen, dass die Daten nicht auf dem neuesten Stand sind. Die anbietende Institution wiederum weiß nur, dass der komplette Datensatz abgerufen wurde, Statistiken über den Zugriff auf einzelne Inhalte darin können nicht erstellt werden.

34 https://www.dbpedia.org/.
35 https://www.wikidata.org/wiki/Wikidata:Main_Page.
36 https://patterns.dataincubator.org/book/follow-your-nose.html.
37 https://www.w3.org/Protocols/rfc2616/rfc2616-sec12.html.

Tab. 2: LOD-Angebote europäischer Nationalbibliotheken

	a) Abruf pro Datensatz	b) "DUMP"	c) Abfrage via SPARQL	d) LOD API	abgefragt am 22.12.2022
BNF (FR)	X	X	X		https://data.bnf.fr/de/semanticweb
DNB (DE)	X	X			https://www.dnb.de/DE/Professionell/Metadatendienste/Datenbezug/LDS/lds_node.html
LIBRIS (SE)	X		X		https://www.kb.se/samverkan-och-utveckling/libris/katalogisering-i-libris/introduktion-till-libris.html
NSZL (HU)	X	X	X		https://old.datahub.io/dataset/hungarian-national-library-catalog
BNE (SP)	X	X	X		https://www.bne.es/es/blog/blog-bne/post-66
DBC (DA)	X	X	X		https://www.dbc.dk/videndeling/bibliografisk-udvikling/om-linked-open-data2
BL (UK)	X	X	X		https://www.bl.uk/collection-metadata/downloads
FENNICA (FI)	X	X	X		https://www.kiwi.fi/display/Datacatalog/Linked+Data
LNB (LV)	X		X		https://dati.lnb.lv/eswc2021/
KB (NL)	X		X		https://www.kb.nl/over-ons/expertises/linked-data-modellering
ÖNB(AT)	X	X	X		https://labs.onb.ac.at/en/dataset/lod/

c) Zugriff via Datenbankabfrage via SPARQL
LOD-Ressourcen können neben den beiden oben genannten Varianten auch über eine standardisierte Abfragesprache in Form eines sogenannten SPARQL[38] Endpoints zur Verfügung gestellt werden. Auf diese Weise können einzelne Datenfelder selektiv abgefragt werden (bspw. alle Titeldatensätze von Autor X aus dem Jahre Y). Diese Zugriffsmöglichkeit bietet die größte Flexibilität für Anwender*innen. Hierfür ist es notwendig, die in RDF transformierten Datensätze über eine Datenbank bereitzustellen. Diese Form der Bereitstellung von Daten verfügt zudem über Aggregatfunktionen, um beispielsweise alle Datensätze zu zählen, die ein bestimmtes Attribut aufweisen, ohne alle Datensätze herunterladen zu müssen. Institutionen, welche Daten auf diese Art und Weise bereitstellen, können durch die Analyse der abgesetzten Abfragen sehr genau feststellen, welche Aspekte der Daten für AnwenderInnen interessant sind. Die Nachteile liegen im Bedarf erweiterter Hard-/Softwareinfrastruktur. Weiters müssen Benutzer*innen das darunterliegende Datenmodell gut kennen und eine entsprechende Dokumentation muss verfügbar sein.

d) Zugriff via LOD API (API = Programmierschnittstelle)
Diese Variante vereint die Vorteile einer einfachen technischen Umsetzung, der hohen Flexibilität bei der Recherche für Anwender*innen sowie einer umfangreichen Analysemöglichkeit der durch diese Nutzer*innen abgesetzten Abfragen. Darüber hinaus entspricht diese Schnittstelle stärker dem „klassischen" Paradigma von Webanwendungen und ermöglicht hiermit den Zugriff auf die Daten auch aus Anwendungen heraus, die nicht auf LOD Prinzipien aufgebaut sind.
Diese Variante ist im Vergleich zur Datenbankabfrage über SPARQL wesentlich ressourcenschonender bezüglich Hardware, da der Aufwand durch die Einschränkung auf bestimmte, bereits vorher optimierte Anfragen entsprechend gesenkt werden kann.
Nutzer*innen des Dienstes müssen keine tiefgehenden Kenntnisse über das darunterliegende Datenmodell besitzen. Bestehende Webanwendungen können zudem die bereitgestellten Daten einfacher integrieren. Für den Download kompletter Datenbestände ist diese Variante weniger gut geeignet.

7 Europäische Nationalbibliotheken und LOD

Ein Vergleich der europäischen Nationalbibliotheken, die LOD zurzeit anbieten (s. Tab. 2), ergibt, dass unter anderem alle die „klassische" Variante d) anbieten und niemand die Variante d). Zusätzlich wird in der Tab. 2 auf die entsprechenden LOD-Angebote bzw. deren Beschreibung verlinkt. LIBRIS und DBC stellen hier einen Sonderfall dar, da es sich bei LIBRIS um einen Zugang zum schwedischen Verbundkatalog handelt, der von der Schwedischen Nationalbibliothek verwaltet wird, und DBC eine eigene, vom Dänischen Staat beauftragte Institution ist, die gemeinsam mit der Dänischen Nationalbibliothek die dänische Nationalbibliografie herausgibt.

