

Zeitschrift für digitale Geisteswissenschaften

Projektvorstellung aus:

Zeitschrift für digitale Geisteswissenschaften, Heft 10 (2025)

Titel:

Karl Kraus im Semantic Web. Zur Integration von Editionsdaten in einen gemeinsamen Wissensgraphen

Autor*in:

Bernhard Oberreither

Kontakt: bernhard.oberreither@oeaw.ac.at

Institution: Austrian Centre for Digital Humanities and Cultural Heritage, Österreichische Akademie der Wissenschaften | Austrian Academy of Sciences, Research Unit LTW – Literatur- und Textwissenschaft

GND: [1036707899](#) ORCID: [0000-0001-8609-2433](#)

DOI des Beitrags:

[10.17175/2025_003](#)

Nachweis im OPAC der Herzog August Bibliothek:

[192008097X](#)

Erstveröffentlichung:

26.03.2025

Lizenz:

Sofern nicht anders angegeben 

Letzte Überprüfung aller Verweise:

17.03.2025

Format:

PDF ohne Paginierung, Lesefassung

GND-Verschlagwortung:

[Datenmodellierung](#) | [Edition](#) | [Die Fackel \(Zeitschrift\)](#) | [Kraus, Karl](#) | [Semantic Web](#) | [Germanistik](#)

Empfohlene Zitierweise:

Bernhard Oberreither: Karl Kraus im Semantic Web. Zur Integration von Editionsdaten in einen gemeinsamen Wissensgraphen. In: Zeitschrift für digitale Geisteswissenschaften 10 (2025). 26.03.2025. HTML / XML / PDF. DOI: [10.17175/2025_003](#)

Bernhard Oberreither

Karl Kraus im Semantic Web. Zur Integration von Editionsdaten in einen gemeinsamen Wissensgraphen

Abstract

Bei allen Unterschieden hinsichtlich des edierten Texts, des editorischen Zugriffs oder der technischen Umsetzung – wenn es um die annotierten Entitäten geht, sehen sich viele digitale Editionen ähnlich: Zu ihren Kernbestandteilen gehören etwa häufig Register für Personen, Texte u. ä. Dabei sind die inhaltlichen Überschneidungen zwischen verschiedenen Projekten technisch oft nicht ausreichend abgebildet. Hier setzte das Projekt SemanticKraus an: Es widmete sich drei digitalen Editionen zu Karl Kraus, mit dem Ziel, deren Projektdaten in einen gemeinsamen Wissensgraphen zu integrieren und im Semantic Web zu publizieren. Der Beitrag beleuchtet die Voraussetzungen, die Vorgehensweise und einige methodische Herausforderungen des Projekts.

Despite all the differences in terms of the edited text, editorial approach or technical implementation – when it comes to the annotated entities, many digital editions look quite similar: their core components often include indexes of, e. g., persons and texts. However, the technical representation of content overlaps between different projects is often inadequate. The SemanticKraus project addressed this issue in the case of three digital editions of Karl Kraus, with the aim of integrating the project data into a common knowledge graph and publishing it on the Semantic Web. The article highlights the prerequisites, the procedures and some of the methodological challenges of the project.

1. Das Semantic Web und die Annotations- und Registerpraxis digitaler Editionen

¹Der Eindruck, den Interessierte aus der Durchsicht digitaler Editionen gewinnen können, ist der einer großen Vielfalt: an inhaltlichen und methodologischen Ansätzen ebenso wie an das Design und die technische Umsetzung betreffenden Lösungen. Zugleich fallen jedoch große Überschneidungen zwischen digitalen Editionen ins Auge. Dies betrifft vor allem die Aufmerksamkeit, die Editionen bestimmten Klassen von Entitäten widmen – ihrer Identifikation, ihrer Annotation im Text und ihrer Aufnahme in Register. Gleich nach der Bereitstellung des Textes in Form von Umschrift und Faksimile zählen Personen-, Orts- und Werkregister erfahrungsgemäß zu den am häufigsten anzutreffenden Modulen digitaler Editionen. Sie stellen Projekterträge dar, die mit großem Aufwand gesammelt und integriert werden. Die Ziele dabei variieren: In einigen Fällen mag die Möglichkeit im Vordergrund stehen, Nutzer:innen neben der linearen Textlektüre auch andere Wege in den Text zu erschließen, ihnen also die Möglichkeit zu bieten, über die Nennungen von Personen, Werken oder anderen Entitäten in den Text einzusteigen; damit bindet die Annotation den Text an die außertextuelle Realität an und hat damit, wie ein Kommentar, eine vermittelnde Funktion inne.² In anderen Fällen mag es umgekehrt darum gehen, die historische Wirklichkeit aus den Entitäten des Textes, den in ihm genannten Ereignissen, Personen, Orten heraus überhaupt erst zu erschließen.³ [1]

Seit Jahren wird immer wieder die Forderung nach der Verknüpfbarkeit und Interoperabilität solcher Registerdaten gestellt. Ein wichtiger Schritt in diese Richtung ist die vielgeübte Praxis der Verlinkung auf Normdatensätze wie etwa die der **Gemeinsamen Normdatei**, was die eindeutige Identifikation der Entitäten [2]

¹ Mit Dank für wertvolle Hinweise an Daniel Elsner.

² Vgl. z. B. Gabler 2010, S. 40.

³ Letztgenanntes entspricht dem, was Georg Vogeler als Zweck einer *assertive edition* definiert hat, und hat seine Domäne tendenziell in den historischen Fächern. Bestimmt gibt es Überschneidungen und stufenlose Übergänge zwischen diesen beiden Zielsetzungen der entitätenbezogenen Annotation: Auch die Registerpraxis genuin literaturwissenschaftlicher digitaler Editionen hat in ihren Bezügen zur außerliterarischen Wirklichkeit »assertive Anteile« – wie auch immer man im Einzelfall die epistemologischen und literaturtheoretischen Hürden zwischen den Sphären der Literatur / der Fiktion und der Realität fassen mag. Vgl. Vogeler 2019.

gewährleistet. Der vom Großteil der Forschungsgemeinschaft getragene Annotationsstandard **TEI** schafft hinsichtlich eines anderen Aspekts des Problems Abhilfe: Durch die Nutzung weitgehend derselben und natürlichsprachlich benannter Kategorien (<person>, <persName>, etc.) stellen die abgelegten Datensätze auch für Außenstehende in der Regel keine ›black box‹ mehr dar.

Das Semantic Web⁴ jedoch bietet darüber weit hinausreichende Möglichkeiten – deren Realisierung dabei von einigen Bedingungen abhängt. Die Kombination von RDF-Graphen⁵ und maschinenlesbaren Vokabularen (Ontologien) ermöglicht eine umfassend ›erläuterte‹ Speicherung von Aussagen: RDF-Triples (Subjekt-Prädikat-Objekt) zerlegen jede Information in atomare Einheiten, deren Glieder in den verwendeten Ontologien ausführlich beschrieben und deren Klassenhierarchie und andere Eigenschaften ebendort menschen- und maschinenlesbar hinterlegt sind. So lassen sich, die entsprechenden Vokabulare vorausgesetzt, Informationen inferieren, die gar nicht expressis verbis in den Daten angelegt sind. Die eindeutige Identifikation von Entitäten durch URIs (*Uniform Resource Identifier*) sowie durch die Verlinkung mit externen Identifiern trägt weiters zur Anbindung der Daten an die Linked Data Cloud bei. Die technischen Grundlagen (die notwendigen Schnittstellen) vorausgesetzt, ist so die kombinierte Abfrage über mehrere Datensets möglich – neben der Generierung inferierten Wissens wahrscheinlich der Goldstandard der Semantic-Web-Nutzung.

[3]

Die Möglichkeiten von Semantic Web und Graphentechnologien im Zusammenhang mit digitalen Editionen wurden schon mehrfach angemerkt, eingefordert, kritisch reflektiert.⁶ Ein rezenter Sammelband widmet sich ausschließlich dieser Thematik.⁷ In methodisch-theoretischer Hinsicht liegt für die digitalen Geisteswissenschaften großes Potenzial in den Möglichkeiten des Semantic Web – immer die bestmögliche Übertragung fachspezifischer Begrifflichkeiten in den Kontext von Semantic-Web-Ontologien vorausgesetzt.⁸

[4]

Die Möglichkeiten des Semantic Web können jedoch unter den Voraussetzungen, unter denen Editionsprojekte geplant und umgesetzt wurden (und vielfach noch werden), oft nicht genutzt werden. Das hat unter anderem mit dem Aufwand der Implementierung bzw. der Prioritätenreihung zu tun. Dem stehen Projekte gegenüber, die sich die Nutzung dieser Technologie von Beginn an auf die Fahnen schreiben, die also in ihrer Umsetzung stets schon RDF und Co. nutzen.⁹ Andernfalls kann man den Schritt ins Semantic Web auch nachträglich machen. So ein Unterfangen hat seine Hürden, aber auch seine Erkenntnispotenziale, und ist Gegenstand der folgenden Ausführungen. Dabei wird zuerst die Ausgangslage erläutert – die digitalen Editionen zum Werk von Karl Kraus, deren Projektdaten Anlass für das Projekt *SemanticKraus: Connecting Kraus-Scholarship to the Semantic Web* waren. SemanticKraus hat sich zum Ziel gesetzt, diese Daten nachzunutzen, indem sie transformiert, in einen gemeinsamen Wissensgraphen eingespeist und menschen- sowie maschinenlesbar publiziert werden. Nach einigen Anmerkungen zur technischen Umsetzung und zum Workflow wende ich mich dem Schwerpunkt der Projektvorstellung zu: der Behandlung der ›konzeptionellen Distanz‹, die zwischen den Ausgangsprojekten selbst sowie zwischen diesen und dem Ziel-Datenformat unter anderem hinsichtlich Datenmodell, Datengranularität und Referenzierbarkeit der Entitäten bestand und die auf dem Weg vom Ausgangsmaterial zum angestrebten RDF-Datensatz überwunden werden musste.

