Graphbasierte Modellierung von Faktenprovenienz als Grundlage für die Dokumentation von Zweifel und die Auflösung von Widersprüchen

Kategorie	Version 1.0
Artikel	14.08.2019

Thomas Efer

Kontakt: efer@informatik.uni-leipzig.de (mailto:efer@informatik.uni-leipzig.de)
Institution: Universität Leipzig, Institut für Informatik
GND: 1125649186 (http://d-nb.info/gnd/1125649186)
ORCID: 0000-0002-8376-3884 (https://orcid.org/0000-0002-8376-3884)

DOI: 10.17175/sb004_011 (http://dx.doi.org/10.17175/sb004_011)

Nachweis im OPAC der Herzog August Bibliothek: 1037074947 (http://opac.lbs-braunschweig.gbv.de/DB=2/XMLPRS=N/PPN?PPN=1037074947)

Erstveröffentlichung: 14.08.2019

Letzte Überprüfung aller Verweise: 14.08.2019

GND-Verschlagwortung: Geschichtswissenschaft (http://d-nb.info/gnd/4020535-6) | Graphdatenbank (http://d-nb.info/gnd/1042198020) | Konzeptionelle Modellierung (http://d-nb.info/gnd/4123555-1) | Wissensrepräsentation (http://d-nb.info/gnd/4049534-6) |

Empfohlene Zitierweise: Thomas Efer: Graphbasierte Modellierung von Faktenprovenienz als Grundlage für die Dokumentation von Zweifel und die Auflösung von Widersprüchen. In: Die Modellierung des Zweifels – Schlüsselideen und -konzepte zur graphbasierten Modellierung von Unsicherheiten. Hg. von Andreas Kuczera / Thorsten Wübbena / Thomas Kollatz. Wolfenbüttel 2019. (= Zeitschrift für digitale Geisteswissenschaften / Sonderbände, 4) text/html Format. DOI: 10.17175/sb004_011 (http://dx.doi.org/10.17175/sb004_011)

Abstract

Ziel dieses Beitrags ist es, die Wichtigkeit einer nachvollziehbaren Herkunft von Aussagen in Wissensbasen der Digitalen Geisteswissenschaften herauszustellen. Neben der Vorstellung genereller Aspekte der Aussagenmodellierung auf abstrakter und beispielgeleiteter Ebene wird das Konzept einer *Faktenprovenienz* entwickelt und in Aussagemodelle integriert. Auf Basis von *Provenienzketten* wird demonstriert, wie eine im System erfasste Herkunftsdokumentation von Einzelaussagen zur Behandlung von Widersprüchen und der Reduzierung von Unsicherheit genutzt werden kann.

This contribution aims to demonstrate the importance of traceable provenance information within knowledge bases in the Digital Humanities. Besides presenting rather general aspects of how to model statements in an abstract and in an exemplary manner, the concept of *fact provenance* is introduced and integrated with statement expression models. Using so-called *provenance chains*, it is shown how provenance information that is captured within an information system can be utilized to handle contradictions and reduce the overall uncertainty of the knowledge base.

1. Motivation und Einführung

Die Geistes- und Sozialwissenschaften befinden sich gegenwärtig unter dem Eindruck und innerhalb der Dynamik einer sich rapide digitalisierenden Gesellschaft vor zahlreichen neuen Herausforderungen.

Wird von *den* Geistes- und Sozialwissenschaften gesprochen, schwingt die gewagte Grundannahme einer Kohärenz der darunter subsumierten wissenschaftlichen Akteure und Aktivitäten mit. Diese lässt sich bei genauerer Betrachtung höchstens für kleine Teilbereiche einzelner Fachrichtungen oder gewisse interdisziplinäre Querschnittsthemen rechtfertigen. Während eine solche bewusst gleichmachende Abstraktion sehr hilfreich für die Initiierung interdisziplinärer Arbeiten ist und darüber hinaus gewinnbringend für die Institutionalisierung und Lobbyarbeit genutzt werden kann, so kann sie problematisch werden: Oft zeigt sich dann erst spät in der Praxis, wie groß und zum Teil unüberwindbar die Differenzen in inhaltlicher und methodischer Dimension tatsächlich sind. Die Forschungsziele der einzelnen Disziplinen und konkreten Einzelforschungsarbeiten verteilen sich (entsprechend dem akademischen Selbstverständnis der Fächer und Forscher*innen) meist sehr breit auf der Skala zwischen dem Erreichen eines konkreten Erkenntnisgewinns und der davon angekoppelten Interpretation bestimmter Sachverhalte in neuen, aktuellen Kontexten. Entsprechend zeigen sich auch deutliche Unterschiede im Umgang mit Quellenmaterial, sekundärer Fachliteratur, vorherrschenden Fächertraditionen, disziplinären Strömungen und einzelnen fachlichen Aussagen im Forschungsprozess – insbesondere im Umgang mit allen Arten von »Daten«.

Im Rahmen der Digital Humanities wird die Unterstützung jeglicher geistes- und sozialwissenschaftlicher Forschungstätigkeit durch »generische« digitale Werkzeuge für diese wichtigen datenzentrierten Forschungstätigkeiten angestrebt. Dafür müssen geeignete Abstraktionen und Verallgemeinerungen gefunden werden. Bei der Überführung bisheriger Forschungstätigkeiten in die digitale Welt werden somit auch digitale Modelle für Daten benötigt, die den Grundbedürfnissen der Disziplinen gerecht werden. Da sich unbestritten ein Großteil der Forschungsarbeiten im Einzugsbereich der DH mit dem Sichten, Sammeln, Erschließen, Bewerten und Verknüpfen von Befunden aus Quellenmaterial beschäftigt (vgl. dazu die Taxonomy of Digital Research Activities in the Humanities (http://tadirah.dariah.eu).[1]), liegt es nahe, in diesem Bereich die gemeinsamen Anforderungen zu ergründen und dafür entsprechende digitale Unterstützung bereitzustellen.

Während die meisten Forschungsdatenbanken ausschließlich die Endprodukte dieser Arbeitsschritte beinhalten – welches bereits eine große Unterstützung darauf aufbauender weiterführender Forschung sein kann –, soll im Rahmen dieses Beitrags eine Herangehensweise vorgestellt werden,

mit deren Hilfe bei einer Weiter- und Nachnutzung tiefer in die Entstehungszusammenhänge dieses Faktenwissens hineingesehen werden kann.

Mit Fakten sind in diesem Fall vom Menschen »gemachte«, als plausibel angesehene Aussagen gemeint – für die Zwecke digitaler Datenbanken speziell solche Aussagen, die sich formalisiert abbilden lassen. Der Faktenbegriff soll hier noch nicht mit absoluten Kategorien wie einem »tatsächlichen Wahrheitsgehalt« gleichgesetzt werden. Auch bei allen als wahr angenommenen Gegebenheiten handelt es sich (speziell im Kontext der Wissenschaft) schließlich immer nur um »Tatsachenbehauptungen«, die analytisch und interpretativ im Forschungsprozess jederzeit angezweifelt, ausgeklammert, zurückgewiesen oder aber übernommen und kombiniert, und damit in weitergreifende Aussagen überführt werden können.

Um die Fakten qualifiziert einschätzen zu können, ist in erster Linie die Kenntnis ihrer genauen Herkunft von Interesse. Die »Entstehungsumstände« des Faktenwissens und die »Überlieferungshistorie« sind dabei die wesentlichen Bestandteile der in diesem Beitrag *Faktenprovenienz* genannten Herkunftsinformationen. Nicht immer sind diese Komponenten in ausreichender Form bekannt oder belegt. Faktenprovenienz ist inhaltlich verwandt, aber bei weitem nicht deckungsgleich zur sogenannten *Datenprovenienz* (wie etwa bei Simmhan et al. beschrieben[2]). Datenprovenienz (auch *Data- Lineage* genannt) beschreibt Anforderungen und Verfahren, um die Transformationen von digital vorliegenden Quelldaten hin zu den in einem System vorgehaltenen oder aus ihm exportierten Enddaten maschinenlesbar zu dokumentieren. Als Vorteile werden verbesserte oder vereinfachte Qualitätssicherung, Attribution von Rechteinhabern und Reproduzierbarkeit angesehen. Ähnliche Ziele verfolgt auch die Berücksichtigung von Faktenprovenienz in Forschungsdatenbanken. Ihr Fokus liegt jedoch nicht allein auf bereits digital vorliegenden Rohdaten, sondern auf der kompletten Historie der enthaltenen Fakten. Nur durch diese erweiterte Sichtweise kann die Arbeit mit Fakten im Forschungsprozess ganzheitlich unterstützt werden.

Bei den digital in einer Datenbank zu erfassenden Aussagen sind im Kontext von fachfragen-orientierten DH-Methoden in erster Linie Aussagen zu Gegebenheiten der Fachdomäne von Interesse und weniger technische oder organisatorische Metainformationen. Im Umgang mit Forschungsdaten der Fachdomänen stellen sich für Datenbanksysteme dabei sehr grundsätzliche Fragen: Welche Daten sollen erfasst werden? Welche Arten von Aussagen sollen im System abbildbar sein? Wie können erfasste Einträge zueinander in Beziehung gesetzt werden (bei der Eingabe und bei der Abfrage)? Wie können externe Daten übernommen und eigene Daten exportiert werden? Für diese Fragen bieten

generische Werkzeuge, Repositoriums- und Datenbanksysteme in der Regel akzeptable bis sehr gute Antworten. Die Provenienz der so in den Datenbanken kodierten Aussagen kann jedoch im Allgemeinen nicht abgebildet werden! Dieser Fehlstelle widmet sich der vorliegende Beitrag.

Für die gemeinsame Speicherung und Abfrage von Fakten und ihrer Provenienz werden Systeme benötigt, mit denen eine sehr flexible Datenmodellierung möglich ist. Ohne Abwertung möglicher Alternativtechnologien soll im Folgenden eine Festlegung auf eine bestimmte Gruppe von dafür geeigneten Systemen getroffen werden.

2. Technologischer Rahmen

Dieser Beitrag bezieht sich auf die Anwendung von Graphdatenbanksystemen, speziell solchen, die das Property-Graph-Datenmodell umsetzen.[3] In diesem Datenmodell stehen für die Repräsentation der zu erfassenden Daten die folgenden einfachen Bausteine zur Verfügung, durch deren Kombination sich komplexere Sachverhalte abbilden lassen:

Die so genannten *Knoten* (*nodes vertices*) können als Repräsentanten für Aussagegegenstände angesehen werden. Sie lassen sich zählen, auflisten und über intern vergebene IDs einzeln adressieren.

