WEMI digital

Johannes Busse [1]

Abstract: Linked Open Government Data (LOGD) hat sich bisher nicht durchgesetzt, obwohl – oder weil? – der zugrunde liegende Standard DCAT-AP.de technisch so einfach gehalten ist, dass die Souveränität der einzelnen Verwaltungen in der Fläche maximal unterstützt wird. Die dezentrale Erfassung von Metadaten durch bibliothekarisch nicht geschulte Mitarbeiter führt allerdings zu Inkonsistenzen und Qualitätsproblemen. These: Statt DCAT sollte eine FRBR-basierte Metadatenhaltung verwendet werden, da diese eine klarere Struktur und bessere Erfassungsmöglichkeiten bietet. Verwaltungen sollten ihre Souveränität in der Metadatenhaltung an ausgebildete Bibliothekare und Bibliotheken übertragen, um LOGD auf Basis von FRBR professioneller zu katalogisieren und so homogener wiederverwendbar zu machen. FRBR-basierte Daten sollten automatisch nach DCAT exportiert werden können; Grundlage dafür ist eine auf LOGD zugeschnitte Ontologie WEMI-digital.

Keywords: WEMI-digital, digitale Souveränität, FRBR, IFLA LRM, DCAT-AP, OpenWEMI

1 Kontext

Der IT-Planungsrat definiert in Anlehnung an [Go17] den Begriff der Digitalen Souveränität als „die Fähigkeiten und Möglichkeiten von Individuen und Institutionen, ihre Rolle(n) in der digitalen Welt autonom, selbstbestimmt und sicher wahrnehmen zu können“ ([IT20]). Insbesondere für die öffentliche Verwaltung sei digitale Souveränität eine wichtige Voraussetzung, um hoheitliche Aufgaben erfüllen zu können (ebd.). Der Gesetzgeber verspricht sich viel von offenen Daten und misst ihnen neben demokratischen Werten auch eine hohe wirtschaftliche Bedeutung bei. Nach §12 E-Government-Gesetz (EGovG) und §11 Informationsfreiheitsgesetz (IFG) sollen Behörden eine Vielzahl nicht personenbezogener Daten als Open Data bereitstellen. Bundesländer wie Hamburg oder Rheinland-Pfalz haben Transparenzgesetze erlassen, die eine Vielzahl von Institutionen zur Veröffentlichung von Open Data verpflichten.

Um in der digitalen Welt autonom, selbstbestimmt und sicher agieren zu können, benötigt ein Akteur verlässliche Metadaten, insbesondere zu Fragen wie: Woher stammen die Daten? Stammen sie aus einer autoritativen Quelle oder handelt es sich nur um eine (hoffentlich unveränderte) Kopie? Was darf man mit den Daten machen, wie darf man sie weitergeben? Und nicht zuletzt: Was bedeuten die Daten? Daten sind nur dann wertvoll, wenn sie gefunden, gelesen, ausgewertet, genutzt, mit anderen Daten integriert werden können. Metadaten mit einer klar definierten Bedeutung sind eine *conditio sine qua non* für digitale Souveränität: Daten ohne verlässliche Metadaten gefährden die digitale Souveränität der Nutzer.

[1] HAW Landshut 1, Fakultät Informatik, Am Lurzenhof 1, 84036 Landshut,
busse@haw-landshut.de, https://orcid.org/0000-0002-7051-6198

Gegenstand dieses Aufsatzes sind Überlegungen zum technischen und institutionellen Metadatenmanagement im Kontext von Datensouveränität. These: Das derzeit in Europa dominierende Metadatenschema DCAT (s.u.) und die damit verbundenen Prozesse und politisch etablierten Strukturen zur Öffnung, Bereitstellung, zum Austausch und zur verlässlichen Metadatenauszeichnung von Regierungsdaten sind nur bedingt geeignet, das Potenzial von Linked Open Government Data (LOGD) auszuschöpfen. Insbesondere die auf DCAT basierende Technik und Institutionalisierung der Metadatenhaltung von „Daten" bleibt weit hinter den insbesondere im Bibliothekswesen etablierten Strukturen der Metadatenhaltung von Dokumenten zurück. Pointiert ausgedrückt: Die derzeitigen politischen und technischen Strukturen richten nicht nur volkswirtschaftlichen Schaden an, sondern verhindern letztlich, dass die Open-Data-Strategien aller bisherigen Bundesregierungen zum „Fliegen" kommen.[2]

Bei der Frage, wie die bestehenden Strukturen der Metadatenhaltung von LOGD verbessert werden können, müssen zum Teil anspruchsvolle bibliothekswissenschaftliche Detailfragen bearbeitet werden. Dies erschwert den Einstieg in die Diskussion aus Sicht der WIF. Umgekehrt müssen auch Bibliothekswissenschaftler Detailfragen aus dem Bereich LOD und Semantic Web lösen, was ebenfalls den Zugang erschwert. Bedenkt man, dass gerade im Bereich der Metadaten für digitale Dokumente auch in der Semantic Web Community bis heute tiefgreifende konzeptionelle Probleme bestehen, verwundert es nicht, dass manche Bibliothekare ihre digitale Kompetenz lieber auf Buch- oder Zeitschriftendokumente beschränken, als sich den komplexeren Herausforderungen der Metadatenhaltung und der Bereitstellung von sogenannten integrativen Web-Ressourcen zu stellen.