[5]

⁴ Siehe dazu grundlegend Berners-Lee et al. 2001.

⁵ **Resource Description Framework**.

⁶ Vgl. etwa Kamzelak 2016; Wettlaufer 2018.

⁷ Spadini et al. (Hg.) 2021.

⁸ Vgl. allgemein Ciotti 2014.

⁹ Auf Grundlage dieser Technologie wurden etwa – um nur drei Beispiele zu nennen – intratextuelle Relationen in einem Notizbuch von Paolo Bufalini, Rechtsfälle aus dem Basler Urfehdebuch oder Entitäten in den Briefen Vespasiano da Bisticcis erschlossen. Vgl. Daquino et al. 2020; Burghartz et al. (Hg.) 2017; Tomasi 2013.

2. Karl Kraus: Eine dicht besiedelte Forschungslandschaft

Karl Kraus' Schriften gehören wahrscheinlich zu den im Web schon am längsten und umfassendsten zur Verfügung gestellten Werkbeständen deutscher Sprache: Schon 2007 ging die [Edition der gesamten Fackel](#)¹⁰ online, der Zeitschrift, die Kraus von 1899 bis zu seinem Tod 1936 herausgab und die er für den größten Teil dieser 37 Jahrgänge alleine verfasste. Die Edition umfasst 415 Hefte mit insgesamt über 22.000 Seiten; seit 2019 wird sie um ein umfangreiches Personenregister mit Verlinkungen auf ca. 130.000 Textstellen ergänzt.¹¹ Im Jahr 2018 wurden auf *Karl Kraus online* Materialien u. a. zu den Rechtsstreitigkeiten des Autors publiziert; 2022 publizierte das [Karl Kraus Rechtsakten](#)-Projekt eine Edition dieser umfangreichen Aktenbestände der für Kraus tätigen Kanzlei Samek.¹² Die digitale [Edition seines Spätwerks Dritte Walpurgisnacht](#) ging 2021 mit einem ersten, zum Jahreswechsel 2023/2024 mit einem zweiten Publikationsschritt online.¹³ Weitere Projekte zu diesem Autor, der die Wiener und deutschsprachige Kulturlandschaft von der Jahrhundertwende bis in die letzten Jahre der österreichischen Ersten Republik kommentierte, kritisierte, maßgeblich mitgestaltete, sind in Planung.

[6]

Für die bereits abgeschlossenen bzw. in weit fortgeschrittenem Stadium befindlichen drei genannten Editionen jedoch gilt: Die Möglichkeiten des Semantic Web waren in der Planung unberücksichtigt geblieben, was teils schlicht mit dem Alter der Editionen, andernteils mit methodischer Ausrichtung oder erwähnter Prioritätenreihung zu tun hatte. In allen drei Projekten wurde beispielsweise auf Normdatensätze verlinkt (im Fall der Fackel-Personendatenbank nachträglich im Zuge der laufenden Aktualisierungsschritte), wobei unter anderem die GND berücksichtigt wurde sowie über die PMB,¹⁴ eine thematisch stärker spezialisierte Datenbank, auch die Verlinkung vom Normdatensatz zurück ins Projekt stattfand. Eine Abfrage über die gesamten Bestände war dennoch nicht möglich – und das, obwohl die Schnittmengen der drei Projekte enorm waren, sein mussten: Karl Kraus befandete sich in seinen Rechtsfällen vielfach mit denselben Personen, die auch in seiner Zeitschrift *Die Fackel* Ziel seiner Angriffe wurden; die *Fackel*-Texte, auf die er in seinem Spätwerk *Dritte Walpurgisnacht* zurückverweist, werden auch in den Schriften seiner Anwaltskanzlei vielfach genannt. Die inhaltlichen Überschneidungen waren also offensichtlich groß, bloß: technisch nicht ausreichend realisiert.

[7]

3. SemanticKraus: Ziele und Voraussetzungen

Diesem Desiderat widmete sich SemanticKraus.¹⁵ Zu den Zielen des Projekts zählte u. a.,

[8]

- die entsprechenden Register- und Textdaten der drei Ausgangsprojekte in ein und dasselbe, CIDOC-CRM-basierte¹⁶ RDF-Datenmodell zu übertragen,
- die resultierenden Datensets gesammelt in einem Triplestore abzulegen,
- diesem ein Userinterface vorzulagern, um die individuelle Erkundung der Daten zu ermöglichen, und
- eine ([SPARQL](#))-Schnittstelle für die maschinelle Abfrage der Daten bereitzustellen.

Sowohl Daten zu Personen als auch zu Texten wurden von allen drei Ausgangsprojekten auf die eine oder andere Weise gesammelt. Diese Klassen an Entitäten standen nun auch im Fokus unserer Transformationsbemühungen. Im Detail umfassten die Daten

¹⁰ Vgl. Biber 2015.

¹¹ Vgl. Biber et al. (Hg.) 2007–.

¹² Knüchel / Langkabel (Hg.) 2022.

¹³ Oberreither (Hg.) 2021.

¹⁴ [PMB – Personen der Moderne Basis](#).

¹⁵ SemanticKraus wurde mit Unterstützung durch das BMBWF durch CLARIAH-AT finanziert. Siehe [Projektseite bei CLARIAH-AT](#).

¹⁶ Vgl. Bekiari et al. (Hg.) 2022.

- im Fall der *Fackel online* die ca. 14.000 Texte der 37 *Fackel*-Jahrgänge, die knapp 130.000 Textpassagen mit Personen-Nennungen und die von dort aus referenzierten ca. 15.000 Einträge der Personendatenbank;
- im Fall der Rechtsakten die annähernd 4.000 juristischen Dokumente inklusive ihrer Bezüge auf 3 Registerdateien, die 1.700 biographische und 850 bibliographische Entitäten verzeichnen;
- im Fall der *Dritten Walpurgisnacht* den Text des Werkes selbst sowie seine ca. 3.000 Bezüge auf ca. 400 Personen und 1.100 Intertexte.

Die Daten unterschieden sich hinsichtlich ihrer spezifischen Modellierung stark voneinander, auch dort, wo dieselbe Technologie und derselbe Standard verwendet wurden.

- Der Volltext der *Fackel online* liegt in einem älteren, hauseigenen XML-Standard vor, der nur entfernte Ähnlichkeit mit TEI aufweist.
- Die dazugehörige Personendaten liegen in einer SQL-Datenbank (aus der sie als TEI-XML ausgeworfen wurden).
- Die Texte sowie die Register der Rechtsakten und der *Dritten Walpurgisnacht* liegen als TEI-XML vor, jedoch in unterschiedlichen Modellierungen.

Bei all den Anforderungen, die diese Voraussetzungen an die Erstellung eines übergreifenden Datenmodells sowie an die technische Transformation der Daten stellten, war SemanticKraus dennoch ein Projekt von sehr kompaktem Zuschnitt. Hingegen waren die institutionellen Voraussetzungen sehr gut: Sämtliche Projekte waren auf die eine oder andere Weise ans ACDH-CH, die ausführende Institution von SemanticKraus, angebunden bzw. liefen noch ebendort. Das sei hier angemerkt, um die ansonsten oft signifikanten Schwierigkeiten, sich in »fremde« Projektdaten einzuarbeiten, durch den Kontrast zu beleuchten: Offene Fragen, dunkle Stellen im jeweiligen Projekt-Datenmodell konnten auf die einfachst mögliche Weise erhellt werden (oft durch den Gang ins benachbarte Büro). Darüber hinaus begrüßten die Stakeholder der verschiedenen Projekte die Unternehmung, vielfach konnte zudem auf vorhandenes Know-How der technischen Säule des ACDH-CH zurückgegriffen werden.¹⁷ Außerdem ermöglichte es die für die Webapplikation genutzte Softwarelösung – ResearchSpace – mit verhältnismäßig niedrigem Aufwand, die erforderlichen Templates für das User-Interface zu erstellen.¹⁸

[9]

Die SemanticKraus Exploration Platform als zentrales Ergebnis des Projekts bietet neben Paratexten zum Projekt und zu den Ausgangsprojekten eine Suchfunktion über die gesamten Daten sowie ein »show case« an Beispiel-Entitäten, das zentrale Personen und Texte von besonderem Interesse als Einstiegspunkte vorschlägt. Durch die Aggregation der Daten lässt sich nun etwa Kraus' Behandlung von historischen Personen durch die *Fackel*, die Rechtsakten sowie das Spätwerk der *Walpurgisnacht* verfolgen. Das Datenmodell wird mittels einer detaillierten Visualisierung vorgestellt. Das auf ResearchSpace aufbauende Userinterface setzt sich aus ca. 20 für die wichtigsten Klassen individuell angepassten Templates zusammen, die verschiedene Visualisierungsmöglichkeiten bieten (vgl. Abbildung 1). Das SPARQL-Interface ermöglicht die Exploration der Daten noch über die Grenzen der Templates hinaus.