Daneben stehen mit den so genannten *Kanten* (*edges*) Konstrukte zur Verfügung, mit denen genau zwei Knoten miteinander verbunden werden können. Kanten unterscheiden dabei zwischen Start- und Zielknoten, so dass die Verbindung »gerichtet« ist. Zwischen beliebigen Knotenpaaren können beliebig viele Kanten in beliebiger Richtung existieren. Den jeweiligen Zielknoten werden die Kanten dabei als »eingehende« Kanten, dem Startknoten als »ausgehende« Kanten zugeordnet. Kanten besitzen genau ein sogenanntes *Label*. Dieses ist eine kategoriale Größe, die verwendet wird, um verschiedene (semantisch oder technisch zu unterscheidende) Arten von Beziehungen zwischen den mit einer Kante verbundenen Knoten auszudrücken. Auch Kanten besitzen interne IDs und lassen sich zählen und auflisten. Darüber hinaus ist es möglich, sie nach ihrem Label zu filtern. Das Label hat dabei üblicherweise eine textuelle Repräsentation, wie ›IST_VATER_VON‹ oder ›FOLLOWS‹. Damit ist es möglich und üblich, die Kante (als technische Verbindung von Knoten) auch als Abbild einer zu modellierenden Beziehung zwischen zwei Aussagegegenständen anzusehen.

Schließlich existieren mit den sogenannten *Properties* noch Konstrukte, mit denen sich »Eigenschaften« von Knoten und Kanten notieren lassen. Diese Eigenschaften sind mit maschinenlesbaren Werten befüllt. Erst dadurch ergibt sich der Datenbankcharakter des Systems. Sie werden in Form von *Schlüssel-Wert-Paaren* (*key-value pairs*) gespeichert. Diese bestehen aus einem Schlüssel, also einer kategorialen Größe, die den in der Property notierten Eigenschaftstyp bestimmt,

und einem Wert, welcher in der Regel in einem primitiven Datentyp, wie etwa *Ganzzahl* oder *Zeichenkette*, vorliegt. Die Property-Schlüssel besitzen (genau wie die Kantenlabels) eine textuelle Repräsentation. Alle Schlüssel-Wert-Paare sind jeweils genau einem Knoten oder genau einer Kante zugewiesen. Losgelöst von diesen Konstrukten können sie nicht existieren.

Damit ist die Palette der verwendbaren Modellierungskonstrukte auch schon vollständig. Im nächsten Unterabschnitt des Artikels wird noch genauer auf die damit umsetzbare Datenmodellierung eingegangen. In der Praxis existieren zahlreiche Nuancen dieses grundlegenden Datenmodells. In der populären und systemübergreifend genutzten Programmierschnittstelle für Graphdatenbanken in Java namens *Tinkerpop* wird erst ab Version 3 erlaubt, einem Knoten oder einer Kante mehrere Properties mit gleichem Schlüssel beizufügen. Damit einher geht auch die Möglichkeit, Properties für Properties zu definieren (welche intern von den Systemen jedoch oft nur mittels der oben beschriebenen Basiskonstrukte »virtuell« umgesetzt wird). Ebenso gibt es unterschiedliche Ansichten darüber, ob im Property-Graph-Modell auch Knoten über ein *Label* verfügen sollten, also einen Typen haben können (oder müssen), wie z. B. im populärsten System, Neo4j, üblich. Da im Rahmen dieses Beitrags nicht allzu tiefgehend auf direkter technischer Ebene mit den Modellierungskonstrukten gearbeitet wird, sollen diese und weitere Feinheiten an dieser Stelle jedoch nicht weiter eruiert werden.

Graphdatenbanken weisen abseits der Spezialisierung auf dieses Datenmodell viele Gemeinsamkeiten mit klassischen, relationalen Systemen auf. Auch sie bewegen sich im Segment der Echtzeitabfragen im so genannten *Online Transaction Processing* (OLTP), unterstützen strukturierte Abfragesprachen und oft auch Transaktionen, also eine Persistierung der Änderungen am Datenbestand nach einem Alles-oder-nichts-Prinzip im logischen Einbenutzerbetrieb. Im OLTP-Betrieb ist die Geschwindigkeit der Beantwortung einer Abfrage von großer Bedeutung. Abfragen in Graphdatenbanksystemen liefern meist Mengen von Knoten zurück, welche entweder direkt über die Werte ihrer Properties ausgewählt werden oder aber indirekt durch das Überspringen von Kanten von einem bereits ausgewählten Knoten aus erreicht werden können. Die Nutzung der Kanten zur Navigation innerhalb des durch sie aufgespannten Knotennetzwerks wird *Traversierung* genannt. Diese Traversierung kann in Graphdatenbanken effizient über sehr viele Zwischenstationen geschehen, wodurch sich meist ein erheblicher Geschwindigkeitsvorteil gegenüber relationalen Datenbanken und deren Tabellenverknüpfung über *Joins* ergibt.[4]

Diesen Geschwindigkeitsvorteil können die Systeme allerdings nur geltend machen, wenn zur Beantwortung der Anfrage ein kleiner, »lokaler« Ausschnitt der Datenbankeinträge (Knoten und Kanten) »besucht« wird. In der Praxis zeigt sich, dass sich nicht wenige Probleme in den

geisteswissenschaftlichen Disziplinen durch begrenzte Umkreissuchen um »interessante« Einträge herum ausdrücken lassen. Oft ist die Analyse aller mit einem Objekt verknüpften anderen Objekte interessanter und zielführender, als die aller nichtverknüpften.[5]

Für »globale« Auswertungen, welche den kompletten Datenbestand oder große Teile davon traversieren, beispielsweise um statistische Kennzahlen zu ermitteln, können keine schnellen Antwortzeiten erwartet werden. Ganz im Gegenteil kann es dazu kommen, dass statt einer Antwort sogar ein durch die Überschreitung vorhandener Systemressourcen (meist des Arbeitsspeichers) hervorgerufener Fehler vermeldet wird. Hierfür werden künftig verstärkt Lösungen im Umfeld der Graphentechnologie benötigt, welche neben OLTP auch ein *Online Analytical Processing* unterstützen und die Systemressourcen unter Aufgabe der Echtzeit-Abfragbarkeit effektiver für Analysezwecke nutzen können. Idealerweise verwenden solche Systeme dasselbe Datenmodell (und dieselben Abfrageschnittstellen und -sprachen) wie Graphdatenbanken und sind ggf. sogar fest mit ihnen verbunden oder in sie integriert. Daher wird im Folgenden nicht weiter auf diese Unterscheidung eingegangen, auch wenn sie für die konkrete Umsetzung in der Praxis sehr relevant ist.

Neben dieser technischen Sichtweise soll die Graphdatenbank hier hauptsächlich als ein Wissensspeicher fungieren. Ähnlich einer *Wissensbasis* (*Knowledge Base*) werden darin einzelne Aussagen erfasst, kontextualisiert und für eine strukturierte Abfrage vorgehalten. Während in semantischen Netzwerken vorwiegend Semantic-Web-Technologien zur Anwendung kommen, wird im Folgenden auf maschinenlesbare Semantik auf Schema-Ebene und auf die Möglichkeiten, ein automatisches logisches Schließen (*Reasoning*) durchzuführen, verzichtet. Dies geschieht mit Hinblick auf Geschwindigkeitsaspekte für die Speicherung, Indizierung und Abfrage großer Datenmengen und um im Speichersystem technische Schemainformationen und Stamm- bzw. Instanzdaten nicht zu vermischen. Eine Überführung von Property Graphs in eine Semantic-Web-Repräsentation ist jedoch jederzeit problemlos möglich, wie beispielsweise die Arbeit von Hartig aus dem Jahr 2014 zeigt.[6] Fehlende technische Möglichkeiten für das generische Reasoning lassen sich in der wissenschaftlichen Anwendungsdomäne zudem leicht verschmerzen, da die so erzielten abgeleiteten Aussagen meist deutlich unspezifischer sind als solche, die durch zielgerichtete Datenbankabfragen mit Kenntnis der Fachdomäne erzielt werden können. Ein solcher Abfragemodus erlaubt einen flexibleren Umgang mit Forschungshypothesen, welche über die nötigen Kontexte (und Konfidenzen), die für eine Folgerbarkeit auf Faktenebene entscheidend sind. Parallel dazu wird ein semantisches Modell des Wissensspeichers in der Regel anwendungsspezifisch durch Festlegung von Geschäftslogik und Interaktionsmöglichkeiten in der Recherchesoftware abgedeckt. Basis aller Folgerung und Interpretation der Daten muss im Kontext von Wissensbasen allgemein und im Kontext

der digitalen Geisteswissenschaften im Besonderen die so genannte *Open World Assumption* sein, wie sie etwa von Moore und Pham beschrieben[7] und genauer analysiert wird. Sie sagt aus, dass nicht enthaltene Aussagen nicht zwingend unwahr sind. Ihr Wahrheitsgehalt ist somit ›unbekannt‹.

Mit den hier umrissenen modelltechnischen Möglichkeiten und Einschränkungen lässt sich nun ein System zur Abbildung von Aussagen entwickeln, welches die benötigten Ankerpunkte und Konstrukte für die Annotation von Faktenprovenienz bereitstellt.

3. Möglichkeiten der Modellierung von Aussagen

Die folgenden Beispiele orientieren sich am zu diesem Beitrag gehörigen, gleichnamigen Vortrag auf der Tagung *Graphentechnologien 2018 – Die Modellierung des Zweifels* in Mainz. Das Ziel hinter der Wahl der eher »informellen« Inhalte ist es, einen vergleichsweise abstrakten und dennoch mit nachvollziehbaren Gegebenheiten verknüpften Zugang zur Aussagenmodellierung zu erreichen – zunächst losgelöst von der wissenschaftlichen Praxis.

Als Ausgangspunkt der Modellierung soll hier zunächst die natürlichsprachlich vorliegende Aussage ›Peter isst eine Banane‹ betrachtet werden. Diese wirft freilich für sich genommen deutlich mehr Fragen auf als sie beantwortet: Welcher Peter ist gemeint? Kann dieser genauer charakterisiert oder gar eindeutig identifiziert werden? Ist es möglich, die mit unbestimmtem Artikel bedachte Banane irgendwie von anderen Bananen zu unterscheiden (oder ist ihr einziges Alleinstellungsmerkmal, von Peter gegessen zu werden)? Was ist der räumlich-zeitliche Kontext der Aussage?

Selbst ohne Kenntnis dieser zusätzlichen Details ist es möglich, die Aussage als einzelnen Fakt in der Datenbank abzubilden. Dazu wird ein Knoten als Repräsentant für ›Peter‹ und ein weiterer Knoten für die in der Aussage referenzierte ›Banane‹ erzeugt. Zwischen beiden wird eine als ›isst‹ typisierte Kante ausgehend von ›Peter‹ eingefügt. Damit ist in der Datenbank immerhin hinterlegt, dass ein Vorgang des ›Essens‹ (Kantentyp ›isst‹) dokumentiert ist.

Diesem Vorgang können nun alle denkbaren Kontextinformationen angefügt werden, beispielsweise der Zeitpunkt, welcher in Form einer Kantenproperty erfasst werden kann. An dem Knoten, der für Peter steht, lässt sich in jedem Fall sein Name als Property erfassen. Eventuelle darüber hinaus bekannte Fakten, wie sein Alter oder die Farbe der Banane ließen sich ebenso als Knotenproperties erfassen. Ebenso kann das implizit oder explizit gegebene Hintergrundwissen, dass Peter ›eine Person‹ und die Banane ›ein Gegenstand‹ ist, als Typenzuweisung für die Knoten abgebildet werden. Dies kann entweder durch die Verwendung von Knotenlabels oder die Erstellung von Repräsentanten für den jeweiligen *Typ* und die Verwendung einer geeignet typisierten Kante (etwa ›is_a‹ oder ›typeof‹ genannt) realisiert werden.

