Projekt Strukturierte Daten
Detailprojekt Datenbank bibliographischer Objekte

Strukturierte Daten

Momentan befindet sich ein Projekt erheblicher Dimension im Wikipedia Konsortium in Planung: Eine Ressource „strukturierter Daten“.

Es handelt sich dabei um das Projekt einer zentralen Datenbank, die beliebige Objektdaten sammelt und diese *strukturiert* verwaltet – ein Projekt gewaltiger Dimension und weitreichender Implikation. „Objekte“ können sein: Der Mond, das Kraftfahrzeug einer bestimmten Baureihe, eine Buch, das über Bibliothekskataloge recherchierbar ist.

Objektdaten sind theoretisch beliebig definiert: Gewicht des Mondes, Daten seiner Erdumlaufbahn, Zahl und Spezifikation der Einzelteile im genannten Kraftfahrzeug, Leistungsdaten seines Motors, Bibliotheksstandorte des recherchierten Buches, Angaben zu Seitenzahl, Format, Inhalt, Genre, Autor, Übersetzer, Vorlage und so fort.

Die Daten wären – als freies Wissen – beliebig nutzbar. Wikipedia, die Enzyklopädie, speiste etwa alle einzelsprachlichen Artikel zum Mond aus der Ressource. Verändert man in der zentralen Ressource ein Datum, wird dieses in allen Artikeln, die dieses Datum nutzen aktualisiert.

Der Nutzen außerhalb der Online-Enzyklopädie ist im Moment vollkommen unabsehbar. Momentane Suchmaschinen und Übersetzungsprogramme arbeiten ohne „Weltwissen“, das heißt sie übersetzen von einer Sprache in die andere, ohne zu wissen, warum es geht, suchen Äquivalenten für Sprachpartikel. Zukünftige Programme sollten erkennen, wovon gesprochen wird, Verständnis-Rückfragen stellen können und dann Antworten aus der zentralen Wissensressource generieren können.

Generelles Anliegen: Das Projekt sollte frühzeitig wissenschaftlich betreut werden

Die deutsche Wikimedia Community sollte versuchen, Fachleute in Kontakt miteinander zu bringen, um Interessen an einer solchen Ressource zu bündeln und frühzeitig an die Entwicklung zu binden. Wichtig wäre es, Vorstellungen davon zu entwickeln, wo die Ressource Nutzung finden könnte und welchen Anforderungen sie dabei genügen sollte, um Entwicklungsspielräume nicht zu verschenken. Mitunter sind Nutzungen einfach, wurden aber schlicht nicht vorgesehen, von denen die an die technische Realisation dachten, und die keine Vorstellung vom späteren Interesse an ihrer Arbeit hatten.

Im Moment sehe ich hier einen mehrseitigen Informationsbedarf – vom Aufbau der neuen Ressource hört man ein Raunen, während es spanend wäre, sich abzeichnende Entwicklungen frühzeitig zu sehen und konstruktiv von den späteren Interessenten aus zu begleiten.

Zweites und konkretes Projekt: Der Aufbau einer zentralen Ressource bibliographisch-archivalischem Wissens

Das nachfolgend näher skizzierte Projekt würde man am effizientesten im Rahmen des skizzierten großen Projektes realisieren. Es gilt dem Detailbereich der momentan weltweit disparat primär von nationalen Ressourcen verwalteten bibliographischen und archivalischen Objektdaten.

Mit dem Aufbau der Bibliotheks-Online-Kataloge wurden Bücher weltweit recherchierbar. „Metakataloge“ führen dieses Wissen derzeit unterschiedlich zusammen:

- Ressourcen wie der KVK – der Karlsruher Virtuelle Katalog – greifen auf verschiedene Onlinekataloge zu und stellen Suchanfragen an sie. Der Benutzer erhält als Ergebnis die Informationen der befragten Ressourcen nach den Anbietern geordnet.
- Dem stehen – weitaus interessanter – Ressourcen gegenüber wie die gegenwärtig im Aufbau befindlichen Nationalkataloge. Bei ihnen werden Kataloge unter erheblichem Arbeitsaufwand zusammengeführt: Man erstellt von einem mehrfach verfügbaren Objekt (Büchern einer Auflage, die heute in vielen Bibliotheken liegen) einen einzigen Datensatz, der unter anderem notiert, welche Bibliotheken den Titel vorrätig haben.