Wirtschaftsinformatik kann in erster Näherung als Schnittmenge von Wirtschaftsinformatik und Informatik verstanden werden, wobei diese Schnittmenge unter Einbeziehung weiterer Bezugswissenschaften sehr differenziert betrachtet wird und so insgesamt eine neue Wissenschaft entsteht. Relevant ist dabei die Richtung, aus der diese Schnittmenge betrachtet wird. Unser Beitrag untersucht eine technische Fragestellung aus dem Bereich des Semantic Web mit dem Bibliothekswesen als Bezugswissenschaft, um darauf aufbauend eine Empfehlung für die politische Dimension des Themas Digitale Souveränität zu geben.

2 Zielsetzung

Trotz teilweise erheblicher Anstrengungen „fliegt" Linked Open Data (LOD) derzeit nicht. Laut dem aktuellen Open-Data-Fortschrittsbericht der Bundesregierung vom März 2025

[2] Beide Reviewer*innen dieses Aufsatzen mahnen an, dass diese Pointierung ohne weiteren fundierten Beleg in einem wissenschaftlichen Artikel nur bedingt angemessen ist. Dem ist einerseits zuzustimmen. Es stellt sich allerdings die Frage nach dem dieser Kritik zugrunde liegenden Ideal von Wissenschaftlichkeit. Die Wirtschaftsinformatik betreibt Design Science Research stets auch in normativen bis hin zu politischen Kontexten, zu denen unser Tagungsthema „Souveränität" definitiv dazugehört. Wertende, pointierte, „griffige" Diskussstrategien gehören hier definitiv zum etablierten Methodenspektrum. Auch der vorliegednde Text versteht sich in Teilen als ein Meinungsaufsatz, jedenfalls dort, wo feingliedrige wissenschaftliche Untersuchungen durch gröbere politische Positionierungen ergänzt werden.

genießt LOD bei den Behörden keine Priorität: Nur 63 von 475 Behörden haben auf die aktuelle Anfrage der Bundesregierung überhaupt geantwortet ([Bu25]). Die Gründe dafür sind vielfältig, es handelt sich um ein komplexes Wirkungsgefüge. Das in diesem Aufsatz beschriebene Artefakt ist eine aktualisierte, fokussierte Definition eines möglichen Teilproblems, einer möglichen Ursache und einer möglichen Lösung.

Die Problemdefinition, die in diesem Aufsatz als Forschungsartefakt angestrebt wird, ist Teil einer umfassenderen Argumentation, die darauf abzielt, die bisherige Strategie und Institutionalisierung des Metadatenmanagements von Linked Open Government Data (LOGD) zu überdenken. Wir spannen den Problemraum auf, indem wir zwei Familien von Metadatenschemata einander gegenüberstellen, die sich im besten Fall unabhängig (oder im schlechtesten Fall konkurrierend und abgrenzend?) voneinander entwickelt haben, nämlich Modellfamilie A: DCAT und Modellfamilie B: WEMI aus FRBR/LRM/RDA (detailliertere Erklärung und Literaturnachweise siehe unten).

Zur Problembeschreibung nehmen wir auch ungelöste (und vielleicht unlösbare?) Standardprobleme auf, die für LOGD besonders relevant sind: Das ungelöste, komplizierte und daher oft ignorierte sogenannte *httpRange-14 Problem*; die Mehrdeutigkeit des Begriffs „Datensatz“; sowie die oft übliche Gleichsetzung von Datensatz und Datei, die im Bereich LOGD oder RDF jedenfalls dann kontraproduktiv ist, wenn man Datenbanken, SPARQL-Endpoints, Messdatenströme etc. betrachtet.

3 Problem

Im Folgenden werden zunächst einige grundlegende vorwiegend technische Probleme dargestellt, die entweder derzeit ungelöst oder sogar grundsätzlich unlösbar sind oder jeweils nur kontextspezifisch gelöst werden können.

Problem: Identifikation von Dokument und Daten. Was ist der Gegenestand unserer Metadaten: (a) klassische seitenorientierte gedruckte Dokumente (Bücher, Aufsätze), die heute auch digital z.B. als pdf, html oder gezippte Website vorliegen? (b) typische „Datendateien“ wie Excel, CSV, JSON, XML, andere? (c) aus RDA-Sicht sog. *integrierende Ressourcen* wie z.B. eine häufiger aktualisierte Website zu einer Veranstaltung? (d) orts- und zeitabhängige Datenströme zu Wetter, Verkehrsfluss, Parkhausbelegung, Abfallentsorgung etc. – die im Idealfall als historische Zeitreihen verfügbar sind und so nah wie möglich an das aktuelle Datum und die aktuelle Uhrzeit herangeführt werden?

Die traditionelle Typisierung von Datenbeständen orientiert sich noch immer am Begriff des Dokuments, der Serialisierung und dem Transport von Daten. Dieses Konzept stammt aus der frühen Phase des Internets, als eine URI eine Datei oder eine Quelle im Internet bezeichnete. Im Bibliothekswesen entspricht dies der Manifestation eines Werks (frbr:manifestation, dcat:Distribution). Zunächst wurden klassische Dokumente wie Bücher oder Aufsätze erfasst. Später kamen „fluide“ Dokumente hinzu, etwa Websites. Schließlich entstanden

Daten-Dokumente, die in unterschiedlichsten „Verpackungseinheiten“ verfügbar sind. Diese Typisierung basiert auf einem dokumentenzentrierten Ansatz: Zunächst standen klassische Dokumente im Fokus, dann fluide Dokumente. Mit der Zeit werden die Grenzen von Dokumenten jedoch zunehmend unbestimmt.