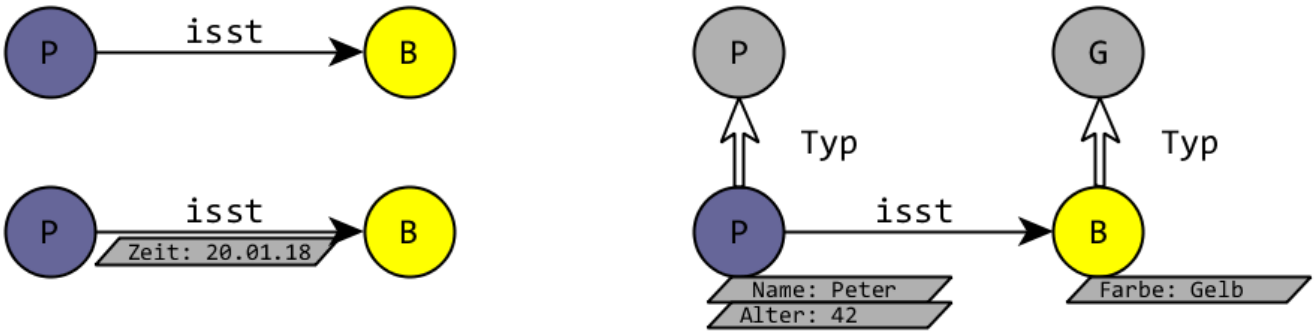

Abb. 1: Modellierung einfacher Aussagen im Graphen. [Efer 2019.]
(http://www.zfdg.de/sites/default/files/medien/faktenprovenienz_2019_001.png)

Abbildung 1 zeigt visuelle Repräsentationen aller bisher verwendeten Modellierungsvarianten für die unterschiedlich stark kontextualisierte Aussage in einer (hoffentlich) intuitiven, jedoch nicht formalisierten oder standardisierten graphischen Notation.

Einfache Aussagen nach diesem Muster lassen sich stets gleich erfassen. Ist klar, dass es sich bei der handelnden Person in einer zweiten Aussage um denselben Peter wie zuvor handelt, so sollte für diesen kein neuer Repräsentant eingefügt, sondern der vorhandene Knoten weiter genutzt werden. Ob innerhalb der Datenbank zur plausiblen Abbildung einer weiteren Aussage dieselbe Banane ein weiteres Mal gegessen werden kann (von Peter oder einer anderen Person), muss dann bereits sehr individuell entschieden werden (indem etwa ›isst‹ auch die Bedeutung ›isst Teile von‹ umfasst). Solche semantischen Probleme sollen hier jedoch zunächst ausgeklammert werden. Stattdessen gilt es, einen weiteren, sehr häufigen und dabei durchaus problematischen Typ von Aussage zu untersuchen:

Lautet die abzubildende Aussage nun nämlich: ›Jürgen sagt, dass Peter eine Banane isst‹, kann der bisherige Ansatz einer ausschließlich direkten Nutzung von Graphkonstrukten nicht mehr verwendet werden. Für Peter, die Banane und die ›isst‹-Kante zwischen ihnen kann noch der selbe Modellierungsansatz wie oben gewählt werden. Auch Jürgen kann einen Knoten als Repräsentanten erhalten. Doch wohin zeigt eine von diesem ausgehende ›sagt-Kante‹? Diese sollte auf die komplette Aussage von oben verweisen. Technisch kann sie jedoch nicht auf ›zwei Knoten und eine Kante‹ verweisen, ebenso wenig auf ›die Kante, an der beide Knoten anliegen‹. Sie kann nur zwei Knoten miteinander verbinden. Es wird also ein Knoten als Repräsentant der obigen »geschachtelten« Aussage benötigt. Die Erstellung eines solchen Knotens wird im Umfeld des Semantic Web

Reifizierung genannt. In Graphdatenbanksystemen ist dafür kein eigenes Modellierungskonstrukt vorgesehen. Das Prinzip lässt sich dennoch anwenden, indem die im Semantic Web übliche Verwendung von *Tripeln* aus Subjekt, Prädikat und Objekt als Abstraktion verwendet wird.

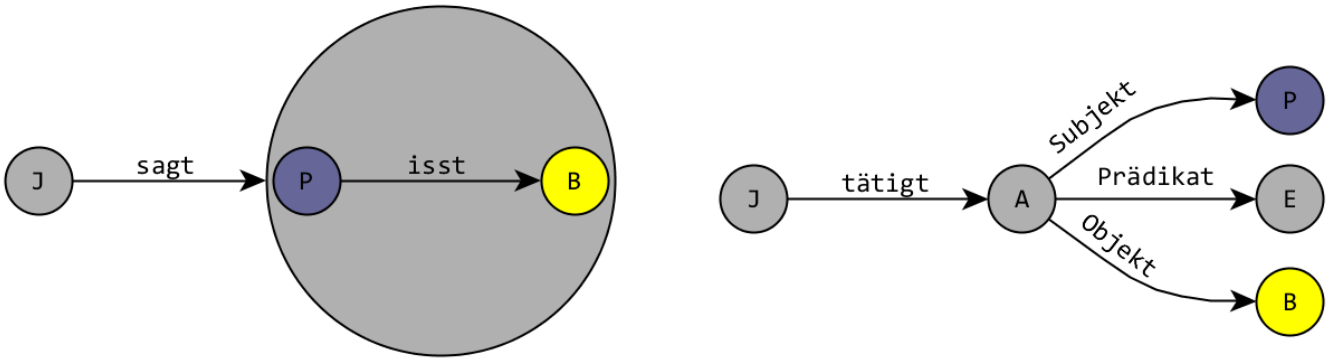

Abb. 2: Modellierung von Aussagen über Aussagen mittels Reifizierung. [Efer 2019.]
(http://www.zfdg.de/sites/default/files/medien/faktenprovenienz_2019_002.png)

Abbildung 2 zeigt, wie ein separater Knoten, welcher für die Aussage (A) steht, mit dem Repräsentanten von Peter, dem der Banane, sowie einem separaten Knoten, der für ›den Vorgang des Essens‹ (E) steht, über entsprechende Kanten (Subjekt, Objekt und Prädikat) verknüpft ist. Nun kann das »Tätigen« der so formell beschriebenen Aussage durch das Verbinden des Repräsentanten für Jürgen mit diesem Aussageknoten abgebildet werden. Eine solche Modellierungsweise ist sehr populär. Da neben dem Tripel Subjekt-Prädikat-Objekt (SPO) mit dem Repräsentanten für die Aussage selbst nun ein vierter Baustein für jede elementare Aussage existiert, werden Systeme, die sich auf die effiziente Speicherung (und bedingt auch Abfrage) von Daten in einem solchen Datenmodell spezialisiert haben, auch *Quadruple Stores* genannt. Das Property-Graph-Modell kann diesen Ansatz sehr effizient und elegant unterstützen. Das Vokabular wird hierbei anstatt im Graphenschema (mit dem Kantentyp ›isst‹) nun als Teil der Daten vorgehalten (mit einem eigenen Knoten als Repräsentant von ›Essen‹).

Nicht alle Aussagen sind sinnvoll mit einem einfachen SPO-Muster erfassbar. Wird etwa der Satz ›Peter kauft im Supermarkt eine Banane.‹ betrachtet, so wird leicht begreiflich, dass ein Verknüpfen von Peter mit der Banane mittels ›kauft_im_Supermarkt‹-Kante nicht sehr weitsichtig ist. Das Vorgehen impliziert, dass künftig auch ›kauft_auf_dem_Wochenmarkt‹- oder ›kauft_im_Internet‹-Kanten benötigt werden, welche nach dem Property-Graph-Modell noch nicht einmal komfortabel in eine Hierarchie von Kantenlabels einsortiert werden können.

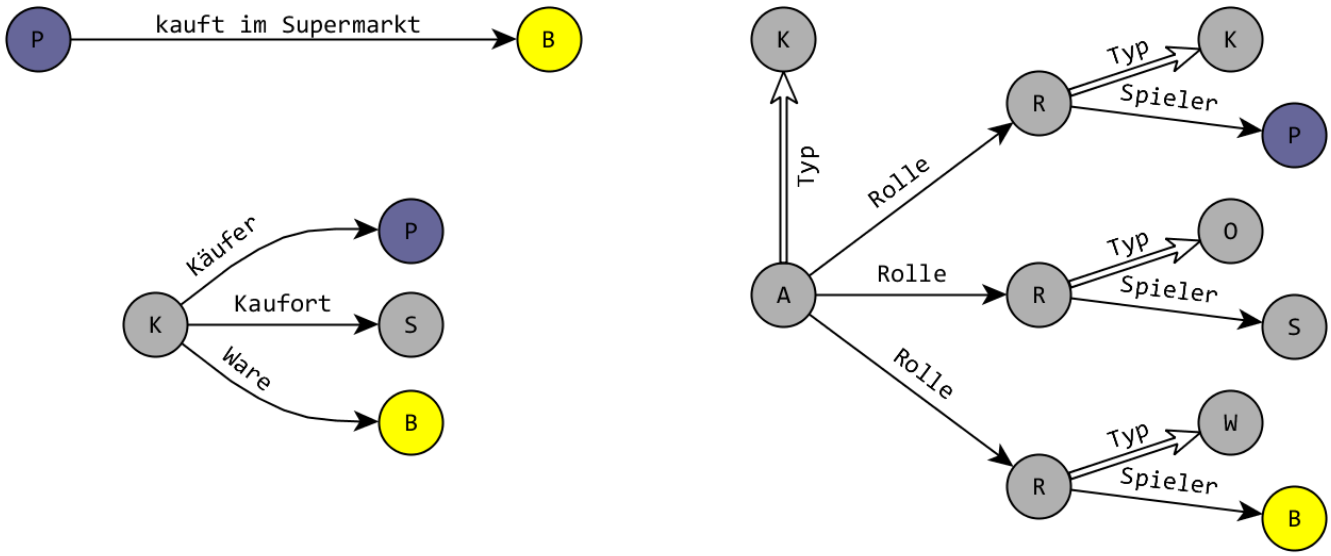

Abb. 3: Modellierung von mehrgliedrigen Assoziationen. [Efer 2019.]
(http://www.zfdg.de/sites/default/files/medien/faktenprovenienz_2019_003.png)

Abbildung 3 zeigt zusätzlich zur so gebildeten Repräsentationsform zwei weitere Ansätze: Zum einen ist es möglich, für die einzelnen Aussage-Teile Knoten zu erzeugen (Peter, der Supermarkt, die Banane) und sie einem »abstrakten« Knoten, der für die »Kaufhandlung« (K) steht. zuzuordnen. Hierbei können nun beliebig viele Aussagegegenstände in Relation gesetzt werden (nicht mehr nur zwei). Dieses Vorgehen ähnelt dabei der bekannten Reifizierung, mit dem Unterschied, dass nun spezielle Kantentypen, ›Käufer‹, ›Kaufort‹ und ›Ware‹, benötigt werden, um die Repräsentanten qualifiziert mit der Kaufhandlung zu verbinden.