[10]

Die Modellierung der Daten als RDF erlaubt verhältnismäßig einfach die zukünftige modulare Erweiterung der Plattform durch die Integration weiterer Forschungs- und Editionsdaten. Die aktuellen Projektdaten sowie die Projektontologie liegen in öffentlichen Git-Repositorien, von wo aus Transformation und Daten-

[11]

¹⁷ Dass bei einem Projekt wie SemanticKraus vielfach auf die Expertise und die Erfahrungswerte der involvierten Research Software Engineers zurückgegriffen wurde, versteht sich von selbst, sei aber dennoch erwähnt. Im Zusammenhang mit dem Folgenden sind insbesondere Peter Andorfer, Andreas Basch, Daniel Elsner und Carl Friedrich Haak hervorzuheben. Eine vollständige Liste der Projektmitarbeiter:innen findet sich unter <https://semantickraus.acdh.oeaw.ac.at/resource/app:Project>.

¹⁸ Tatsächlich erfordert die Anpassung der HTML-Templates in ResearchSpace nicht unbedingt einen technischen Hintergrund; HTML- und SPARQL-Grundkenntnisse sind notwendig, letztere bedürfen je nach Anwendung punktueller Vertiefung.

Ingest in den der Web-Applikation zugrundeliegenden GraphDB-Triplestore mittels CI/CD-Workflows¹⁹ automatisiert erfolgen. Die Daten umfassen derzeit u. a. 15.900 Personen (konsolidiert von insgesamt 16.600 Personendatensätzen in den Ausgangsdaten) sowie 19.700 Texte mit 18.900 intertextuellen Relationen und 145.800 Personennennungen.

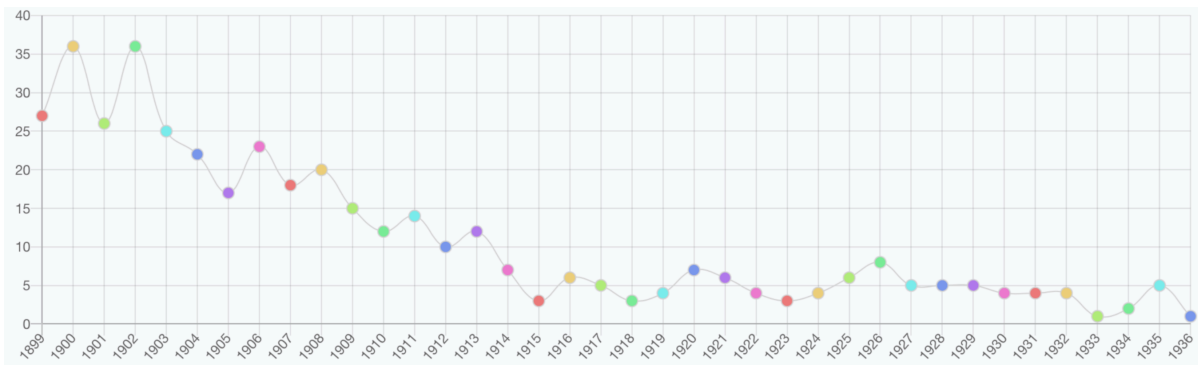


Abb. 1: Eine Visualisierung des Erscheinungsverlaufs der *Fackel*, basierend auf dem transformierten Text-Index der *Fackel online*, erstellt mittels einer ResearchSpace-Komponente und abrufbar unter der [URI der Fackel in SemanticKraus](#). [Screenshot: Bernhard Oberreither 2025]

Diese Projektvorstellung widmet sich im Folgenden vorrangig der eingangs erwähnten »konzeptionellen Distanz« zwischen den TEI-XML-Daten, die am Anfang standen, und dem resultierenden RDF: RDF bzw. Linked Data zeichnet sich ja prinzipiell durch einen großen Grad an Eindeutigkeit aus (alle Entitäten sind eindeutig identifiziert); darüber hinaus – jedenfalls innerhalb der CIDOC-CRM-Familie – durch einen großen Grad an Feinkörnigkeit und damit einhergehend an Explizitheit (durch die präzisen Definitionen der Klassen und Relationen in den Ontologien). Ausgangspunkt meiner Überlegungen ist die recht basale Beobachtung, dass das bei (TEI-)XML nicht in diesem Maß der Fall ist, nicht in der üblichen Praxis, nicht in unseren konkreten Ausgangsprojekten. Vorerst lässt sich das Problemfeld der Konversion von TEI-XML in RDF / CIDOC CRM wohl mit diesen – zugegeben: stark vereinfachenden – Befunden umreißen:

[12]

- Ungenaue Daten: Den Daten, die das TEI-XML enthält, mangelt es im Vergleich zum Zieldatenmodell oft an differenzierten Klassifikationen für die Entitäten.
- Zu wenige Daten: Das TEI-XML enthält meist weniger explizite Information (weniger »Datenfelder«), als das Ziel-Datenmodell voraussetzt.
- Uneindeutige Daten: Den Entitäten der Ausgangsdaten fehlen oftmals eindeutige Identifier.
- Die falschen Daten: Die Daten, die das Ausgangs-XML annotiert, entsprechen nicht exakt denen, die man extrahieren und in das Ziel-Datenmodell übertragen will.

Der vierte Punkt ist womöglich weniger verallgemeinerbar als die anderen, er spielte jedoch im Projektkontext eine große Rolle (vgl. Abschnitt 6.2 zum *Fackel*-Textindex) und wird deshalb hier berücksichtigt. In den kommenden Abschnitten werden exemplarisch Ausgangslage und Arbeitsschritte vorgestellt, die manchmal einen, meist mehrere dieser Befunde betreffen. Es versteht sich übrigens von selbst, dass diese Liste keinerlei Kritik impliziert – weder an der gängigen TEI-Praxis noch an unseren drei Ausgangsprojekten (von denen ich selbst eines verantworte). Die Differenz zwischen TEI und CIDOC CRM liegt schlicht in der Natur der Sache, im Unterschied nämlich zwischen einer Text-Auszeichnungssprache und einer ursprünglich dem Museumskontext entstammenden, ereigniszentrierten Ontologie.²⁰

[13]

¹⁹ CI/CD – *continuous integration / continuous delivery* – steht für die automatisierte Übernahme von Änderungen und Ergänzungen im Code bzw. in den Daten in die Zielapplikation. So löste beispielsweise jede Änderung in den XML-Daten automatisch deren erneute Transformation in RDF sowie den Upload in den Triplestore aus. Siehe den GitHub-Artikel [CI/CD: The what, why, and how](#).

²⁰ Diese Distanz zu überbrücken war schon Thema umfangreicher Bemühungen, siehe zum Beispiel Ore / Eide 2009; Eide 2015; Ciotti / Tomasi 2016.

Am Beispiel der Transformation intertextueller Verweise und bibliographischer Daten lassen sich die Erkenntnispotenziale solch eines Projekts am besten illustrieren. Dies liegt zum Teil daran, dass hier die Modellierungsunterschiede in unseren Ausgangsprojekten relativ groß waren. Darüber hinaus war die Modellierung solcher Daten in unserem RDF-Datenmodell deutlich komplexer als beispielsweise die biografischen Daten. Zudem traten in diesem Bereich die Folgen des, wie man es nennen kann, ›analogen Erbes‹ der verschiedenen Standards, nach denen solche Daten strukturiert sein können, auf interessante Weise zu Tage. Im Rückblick lässt sich jedenfalls feststellen: Datentransformation ist der beste Weg, Daten – und zwar auch die eigenen – noch einmal ganz neu kennenzulernen.

[14]

Zunächst jedoch ein paar Worte zu den anderen Eckpfeilern des Projekts: zu Workflow und technischem Setup sowie – was in einem Projekt, das fremde Daten akkumuliert, von besonderer Bedeutung ist – zur Darstellung der Datenprovenienz.