Die zweite Form der Kataloge ist zukunftsweisend, da in ihr Informationen sinnvoll zusammengefügt werden. Deutlich wird das, wenn man auf die Fragen sieht, die sich erst an Kataloge des zweiten Typs stellen lassen: Man kann den ESTC (den English Short Title Cataloge) für alle auf Englisch respektive im englischen Sprachraum zwischen 1473 und 1800 gedruckten Titel statistisch auswerten. Wie viele Titel wurden im Jahre 1700 gedruckt? Wie viele davon lassen sich als Romane einstufen? Wieviele Bücher wurden in Edinburgh gedruckt? Wie viele in London? Jede einzelne Auflage eines Titels ist im ESTC nur einmal notiert. Eine Ressource wie der KVK listet dagegen den einen gesuchten Titel so oft, wie sie positive Rückmeldungen von konsultierten Katalogen erhielt – die Treffermenge entscheidet sich über die Zahl ausgewählter Kataloge. Präzise Fragen lassen sich an den menschlich erstellten Katalog stellen: Wie viele Verleger arbeiteten an einem Ort zum beliebigen Zeitpunkten simultan? Wie viele Bücher druckte ein Verleger durchschnittlich etc.

- Kataloge wie der englische ESTC, der niederländische STCN, die deutschen Jahrhundertkataloge VD16, VD17 und, soeben in Konstruktion, VD18 werden soeben in nationalen Interessen erstellt. Man will mit ihnen mehr über das Schrifttum der eigenen Nation und Sprache erfahren.
- Die Produktion der Gesamtkataloge geht derzeit von Bibliothekskooperationen aus – entsprechend bibliothekarisch sind die Anfragemöglichkeiten formuliert.
- Benutzer sind bislang durchweg von Mitarbeit an diesen Ressourcen ausgeschlossen.[1]

Der Ausschluss der Benutzer hat gravierende Konsequenzen. Um Korrekturen (man hielt etwa den Titel in der Hand, sah, dass das Titelblatt falsch abgeschrieben wurde, kann den Titel präziser datieren, weiß, wer der Autor ist) kann man per e-mail einreichen, sein Wissen vergibt man dabei ohne weitere Spuren zu hinterlassen. Fachwissen kann in die Ressourcen nicht einfließen: Sie sehen nicht vor, dass Informationen begründet werden – man erhält etwa verschiedenen Angaben zu möglichen Autoren, nicht aber in Fußnoten die Gründe für die divergierenden Mutmaßungen. Fachliteratur muss man in anderen Ressourcen recherchieren etc.

Komplexes Wissen bleibt in diesen Katalogen aus: man würde sich bei einer Recherche wie „Robinson Crusoe“ eine strukturierte Publikationsgeschich-

[1] Ich sprach jüngst mit dem Leiter der Bibliothek Wolfenbüttel, und erfuhr dass man hier derzeit an eine Öffnung denkt, an eine Kooperation mit Wikipedia dachte man bislang dort nicht.

te wünschen: Was ist die Erstausgabe, wie liegt die exakte Chronologie ermittelter Folgeauflagen, welche Raubdrucke und Übersetzungen folgten welchen Vorlagen – die Kataloge geben allenfalls einen maschinell chronologische geordneten Überblick, den sie aus den internen Daten generieren, nicht die tatsächliche Publikationsgeschichte des Titels.

Prekär sind die nationalen Grenzziehungen: Der ESTC nennt die englischen *Robinson Crusoe* Ausgaben grob geordnet, er erlaubt keinen Blick auf die europäische Verbreitung. Sie muss in den anderen Nationalkatalogen recherchiert werden.

Speziell bei statistischen Auswertungen lassen die aktuellen Kataloge Wünsche offen – die Daten sind vorhanden, dass jedoch jemand wissen will, wie sich der Buchdruck an einem Ort entwickelte, sah man nicht vor – die komplexeren Suchanfragen lassen sich oft nicht stellen, obwohl alle Detaildaten im System liegen.