Problem Dataset. Eine moderne Typisierung geht abstrakter vor. Mit Metadaten will man typischerweise „data sets“ problemadäquat beschreiben. Bei genauerem Hinsehen zeigt sich, dass dieser Begriff höchst mehrdeutig ist. Aus technischer Sicht üblich sind Verwendungen im Sinne von CSV-Zeile, Record/Struct, CSV-Tabelle (Pandas DataFrame), Sammlung von Tabellen. Noch abstrakter kann man ein Schlagwort- oder Stichwortverzeichnis, eine Terminologie, Taxonomie, Klassifikation, Ontologie als Datensatz bezeichnen – und zwar unabhängig von der medialen Repräsentation. Spätestens hier hat man die Ebene des Dokuments verlassen und bewegt sich im Raum abstrakter, mehr oder weniger formal darstellbarer Ausprägungen komplexer Ideen.

Für jeden Datentyp stellen sich z.B. folgende Fragen: Welche Metadatenmodellfamilie beschreibt diesen Typ in seiner aktuellen Form angemessen? Welche Beschreibungslücken bestehen? Wie würde ein adäquateres Beschreibungsmodell aussehen? Wie einfach lässt sich die betreffende Familie erweitern, um die Beschreibungslücke zu schließen? Kann das adäquatere Beschreibungsmodell logisch anschlussfähig durch Verfeinerung erweitert werden, oder wäre eine „logisch disruptive“ (und damit in der Praxis schwer vorstellbare) Strukturänderung notwendig? Offensichtlich unterscheiden sich die Familien A und B grundlegend in ihrer Komplexität. Hier stellt sich die Frage: Wie wollen wir beurteilen, welche Beschreibungskomplexität angemessen ist?

In der Welt von Linked Open Data und RDF stehen Graphdatenbanken im Mittelpunkt. Hier stellt sich die Frage: Ist schon ein ein einzelner Named Graph innerhalb eines RDF Data Set ein Dataset oder nur ein „RDF Data Set“ als eine Menge von Datasets? Ist ein Teilgraph innerhalb eines RDF-Graphen, der ein Objekt als komplexes „Molekül“ beschreibt (entspricht einem Record in einer relationalen Datenbank), ein Dataset, oder gar ein einzelner Knoten mit seinen ausgehenden Kanten (was einer Zeile in einer CSV-Datei entsprechen könnte)?

Das Verwaltungsdatenportal *GovData.de* der Bundesrepublik Deutschland gibt im Juni 2025 an, über ca. 136.000 Datensätze zu verfügen. Dabei stellt sich die Frage, ob z.B. zwei separate Dateien zur Belegung aller Parkhäuser z.B. in Ulm (BW) und Neu-Ulm (BY) als eigenständige Datensätze oder lediglich als zwei Teile eines einzigen Datensatzes zu betrachten sind. Könnte man die Anzahl der berichteten Datensätze nicht publikumswirksam erhöhen, indem man die Parkhausbelegung für jedes Parkhaus einzeln als Datensatz zur Verfügung stellt? Verzwölffachen, wenn man von jährlicher auf monatliche Berichterstattung umstellt? Wir plädieren für den umgekehrten Ansatz: In Modellfamilie B würden wir alle gleich strukturierten Zeitreihen zur Parkhausbelegung in ganz Deutschland als einen einzigen verteilten Datensatz betrachten.

Problem: httpRange-14. Das konzeptuelle Problem von httpRange-14 wird in der Wikipedia als ein „langjähriges logisches Rätsel oder Designproblem im semantischen Web“ beschrieben. In den frühen Tagen des Internets bezeichnete *resource* eindeutig ein digitales Objekt aus dem Internet. Dies ändert sich mit RFC 2396, der von RFC 3986 im Jahr 2005 als *RFC-Internet Standard* weitergeführt wurde. Gemäß RFC 3986 kann ein URI nicht nur Dateien (in den Texten oft informell *information resources* genannt), sondern Beliebiges identifizieren[3]. Für LOD wird dieses Problem relevant, wenn wir verschiedene Aspekte eines Datensatzes unterscheiden müssen, darunter den eigentlichen Inhalt des Datensatzes (z. B. eine Zeitreihe zur Parkhausbelegung), den Datensatz als technisches Artefakt (inklusive Format und Schema) sowie den Download-Ort des Datensatzes. Schon 2009 diskutiert [Jo09], wie FRBR dieses Problem lösen könnte; auch [Po18] weist 2023 wieder auf die Bedeutung von RDA insbesondere für LOD hin.

Problem: Daten oder Datenquelle? Bei der Beschreibung von Metadaten stellt sich die Frage, ob sich diese primär auf die Daten selbst oder auch auf sogenannte „Endpoints“ beziehen, also auf Online-Zugänge zu Datenbanken, die bei einer Anfrage z.B. über eine entsprechende API oder ein REST-Protokoll Daten zur Verfügung stellen. Viele interessante Metadaten beziehen sich nicht mehr primär auf klassische Dokumente, sondern immer öfter auf Datenquellen, die „on demand“ konventionelle Dateien oder kontinuierliche Datenströme liefern. Dementsprechend unterscheiden sich auch die Anforderungen an ihre Beschreibung. Das DCAT-Modell trägt dieser Unterscheidung Rechnung, indem es zwischen *dcat:Dataset* für Datensätze und *dcat:DataService* für Datenzugriffsdienste unterscheidet; in RDA besteht hier eine Lücke; man könnte aber an einen *mit Manifestation in Beziehung stehenden Akteur* in Verbindung mit einer *Verantwortlichkeitsangabe* denken.

Diese Grundlagenprobleme bilden den Kontext für die hier behandelte Gegenüberstellung der Metadatenmodelle A und B.