Auch solche unterstützenden Kantentypen werden mit der Aufnahme weiterer Aussagen in die Datenbank perspektivisch in immer neuen Ausprägungsformen vorkommen. Sie werden dringend benötigt, um die *Rolle* der an der Aussage beteiligten Knoten zu definieren. Ohne eine solche Unterscheidung könnte die modellierte Aussage ebenso gut als ›Ein Supermarkt kauft in der Banane Peter.‹ interpretiert werden. Sollen diese Rollendefinitionen nicht Teil des Graphenschemas werden, so kann als weitere Abstraktion das Assoziationsmodell der semantischen Technologie *Topic Maps* (ISO 13250) verwendet werden. Abbildung 3 zeigt auf der rechten Seite die dafür angelegte Struktur. In dieser existiert ein generischer abstrakter Knoten (A) für die *Assoziation* (also die Gesamtaussage) sowie einzelne generische abstrakte Knoten für alle Rollen (R). Diesen abstrakten Knoten ist ein Typ zugeordnet, hier ›Kaufhandlung‹ als Assoziationstyp sowie ›Käufer‹, ›Kaufort‹ und ›Ware‹ als Rollentypen. Die Rollen-Knoten sind darüber hinaus mit ihren *Rollenspielern* , also den konkreten Aussagegegenständen verbunden. So ist das komplette Domänenvokabular innerhalb der

Instanzdaten abgebildet. Die Zahl der Kantentypen bleibt auch bei Erweiterung um neue Aussagentypen konstant. Dieses Prinzip der *abstrakten Stellvertreterknoten* wird als *Indirektion* im nächsten Unterabschnitt noch näher vorgestellt.

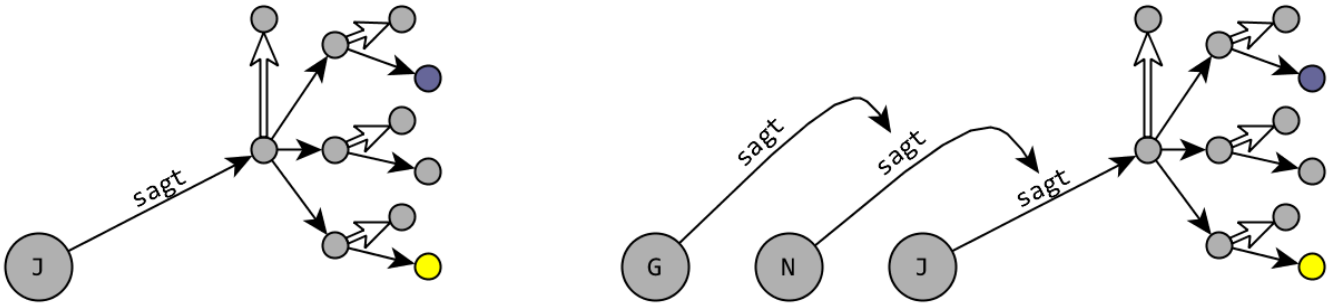

Abb. 4: Problematik rekursiver Reifizierung mittels Assoziationsmodellen. [Efer 2019.]
(http://www.zfdg.de/sites/default/files/medien/faktenprovenienz_2019_004.png)

Abbildung 4 zeigt, dass auch referenzierte Aussagen (im Stil von ›Jürgen sagt, dass…‹) mit dem Assoziationsmodell problemlos aufgegriffen werden können. Allerdings ist es nach wie vor nicht möglich, eine solche »indirekte Rede» selbst zu referenzieren. Beginnt ein Satz also mit ›Laut Günther hat Nora erzählt, Jürgen würde behaupten, dass Peter…‹ so würde für jede Ebene der Meta-Aussage wieder eine eigene Assoziation gebildet werden müssen, wodurch das eigentliche Domänenwissen (über Kauf- und Ernährungsgewohnheiten von Menschen) in der Datenbank sehr stark von technischen Modellierungskonstrukten überschattet würde.

Der nächste Unterabschnitt baut nun auf den eingangs besprochenen Überlegungen zur Notwendigkeit der Modellierung von Faktenprovenienz und den bis hierhin vorgestellten Modellierungsansätzen auf, um ein einfaches System zu entwickeln, in dem sich lange Folgen von Überlieferung einzelner Fakten effizient abbilden und für die Forschung nutzbar machen lassen.

4. Provenienzketten und Indirektion

Für die weitere Veranschaulichung soll von hier an (ebenfalls analog zum Vortrag) ein neues, eventuell forschungsnäheres Beispiel eingeführt werden. Es soll der übliche Fall der Erfassung historischer Aussagen aus digitalen Versionen älterer Quellen betrachtet werden (ohne hierbei allzu realistische Fach- und Forschungsfragen zu beachten).

Bei der Durchsicht einer ins Englische übersetzten Internetversion von Herodots Historien könnte beispielsweise die folgende Textstelle den Autor dieses Artikels zur Aufnahme eines neuen Fakts in eine Forschungsdatenbank zur Beziehung von Völkern und Stämmen im antiken Mittelmeerraum animieren:

> »[…] they further had a huge vase made in bronze, […] which they sent to Croesus as a return for his presents to them. The vase, however, never reached Sardis. […] The Lacedaemonian story is that when it reached Samos, on its way towards Sardis, the Samians having knowledge of it, put to sea in their ships of war and made it their prize.«[8]

Es könnten sehr einfach Knoten für eine Vase, die Spartaner (Lakedaimonier), die Samier und Krösus (bzw. allgemein die Lydier in Sardis) erstellt werden. Dazu gesellen sich eine Schenkungs-Assoziation und eine Raub-Assoziation. Für letztere ist es besonders interessant, die Faktenprovenienz zu kennen. Denn es ist bei weitem nicht sicher, dass der so dokumentierte Fakt der historischen Wahrheit entspricht – und latenter Zweifel daran ist bereits in der Textquelle enthalten.

Für eine qualifizierte Angabe der Faktenprovenienz sollten an dieser Stelle alle bekannten Quellen und Überlieferungsschritte identifiziert, abgegrenzt und geordnet erfasst werden, von den ältesten Belegen bis hin zur letzten Instanz vor der Eingabe in die Datenbank.

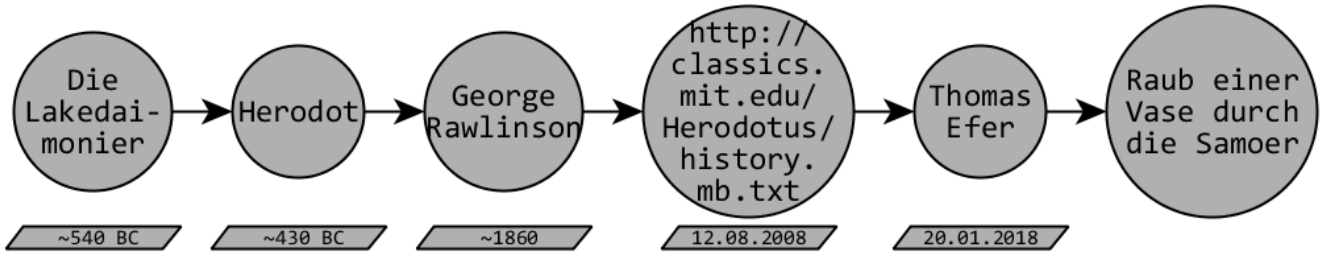

Abb. 5: Beispielhafte Provenienzkette. [Efer 2019.]
(http://www.zfdg.de/sites/default/files/medien/faktenprovenienz_2019_005.png)

Abbildung 5 zeigt, wie sich daraus eine Kette von Knoten in der Datenbank ergibt, welche über einen geeigneten Kantentyp miteinander verbunden sind, bis zur Aussage (bzw. dem dafürstehenden Assoziationsknoten). Die einzelnen Zeitpunkte der Überlieferung (falls bekannt), können direkt an dem Überlieferungsknoten notiert werden. Anders als bei der indirekten Wiedergabe (›Laut Günther hat Nora erzählt, …‹) steht hier das von der Originalquelle am weitesten entfernte Element am

nächsten an der Aussage. Denn nur über dieses Element »erfährt« das System von der Existenz der anderen Kettenglieder. Das entspricht der typischen Angabe von Provenienzen bei Objekten, wo auch die »letzten« Besitzer*innen und Aufenthaltsorte zuerst genannt werden.

Diese Überlieferungskette gibt dem eigentlichen Fakt wertvollen Kontext: Er sollte nur dann in der Forschung direkte Beachtung finden, wenn davon auszugehen ist, dass er sich nicht aufgrund von Fehlinterpretation des Eingebenden in die Datenbank, inkonsistenter Textwiedergabe in der Internetquelle, falscher Übersetzung durch George Rawlinson, Fiktion des griechischen Historikers oder falscher Bezichtigung durch die Spartaner entstanden ist. Die Provenienzkette macht diese logischen und quellenkritisch relevanten Abhängigkeiten für die Forschung erstmals explizit und ihre (zum Teil auch mangelhafte) Berücksichtigung im Forschungsprozess für externe Betrachter*innen endlich transparent nachvollziehbar.

Aus Modellierungssicht sollte das bisher sehr einfache Konstrukt der Provenienzkette noch etwas verfeinert werden. Zur Demonstration der Notwendigkeit kann eine zufällige weitere Textstelle aus derselben Onlinequelle herangezogen werden:

> »[…] Archias, a man named Archias like his grandsire […] told me that his father was called Samius, because his grandfather Archias died in Samos so gloriously, [that he] was buried with public honours by the Samian people.«[9]

Hieraus kann neben prosopographischen Abhängigkeiten auch leicht der Fakt gewonnen werden, dass Archias (der Großvater) ein Ehrenbegräbnis erhalten hat. Dieser kann mit den üblichen Mitteln als Assoziation in der Datenbank erfasst werden. Die Faktenprovenienz könnte nun in einer eigenen, von der obigen unabhängigen Kette erfasst werden. Doch dann gäbe es beispielsweise für George Rawlinson zwei Repräsentanten in der Datenbank. Dies ist in Wissensbasen generell sehr unerwünscht, kann Inkonsistenzen und Mehrarbeit bei der Pflege der Daten hervorrufen und erschwert eine gemeinsame Betrachtung der Überlieferungslage aller Fakten. Stattdessen sollten bedeutungstragende Knoten, welche für Entitäten stehen, stets nachgenutzt werden.