[15]

4. Workflow, Datenformat, technisches Setup und Datenprovenienz

Die zentralen Work Packages des Projekts umfassten:

[16]

- Datensichtung
- Erstellung des Datenmodells
- Erstellung des Test-Datensatzes
- Einrichtung der GraphDB-Instanz
- Datenanreicherung / Erstellung von Hilfsdateien
- Datentransformation (Python, XSLT)
- Erstellung des Userinterfaces
- Qualitätskontrolle in mehreren Iterationen

Einige der Work Packages liefen parallel, um enge Feedbackschleifen zu ermöglichen. Zeitgleich mit der Erstellung und Verfeinerung des Datenmodells begann die Erstellung des Test-Datensatzes, einer RDF-Datei im Turtle-Format (.ttl), die das gesamte Datenmodell mit jeweils ein bis zwei Entitäten abdeckte.²¹ Auf Grundlage dieses Test-Datensatzes wurde schon sehr früh im Projekt mit der Erstellung der Web-Applikation begonnen. Zugleich ermöglichte dieser Testdatensatz das Auffinden von Verbesserungsmöglichkeiten im ursprünglichen Datenmodell. Die frühe Erstellung der grundlegenden ResearchSpace-Applikation erleichterte auch im Folgenden die laufende Qualitätskontrolle der Daten, umso mehr, als dort sowohl die händische Datenexploration als auch SPARQL-Abfragen komfortabel möglich sind. Im Weiteren erfolgten Datentransformation und -Ingest nach jedem Korrekturdurchgang automatisch mittels CI/CD-Workflows; parallel zu den letzten Korrekturdurchgängen in den RDF-Daten wurde auch das User-Interface fertiggestellt.

[17]

Weil SemanticKraus die Daten dreier Ausgangsprojekte akkumuliert und diese Ausgangsprojekte wiederum selbst teils Quellen für die bereitgestellten Informationen angeben, wurde die Frage der Datenprovenienz auf zwei Ebenen behandelt: auf der Ebene individueller (etwa Personen-)Einträge und auf der übergeordneten Ebene des Projektdatensatzes. Um für individuelle Einträge die Datenherkunft anzugeben, bieten sich unterschiedliche Methoden an: Am feinkörnigsten ist sicherlich die Angabe der Quelle zur jeweils einzelnen Aussage, zum einzelnen RDF-Triple (was u. a. durch dessen Reifizierung²² erreicht werden kann²³). Einen

[18]

²¹ Zur Erstellung dieser Testdaten wurde die Datenintegrations-App **Karma** verwendet, mit der wir auf Grundlage tabellarischer Dummy-Daten RDF produzierten.

²² Dabei steht eine neue URI anstelle des ursprünglichen Triples; Subjekt, Prädikat und Objekt werden über die Properties `rdf:subject`, `rdf:predicate`, `rdf:object` ergänzt, über andere Properties können die Metadaten des Triples angegeben werden.

mittleren Grad an Feinkörnigkeit stellt die Angabe einer Quelle zur (gesamten) betreffenden Entität dar, womit sich die Quellenangabe auf sämtliche Informationen z. B. zur Person bezieht. Am wenigsten exakt wäre eine Angabe auf Höhe des gesamten Registers, die klärt, welche Quellen insgesamt für die Erhebung biographischer Angaben genutzt wurden, ohne genauere Zuordnung. In SemanticKraus wurde für Daten dieser Art der zweitgenannte Weg gewählt: Beispielsweise geben drei Viertel der <person>-Elemente des Registers der *Dritten Walpurgisnacht* eine Quelle an; diese Quellenangabe wurde mit PROV-O²⁴ modelliert und mit der jeweiligen Person verknüpft.

Auf Ebene der Projekte hat die Angabe der Datenprovenienz zusätzlich zum Modellierungs- auch einen technischen Aspekt: Als Datenformat für SemanticKraus wurde **TriG** (.trig) gewählt, eine Syntax zur Darstellung von *Named Graphs*. Dabei handelt es sich um RDF-Datensätze, die in ihrer Gesamtheit ebenfalls durch eine URI vertreten werden, sodass der Datensatz mit Metadaten ausgestattet werden kann – etwa zur Datenprovenienz. Diese Methode wurde im Projekt SemanticKraus angewendet, unter Einbeziehung von Angaben in Prosa zum Ausgangsprojekt und zur Konvertierung (vgl. Abbildung 2).

[19]

```
@prefix sk: <https://sk.acdh.oeaw.ac.at/> .
@prefix crm: <http://www.cidoc-crm.org/cidoc-crm/> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix ed: <https://sk.acdh.oeaw.ac.at/project/> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix sd: <http://www.w3.org/ns/sparql-service-description#> .

sk:provenance {
  ed:dritte-walpurgisnacht a sd:NamedGraph ;
  rdfs:label "Dritte Walpurgisnacht"@en ;
  dcterms:bibliographicCitation "Karl Kraus: Dritte Walpurgisnacht. [...].";
  dcterms:contributor "Research – concerning persons mentioned and [...]."@en ;
  dcterms:source "https://kraus1933.ace.oeaw.ac.at" ;
  dcterms:title "Dritte Walpurgisnacht"@de ;
  rdfs:comment "The digital edition of 'Dritte Walpurgisnacht' [...]."@en ;
  sd:name "Dritte Walpurgisnacht" .
}

ed:dritte-walpurgisnacht {
  [...]

  sk:DWpers0125 a crm:E21_Person ;
  rdfs:label "Kerr, Alfred"@und ;
  crm:P100i_died_in <https://sk.acdh.oeaw.ac.at/DWpers0125/death> ;
  crm:P14i_performed <https://sk.acdh.oeaw.ac.at/DWpers0125/occupation/01>,
    <https://sk.acdh.oeaw.ac.at/DWpers0125/occupation/02>,
    <https://sk.acdh.oeaw.ac.at/DWpers0125/occupation/03> ;
  crm:P1_is_identified_by <https://sk.acdh.oeaw.ac.at/DWpers0125/appellation/0>,
    <https://sk.acdh.oeaw.ac.at/DWpers0125/appellation/2>,
    <https://sk.acdh.oeaw.ac.at/DWpers0125/identifier/DWpers0125>,
    <https://sk.acdh.oeaw.ac.at/DWpers0125/identifier/idno/0>,
    <https://sk.acdh.oeaw.ac.at/DWpers0125/identifier/idno/1> ;
  crm:P98i_was_born <https://sk.acdh.oeaw.ac.at/DWpers0125/birth> ;
  prov:wasDerivedFrom sk:DWSource0026 .

  [...]

  sk:DWSource0026 a dcterms:BibliographicResource,
    prov:Entity ;
  rdfs:label "Killy"@de ;
  dcterms:bibliographicCitation "Walter Killy (Hg.): Literaturlexikon. [...].";

  [...]
}
```

Abb. 2: Datenprovenienz in SemanticKraus, auf Ebene des Projektdatensatzes (als Named Graph im TriG-Format innerhalb von geschweiften Klammern) sowie auf Ebene einzelner Entities (unten, die Person wird durch prov:wasDerivedFrom mit der Quelle verbunden). [Grafik: Bernhard Oberreither 2025]

²³ Einen Überblick über die Darstellung von Datenprovenienz – u. a. durch Named Graphs – bieten Sikos / Philp 2020.

²⁴ Vgl. Lebo et al. 2013.

5. Das Datenmodell

Im Bereich der Datenmodellierung fiel die Wahl eines Metamodells schnell auf CIDOC CRM und die damit kompatiblen Modelle – da CIDOC CRM hinsichtlich der Größe seiner Domäne, seiner Verbreitung in der Forschungsgemeinschaft sowie seiner Flexibilität die beste Option darstellt, zudem ausreichend Klassen für die Modellierung der biographischen Daten bereitstellt und das kompatible Modell FRBRoo²⁵ auch die Modellierung von bibliographischen Daten in großer Trennschärfe ermöglicht. Wo FRBRoo aufhörte, dort nämlich, wo es um die Modellierung einzelner Textpassagen mit intertextuellen Verweisen bzw. Personennennungen ging, schloss das ebenfalls CIDOC-kompatible INTRO²⁶ an. CIDOC CRM ist event-zentriert, das heißt: Personen, Objekte etc. werden um die sie verbindenden Ereignisse herum gruppiert; diese Ausrichtung wurde dort, wo es um bibliographische Daten ging, beibehalten: Auch FRBRoo modelliert zum Beispiel das Verhältnis zwischen einem Text und einem Autor als Ereignis (nämlich als Schaffung des Texts). Wo es um (etwa: semantische) Textmerkmale ging, trat die ereigniszentrierte Sichtweise in den Hintergrund zugunsten einer Modellierung, in deren Zentrum das aus INTRO bezogene konzeptuelle Objekt ›Intertextueller Verweis‹ (bzw. im Fall der Personennennung: ›Referenz‹) steht.