Metadaten-Familien DC. Der 1995 in Dublin (USA) entwickelte *Dublin Core* [AAS18] ist ein Metadatenmodell, das ursprünglich auf einer einzigen Tabelle basiert. In der ersten Version umfasste dieses Modell 14 Spalten, während die erweiterte Variante *DCTERMS* deutlich mehr Felder enthält. Mehrere Einträge pro Datenfeld sind dabei zulässig, sodass mehrwertige Attribute unterstützt werden. Aus datenbanktechnischer Sicht entspricht dieses Modell der *Normalform Null*, da es keine weitere Strukturierung der Daten vornimmt. In der bibliothekarischen Fachliteratur wird diese Form auch als „flaches Modell“ bezeichnet. Strukturell ähnelt es einem *BibTeX*-Eintrag oder einem Datensatz im Open-Source-Literaturverwaltungssystem *Zotero*. Der Begriff „Datensatz“ wird in diesem Kontext

[3] *This specification does not limit the scope of what might be a resource; rather, the term “resource” is used in a general sense for whatever might be identified by a URI. Familiar examples include an electronic document, an image, a source of information with a consistent purpose (e.g., “today’s weather report for Los Angeles”), a service (e.g., an HTTP-to-SMS gateway), and a collection of other resources. A resource is not necessarily accessible via the Internet; e.g., human beings, corporations, and bound books in a library can also be resources. Likewise, abstract concepts can be resources, such as the operators and operands of a mathematical equation, the types of a relationship (e.g., “parent” or “employee”), or numeric values (e.g., zero, one, and infinity).* (https://datatracker.ietf.org/doc/html/rfc3986)

im Sinne einer einzelnen Zeile innerhalb einer einzigen großen, nicht normalisierten Tabelle verwendet.

DCAT-AP.de, die Deutsche Adaption des „Data Catalogue Application Profile“ (DCAT-AP) für Datenportale in Europa, wurde in Version 2.0 im Jahr 2022 publiziert; derzeit aktiv in Entwicklung befindet sich Version 3. DCAT wird als *der* zentrale Standard für den Austausch von Metadaten in Europa propagiert: “The (meta)data model used in data.europa.eu is DCAT-AP, the application profile for data portals in Europe. The specification of the DCAT-AP was a joint initiative of the Directorate-General for Communications Networks, Content and Technology, the Publications Office of the European Union and the interoperable Europe programme. The specification was elaborated by a multi-disciplinary working group with representatives from 16 EU Member States, some European institutions, and the United States.” (https://op.europa.eu/de/web/eu-vocabularies/dcat-ap). DCAT steht in der Tradition von Dublin Core und verwendet neben eigenen Attributen auch viele Attribute aus Dublin Core. Ein `dcat:Dataset` wird als „logische Entität, die die veröffentlichte Information repräsentiert, eine `dcat:Distribution` als „physische Verkörperung/Repräsentation des Datensatzes in einem spezifischen Format“ definiert. Der Unterschied zwischen Resource und Distribution wird in der Version 3 von DCAT wie folgt erläutert: “The original Recommendation [VOCAB-DCAT-1] published in January 2014 provided the basic framework for describing datasets. It made an important distinction between a dataset as an abstract idea and a distribution as a manifestation of the dataset.” (https://www.w3.org/TR/vocab-dcat-3/#motivation). Es stellt sich die Frage, ob der Vergleich mit Werk/Ausdruck und Erscheinungsform/Item aus derFRBR_Welt tatsächlich so geradlinig gezogen werden kann. Alternativ könnte man Dcat-Dataset und dcat-Distribution auch auf Schema und Instanz-Set, semantisch auf eine *OWL T-Box* und eine *OWL A-Box* beziehen; in der FRBR-Welt würde man dann zwei eigenständige Ideen erkennen, die jeweils nach einer eigenen, unabhängigen WEMI-Beschreibung verlangen.

Das Modell **Functional Requirements for Bibliographic Records (FRBR)** entstand um 1996 als Ergebnis einer systematischen Untersuchung der funktionalen Anforderungen von Bibliotheksbenutzern[4][Co; Co14; Co15; WH17]. Ein Dokument wird innerhalb der FRBR-Familie in vier über Fremdschlüssel verknüpften Tabellen erfasst. Ein vollständiger *WEMI*-Datensatz besteht laut FRBR aus mindestens einem Eintrag in jeder dieser Tabellen. Konventionell wird FRBR als eine Kette von disjunkten Klassen (*W-E-M-I*) dargestellt, die

[4] FRBR wurde 2014 unter dem Titel *IFLA Library Reference Model* (LRM) ([RBŽ17]) als konzeptuelles Modell zur Modellierung bibliographischer Daten weitergeführt. In seiner neuesten Version 1.0 aus dem Jahr 2024 [ARŽ24] liegt es u.a. auch in einer objektorientierten Version vor ([Be21]) und ist mit dem *CIDOC Conceptual Reference Model* (CIDOC CRM) abgestimmt. Das LRM bildet die Grundlage für die Implementierung des Regelwerks *Resource Description and Access* (RDA), das seit 2016 als Standardmodell im europäischen Bibliothekswesen gilt. Für Bibliothekare, die mit RDA und damit IFLA LRM arbeiten, gehört die WEMI-Unterscheidung zu den Grundlagen ihrer Expertise. RDA unterscheidet im digitalen Bereich nicht zwischen HTML- und PDF-Dokumenten, bleibt aber flexibel und erweiterbar. Eine Weiterentwicklung des FRBR-Modells für digitale Dokumente unter dem Titel *Functional Requirements for Information Resources* (FRIR) ([Mc12]) aus dem Jahr 2012 wurde nicht in RDA übernommen, kann aber bei Bedarf konzeptuell diese Lücke füllen.