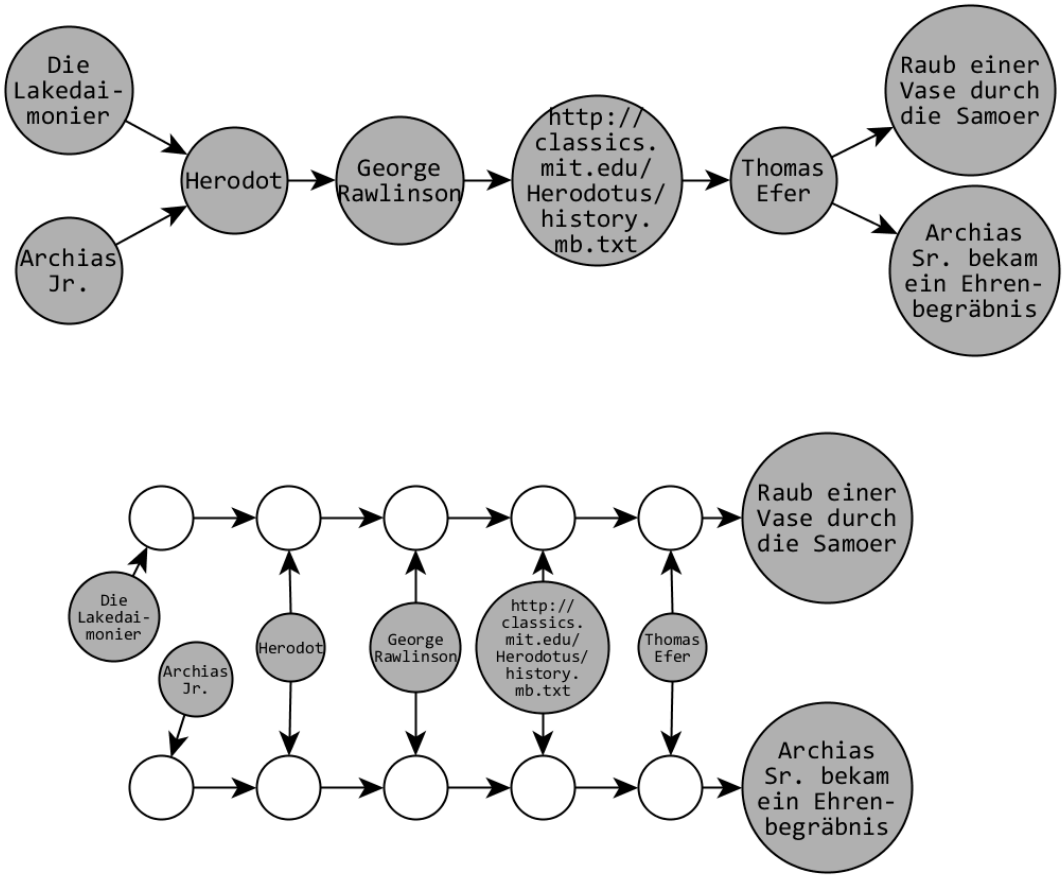

Abb. 6: Trennung von überlappenden Provenienzketten mittels Indirektion. [Efer 2019.] (http://www.zfdg.de/sites/default/files/medien/faktenprovenienz_2019_006.png)

Abbildung 6 zeigt oben zunächst, wie eine kombinierte Provenienzkette unter Knotennachnutzung aussehen könnte. Es zeigt sich, dass beide Aussagen auf den gleichen Eintragenden, die gleiche Onlinequelle, etc. zurückgehen. Was zunächst wie eine elegante Lösung erscheint, birgt nun allerdings bei der Abfrage das Problem, dass die Zuordenbarkeit der Ursprungsquellen Herodots zu beiden Aussagen nicht mehr gegeben ist. Über eine Graphenabfrage kann nicht mehr rekonstruiert werden, ob die Spartaner oder Archias (der Enkel) vom Raub der Vase berichtet haben.

Deshalb ist es notwendig, für die gemeinschaftliche Abbildung mehrerer Provenienzketten, welche sich in den beteiligten Entitäten überlappen, eine Stufe der *Indirektion* in der Modellierung zu verwenden. Es werden abstrakte, »leere« Knoten als Platzhalter für die *Position* in der Kette erzeugt. Diese sind dann mit den eigentlichen Entitäten verknüpft – grob vergleichbar mit den oben vorgestellten Rollen-Knoten und Rollenspielern im Assoziationsmodell von Topic Maps.[10] Abbildung 6 zeigt, wie mit diesem Ansatz beide Ketten sauber getrennt und dennoch auf der Ebene der Entitäten unifiziert sind.

Provenienzketten sind ein erster Vorschlag zur Dokumentation der Faktenherkunft. Eventuell lassen sich für bestimmte Anwendungsgebiete andere, geeignetere Repräsentationsformen finden. Worin genau besteht der Vorteil von Ketten? Was sagt die Reihenfolge aus und lässt sie sich immer genau definieren? Diese Fragen weisen auf viele praktische Probleme hin, die die Erfassung und Verwaltung der Provenienzinformation mit sich bringen kann. Alternativ zur Kette ließen sich einzelne Überlieferungsstationen auch als *Menge* wiedergeben. Mit angegebenen und gegebenenfalls approximierten Zeitinformationen ließe sich eine (oder mehrerer hypothetische) Ketten rekonstruieren, unter Berücksichtigung von Unsicherheiten bei der Datierung. Allerdings erschwert ein solcher Ansatz die Abfrage im Graphenmodell. Ein weiterer damit verknüpfter und noch ungeklärter Punkt betrifft die Fragen: Woher stammen die Provenienzangaben und muss eine Provenienz der Provenienz abgebildet werden (oder ist diese identisch oder zumindest sehr stark verschränkt mit der Faktenprovenienz)?

Im Lichte dieser Unklarheiten und da die vorgeschlagene Umsetzung bisher nur auf einfachster technischer Ebene mit den Basiskonstrukten des Property-Graph-Modells operiert, welche sich vielseitig verstehen und verwenden lassen, ist der Wunsch nach einem gemeinsam zu nutzenden Rahmenwerk für eine formale semantische Interpretation von Provenienzketten (oder -mengen) mehr als nachvollziehbar. Bisherige Arbeiten in verwandten Bereichen sind etwa die PROV-Ontologie des W3C. Diese umfasst ein umfangreiches Vokabular und enthält viele gute Überlegungen auf der Abstraktionsebene von *Entity*, *Activity* und *Agent*. Zudem wartet es mit einem eigenen (vergleichsweise komplizierten) RDF-basierten Datenmodell auf, wie im Standard von Moreau und Missier beschrieben.[11] Damit mutet es allerdings ähnlich schwerfällig an, wie allgemeine semantische Basis-Referenzmodelle, etwa das CIDOC-CRM, wie von Ore, Doerr und anderen definiert.[12] Letzteres ließe sich konzeptionell sehr gut als Grundlage für die Definition von *Events* im Sinne von Überlieferungsereignissen nutzen. Insbesondere für die semantisch sinnvolle Charakterisierung von *Aktionen* als Vorgängen der Überlieferung oder aktiven Bewahrung von Fakten besteht noch weiterer Forschungsbedarf, der sich auch nur im engen Dialog mit den Fachdisziplinen der Geistes- und Sozialwissenschaften auflösen lässt. Es stellt sich nun die Frage, ob Provenienzketten bereits in ihrer bisherigen Form nutzbringend für die Forschung sein können.

In der arbeitsteiligen Forschung wurde die Frage nach der Herkunft von Faktenwissen bisher oft implizit durch aufwendige Editionsprozesse von der oder dem Forschenden ferngehalten. Glaubwürdigkeit und Plausibilität von Fakten wurden unter Einhaltung guter wissenschaftlicher Praxis vorab (nach Ermessen der Editor*innen und geknüpft an ihre Reputation) geprüft. Die Herkunft der enthaltenen (und ausgesparten) Fakten sind für die Editor*innen nachverfolgbar und werden meist allenfalls ausschnittsweise in Form eines kritischen Apparats oder durch Begleittexte

kommuniziert. Digital erfasste Provenienzketten ersetzen diese Praxis nicht. Sie erlauben jedoch die Introspektion und falls nötig auch ein qualifiziertes Abweichen von der durch sie gefestigten »Lehrmeinung«. Fakten, die erfasst werden und deren Herkunft dokumentiert wird, sind dabei bisher explizit noch nicht an Wahrheitsgehalte geknüpft. Denn oftmals ist es in der Forschung auch von großem Interesse, woher »falsche« Informationen stammen, ebenso wem sie wann vorlagen und wessen Urteile sie eventuell beeinflussten. Diese wissenschaftsgeschichtlich spannenden Fragen setzen sich bis in heutige Theoriegebilde zahlreicher Disziplinen fort. Ihre Offenlegung und damit auch die kritische Hinterfragung des aktuellen Forschungsstandes ist eine wesentliche Aufgabe moderner (digital unterstützter) Forschungstätigkeit.

Bevor die Nutzung von Provenienzketten für den Umgang mit Zweifel vorgestellt wird, soll an dieser Stelle noch ein kurzer Exkurs zu den möglichen *Kettengliedern* der Überlieferung, Modi der Übernahme und Extraktion von Fakten sowie ihre digitale Repräsentation im Graphen eingeschoben werden.

5. An Faktenprovenienz beteiligte Entitäten

Entitäten sind konkrete oder abstrakte, belebte oder unbelebte »Dinge«, die in der Regel benannt werden können. Besitzen sie keinen Namen, so können sie doch insoweit durch Nummerierung oder über ihre Relation zu anderen Entitäten charakterisiert werden, dass man über sie (und nur sie) mit anderen kommunizieren kann. Entitäten benötigen in einer digitalen Datenbank eindeutige Identifikationsmerkmale, so genannte *Identifier*, damit auf sie verwiesen werden kann. Projektintern kann dies über einfache Datenbank-IDs geschehen, beim Datenaustausch über Projekte hinweg bieten sich global eindeutige Identifier an, wie sie beispielsweise im Semantic Web über Webadressen (hinter denen üblicherweise Informationsressourcen hinterlegt sind) realisiert wird. Über die Prinzipien von *Linked Open Data* (LOD) können Anbieter und Konsumenten offener Datensammlungen verteilt dieselben Entitäten referenzieren. Über diese technologischen und organisatorischen Möglichkeiten ist zudem ein Übertrag zwischen verschiedenen Identifier-Systemen möglich. Zahlreiche Gremien und Autoritäten der kulturellen und staatlichen Domäne werden für Entitäten von übergeordneter Bedeutung Normdatensätze erzeugt. Heutzutage werden diese durchweg in maschinenlesbarer Form angeboten.

Da Benennungen an sich keine gute Identifier sind (gleiches kann unterschiedlich benannt werden und unterschiedliches gleich) muss für eine saubere Nutzung von Normdaten und für die aussagefähige Modellierung der Domänendaten eine Disambiguierung und eine Zusammenführung

von Entitäten und den mit ihnen verbundenen Datensätzen stattfinden. Für die Faktenprovenienz sind verschiedene Typen von Entitäten relevant. Diese können (und werden es in vielen Fällen auch) Teil der erfassten Domänendaten sein.

Personen sind im historischen Kontext nicht selten schwer zu greifen und zuweilen nur schwer eindeutig zu identifizieren. Die Informationen, die sich über sie aus den Quellen gewinnen lassen, sind nicht immer ausreichend, um eine Verknüpfung zu Normdaten zu ermöglichen. Oft ist die Forschung auch so spezifisch, dass noch gar keine ausreichenden Normdaten für die besondere Domäne im genau passenden geo-temporalen Kontext existieren. Dazu kommt, dass zuweilen fiktionale, mythische Charaktere attribuiert werden oder idealisierte Variationen von realen historischen Personen beschrieben werden. Das Herausarbeiten solcher Identität(en) ist (auch ohne explizite Modellierung von Faktenprovenienz) eine wichtige Forschungstätigkeit und ein Schlüssel zum Verständnis des Materials.