[20]

Die Modellierung von Personennennungen und intertextuellen Verweisen berührt aufgrund der Nähe zur Textannotation das Thema des Verhältnisses von TEI zu CIDOC CRM. SemanticKraus baut auf einer komplementären Ergänzung von TEI-XML und RDF auf: So werden die TEI-Elemente aus den Registern in RDF rekonstruiert und diese Rekonstruktionen mit Textpassagen im publizierten TEI verlinkt. Letzteres geschieht über IDs, die von TEI grundsätzlich unabhängig sind. Die Wiedergabe der Textstruktur durch TEI bleibt unangetastet, unberücksichtigt bleiben auch die Unterschiede zwischen inline-Elementen im TEI-XML (<quote> und <rs> etwa wurden beide schlicht als ›Textpassage‹ erfasst), etc. Diese Form der gegenseitigen Agnostik zwischen TEI-XML und RDF hat sich als durchaus praktikabel erwiesen.²⁷

[21]

6. Differierende Modellierungen bibliographischer Angaben

6.1 Anreicherung als Strukturierung

Der Weg von der traditionellen bibliographischen Angabe zum in FRBRoo ausmodellierten Datensatz – ein Weg, der letztlich in der Explikation von Informationen besteht, die in der analogen akademischen Praxis implizit blieben – wurde im Vorfeld von SemanticKraus, nämlich im TEI-Code der Ausgangsprojekte ja schon zum Teil zurückgelegt. In TEI lässt sich natürlich so gut wie alles explizieren; disziplinäre Bräuche einerseits, Aufwandsabwägungen andererseits führen jedoch dazu, dass diese Möglichkeiten kaum voll genutzt werden (was auch immer ›volk im jeweiligen Fall heißen mag). Die Datenpräsentation in der Webapplikation einer digitalen Edition steht oft auch unter dem Einfluss traditioneller Standards – mit gutem Grund, ist das doch der Standard des (erwarteten) Publikums. In genau dem Maß jedoch, in dem die Modellierung dieser Daten auf deren traditionelle Präsentation hin optimiert ist, wird die Implizitheit von Information in diesen hergebrachten Zitationsstandards schlagend. So kann ein TEI-<bibl>-Element eine nicht weiter strukturierte bibliographische Angabe nach traditionellem Zitationsmuster enthalten, was für die Anzeige in der Infospalte neben dem annotierten Text oder im Register einer Edition schon ausreichend sein kann – dabei aber ein Extrem an Implizitheit von Information darstellt. Dieses Extrem berührten die Ausgangsdaten von

[22]

²⁵ Vgl. Bekiari et al. (Hg.) 2015. FRBRoo wurde mittlerweile durch das etwas vereinfachte Modell LRMoo abgelöst, vgl. Bekiari et al. (Hg.) 2024.

²⁶ Vgl. Oberreither 2023.

²⁷ Øyvind Eide unterscheidet drei Modi der Verbindung von TEI-XML und RDF. In zweien davon vermitteln (einmal TEI-, einmal Non-TEI-) Elemente im <teiHeader> zwischen dem externen RDF und den Elementen des <body>. SemanticKraus folgt indes der anderen genannten Möglichkeit, der direkten Verknüpfung des externen RDF-Datensatzes mit der jeweiligen Textpassage. Vgl. Eide 2015.

SemanticKraus nicht; ein Teil musste dennoch in größerer Granularität annotiert werden (ein Arbeitsschritt, der größtenteils mit RegEx und Such- und Ersetz-Durchgängen im jeweiligen Register-XML erledigt werden konnte).

6.2 Ausdifferenzierung der ontologischen Ebenen von ›Text‹

Grundsätzlich erwies sich die Repräsentation textbezogener – bibliographischer und intertextueller – Daten als die deutlich komplexere Aufgabe als die Modellierung der biographischen Daten. Dies liegt vor allem daran, dass annotierte Personennennungen meist schlicht eine reale Person referenzieren: »Martin Heidegger« bezieht sich meist auf exakt ein reales Objekt. Bei *Sein und Zeit* beispielsweise sieht das schon ganz anders aus: Dieser Gegenstand lässt sich ganz unterschiedlich kategorisieren, unter Heranziehung von Begriffen wie Werk, Text, Ausgabe, etc. In Hinblick auf diese Ebenen bibliographischer ›Seinsweisen‹ griff SemanticKraus unter anderem auf die FRBRoo-Klassen ›F22 Self-Contained Expression‹ und ›F24 Publication Expression‹ (in etwa: Text und publizierte Fassung des Texts) zurück.²⁸

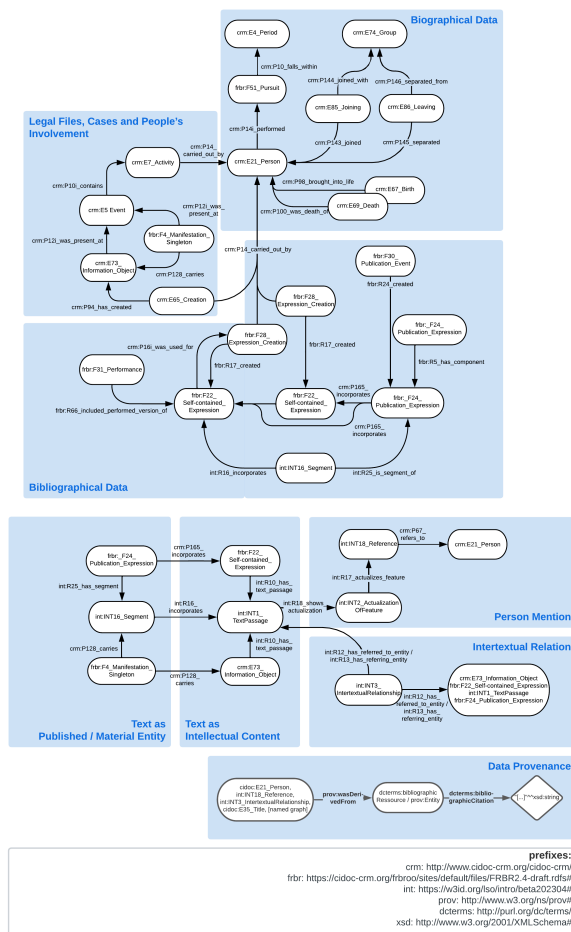


Abb. 3: Das Datenmodell im Großen; ausgespart sind Zeit- und Ortsangaben zu sämtlichen Ereignissen, die als E52 Time-Span bzw. E53 Place modelliert wurden, sowie E42 Identifier, E33 Linguistic Object, E41 Appellation, rdfs:label und rdf:value; zur Datenprovenienz siehe auch [Abbildung 2](#). [Grafik: Bernhard Oberreither 2025]

²⁸ Aufwändiger ist die Verknüpfung dieser Konzepte untereinander: Die Teile auch der vermeintlich einfachsten »traditionellen« bibliographischen Angabe sind in ihrer FRBRoo-Modellierung immer wieder durch überraschend viele Zwischenschritte voneinander getrennt.

Die Differenz zwischen traditionellem Anspruch und den Anforderungen der RDF-Konvertierung zeigt sich hier auf recht subtile Art: Verweist etwa eine Textpassage auf eine (auch: strukturierte) bibliographische Angabe, ist den User:innen intuitiv oder aus dem Kontext klar, was gar nicht zwangsläufig spezifiziert ist: ob diese Angabe sich auf einen Text in einer spezifischen Fassung (einer Publikation) oder bloß den Text ›im weiteren Sinn‹ (z. B. als ›Werk‹) bezieht; oder ob mit dem Verweis auf eine Zeitschrift eine Ausgabe der Zeitschrift, einen Artikel daraus oder die Zeitschrift als das der einzelnen Ausgabe hierarchisch übergeordnete ›Werk‹ referenziert wird. Alles dies lässt sich mit TEI explizieren, in der Praxis wird es aber mitunter offen bleiben: vielleicht aus Gründen der Effizienz, oder weil die Datengrundlage nicht ergiebig genug ist, um solche Fragen abschließend zu klären, oder weil der dazu oft nötige Interpretationsschritt im Editions-kontext aufgrund methodologischer Überlegungen bewusst vermieden wird.

[24]

Im Zuge unserer Datenanreicherung mussten bibliographische Einträge dementsprechend verschiedentlich kategorisiert werden, je nachdem, wie die resultierenden RDF-Entitäten zu klassifizieren waren. Abhängig von Aufwand und Machbarkeit geschah dies entweder im Rahmen der Transformation oder schon im Vorhinein als Anreicherung der Ausgangsdaten mit entsprechenden Attributen. <bibl>-Elemente aus den *Rechtsakten* wurden beispielsweise im @subtype grob in Texte, publizierte Texte, Periodika, Periodika-Ausgaben und Texte in Periodika ausdifferenziert. Diese Unterscheidungen explizieren, dass die bibliographische Angabe ganz unterschiedliche Modellierungen unter Einbezug ganz unterschiedlicher ontologischer Ebenen von Text und damit in Zusammenhang ganz unterschiedlicher Ereignisse erfordert, die wiederum unterschiedlichen Möglichkeiten der Anbindung an Personen, Orte und Zeitpunkte bieten (vgl. Abbildung 4).