durch assoziative *m:n-Relationen* verbunden sind. Die Tabelle *Work* (W) beschreibt die intellektuelle Idee, *Expression* (E) die konkrete Umsetzung, *Manifestation* (M) die spezifische Codierung oder Formatierung und *Item* (I) die physische oder digitale Instanz mit ihren Zugriffsmöglichkeiten. Einem Werk können mehrere Ausprägungen, einer Ausprägung verschiedene Erscheinungsformen und einer Erscheinungsform wiederum verschiedene Items zugeordnet werden. Zusätzlich werden innerhalb jeder Klasse klasseninterne Beziehungen definiert, wie z.B. *has-part* oder *is-successor-of*. Damit können auch zusammengesetzte Werke, Teilausprägungen oder Erscheinungsformen, die aus mehreren Komponenten bestehen, präzise beschrieben werden. FRBR wird in der Regel als *Entity-Relationship*-Modell beschrieben, so dass be einer „sauberen“ Modellierung die Menge der daraus abgeleiteten Tabellen automatisch der Datenbanknormalform NF3 genügt.[5]

Das FRBR-Modell ist wesentlich komplexer als das Dublin Core Modell (DC). Technisch gesehen kann FRBR als eine normalisierte und stark erweiterte Version von DC interpretiert werden. FRBR bringt daher sowohl die Vorteile als auch die Herausforderungen mit sich, die sich aus der Normalisierung ergeben. Die Erfassung von Metadaten nach dem FRBR-Modell erfordert nicht nur einen deutlich höheren Aufwand als nach dem DC-Modell, sondern auch ein wesentlich tieferes Fachwissen, um ein Dokument adäquat zu beschreiben. Im Allgemeinen steht solches bibliothekswissenschaftliches Fachwissen den Produzenten von LOGD nicht zur Verfügung.

Open WEMI. Auch innerhalb der Dublin-Core-Familie wird eine Erweiterung des flachen DC-Modells in Richtung FRBR vorgeschlagen. Im August 2024 wurde ein OpenWEMI-Modell prominent auf der Dublin-Core-Homepage zur Diskussion gestellt (https://www.dublincore.org/blog/2024/announcing-openwemi/). Das in 2024 publizierte Modell lässt auf formaler Sicht noch Fragen offen (https://github.com/dcmi/openwemi/issues/123), die sich in naher Zukunft allerdings ausräumen lassen sollten. Eine andere Eigenschaft von Open-WEMI stimmt nachdenklicher: Die Autoren legen explizit auf die Feststellung Wert, dass die vier WEMI-Klassen als Rollen interpretiert werden müssen und daher *nicht* disjunkt sind. Daraus folgt unseres Erachtens, dass Open-WEMI trotz gleicher Bezeichnung kein Modell in der Tradition von FRBR-WEMI ist und entsprechende Open-WEMI-Modelle nur bedingt an die einschlägigen Modelle der FRBR-Familie anschlussfähig sind.

[5] Die Tabelle *Work* enthält z.B. die Spalten *Autor*, *Schlagwort* und *Titel*, die in FRBR selbst wieder als Fremdschlüssel auf entsprechende weitere Tabellen realisiert sind. Die Tabelle *Expression* enthält unter anderem die Spalte *Sprache*. In der Tabelle *Manifestation* werden Attribute wie *Format* gespeichert, während die Tabelle *Item* Informationen über *Zugriffsort* und *Zugriffsrechte* enthält. Der aktuelle Standard sieht vor, dass die Inhalte dieser Tabellen – etwa Autorennamen, Schlagwörter oder Sprachen – idealerweise nicht als Freitexte, sondern aus öffentlich zugänglichen, kontrollierten Vokabularen entnommen werden; technisch entspricht dies Fremdschlüsseln in weiteren, hier zentral verwalteten Tabellen. Neben den vier zentralen Entitäten *Work (W)*, *Expression (E)*, *Manifestation (M)* und *Item (I)* definieren die aktuellen Nachfolger von FRBR weitere zentrale Klassen. Dazu gehören beispielsweise *Subject* (das Thema oder die „Aboutness“ eines Werks), *Nomen* sowie Akteure und Hersteller.

4 Lösungsansatz

Die Welt hat sich verändert, die Modelllandschaft hat sich weiterentwickelt. Vor 10 Jahren hat man sich – vermutlich aus sehr guten Gründen – für ein Metadatenformat aus der Modellfamilie A entschieden. Inzwischen bietet sich prinzipiell auch ein Modell der Modellfamilie B an. Die Frage ist: Gibt es genügend Hinweise oder treten Probleme auf, die gravierend genug sind, um über B neu nachzudenken?

Eine zwar nicht hinreichende, aber doch notwendige Grundlage sind – zumindest aus Sicht der Informatik – zunächst technische Überlegungen, verbunden mit den in den Strukturwissenschaften immer auch relevanten Urteilen über die formale Schönheit und Eleganz eines Modells. Es lässt sich zeigen, dass auch Modell B jedenfalls eine effiziente Lösung für die Metadatenmodellierung darstellt ([Jo09; Po18]). Das Urteil über die Effizienz muss ergänzend auch nicht-technische, insbesondere politische Entscheidungen einbeziehen: Wo will man die Hoheit über die eigene Metadatenhaltung behalten und wo will man sie abgeben – an wen, an welche Institutionen, mit welchen Ressourcen? Hier kommen administrative Anforderungen ins Spiel, u.a. hinsichtlich der Komplexität der Lösung und der Anschlussfähigkeit an bestehende institutionalisierte Strukturen. Effizienz bewertet in bester gestaltungswissenschaftlicher Tradition die Passung von Problem und Aufgabenstellung, insbesondere auch im Hinblick auf politische Ziele wie die Unterstützung von Verwaltung, Wirtschaft, Wissenschaft, Politik, Demokratie sowie spezifisch in unserem Kontext digitale Souveränität.