Falls eine Überlieferung oder originale Bekundung eines Fakts in Form von Dokumenten stattgefunden hat, gilt es, eine geeignete Granularitätsstufe für die Referenzierung zu finden. Zwischen dem abstrakten *Werk* eines Autors oder einer Autorin und einem konkreten physischen Buch in einer spezifischen Auflage können mehrere konzeptionelle Ebenen liegen, wie sie etwa in den *Functional Requirements for Bibliographic Records* (FRBR) unterschieden werden.[13] Auch Paratexte auf Textträgern können als Quellen dienen und sollten damit als eigene referenzierbare Entitäten vorliegen (welche im Graphen freilich mit ihren übergeordneten Einheiten verbunden werden können). Für unumstrittene, bekannte und edierte Werke bieten sich eher Referenzierungsschemata für kanonische Texte an, für Handschriften kann über die Verwendung von Identifiern für den physischen Träger des Textes nachgedacht werden, da dieser direkt mit ihm verbunden ist und auch im Forschungskontext oft als Einheit verstanden wird.

Bei Dokumenten aus dem Web gilt zu beachten, dass diese selbstverständlich über ihre Webadresse eindeutig identifiziert werden können. Jedoch: Im Semantic Web werden alle Arten von Ressourcen durch Webadressen referenziert! Wird eine Person beispielsweise über die URL ihrer Instituts-Webseite identifiziert, so ist bei einer Aussage ›laut der Ressource mit dieser URL‹ nicht mehr zu erkennen, ob sie ›laut der Person‹ oder ›laut der Webseite‹ getroffen wurde. Im Referenzmodell von Topic Maps wird daher zwischen *Identifier* und *Locator* einer Ressource unterschieden. Dies ist eventuell auch für die Entitäten in Provenienzketten sinnvoll.

Neben den prinzipiell klar umreißbaren Entitätentypen ›Person‹ und ›Dokument‹ existieren auch diffusere Glieder in Überlieferungsketten. Beispiele können Gruppen (wie ›die Illuminaten‹), Ethnien, Sammelidentitäten (etwa ›der Senat‹) oder Institutionen sein. Diese können zu unterschiedlichen

Zeitpunkten (was abgeschwächt auch für andere Entitätentypen gilt) ganz unterschiedlichen Charakter besitzen und sich in ihrer Zusammensetzung, Ausrichtung, Wirkmächtigkeit und nicht zuletzt Glaubwürdigkeit (allgemein und in Bezug auf spezielle Themenbereiche) sehr unterscheiden. Auch gilt es hier, bei der Referenzierung ein angemessenes Granularitätsniveau zu wählen: Ist nun die Pressesprecherin persönlich, die Presseabteilung oder die Institution selbst in die Kette zu übernehmen? Warum nicht alle drei in dieser Reihenfolge? Eng damit verbunden ist die Frage einer Rollenidentität (also eine bestimmte Person ›als Vater einer anderen Person‹ oder ›als König‹ oder ›als Zeuge in einem Prozess‹). Möglicherweise sollten für solche Ausdifferenzierungen einer Entität stets spezielle Repräsentanten erstellt und mit der Hauptentität verknüpft werden.

Schließlich kann es auch sinnvoll sein, Werkzeuge zur Faktenextraktion, etwa physikalische Messungen zur Datierung oder Programme bzw. von ihnen verwendete Algorithmen an passender Stelle in die Provenienzkette aufzunehmen. Verfahren, die Fakten generieren, sind als initiales Element einzusetzen, während Verfahren, die Informationen verarbeiten (und nicht nur einer unveränderten Weitergabe dienen), entsprechend später in die Kette eingefügt werden sollten. Jede Form der automatischen oder manuellen Veränderung, Edierung, Übersetzung oder Interpretation sollte nachvollziehbar sein.

Wenn durch Identifizierung, Disambiguierung und Referenzierung eine vollständige Kette aus einzelnen beteiligten Entitäten gebildet ist, kann diese im nächsten Schritt für Umgang mit unklarer Faktenlage verwendet werden.

6. Unsicherheit, Zweifel und Widersprüche

Bis hierhin wurde ein System entwickelt, mit dem sich die Überlieferungsgegebenheiten für Fakten auf der Basis von beteiligten Entitäten abbilden lassen. Ob die dabei beschriebenen Aussagen (über Aussagen, über Aussagen… usw.) tatsächlich getätigt wurden, oder nicht, lässt sich im Nachhinein nicht ermitteln. Die Open-World-Assumption suggeriert zudem, dass weitere unterstützende oder auch gegenteilige Aussagen existieren können.

Die Faktenlage muss daher auf ihre Plausibilität hin und ihre Konsistenz untersucht werden. Für eine höhere Sicherheit über den Wahrheitsgehalt sorgt das Sammeln von Belegen. Zur guten Praxis der Wissenschaft gehört es jedoch genauso, als wahr angenommene Fakten (insbesondere im historischen Kontext) immer wieder anzuzweifeln. Nichts sollte von Forscher*innen diskurslos akzeptiert werden. Viele der Aktivitäten in den Geisteswissenschaften zielen auf genau diese kritische (Re-)Kontextualisierung bekannter Fakten ab. Statt absoluter Sicherheit muss mit unterschiedlichen Graden von Unsicherheit umgegangen werden. Wie kann diese Unsicherheit abgebildet werden?

Unsicherheit auf Eigenschaftsebene, die sich aus Vagheit von Wertezuweisungen oder einer bloßen Angabe von Schranken für Werte (anstatt der Werte selbst) ergibt, soll hier im Weiteren nicht behandelt werden. Quellen für diese Unsicherheiten können divers sein, wie etwa Toleranzen und die Unschärfe von Messungen, Subjektivität, Erinnerungslücken oder mögliche Divergenzen, die sich aus Übersetzung und Datenüberführung zwischen konzeptionell unterschiedlichen Maßeinheiten oder diskretisierten Einteilungen einer Größe ergeben.

Unsicherheit über die Tatsächlichkeit von Begebenheiten, also der generelle Zweifel am Wahrheitsgehalt eines Fakts in der Wissensbasis ist dagegen stets gleichbedeutend mit dem Zweifel an Zeugen, Quellen oder einzelnen Überlieferungsschritten.

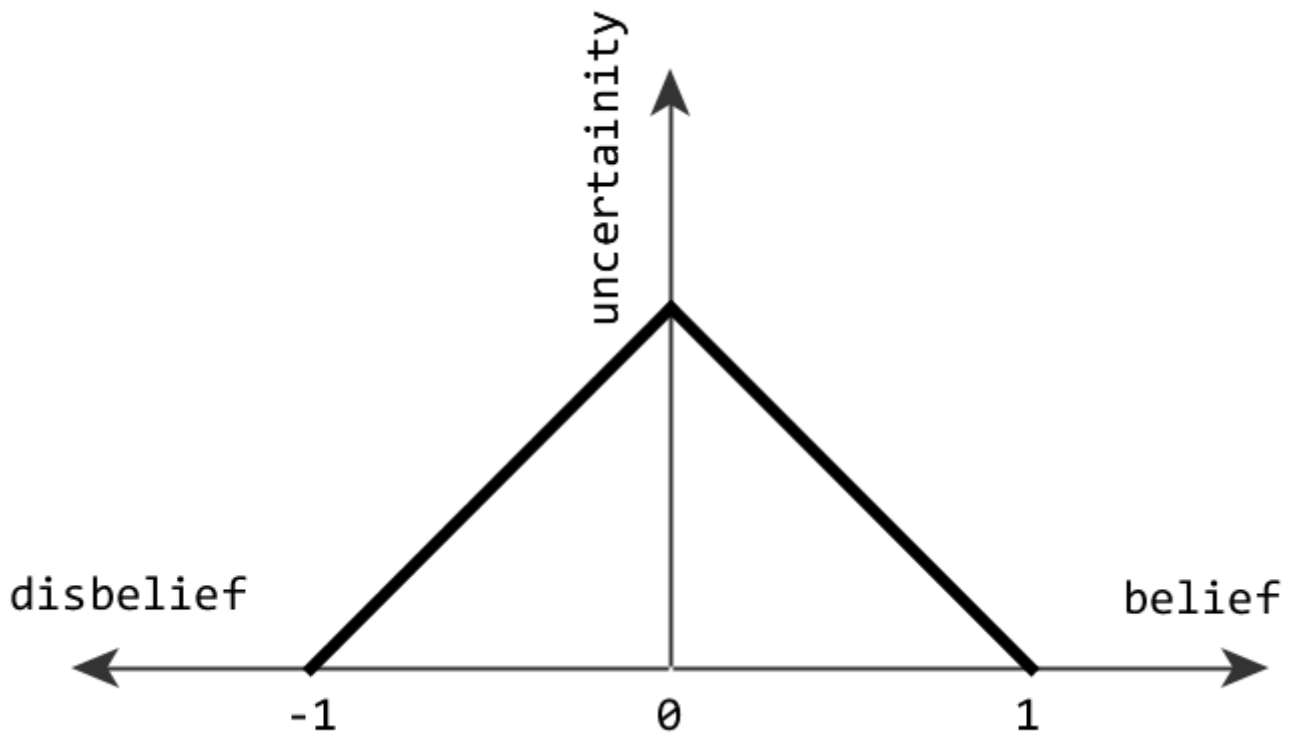

Abb. 7: Unsicherheit im Spannungsfeld zwischen ›belief‹ und ›disbelief‹, nach Hartig 2009. [Efer 2019.]
(http://www.zfdg.de/sites/default/files/medien/faktenprovenienz_2019_007.png)

Abbildung 7 zeigt eine Übersicht, in der sich Unsicherheit als Zweifel an konkreten Aussagen interpretieren lässt. Der Bereich zwischen ›belief‹ und ›disbelief‹ gegenüber einem Fakt (hervorgerufen z. B. durch mehrere widersprüchliche Aussagen oder unbeweisbarem Misstrauen gegenüber einzelnen Quellen) weist die höchste *uncertainty* auf.

Die Spezifikation von Unsicherheit innerhalb der Provenienzkette berührt wieder das Themenfeld der »Provenienz der Provenienz«. Sie kann im Modell prinzipiell auf verschiedene Arten und Weisen ausgedrückt werden, kann dabei oft nur den Charakter eines »Kommentars« für menschliche Bearbeiter*innen haben, denn als maschinell zu behandelnde Kategorie gelten. Davon ausgenommen sind zwei gut abzubildende Fälle: Zum einen können offensichtliche Fehler und Ungenauigkeiten bei der Übernahme von Fakten direkt an die passende Stelle notiert werden, indem eine entsprechende Property an die korrespondierende Kante innerhalb der Kette hinzugefügt wird. Zweitens kann die generelle Unglaubwürdigkeit einzelner Entitäten ebenfalls direkt an diesen notiert werden. Dies bietet sich etwa an bei gefälschten Dokumente, identifizierten Hochstapler*innen., aber auch integren Forscher*innen. die jedoch ihre Fakten erzeugende Interpretation von Quellen auf Basis eines mittlerweile überholten Forschungsstandes vorgenommen haben.