[25]

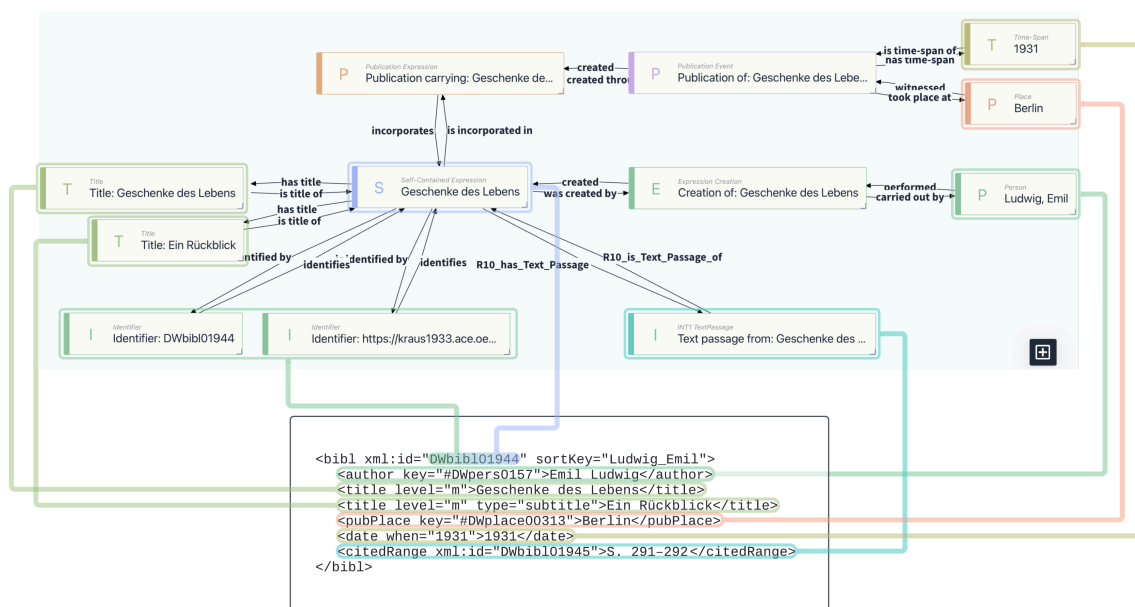


Abb. 4: Die in TEI modellierte bibliographische Angabe eines publizierten Textes und seine ereigniszentrierte RDF-Modellierung darüber (letzte als Visualisierung in ResearchSpace). Der Text selbst ›hat‹ nur Titel und Identifier, die übrigen bibliographischen Angaben liegen in seiner Peripherie, je nach Art des Textes verbunden durch in der TEI-Codierung implizite Ereignisse, aus denen der Text sowie seine publizierte Fassung hervorgegangen sind. [Grafik: Bernhard Oberreither 2025]

Eine andere Ausformung der Differenz zwischen verschiedenen Präsentationsformen bibliographischer Daten abhängig von ihrem Verwendungszweck stellten die ca. 13.000 Texte der *Fackel* dar. Diese waren nicht als solche erfasst. Was vorlag, war ein umfangreiches XML-File, das ein Inhaltsverzeichnis der gesamten 37 Jahrgänge beinhaltete. Die Ähnlichkeiten zwischen einem Inhaltsverzeichnis und einem Textverzeichnis sind groß, die Unterschiede subtil, aber folgenreich. Dazu gehören allgemeine Merkmale von Inhaltsverzeichnissen wie die Tatsache, dass ihre Einträge zwar mit Von-, nicht aber mit Bis-Seitenzahlen

[26]

ausgestattet sind: Inhaltsverzeichnisse geben Wegmarken an, »Pfade« zu einem Text, ohne damit notwendigerweise dessen Ausdehnung zu beschreiben (was hingegen ein Textverzeichnis leisten soll). Darüber hinaus mussten die Textgrenzen nicht nur seiten-, sondern absatzgenau erhoben werden, da die Personennennungen sonst nicht treffsicher den jeweiligen Texten zugeordnet werden konnten. Spezifisch für das *Fackel*-Inhaltsverzeichnis war zudem die Tatsache, dass Autor:innen nicht erfasst (sondern ggf., wenn auch unregelmäßig, Teil des Titels) waren. Wenn man darüber hinaus berücksichtigt, dass der Aufbau der *Fackel* sich im Lauf der Jahrgänge wiederholt änderte, wie es auch die inhaltliche Ausrichtung der Zeitschrift tat,²⁹ dass die Relationen zwischen den einzelnen Texten dieser Publikation über die bloße Aufeinanderfolge hinaus vielfache Verschachtelungen aufweisen, dass die Texte sich gegenseitig kommentieren und rahmen, was in typographisch nicht einheitlich markierten Hierarchien resultiert – dann wird der Aufwand bei der Überarbeitung dieses XML-Inhaltsverzeichnisses in ein Textverzeichnis deutlich. Er erschien uns insofern gerechtfertigt, als dieses Textregister im Zentrum des Transformationsprojekts SemanticKraus steht: Sowohl die *Dritte Walpurgisnacht* als auch die *Rechtsakten* beziehen sich in unzähligen Fällen auf die Texte dieser Zeitschrift.

6.3 URI-Sourcing

Die Abbildung von Daten im Semantic Web zeichnet sich, wie erwähnt, auch noch durch eine weitere Anforderung aus: Jede modellierte Entität muss durch eine URI repräsentiert werden. Das Design dieser URIs kann verschiedenen Maßgaben folgen.³⁰ Im Fall von SemanticKraus waren die folgenden Überlegungen richtungsweisend:

[27]

- Eine URI sollte bei jeder Iteration der Datentransformation identisch wiederhergestellt werden können. So kann theoretisch über verschiedene Versionen des Datensatzes hinweg die Verbindung zwischen identischen Entitäten erhalten bleiben. Zufällig generierte URIs³¹ waren damit ausgeschlossen.
- Eine URI soll möglichst auch dem bloßen Auge ein Mindestmaß an Informationen über die betreffende Entität liefern. Das erleichtert die Lesbarkeit nicht nur während der zahlreichen Datenkontrolldurchgänge, sondern auch danach in der Nutzung. Idealerweise beinhaltet dieses Mindestmaß zum Beispiel einen Hinweis auf die nächstgelegene / -zentralere Entität im Datenmodell (im Fall von SemanticKraus: auf eine Person bzw. einen Text).
- Dennoch sollte die URI keine konkreten Informationen über das angewendete Klassifizierungsschema (die Ontologie) beinhalten – etwa für den Fall des Wechsels des Referenzmodells oder der doppelten Klassifizierung einer Entität.
- Die URIs sollten möglichst ohne großen Aufwand und unabhängig vom Ausgangsdatsatz nach gleichbleibendem Muster erstellt werden.

Ausgangspunkt für das Design der URIs war stets die projekteigene @xml:id der jeweiligen Entität. Zur »best practice« in TEI gehört ohnehin die Ausstattung möglichst jeder Entität mit einer einzigartigen ID. Dabei stellt sich allerdings die Frage: Wo beginnt, wo endet der Begriff »Entität«? Eine Person, ein Ort, ein Text – hier ist die Sache einfach, hier handelt es sich jeweils um eine Entität, die jeweils eine ID erhält, dazu nach Möglichkeit Verweise auf Identifier in Normdatensätzen. Die Namen bzw. Titel dieser Personen / Orte / Texte hingegen – handelt es sich auch dabei um eigenständige Entitäten? Die entsprechenden Elemente erhalten in TEI-XML jedenfalls meist keine IDs, auf denen eine URI aufbauen könnte. Auch in RDF ist ihre Eigenständigkeit (d. h.: die Darstellung über eine eigene URI) optional, weil dort Namen und dergleichen theoretisch auch über Datenproperties (wie `rdfs:label`), also als bloße Datenstrings angegeben werden können. In Normdatensätzen erhalten sie oft keine eigenständige URI.³²

[28]

²⁹ Vgl. z. B. Stieg 2022.

³⁰ Für Grundsätzliches und Polemisches dazu siehe schon Berners-Lee 1998.

³¹ Etwa UUIDs.

³² Bsp. GND: Namen werden hier durch *blank nodes* repräsentiert, um nur im lokalen Kontext näher bestimmt werden zu können.

Das Datenmodell von SemanticKraus schrieb im Regelfall die Kreierung einer eigenständigen Entität vor. In vielen Fällen mussten also Entitäten URIs erhalten, die in der üblichen TEI-Modellierung oft implizit bleiben und / oder selbst keine ID haben. Ob eine URI nun durch die nachträgliche Anreicherung der XML-Daten um weitere IDs oder durch die mehrfache Nutzung vorhandener IDs erstellt wurde, hing vom jeweiligen Ort im Datenmodell ab: beispielsweise davon, ob auch anderswo in den Ausgangsdaten auf die betreffende Entität verwiesen wurde. Die Datumsangabe (<date>) eines Zeitungsartikels etwa repräsentiert in TEI den Zeitpunkt des Erscheinens und hat keine eigene ID. Impliziert ist hier jedoch oft die Ausgabe dieses Datums, in der zudem auch andere Zeitungstexte des Datensatzes erschienen sein können. Das RDF benötigt darum eine URI – in unserem Datenmodell: für die ›Ausgabe als Text‹ sowie die ›Ausgabe als Publikation‹, ebenso für das Publikationsereignis und zuletzt für den Zeitpunkt dieses Ereignisses. Zuerst wurde also das entsprechende <date>-Element um ein @key angereichert, dessen Wert für die implizierte Ausgabe der Zeitung stand und aus dem dann die URIs für all die genannten Entitäten generiert wurden – eine gelinde Form des ›tag abuse‹, wenn man so will. Konkret erhielt die ›Ausgabe als Text‹ die URI auf Grundlage des Attributwertes, während die der anderen genannten Entitäten durch entsprechende Erweiterungen erstellt wurden (z. B. »/published-expression«; vgl. Abbildung 5). So ergab sich die oben erwähnte angestrebte Lesbarkeit der URIs und deren Bezugnahme auf benachbarte, zentralere Entitäten ganz von selbst. Zudem wurden so diejenigen <bibl>-Elemente, die Zeitungstexte aus derselben Ausgabe repräsentierten, verbunden – über die aus dem <date> neu geschaffene Ausgabe. (Ähnlich wird bei der Modellierung des Periodikums vorgegangen, das durch den @key im <title>-Element mit @level »j« repräsentiert wird.)