Wir skizzieren kurz, wie WEMI aus der Tradition der Modellfamilie B zu einem *WEMI-digital* weiterentwickelt werden kann. Dabei setzen wir voraus, dass die WEMI-Idee bekannt ist, eine ausführliche Einführung bietet z.B. [WH17]. Unser Anwendungsbeispiel aus dem Bereich LOGD oder Smart City wäre die aktuelle und historische Belegung von Parkhäusern in Deutschland, d. h. die Anzahl der belegten und freien Stellplätze eines beliebigen Parkhauses, das durch eine Parkhaus-ID identifiziert werden kann. Das wemi:Werk wäre „Parkhausbelegung“ als gedankliche Vorstellung. Im einfachsten Fall konstruieren wir als logisches Datenmodell eine einzige Relation mit den Spalten Parkhaus-ID, Zeitstempel, belegt, frei. Dieses Schema beschreiben wir in normaler Sprache, fügen ein UML-Diagramm und eine OWL-Ontologie hinzu. Damit erhalten wir einige unterschiedliche Teilmodelle, die sich teilweise ergänzen, teilweise konkurrieren, aber insgesamt die Idee der Parkhausbelegung recht gut beschreiben. Unter anderem fordern wir, dass die Parkhausbelegungsdaten schemakonform (abstraktes Datenmodell) im CSV-Format (das als anwendungsspezifisches konkretes Datenmodell interpretiert werden kann) geliefert werden. Im Idealfall liefern uns die Parkhausbetreiber ihre Daten dynamisch z. B. über SPARQL-Endpoints als orts- und zeitkonfigurierbare Zeitreihen inkl. Echtzeitdaten. Mit WEMI können wir diese Daten interpretieren:

Werk: Die geistige Vorstellung, die menschliche kulturelle Konstrution „Parkhausbelegung“ besteht aus dem Teilwerk „Parkhausbelegung-Schema“ (in der Welt des Semantic Web die Terminologie, die TBox) und dem Teilwerk „Parkhausbelegung-Daten“. (in der

RDF-Welt die Assertions, die ABox). Bei WEMI sprechen wir hier von einem einzigen (!) Werk „Parkhausdaten“ mit den Teilwerken „Parkhausstammdaten“ und „Parkhausbelegungsdaten“. – **Expression:** Die gedankliche Idee „Parkraumbelegung“ findet ihren Ausdruck in einer einzigen (!) wemi:Expression, nämlich der hypothetischen gedanklichen Vereinigung aller Parkraumbelegungsdaten europaweit zu allen Zeiten. – **Manifestation:** Entsprechend können wir auch eine hypothetische, gedankliche „große“ wemi:Manifestation als Vereinigungsmenge aller verfügbaren Belegungsdaten konstruieren. Diese Manifestation wird natürlich niemals in ihrer überwältigenden Gesamtheit real vorliegen. Was es aber gibt, sind sehr viele einzelne CSV-Datenpakete, die man in WEMI als Teilausprägungen dieser großen Ausprägung interpretieren kann. – **Item:** Ein einzelnes Datenpaket kann heruntergeladen, kopiert oder durch eine Datenbankabfrage (in LOD typischerweise an einen SPARQL-Endpoint) erzeugt und lokal gespeichert werden: Dadurch wird ein neues Element desselben Datenpakets erzeugt.

Unser Gedankenexperiment zeigt, dass die Metadatenmodellfamilie B eine Reihe von eingeführten feinen Unterscheidungen bereitstellt, die auch im digitalen Bereich leicht weiter ausdifferenziert werden können. Offen ist die Frage, welche Rolle nun die oben als relevant deklarierten Gundlagenprobleme spielen.

WEMI erscheint uns bei entsprechender Weiterentwicklung zu *WEMI-digital* geeignet, auch das httpRange-14-Problem zu entschärfen. Einen Hinweis gibt der bereits oben zitierte Wikipedia-Eintrag zum httprange-14-Problem: “[. . .] The impact of the issue (more correct the impact of confusion around the issue) is greatest in semantic web communities whose models involve large numbers of abstract concepts which cannot be serialized, such as the FRBR community”: Die hier zitierte Konfusion bietet auch die Chance für eine neue Lösung. Wir glauben, dass die Modellfamilie B und insbesondere *WEMI digital* generell als Basisontologie überall dort geeignet ist, wo etwas schriftlich beschrieben, dokumentiert, erfasst werden soll – also als zentrale Ontologie insbesondere der digitalisierten Welt.

Sehr verkürzt stellt sich der Zusammenhang wie folgt dar: Wann immer ein URI ohne Fragment-Identifier (ein sogenannter Slash-URI) verwendet wird, ist damit ein *Item* gemeint. Der URI ist dann zu verstehen als eine URL, ein *Locator*, eine *Adresse*, die auf eine entsprechende http-Anfrage einen Datenstrom zurückliefert. Die relevante Operation ist hier *retrieve*. Alle Ergebnisse von http-Abfragen (und damit auch Datenbankabfragen) mit derselben (z.B.) sha256sum sind Items derselben Teilmanifestation. – Wann immer man dagegen etwas über diesen Datenstrom aussagen will, verwendet man dafür sogenannte Hash-URIs, also einen URI mit Fragment-Identifier: Zu einem gegebenen Slash-URI erzeugt man z.B. durch Skolemisierung immer drei weitere Hash-URIs, mit denen man genauere Angaben zu dem jeweiligen WEMI-Aspekt machen kann. Die *Manifestation* hat etwas mit Format und Kodierung zu tun; ein relevantes Attribut ist Mimetype und sha256sum; die interessante Operation ist hier *open*. Die *Expression* hat etwas mit Sprache zu tun, die interessante Operation ist *read*. Auch für RDF-Modelle gibt es Hash-Funktionen, die formatunabhängig einen „abstrakten“ (z.B.) SHA-3 Hash für einen abstrakten RDF-Graphen erzeugen.