7. Mögliche Auflösungsmechanismen

Die Einschätzung einer Entität als (möglicherweise) unglaubwürdig hat nun im Modell den Effekt, dass das betroffene Kettenglied bei der Auflistung der Faktenprovenienz nicht einfach ausgelassen oder übersprungen werden darf, da alle vorherigen Kettenglieder nur durch die Annahme seiner Integrität im Graphen »erreichbar« sind.

Angenommen, der Autor dieses Artikels würde behaupten, dass bereits Herodot sagte: ›Peter isst eine Banane‹. In diesem abstrusen Fall würde es sich lohnen, auch alle weiteren Fakten erneut zu überprüfen, in deren Provenienzkette er prominent vorkommt.

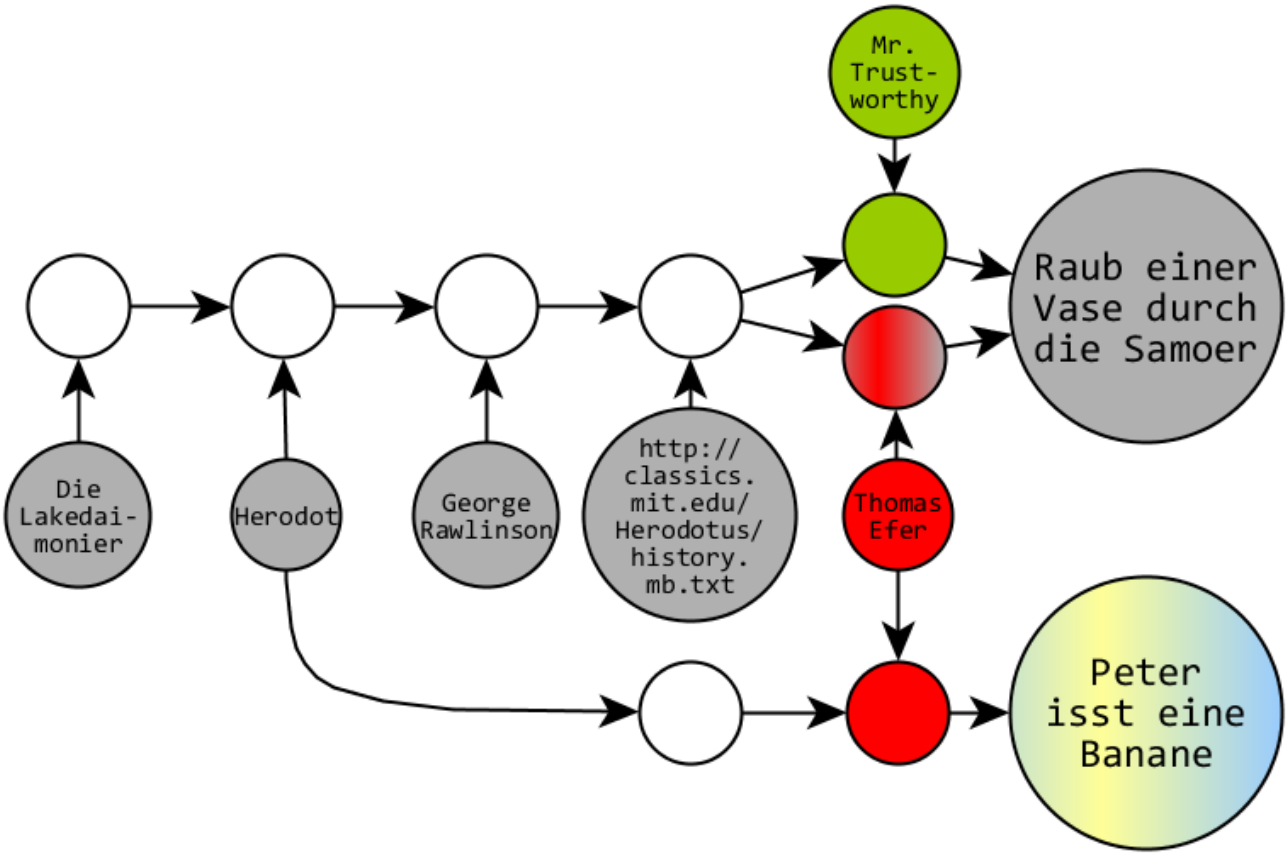

Abb. 8: Invalidierung unglaubwürdiger Kettenglieder und Ausbesserung durch Nachrecherche. [Efer 2019.]
(http://www.zfdg.de/sites/default/files/medien/faktenprovenienz_2019_008.png)

Abbildung 8 zeigt in Rot die unglaubwürdige Überlieferungsbekundung und die dadurch unglaubwürdig gemachte Entität. Wenn dieser Entität nun nicht mehr getraut wird, fallen auch glaubhafte Belege für den Fakt des Raubes der Vase durch die Samoer weg. Ein Mitglied der fachwissenschaftlichen Community müsste nun selbst eine Recherche durchführen (z. B. durch Aufruf der Online- Ressource oder Konsultation der Printausgabe oder einer unübersetzten griechischen Edition) und durch diese Rechercheleistung die Fehlstelle »auffüllen«. Diese vertrauenswürdige Person kann nun als Entität (mit eigenem Stellvertreterknoten) in die Provenienzkette eingefügt werden. Aus der Kette wird gewissermaßen ein Strang (wie bei aller Form der mehrfachen »Absicherung« der Provenienz, z. B. durch mehrere Überlieferungen in Manuskriptform oder Ähnliches).

Nicht immer lassen sich Fakten als falsch widerlegen oder mit überwältigender Evidenz unterfüttern. Die Existenz unauflösbarer Widersprüche kann durchaus plausibel sein und ist grundsätzlich für die Nutzung einer Wissensbasis im Rechercheprozess nicht schädlich. Die explizite Dokumentation solch

strittiger und unklarer Punkte erzeugt Transparenz und erleichtert zudem die Ursachenforschung und im Zuge dessen ggf. sogar das Aufspüren neuen Quellenmaterials im Umfeld der bisher konsultierten Überlieferungszeugen.

Die Notation der Unglaubwürdigkeit einzelner Quellen und Interpretationen kann in einem solchen Modell unabhängig von der Inklusion einzelner Fakten in die Wissensbasis erfolgen. Die kritische Beschäftigung mit Faktenprovenienz kann somit in einer zielgerichteten *Stand-Off-Annotation* der einzelnen Aussagen und ihrer Herkunft münden. Permanente und verlustbehaftete Filterung, Löschung, Korrektur und Edition an Ort und Stelle sind so nicht mehr nötig und versperren künftigen Forscher*innen nicht mehr die Sicht auf die (komplexe) Quellenlage. Gleichzeitig kann die »Übernahme« fremder Einschätzungen zur *Fakten-Glaubwürdigkeit* effizient und transparent innerhalb eines gemeinsamen Referenzsystems erfolgen.

Zum wichtigen Werkzeug zur Unterstützung von Recherchearbeit könnte eine (semi-)automatische Identifikation »schwacher« Provenienzketten werden. Dafür könnten beispielsweise die Zeiträume zwischen den einzelnen Kettengliedern beleuchtet werden: Mehrere hundert Jahre Abstand sind unplausibel für eine direkte Überlieferung zwischen einzelnen Personen. Hier könnte eine genauere Recherche weitere Zwischenstationen zu Tage fördern. Weiterhin könnten (für Zeiten ohne briefliche oder fernmündliche Kommunikationsmöglichkeiten) große räumliche Distanzen zwischen den Wirkungsorten von überliefernden Personen als Schwachstelle identifiziert werden. Diese machen eine direkte Faktenüberlieferung unwahrscheinlicher, außer es existieren Belege für ein persönliches Treffen. Lange Zeiträume sind bei physischen Trägern der Fakten zunächst nicht unüblich, wie z. B. bei handgeschriebenen Autographen und lange aufbewahrten Archivalien. Wenn allerdings Kopien, Teil-Abschriften und Kompilationen beteiligt sind, gilt es dort, mit den üblichen Herangehensweise der Quellenkritik (entsprechend des genauen Entitätentyps) vorzugehen.

Eine solch differenzierte Auswertung benötigt letztlich einen geeigneten ontologischen Zugang. Dabei ist noch unklar, ob dies generisch, wie bei den bereits genannten Vokabularen, geschehen kann, oder auf die spezifische Fachkultur und Arbeitsweise einzelner Fachgebiete sowie deren jeweilige Quellenlage zugeschnitten sein muss, insbesondere bezüglich Art, Zustand, Umfang, Alter und Heterogenität der Quellensammlungen. Spezielle Anforderungen ergeben sich dort aus den Berührungspunkten der Forschung mit dem Sektor des kulturellen Erbes. Für die Provenienzdokumentation von Fakten sind auch gegenständliche Provenienzen von größtem Interesse und es ist leicht ersichtlich, dass hierbei eine Fülle von kuratorischen Gepflogenheiten, materiellen Besonderheiten und organisatorischen Standards zu berücksichtigen ist: Transkripte von

Wachszylindern, Fachartikel zur Grabungsdokumentation archäologischer Quellen, digitale Korpussammlungen und Forschungsdatenbanken, Sammlungen zur Oral History – all diese Überlieferungswege müssen adäquat abbildbar sein.

8. Kritik, Desiderate und Ausblick

Die hier vorgestellten Ansätze können (so die Hoffnung des Autors) zu einem gewissen Grad helfen, Forschungsergebnisse und ihre Herleitung übersichtlicher und für andere anschlussfähiger zu erfassen. Dem gewählten Konzept wohnt (wie jedem Modell) eine bewusste Simplifizierung inne. Damit greift es für spezielle Anwendungsfälle sicher in vielerlei Hinsicht noch kurz, zumal viele Aspekte insbesondere der Abbildung von Identität und normierten Ankerpunkten für Entitäten bislang sehr unterspezifiziert sind. Das System wie hier beschrieben ist in der Praxis bisher nicht systematisch erprobt worden., zumindest eine Nachjustierung einzelner Teile ist im Zuge dessen zu erwarten.

Wie im Abstract erwähnt, liegt dem Beitrag der Wunsch zugrunde, eine Diskussion über die Inklusion von Provenienzinformationen in Forschungsdatenbanken und Recherchessysteme zu initiieren. Auch wenn diese Idee auf fachwissenschaftlichen Zuspruch stoßen sollte, bleibt zu untersuchen, wie bestehende Forschungsmethodik und eine explizite, digitale und transparente Modellierung von Faktenprovenienz zusammenpassen und welche Implikationen sich für Forschungsprozesse ergeben. Insbesondere die Identifikation von Grenz- und Sonderfällen des Modells kann nur sinnvoll aus praxisnahen Erwägungen und durch realitätsnahe Testfälle geschehen.