[29]

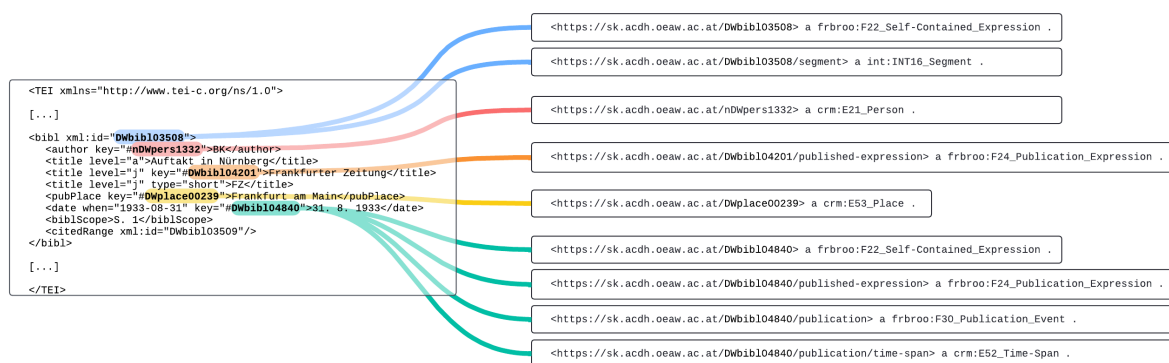


Abb. 5: Die TEI-Modellierung der bibliographischen Angabe eines Zeitungstextes und die Entsprechungen der IDs zu den URIs in der RDF-Modellierung. Gut ersichtlich ist die Generierung zahlreicher Entitäten aus dem @key in <date> zur Repräsentation der Ausgabe eines Periodikums. [Grafik: Bernhard Oberreither 2025]

7. Abschließende Überlegungen

Die bisher angesprochenen Punkte aus dem Projektkontext von SemanticKraus illustrieren auf die eine oder andere Art die eingangs etwas salopp angeführten Problemfelder bei der Transformation von TEI-XML zu RDF. Abschließend können wir das nun – etwas stärker strukturiert – zu einer kurzen Liste von Merkmalen zusammenfassen, Parametern, die für so ein Transformationsvorhaben relevant sind. Diese Liste ist weder vollständig noch allgemeingültig, sie beschränkt sich (zumeist) auf die Ebene der Datenmodelle und setzt zudem ein feinkörniges Ziel-Datenmodell (das etwa auf CIDOC CRM basiert) voraus. Einiges von dem, was nun folgt, ist zudem durch die best practice in der TEI- bzw. der Editor:innen-Community ohnehin abgedeckt; anderes scheint spezifischer für Datentransformations-Vorhaben zu sein. Zu lesen sind diese

[30]

abschließenden Beobachtungen und Überlegungen als Orientierungshilfe bei der Aufwandsabschätzung vor einer Transformation; womöglich auch als Anregung für Projekte, die sowohl TEI-XML als auch RDF publizieren und dabei Letzteres aus Ersterem gewinnen wollen.³³

Granularität der Modellierung. Der Weg vom XML zum tendenziell feinkörnigeren CIDOC-CRM-Modell umfasst in der Regel ohnehin die Erschaffung zusätzlicher Entitäten in zusätzlichen Klassen; dieser Weg ist umso kürzer – und die Transformation in der Regel umso einfacher –, je feinkörniger die Ausgangsdaten strukturiert sind. [31]

Flache vs. verschachtelte Datenstruktur. Die hierarchische Baumstruktur in XML ist nicht von sich aus geeignet zur Abbildung von n-zu-n-Relationen (beispielsweise: mehrere Texte in einer Publikation, mehrere Publikationen desselben Texts). Nicht, dass TEI nicht Möglichkeiten bieten würde, dieses Problem zu umgehen: Verschiedene Attribute können Verbindungen über die Elementhierarchie hinweg herstellen. Es ist allerdings zu vermuten, dass die hierarchische Struktur in ihrer Anwendung die Praxis der Anwender:innen prägt, sodass eine bibliographische Angabe viel eher als Hierarchie ausgehend von einem Text konzeptualisiert wird denn als Netzwerk gleichberechtigter Entitäten. Im Zusammenspiel mit traditionellen Zitierstandards führt das dazu, dass vermeintlich »untergeordnete« Entitäten – wie die Ausgabe eines Periodikums, die den in Frage stehenden Text enthält – als bloße Anhängsel betrachtet und in der Modellierung vernachlässigt werden. Hierarchischen Verhältnissen dieser Art muss in der Transformation irgendwie begegnet werden. Um dem abzuhelpen, bietet TEI selbst das <relation>-Element, in dem eine Triple-Struktur mithilfe von Pointern, die auf anderswo deklarierte Elemente verweisen, nachgeahmt werden kann. Eine Alternative als gewissermaßen »entitätenbewusste« Modellierung läge in einer verflachten Datenstruktur aufbauend auf der Verknüpfung von hierarchisch nebengeordneten Elementen mittels @corresp oder ähnlichen Attributen. Die Aufteilung beispielsweise der Angabe eines Zeitungsartikels auf drei Geschwister-(<bibl>-)Elemente, die jeweils Text, Ausgabe und Periodikum repräsentieren, würde das Hierarchieproblem entschärfen. [32]

Differenziertheit der Datenfelder. Dieser Punkt beleuchtet die Frage der Granularität noch aus einer anderen Perspektive: Während man in TEI nicht immer veranlasst ist, die im selben Element – etwa <bibl> – angelegten Informationen noch weiter auszudifferenzieren, ist gerade das eine gute Voraussetzung für Transformationsszenarien; @type und @subtype bieten hierfür ausreichend Möglichkeiten. [33]

Identifizierbarkeit der Entitäten. Für die eindeutige Identifikation von Entitäten wird in TEI-XML die Vergabe von eindeutigen Projekt-IDs empfohlen. Aus SemanticKraus-Perspektive ist zu ergänzen: je großzügiger, desto besser. Diese Maßnahme trägt zukünftigen Möglichkeiten Rechnung, auf Grundlage dieser IDs URIs zu vergeben. Darüber hinaus hat sich schon lange die Einbindung von Normdaten-Identifiern durchgesetzt. Diese erleichtern es, auch im Nachhinein die inhaltlichen Schnittmengen zwischen verschiedenen Projektdatensätzen zu erfassen und doppelt vorkommende Entitäten beispielsweise über owl:sameAs zu kennzeichnen (wie das in SemanticKraus gehandhabt wurde). Eine Orientierung an thematisch benachbarten Projekten kann bei der Vergabe von Normdaten-Identifiern darüber hinaus auch dazu anleiten, neben den weitverbreiteten ggf. auch themenspezifischere Normdatensätze zu referenzieren. In Hinsicht auf zukünftige Interoperabilität ist dies besonders ratsam (im Fall von SemanticKraus war es vor allem die PMB, deren Identifier im Zuge der Datenanreicherung ausgegeben und dann bei der Aggregation berücksichtigt wurden). [34]

³³ Eine weitere Möglichkeit für solche Fälle soll hier nicht unerwähnt bleiben: Während der Datenbestand der Ausgangsprojekte von SemanticKraus ausgesprochen divers und die individuelle Erstellung von Skripten und Hilfsfiles darum am vielversprechendsten war, liegen für anders gelagerte Fälle bereits Tools vor, die innerhalb bestimmter Grenzen automatisiert RDF aus TEI-XML generieren. Das TEI-Modell von Beginn an auf die Spezifikationen dieser Tools abzustimmen, könnte sich als große Arbeitersparnis erweisen. So zieht etwa das Tool LIFT mittels eines Sets an – anpassbaren – Python-Skripten RDF aus TEI-XMLs, solange letztere einem gewissen TEI-Modell entsprechen; das Tool ist ausführlich dokumentiert (Giovanetti / Tomasi 2022).

Referenzierbarkeit der Entitäten. Zu den Benchmarks eines RDF-Datensatzes gehören resolvende URIs.

[35]

Gerade bei der Transformation von Projektdaten, die auch eine andere Repräsentation als RDF haben, stellt sich darüber hinaus die Frage nach der Möglichkeit, aus dem RDF auf diese Repräsentation zurück zu verweisen, idealerweise via Permalinks. Die Existenz solcher Permalinks entscheidet letztlich darüber, ob die RDF-Datensätze auch Links etwa zu Einträgen in Personen- oder Textregistern des Ausgangsprojekts anbieten können. Im Fall von SemanticKraus war von Vorteil, dass ein Teil der Applikationen (konkret: *Dritte Walpurgisnacht*) noch in Arbeit war, eine andere (*Fackel online*) regelmäßige Updates erhält, sodass die nachträgliche Einführung von Permalinks für Register-, aber auch inline-Elemente keine Hürde darstellte. So konnten fast alle Entitäten der zentralen Klassen mit E42 Identifiern ausgestattet werden, deren `rdf:value` einen Link ins Ausgangsprojekt beinhaltete.