5 Zusammenfassung und Ausblick

Wir haben zwei konkurrierende Modellfamilien für die Metadatenhaltung von LOD vorgestellt. Ein ausführlicher, hier im Detail aus Platzgründen nicht vollständig nachweisbarer Vergleich zeigt, dass Metadatenmodelle aus Familie B möglicherweise technisch besser geeignet sind, LOD zu verwalten, als Metadatenmodelle aus Familie A. Die Entscheidung, Familie A statt B zu verwenden ist historisch nachvollziehbar, aber aus heutiger Sicht möglicherweise suboptimal. Eine Migration von A nach B ist informationslogistisch möglich, aber nicht trivial.

Die Metadatenverwaltung von GovData.de folgt dem DCAT-Modell, einem Mitglied der Metadatenfamilie A. Familie A ist im Vergleich zu Familie B weniger komplex, weniger differenziert und auch für interessierte Laien leicht anwendbar. Sie erlaubt die Erstellung von Metadaten auch in den Verwaltungen selbst. Unsere föderale Infrastruktur in Deutschland unterstützt die Souveränität der einzelnen Bundesländer und Verwaltungen. Die einfacheren Standards der Familie A machen die Verwaltungen unabhängiger und ermöglichen ihnen eine eigene, optimierte Metadatenhaltung. Die Metadatenhaltung der Metadatenfamilie B ist komplexer und erfordert ausgebildete Bibliothekare, die sich zusätzlich auf LOD spezialisiert haben.

Ist ein Wechsel von A zu B überhaupt eine realistische Option? Die Frage, ob B grundsätzlich besser wäre als A, ist nur dann relevant, wenn eine Migration von A nach B technisch und organisatorisch überhaupt machbar ist. Aus technischer Perspektive wäre zu prüfen, ob ein Modellwechsel effektiv möglich ist und effizient realisiert werden kann. Ein wesentlicher Aspekt ist dabei die *Informationslogistik*: Enthält das ausdrucksschwächere Modell A genügend Informationen, um eine automatische Übersetzung in Modell B zu ermöglichen? Eine Migration von DCAT nach WEMI würde zusätzliche Informationsquellen erfordern, um die Informationslücken zu schließen. – Der Begriff „Modellwechsel“ ist in diesem Zusammenhang stark zu relativieren: Natürlich ist es undenkbar, einen etablierten EU-weiten Standard wie DCAT durch einen anderen Standard ersetzen zu wollen. Wohl aber ist es möglich, Metadaten lokal in einem gegenüber DCAT ausdrucksstärkeren Modell vorzuhalten, das bei Bedarf automatisch auf DCAT abgebildet werden kann.

Aus politischer Sicht wäre eine Migration von A nach B ein äußerst sensibles Thema, da sie tief in die Kompetenzverteilung zwischen Bund und Ländern eingreifen würde. Sie hätte zudem weitreichende Auswirkungen auf die Aufgaben und Zuständigkeiten des Bibliothekswesens – und damit auf einen zentralen Aspekt der digitalen Souveränität. Stark verkürzt lautet unsere auf technischer Grundlage gestützte Botschaft an die Politik: Wenn – wie in der föderalen Struktur in Deutschland – jede Verwaltung und jedes Bundesland sein eigener Souverän ist, dann wird die Koordination zwischen den mächtigen Akteuren immer schwieriger. Die gemeinsame Nutzung von Daten und Ressourcen wird durch eine föderale Fragmentierung behindert, in der jeder Eigentümer von Daten und jeder Hersteller von Verwaltungs-Informationssystemen seinen Datenexport nach eigenem Belieben strukturieren darf. Nach DCAT gestaltete technische und Verwaltungs-Infrastrukturen sind so aufgebaut,

dass sie die Souveränität einzelner Verwaltungen fördern. Wenn in solchen Kontexten „Souveränität“ positiv konnotiert verstanden wird, sehen wir das kritisch.

Die Ressourcen, die in den Aufbau der Infrastruktur und in die Metadatenhaltung nach DCAT fließen, sollten unseres Erachtens besser in die Entwicklung von Metadatenstandards nach LRM investiert werden. Eine Konkurrenz zwischen DCAT und FRBR kann abgemildert werden, wenn ein WEMI-digital so gestaltet wird, dass DCAT-Daten daraus automatisch erzeugt werden können. Unser hier lediglich skizzenhaft vorgestellter Entwurf für WEMI-digital auf der Basis von FRBR entzieht als Nebenwirkung den Ländern und einzelnen Verwaltungen die Hoheit und überträgt sie in die Zuständigkeit des Bibliothekswesens.

Die Autoren dieses Aufsatzes werden die technische Seite von WEMI-digital und die Migration von DCAT aus wissenschaftlichem Interesse weiter verfolgen. Ob diese Forschung im Sinne der Design Sciences „relevant“ ist, bedeutet hier „denkbar“, „gewollt“, „machbar“ etc. – Das sind alles politische Fragen im Kontext der digitalen Souveränität, über die zumindest nachzudenken wir politisch für sinnvoll halten. Die technischen Voraussetzungen zu klären lohnt sich jedenfalls.

Literatur

[AAS18] Arakaki, F.; Alves, R.; Santos, P.: Dublin core: State of art (1995 to 2015). Informacao e Sociedade 28/, S. 7–20, 2018.