38 https://www.w3.org/TR/sparql11-overview/.

7.1 LOD-Angebot der Österreichischen Nationalbibliothek

Abschließend noch ein Blick auf das LOD-Angebot der Österreichischen Nationalbibliothek. Im Rahmen einer Linked-Open-Data-Strategie wurden 2018 LOD-Datenschnittstellen eingerichtet und in die ÖNB Labs integriert.[39] Das Linked-Open-Data-Set der ÖNB Labs beinhaltet Metadaten zu den historischen Zeitungen (ANNO[40]) und Postkarten (AKON[41]) sowie Katalogdaten. Insgesamt werden Metadaten zu über 1 380 000 Zeitungsausgaben und 42 800 Periodika über SPARQL als Download per Datensatz und Komplettdownload („Dump“) bereitgestellt. Das Selbige gilt für die über 38 800 Ansichtskarten in AKON. Als primäres Datenformat für ANNO und AKON wird das Europeana Data Model (EDM[42]) verwendet. Der Abruf pro Datensatz für Metadaten aus dem Katalog ist zudem über die LOD-Schnittstelle der Bibliotheksdienstplattform ALMA[43] möglich.

Für die nächsten Jahre ist ein kontinuierlicher Ausbau bestehender LOD-Sets sowie der Aufbau weiterer neuer Sets geplant. Unterschiedliche Arten von Ressourcen aus den vielfältigen und einzigartigen Sammlungen der Österreichischen Nationalbibliothek zu unterschiedlichen Themen- und Fachbereichen sollen dadurch möglichst niederschwellig für die Forschung und interessiere Nutzer*innen zur weiteren Bearbeitung und Analyse zur Verfügung gestellt werden. Damit leistet die Österreichische Nationalbibliothek einen bedeutenden Beitrag zur weiteren Öffnung ihrer Informationsstruktur und trägt durch die Qualität ihrer Daten und der Erfüllung internationaler Standards zur Beständigkeit und Zuverlässigkeit der Linked-Data-Cloud bei.

7.2 URL-Schemata und weiterführende Informationen

Download per Datensatz für AKON: http://data.onb.ac.at/AKON/{akon-id}.rdf

Download per Datensatz für ANNO: http://data.onb.ac.at/ANNO/{anno-id}.rdf.

BIBFRAME über URL: https://open-na.hosted.exlibrisgroup.com/alma/{INSTITUTION-CODE}/bf/entity/instance/{MMS-ID}

JSON-LD über URL: https://open-na.hosted.exlibrisgroup.com/alma/{INSTITUTION-CODE}/bibs/{MMS-ID}

RDA/RDF über URL: https://open-na.hosted.exlibrisgroup.com/alma/{INSTITUTION-CODE}/rda/entity/manifestation/{MMS-ID}.rdf

API-Beschreibung der EDM-Identifikatoren für ANNO und AKON: https://iiif.onb.ac.at/api#_digitization_projects

Literaturverzeichnis

Berners-Lee, Tim; Hendler, James; Lassila, Ora (2001): The Semantic Web. In: *Scientific American*. DOI:10.1038/scientificamerican0501-34.

Bruno Bauer; Guido Blechl; Christoph Bock; Patrick Danowski; Andreas Ferus; Anton Graschopf; Thomas König et al. (2015): Empfehlungen für die Umsetzung von Open Access in Österreich. In: *Recommendations for the Transition to Open Access in Austria/Empfehlungen für die Umsetzung von Open Access in Österreich*. DOI:10.5281/zenodo.33178.

Koster, Lukas; Woutersen-Windhouwer, Saskia (2018): FAIR Principles for Library, Archive and Museum Collections: A proposal for standards for reusable collections. In: *The Code4Lib Journal* (40). Verfügbar unter https://journal.code4lib.org/articles/13427, zugegriffen am 22.12.2022.

Wilkinson, Mark D.; Dumontier, Michel; Aalbersberg, IJsbrand Jan; Appleton, Gabrielle; Axton, Myles; Baak, Arie; Blomberg, Niklas et al. (2016): The FAIR Guiding Principles for scientific data management and stewardship. In: *Scientific Data*, 3 (1), 160018. DOI:10.1038/sdata.2016.18.

Mag. DI Georg Petz
Forschung und Datenservices
Österreichische Nationalbibliothek
Josefsplatz 1
A-1015 Wien
Österreich
georg.petz@onb.ac.at
https://orcid.org/0000-0002-7843-3397

39 https://www.onb.ac.at/ueber-uns/presse/pressemeldungen/jahrespressekonferenz-oesterreichische-nationalbibliothek.
40 https://anno.onb.ac.at/.
41 https://akon.onb.ac.at/.
42 https://pro.europeana.eu/page/edm-documentation.
43 https://developers.exlibrisgroup.com/.