Einbezogene Editionsprojekte

Hanno Biber / Evelyn Breiteneder / Heinrich Kabas / Karlheinz Mörrth (Hg.): AAC – Austrian Academy Corpus: »Die Fackel. Herausgeber: Karl Kraus, Wien 1899-1936« (= AAC Digital Editions, 1). 2007-. [\[online\]](#)

Johannes Knüchel / Isabel Langkabel (Hg.): Rechtsakten Karl Kraus. Rechtsakten der Kanzlei Oskar Samek. Digitale Edition. Hg. auf der Grundlage von Vorarbeiten von Katharina Prager, in Zusammenarbeit mit Laura Untner u. a. Wien 2022. [\[online\]](#)

Bernhard Oberreither (Hg.): Karl Kraus, Dritte Walpurgisnacht. Historisch-kritische Edition. Publikationsschritt 1: Annotierte Lesefassung. Wien 2021. [\[online\]](#)

Bibliografische Angaben

Chrysoula Bekiari / Martin Doerr / Patrick Le Boëuf / Pat Riva (Hg.): FRBR – Object-Oriented Definition and Mapping from FRBRer, FRAD and FRISAD. Version 2.4 vom November 2015. PDF. [\[online\]](#)

Chrysoula Bekiari / George Bruseker / Erin Canning / Martin Doerr / Philippe Michon / Christian-Emil Ore / Stephen Stead / Athanasios Velios (Hg.): Definition of the CIDOC Conceptual Reference Model. Version 7.1.2. vom Juni 2022. PDF. [\[online\]](#)

Chrysoula Bekiari / Martin Doerr / Patrick Le Boëuf / Pat Riva (Hg.): LRMoo – Object-Oriented Definition and Mapping from the IFLA Library Reference Model. Version 1.0 vom April 2024. PDF. [\[online\]](#)

Tim Berners-Lee: Cool URIs Don't Change. In: W3C Style. Style Guide for Online Hypertext. 1998. HTML. [\[online\]](#)

Tim Berners-Lee / James Hendler / Ora Lassila: The Semantic Web. In: Scientific American 284 (2001), H. 5, S. 34–43. 01.05.2001. PDF. DOI: 10.1038/scientificamerican0501-34

Hanno Biber: AAC-Fackel. Das Beispiel einer digitalen Musteredition. In: Constanze Baum / Thomas Stäcker (Hg.): Grenzen und Möglichkeiten der Digital Humanities (= Zeitschrift für digitale Geisteswissenschaften / Sonderbände, 1). Wolfenbüttel 2015. 19.02.2015. DOI: [10.17175/sb001_019](#)

Susanna Burghartz / Sonia Calvi / Georg Vogeler (Hg.): Urfehdebücher der Stadt Basel – digitale Edition. Basel u. a. 2017. HTML. [\[online\]](#)

Fabio Ciotti: Digital Literary and Cultural Studies: State of the Art and Perspectives. In: Between 4 (2014), H. 8. HTML. DOI: [10.13125/2039-6597/1392](#)

Fabio Ciotti / Francesca Tomasi: Formal Ontologies, Linked Data, and TEI Semantics. In: Journal of the Text Encoding Initiative, H. 8 (September 2016 – Dezember 2017). 24.09.2016. HTML. DOI: [10.4000/jtei.1480](#)

Marilena Daquino / Martina Dello Buono / Francesca Giovannetti / Francesca Tomasi: Paolo Bufalini, Appunti (1981-1991) [Semantic Scholarly Digital Edition]. Bologna 2020. DOI: [10.6092/unibo/amsacta/6415](#)

Øyvind Eide: Ontologies, Data Modeling, and TEI. In: Journal of the Text Encoding Initiative, H. 8 (Dezember 2014 – Dezember 2015). 09.04.2015. HTML. DOI: [10.4000/jtei.1191](#)

Hans Walter Gabler: Theorizing the Digital Scholarly Edition. In: Literature Compass 7 (2010), H. 2, S. 43–56. PDF. DOI: [10.1111/j.1741-4113.2009.00675.x](#)

Francesca Giovanetti / Francesca Tomasi: Linked Data from TEI (LIFT): A Teaching Tool for TEI to Linked Data Transformation. In: Digital Humanities Quarterly 16 (2022), H. 2. HTML. [\[online\]](#)

Roland S. Kamzelak: Digitale Editionen im *semantic web*. Chancen und Grenzen von Normdaten, FRBR und RDF. In: Kristina Richts / Peter Stadler (Hg.): »Ei, dem alten Herrn zoll' ich Achtung gern«. Festschrift für Joachim Veit zum 60. Geburtstag. München 2016, S. 423–435. PDF. [\[online\]](#)

Timothy Lebo / Satya Sahoo / Deborah McGuinness: PROV-O: The PROV Ontology. 20.04.2013. HTML. [\[online\]](#)

Bernhard Oberreither: INTRO - an Intertextual Relationships Ontology for Literary Studies. Version Beta 202304. GitHub. 2023. [\[online\]](#)

Christian-Emil Ore / Øyvind Eide: TEI and Cultural Heritage Ontologies: Exchange of Information? In: Literary and Linguistic Computing 24 (2009), H. 2, S. 161–172. 04.05.2009. HTML. DOI: [10.1093/lilc/fqp010](#)

Leslie F. Sikos / Dean Philp: Provenance-Aware Knowledge Representation: A Survey of Data Models and Contextualized Knowledge Graphs. In: Data Science and Engineering (2020), H. 5, S. 293–316. DOI: [10.1007/s41019-020-00118-0](#)

Elena Spadini / Francesca Tomasi / Georg Vogeler (Hg.): Graph Data-Models and Semantic Web Technologies in Scholarly Digital Editing. Norderstedt 2021. [\[Nachweis im GVK\]](#)

Gerald Stieg: Die Fackel. In: Katharina Prager / Simon Ganahl (Hg.): Karl Kraus-Handbuch. Leben – Werk – Wirkung. Berlin 2022, S. 103–122. [\[Nachweis im GVK\]](#)

Francesca Tomasi: Vespasiano da Bisticci, Lettere. A Semantic Digital Edition. Bologna 2013. Version 3.0 vom Mai 2020. DOI: [10.6092/unibo/vespasianodabisticciletters](#)

Georg Vogeler: The »Assertive Edition«. On the Consequences of Digital Methods in Scholarly Editing for Historians. In: International Journal of Digital Humanities 1 (2019), H. 2, S. 309–322. DOI: [10.1007/s42803-019-00025-5](#)

Jörg Wettlaufer: Der nächste Schritt? Semantic Web und digitale Editionen. In: Roland S. Kamzelak / Timo Steyer (Hg.): Digitale Metamorphose. Digital Humanities und Editionswissenschaft (= Zeitschrift für digitale Geisteswissenschaften / Sonderbände, 2). Wolfenbüttel 2018. 15.03.2018. DOI: [10.17175/sb002_007](#)

Abbildungsverzeichnis

Abb. 1: Eine Visualisierung des Erscheinungsverlaufs der *Fackel*, basierend auf dem transformierten Text-Index der *Fackel online*, erstellt mittels einer ResearchSpace-Komponente und abrufbar unter der [URI der Fackel in SemanticKraus](#). [Screenshot: Bernhard Oberreither 2025]

Abb. 2: Datenprovenienz in SemanticKraus, auf Ebene des Projektdatensatzes (als Named Graph im TriG-Format innerhalb von geschweiften Klammern) sowie auf Ebene einzelner Entities (unten, die Person wird durch prov:wasDerivedFrom mit der Quelle verbunden). [Grafik: Bernhard Oberreither 2025]

Abb. 3: Das Datenmodell im Großen; ausgespart sind Zeit- und Ortsangaben zu sämtlichen Ereignissen, die als E52 Time-Span bzw. E53 Place modelliert wurden, sowie E42 Identifier, E33 Linguistic Object, E41 Appellation, rdfs:label und rdfs:value; zur Datenprovenienz siehe auch [Abbildung 2](#). [Grafik: Bernhard Oberreither 2025]

Abb. 4: Die in TEI modellierte bibliographische Angabe eines publizierten Textes und seine ereigniszentrierte RDF-Modellierung darüber (letztere als Visualisierung in ResearchSpace). Der Text selbst ›hat‹ nur Titel und Identifier, die übrigen bibliographischen Angaben liegen in seiner Peripherie, je nach Art des Textes verbunden durch in der TEI-Codierung implizite Ereignisse, aus denen der Text sowie seine publizierte Fassung hervorgegangen sind. [Grafik: Bernhard Oberreither 2025]

Abb. 5: Die TEI-Modellierung der bibliographischen Angabe eines Zeitungstextes und die Entsprechungen der IDs zu den URIs in der RDF-Modellierung. Gut ersichtlich ist die Generierung zahlreicher Entitäten aus dem @key in <date> zur Repräsentation der Ausgabe eines Periodikums. [Grafik: Bernhard Oberreither 2025]