Desiderat: Der von Benutzern betreute und fortentwickelte Metakatalog

Interessant wäre aus Sicht der Forschung ein bibliographisches Recherchewerkzeug, das von den Benutzern mitgestaltet würde und die Informationen der nationalen Projekte konstruktiv zusammenführte. Es sollte möglich sein, in dieser Ressource

- Daten manuell zu verknüpfen und zu strukturieren (man erhält etwa bei der Erstausgabe die Publikationsgeschichte samt Links in die Folgeausgaben, Übersetzungen und Digitalisate).
- Daten zu korrigieren.
- Daten zu kommentieren und der kritischen kollektiven Überprüfung auszusetzen.
- Wissen zu Objekten einzuspeisen (Inhaltsangaben, Exzerpte, Verweise auf Fachliteratur sind hie wertvoll).
- Forschung an der Datenbank in einem eigenen Projektnamensraum an sie anzubinden

Der letzte Punkt mag mit zwei Beispielen illustriert sein. Promovenden, die mit Büchern arbeiten geben über ihre Projekte bislang auf Universitätsseiten und Seiten der Stipendiengeber Auskunft. Sinnvoll wäre es, sie vernetzten ihre Arbeit mit Materialien in der Ressource selbst mit ihrem Projekt. Spannend wäre die Interaktion mit den Forschungsprojekten: Momentan versuchen verschiedene Projekte Perspektiven auf die Briefzirkulation unter Gelehrten der frühen Neuzeit zu geben. Die Briefe liegen in verschiedenen Bibliotheken und Archiven, man katalogisiert sie, wertet sie inhaltlich

aus, vernetzt sie und versucht Einblick in die Netzwerkstrukturen und deren dynamische Entwicklung zu geben.[2] Die verschiedenen Forschergruppen bauen sich dazu jeweils eigene Datenbanken auf, Insellösungen, die rasch veralten und deren Informationen aufgrund der insolierten technischen Standards nicht mehr in Nachfolgeprojekte fließen. Spannend wäre die Ressource die zentral die Datenlage produziert und die Arbeit *mit* der Datenlange *an* sie anbindet.

Kommunikationsangebote, die von uns ausgehen müssten

Nötig wäre es, zügig auf Kooperationspartner zuzugehen, die Daten zur Verfügung stellten. In Gesprächen die ich im Vorfeld des VD18 führte, erhielt ich Signale einer Bereitschaft, die Daten mit einem Projekt wie Wikipedia zu teilen, es ist unklar, wer heute mit uns kooperieren würde.

Interessant wäre es, das bibliographische Projekt gemeinsam mit Google Books zu realisieren. Google Books ging breite Kooperationen mit Institutionen wie der Bayerischen Staatsbibliothek ein. Der Zugriff auf die Digitalisiate geschieht derzeit über eine Datenbank, die mehr an Youtube und Amazon erinnert, denn an einen sinnvollen Katalog. Man sucht Surfer, die sich vielleicht auch für das nächste ähnlichen Link interessieren. Man hofft wie bei Amazon auf Benutzer, die eine „Review“ schreiben, eine Kurzbeurteilung durch den Kunden – ein Kuriosum in einem Medium, das Wissenschaftlern Bücher aus der frühen Neuzeit zur Verfügung stellt. Interessant wäre es, Google-Books anzubieten, den Zugriff auf die Digitalisate durch eine interaktive bibliographische Community-gestützte Ressource zu optimieren.

Nächstliegendes Antragsziel

wäre nach dem gesagten eine Konferenz auf der Fachkreise über die laufende Planung informiert würden und in einem Workshop Ideen generierten, in welchen Richtungen das Projekt Entwicklungsspielräume gewinnen sollte.

Kostenfaktoren wären hier die Organisation eines Konferenzortes (etwa Bibliothek Göttingen), die Kommunikation mit Teilnehmern im Vorfeld (Recherche, Kontaktaufnahmen, Vorabgespräche), die Finanzierung von Reisen und Übernachtungen.

[#akuter Beratungsbedarf meinerseits o.s.]

[2] Siehe etwa das Stanford-Projekt: http://shc.stanford.edu/intellectual-life/video-podcasts/detail/tracking-18th-century-social-network-through-letters