Jede im Forschungsalltag zusätzlich zu erfassende Information ist mit erhöhten Arbeitsaufwänden verbunden. Hinzu kommen weitere Aufwände bei der kontinuierlichen Pflege und Aktualisierung der Daten. Solange sich im Arbeitsalltag kein dauerhafter, sichtbarer Nutzen aus der Erfassung von Provenienzinformationen gesammelter Fakten ergibt, ist der Wunsch verständlich, die dafür aufzuwendende Arbeitszeit lieber für die »eigentliche« Forschung zu verwenden. Hier können sich ähnliche Reibungspunkte ergeben, wie sie für die Dokumentation von Programmquelltext existieren. Welche Dokumentationsaufwände sind mindestens nötig? Welche angemessen? Welchen Grad der öffentlichen oder internen Nachnutzung der Fakten und ihrer Provenienzinformationen wird es voraussichtlich geben? Insbesondere bei öffentlicher Verfügbarmachung ist zudem der subjektive Eindruck der eintragenden Nutzer*innen nicht zu unterschätzen, eventuell wertvolle Recherchearbeit »für Andere« zu leisten, ohne dabei direkte akademische Reputationsvorteile zu erhalten. Als Gegenargument könnte angemerkt werden, dass die Provenienzketten eindeutig den jeweiligen Bearbeiter, die jeweilige Bearbeiterin als letztes Glied enthalten können und damit die kleinen

individuellen Beiträge zum kollektiven Wissen sogar besser als bisher herausgestellt werden können. Solange dies jedoch keinen anerkannten »Wert« in der Fachwelt darstellt ist, stellt dies wohl nur einen schwachen Trost dar.

Aus der eher kollektiven Sicht heraus lässt sich feststellen, dass sich Provenienzketten für die Qualitätskontrolle einer Wissensbasis nutzen lassen und darüber hinaus einen transparenteren Forschungsprozess befördern können. Neben den vielen positiven Aspekten für die informiertere und fundiertere Ableitung von Aussagen bestehen dabei auch Gefahren in der Praxis: Durch die Nutzungsmuster der Fakten wird ggf. sichtbar, wer welche (noch lebenden) Forscherkolleg*innen als »unglaubwürdig« einstuft. Dies birgt großes Konfliktpotential, zumindest bis durch die weite Adaption solcher Forschungsmethodik eine Versachlichung der Debatten und eine kollaborative Konfliktlösung eintritt, falls dies realistisch ist.

Ein weiterer Kritikpunkt an den vorgestellten Lösungsansätzen für die Modellierung von Faktenprovenienz kann die Vermischung von primären Domänendaten und (sekundären) Provenienzdaten in der Wissensbasis sein. Eine solche komplexitätserhöhende Anreicherung von diversen Informationen ist nicht immer gewünscht und nicht immer praktisch für die Erstellung von Präsentations- und Fachanwendungen. Eine logische Trennung aller erfassten Daten durch speziell zugewiesene Properties, einer knotenweisen Verknüpfung zu *Data Collections* oder die Nutzung logischer Subgraphen (falls vom Datenbanksystem unterstützt) löst dieses Problem, bedeutet allerdings zusätzlicher Aufwand bei Konzeption und Umsetzung des Systems, sowie bei der Formulierung und Abarbeitung von Anfragen. Hier existiert definitiv ein Bedarf für weitere Überlegungen und praktische Untersuchungen.

Als weiteres Desiderat kann die Konzeption (standardisierter) Interaktionsweisen und Nutzeroberflächen für die Visualisierung und schnelle Introspektion, aber auch für die unterstützte Dateneingabe von Provenienzketten angesehen werden. Daneben ist zu untersuchen, wie sich die Workflows mit weiteren externen Softwareprodukten verknüpfen lassen. Für die Nachrecherche von Fakten wäre es beispielsweise wünschenswert, die entsprechenden zusätzlichen parallelen Kettenglieder von dem oder der Forschenden zur konsultierten Quelle in der Provenienzkette automatisch hinzuzufügen, wenn die damit verknüpfte Rechercheaufgabe in einem Ticketsystem als erfolgreich abgearbeitet markiert wird.

Die hier vorgestellten Konzepte werden derzeit im Langzeitvorhaben Bibliotheca Arabica (https://www.saw-leipzig.de/de/projekte/bibliotheca-arabica) der Sächsischen Akademie der Wissenschaften zu Leipzig auf ihre Praxistauglichkeit hin untersucht und entsprechend weiterentwickelt. Das Projekt beschäftigt sich mit der reichhaltigen Produktion, Rezeption und

Transmission von Literatur im bisher als *post-klassisch* bezeichneten Zeitalter der Manuskriptkultur in der arabischsprachigen Welt. In diesem Rahmen ist die Erstellung einer graphbasierten bio-bibliographischen Datenbank, in welche unter anderem die Daten von über 100 gedruckten Handschriftenkatalogen eingespeist werden sollen, vorgesehen. Durch sie wird eine digitale Arbeitsweise unterstützt, welche nicht nur neuartige Arbeitsweisen bei der Analyse der Manuskriptdaten ermöglicht, sondern insbesondere auch die Bereitstellung von Provenienzinformationen für Fakten zur Kontextualisierung der abgebildeten kodikologischen, bio- und bibliographischen und aller weiteren übernommenen Aussagen erfordert. Aus diesem Projekt werden also neue Impulse für die graphbasierte Modellierung von Faktenprovenienz hervorgehen.

Allgemein wird angestrebt, für die hier beschriebenen Ideen eine generische webbasierte Datenbanklösung als Mischung aus dokumentierter Referenzimplementierung und einfach nutzbarer Endanwendersoftware zur Verfügung zu stellen. Diese sollte generische Oberflächen mit entsprechenden individuellen Anpassungsmöglichkeiten, etwa im Stil der beliebten MyCoRe-Repositoriumssoftware (http://www.mycore.de/) bieten. Ein realistischer Zeitplan für die Umsetzung dieses Vorhabens ist vorerst jedoch noch nicht abzusehen.

Fußnoten

[1] Borek et al. 2016

[2] Simmhan et al. 2005, S. 31–36.

[3] Vgl. dazu Robinson et al. 2013.

[4] Vgl. Rodriguez / Neubauer 2011, S. 29–46.

[5] Vgl. Efer 2017.

[6] Hartig 2014.

[7] Moore / Van Pham 2015.

[8] Herodotus / Rawlinson 1994-2009.

[9] Herodotus / Rawlinson 1994-2009.

[10] ISO/IEC 13250 International Organization for Standardization 2003, Stage 90.92.

[11] Moreau / Missier 2013.

[12] Ore et al. 2018.

[13] International Federation of Library Associations and Institutions 2009.

Bibliographische Angaben

Luise Borek / Quinn Dombrowski / Jody Perkins / Christof Schöch: TaDiRAH: a Case Study in Pragmatic Classification. In: Digital Humanities Quarterly 10 (2016), H. 1. [online (http://www.digitalhumanities.org/dhq/vol/10/1/000235/000235.html)]

Thomas Efer: Graphdatenbanken für die textorientierten e-Humanities. Leipzig, 2017. URN: urn:nbn:de:bsz:15-qucosa-219122 (http://nbn-resolving.de/urn:nbn:de:bsz:15-qucosa-219122) [Nachweis im GVK (http://gso.gbv.de/DB=2.1/PPN?PPN=1556056575)]

Olaf Hartig: Querying Trust in RDF Data with tSPARQL. In: The Semantic Web: Research and Applications. Hg. von Lora Aroyo et al. (ESWC: 6, Heraklion, 31.05.-04.06.2009) Berlin u.a. 2009, S. 5-20. DOI: 10.1007/978-3-642-02121-3_5 (https://doi.org/10.1007/978-3-642-02121-3_5) [Nachweis im GVK (http://gso.gbv.de/DB=2.1/PPN?PPN=600606880)]

Olaf Hartig: Reconciliation of RDF* and Property Graphs. In: arXiv.org. Technical Report vom 11.09.2014. [online (https://arxiv.org/abs/1409.3288)]

Herodotus: The History of Herodotus. Translated by George Rawlinson. In: The Internet Classics Archive. 1994-2009. [online (http://classics.mit.edu/Herodotus/history.mb.txt)]

Functional Requirements for Bibliographic Records. Hg. von IFLA Study Group on the Functional Requirements for Bibliographic Records. In: ifla.org. Version von 02.2009. [online (http://www.ifla.org/VII/s13/frbr/)]

Information technology – SGML applications – Topic maps / ISO. Hg. Von International Organization for Standardization. ISO/IEC 13250 Stage 90.92. Second Edition vom 23.10.2003. [online (https://www.iso.org/standard/38068.html)]

Philip Moore / Hai Van Pham: On Context and the Open World Assumption. In: 2015 IEEE 29th International Conference on Advanced Information Networking and Applications Workshops. Hg. von Leonard Barolli. (IEEE AINA: 29, Gwangiu, 24.-27.03.2015) Piscataway, NJ. 2015. [Nachweis im GVK (http://gso.gbv.de/DB=2.1/PPN?PPN=832059382)]

Luc Moreau / Paolo Missier: PROV-DM: The PROV Data Model. In: w3.org. W3C Recommendation vom 30.04.2013. [online (https://www.w3.org/TR/prov-dm/)]

Definition of the CIDOC Conceptual Reference Model. Hg. von Christian Emil Ore / Martin Doerr / Patrick LeBœuf / Stephen Stead. In: cidoc-crm.org. Version 6.2.3. von 05.2018. [online (http://www.cidoc-crm.org/Version/version-6.2.3)]

Ian Robinson / Jim Webber / Emil Eifrem: Graph Databases. Beijing u.a. 2013. [Nachweis im GVK (http://gso.gbv.de/DB=2.1/PPN?PPN=743805844)]

Marko A. Rodriguez / Peter Neubauer: The Graph Traversal Pattern. In: Graph Data Management: Techniques and Applications. Hg. von Sherif Sakr / Eric Pardede. Hershey, PA 2012, S. 29–46. [Nachweis im GVK (http://gso.gbv.de/DB=2.1/PPN?PPN=897738977)]

Graph Data Management: Techniques and Applications. Hg. von Sherif Sakr / Eric Pardede. Hershey, PA 2012, S. 29–46. Siehe auch [Nachweis im GVK (http://gso.gbv.de/DB=2.1/PPN?PPN=897738977)]

Yogesh L. Simmhan / Beth Plale / Dennis Gannon: A survey of Data Provenance in e-Science. In: ACM SIGMOD Record 34 (2005), H. 3, S. 31–36. [Nachweis im GVK (http://gso.gbv.de/DB=2.1/PPN?PPN=129615544)]

Abbildungslegenden und -nachweise

Abb. 1: Modellierung einfacher Aussagen im Graphen. [Efer 2019.]

Abb. 2: Modellierung von Aussagen über Aussagen mittels Reifizierung. [Efer 2019.]

Abb. 3: Modellierung von mehrgliedrigen Assoziationen. [Efer 2019.]

Abb. 4: Problematik rekursiver Reifizierung mittels Assoziationsmodellen. [Efer 2019.]

Abb. 5: Beispielhafte Provenienzkette. [Efer 2019.]

Abb. 6: Trennung von überlappenden Provenienzketten mittels Indirektion. [Efer 2019.]

Abb. 7: Unsicherheit im Spannungsfeld zwischen ›belief‹ und ›disbelief‹, nach Hartig 2009. [Efer 2019.]

Abb. 8: Invalidierung unglaubwürdiger Kettenglieder und Ausbesserung durch Nachrecherche. [Efer 2019.]

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung
(http://www.bmbf.de)

MARBACH WEIMAR WOLFENBÜTTEL FORSCHUNGS VERBUND
(http://www.mww-forschung.de/)

DHd digital humanities im deutschsprachigen raum
(http://www.dig-hum.de)