[ARŽ24] Aalberg, T.; Riva, P.; Žumer, M.: LRMOO object-oriented definition and mapping from the IFLA Library Reference Model. en,/Version 1.0, 2024.

[Be21] Bekiari, C.; Doerr, M.; Le Bœuf, P.; Riva, P.: LRMoo_V0.7(draft 2021-06-29).pdf, Techn. Ber., International Working Group on LRM, FRBR und CIDOC CRM Harmonisation, 2021, URL: https://www.cidoc-crm.org/frbroo/sites/default/files/LRMoo_V0.7%28draft%202021-06-29%29.pdf, Stand: 03. 04. 2024.

[Bu25] Bundesregierung: Zweiter Bericht der Bundesregierung über die Fortschritte bei der Bereitstellung von offenen Daten und Evaluierung der Wirkungsziele des § 12a des E-Government-Gesetzes, Techn. Ber. Bundestag Drucksache 20/15020, 2025, URL: https://dserver.bundestag.de/btd/20/150/2015020.pdf, Stand: 01. 03. 2025.

[Co] Coyle, C.: http://kcoyle.net/FRBR20.pdf, URL: http://kcoyle.net/FRBR20.pdf, Stand: 02. 11. 2024.

[Co14] Coyle, C.: Creating the Catalog, Before and After FRBR, 2014, URL: http://kcoyle.net/mexico.html, Stand: 30. 10. 2024.

[Co15] Coyle, K.: FRBR, Twenty Years On. en, Cataloging & Classification Quarterly 53/3-4, S. 265–285, 2015, ISSN: 0163-9374, 1544-4554, URL: http://www.tandfonline.com/doi/full/10.1080/01639374.2014.943446, Stand: 02. 11. 2024.

[Go17] Goldacker, G.: Digitale Souveränität, de, Techn. Ber., Fraunhofer-Institut für Offene Kommunikationssysteme FOKUS, 2017, URL: https://www.oeffentliche-it.de/documents/10181/14412/Digitale+Souver%C3%A4nit%C3%A4t.

[IT20] IT-Planungsrat: Stärkung der Digitalen Souveränität der Öffentlichen Verwaltung. de, Beschluss Nr.: 2020/19 des IT-Planungsrats vom 04.05.2020/, 2020, URL: https://www.cio.bund.de/SharedDocs/downloads/Webs/CIO/DE/digitale-loesungen/eckpunktpapier-digitale-souveraenitaet.pdf?__blob=publicationFile&v=4.

[Jo09] Johnston, P.: httpRange-14, Cool URIs & FRBR, 2009, URL: https://efoundations.typepad.com/efoundations/2009/02/httprange14-cool-uris-frbr.html, Stand: 10. 04. 2024.

[Mc12] McCusker, J. P.; Lebo, T.; Graves, A.; Difranzo, D.; Pinheiro, P.; McGuinness, D. L.: Functional Requirements for Information Resource Provenance on the Web. In (Groth, P.; Frew, J., Hrsg.): Provenance and Annotation of Data and Processes. Bd. 7525, Series Title: Lecture Notes in Computer Science, Springer Berlin Heidelberg, Berlin, Heidelberg, S. 52–66, 2012, ISBN: 978-3-642-34221-9 978-3-642-34222-6, URL: https://link.springer.com/10.1007/978-3-642-34222-6_5, Stand: 10. 04. 2024.

[Po18] Possemato, T.: How RDA is essential in the reconciliation and conversion processes for quality Linked Data. en, JLIS/1, 2018, ISSN: 20381026, URL: https://doi.org/10.4403/jlis.it-12447, Stand: 30. 10. 2024.

[RBŽ17] Riva, P.; Bœuf, P. L.; Žumer, M.: IFLA Library Reference Model A Conceptual Model for Bibliographic Information, en, Techn. Ber., 2017.

[WH17] Wiesenmüller, H.; Horny, S.: Basiswissen RDA: Eine Einführung für deutschsprachige Anwender. De Gruyter Saur, Berlin ; Boston, 2017, ISBN: 978-3-11-053868-7.

Ergänzungen online: Ergänzende Materialien zu diesem Aufsatz – darunter insbesondere auch die zugehörige Präsentation auf dem AKWI 2025 – finden sich unter https://www.jbusse.de/akwi2025/.

Generative KI als Schreibwerkzeug: Alle Absätze dieses Textes wurden ohne Verwendung von generativer KI vom Autor eigenhändig in korrektem (aber nicht immer sehr schönem) Deutsch formuliert, dann absatzweise mit Hilfe von Deepl Write in Bezug auf Rechtschreibfehler und Grammatik glattgezogen und posteditiert. Fünf Absätze wurden stichwortartig in Form von Nominalphrasen formuliert und an ChatGPT übergeben, danach (eher aufwändig) posteditiert. Der Prompt lautete hier: „Ich gebe dir einen Text, der in Nominalphrasen geschrieben ist. Bitte mache einen grammatikalisch flüssigen Text daraus, der den Sinn nicht ändert“. In zwei dieser Fälle wurde das Ergebis von ChatGPT verworfen und der Absatz in eigenen Worten neu formuliert.

Dank: Dieser Aufsatz beruht insbesondere auf Diskussionen in dem Dagstuhl Research Meetings „Ontologie, Linguistik, Terminologie, Logik“ (2023: https://www.dagstuhl.de/23144; 2024: https://www.dagstuhl.de/24014) sowie den Research Meetings „Applied Machine Intelligence“ (2022: https://www.dagstuhl.de/22173; 2023: https://www.dagstuhl.de/23263)