Der Datengarten - Kollaborative Pflege von Norm- und Metadaten

Data Gardening – Collaboratively Curated Authority and Meta Data

Mathias Schindler
Wikimedia Deutschland e.V.
Eisenacher Straße 2
10777 Berlin
mathias.schindler@wikimedia.de

Zusammenfassung:
Der Dartengarten ist die unscharfe Bezeichnung für den kreativen und experimentierfreudigen Umgang mit Personendaten, um die Zielsetzung von Wikipedia besser zu erfüllen. Das gewählte Lizenzmodell erlaubt die – auch kommerzielle – Weiternutzung von Wikipedia-Inhalten inklusive der in Wikipedia strukturiert gespeicherten Daten

Abstract:
The data garden is a metapor to describe the creative and experiment-driven approach on authority files in order to fulfill the mission of Wikipedia. The chosen licensing model allows commercial re-use.

Wikipedia entstand 2001 als Konsequenz eines grandiosen Fehlschlages. Der Unternehmer Jimmy Wales hatte mit seinem Projekt Nupedia versucht, eine Internetenzyklopädie auf die Beine zu stellen, bei der unbezahlte, ehrenamtliche Freiwillige nach einem aufwändigen Akkreditierungsprozess Artikel genau so schreiben sollten, wie man es dem Organigramm aus den Informationsblättern gedruckter Nachschlagewerke entnommen hatte: Die Freiwilligen sollten ihre Qualifikation und Identität nachweisen, sich um ein noch ungeschriebenes Lemma bewerben, ihr Ergebnis in einen Peer-Review eingeben und die Rückmeldungen zu Stil und Form einarbeiten. Nupedias Stärke war das umfängliche Demotivieren von Freiwilligen, mittels Struktur den Weg zur eigentlichen Arbeit zu verstellen.

Um das darbende Projekt etwas zu entlasten, sollte Wikipedia als vorgeschalter Schmierzettel einer größeren Gruppe von ebenfalls Freiwilligen erlauben, als Zuträger für die weiterhin mit festen Strukturen arbeitenden Artikelschreiber zu Hilfe zu schreiten. Die Autoren stimmten mit ihren Füßen ab, exportierten die wenigen Nupedia-Inhalte nach Wikipedia, begannen mit der Sammlung neuer Einträge und verzichteten fortan auf jeden Versuch, sich die eigene Freizeit weiter durch Bürokratie zu versauen.

Dieses doppelte Scheitern – der nicht erfolgreiche Versuch der Wiederbelebung des untauglichen Nupedia-Ansatzes – ist der Grundstein für eine Online-Enzyklopädie mit derzeit (Stand: Oktober 2011) 20 Millionen Einträgen in 270 Sprachausgaben und einer Leserschaft von monatlich weltweit 409 Millionen Nutzern (Stand: Juni 2011, Quelle: ComScore).

Teil der Geschichte von Wikipedia sind unendliche Debatten zur Zuverlässigkeit von Wikipedia, zur Zuverlässigkeit anderer Werke und zur Zuverlässigkeit einer Textgattung allgemein, die bislang mehrheitlich auf Einzelbelege verzichten durfte. Wikipedia-Autoren werden hier je nach Konstellation entweder auf entsprechende Einzeluntersuchungen verweisen, die die Ebenbürtigkeit Wikipedias mit inzwischen teilweise verblichenen anderen Nachschlagewerken behaupten oder alternativ auf die großen methodischen Mängel eben jeder Untersuchungen hinweisen, die jede inhaltliche Aussage zur Qualität von Wikipedia eher in den Bereich der Anekdote schieben wird.

Grundpfeiler der Inhalte von Wikipedia sind das verwendete Lizenzmodell, Creative Commons mit den Lizenzbausteinen Attribution und ShareAlike. Dies bedeutet, dass jedermann die Inhalte beliebigen, auch zu kommerziellen Zwecken nutzen kann, sofern gewisse Nutzungsbedingungen eingehalten werden. Diese sind im Kern die Nennung des Urhebers, die Nennung der Quelle und die Nennung der Lizenz selbst. Sharealike verpflichtet darauf, abgeleitete Werke ebenfalls wieder unter gleichen Lizenzbedingen freizugeben.

Dass derzeit für die Erstellung und Pflege der Inhalte ein Wiki eingesetzt wird, ist kein konstituierendes Element der Wikipedia – wenn es morgen eine Software gäbe, die die zu erledigende Arbeit einfacher gestaltet, stünde einem Wechsel auf diese Plattform nichts im Wege.

Bedingt durch die selbst auferlegte Verpflichtung zur Nennung von Einzelbelegen ist die Wikipedia ein Sprungbrett zu anderen – vorwiegend im Netz publizierten – Inhalten. Wikipedianer pflegen in redaktioneller Kleinstarbeit die jeweiligen Linklisten pro Artikel und kämpfen gegen den „bit rot“ an, das Verrotten von Hyperlinks durch Änderung der Organisationsstruktur vieler Webseiten.

An dieser Stelle setzen Normdaten ein.

2004 kam es zu einer kleinen ungeplanten Fingerübung bei der Integration von Normdaten in Wikipedia, als ein Berliner Verlag eine CD-ROM (später: DVD) mit den Inhalten der deutschsprachigen Wikipedia auf den Markt brachte. Um die Softwarefunktion nach der werküber-greifenden Personensuche zu ermöglichen, fügten Freiwillige in Form von per CSS unsichtbar markierten Info-Tabellen die für die Identifikation nötigen Personendaten in die jeweiligen Wikipedia-Artikel. Die Vorlage:Normdaten existiert in dieser Weise noch heute. Auf Initiative zweier Bibliothekare und Wikipedianer folgte, sofern vorhanden, im Sommer 2005 ein Link auf die Personennamendatei der Deutschen Nationalbibliothek.

Mit Stichtag 25. Oktober 2011 besteht die deutschsprachige Wikipedia aus 1,3 Millionen Artikeln (ausgenommen Weiterleitungen, Begriffsklärungsseiten und, darunter 400.000 Personenartikeln. Ungefähr 40% dieser Artikel enthalten ebenfalls eine PND-ID. Der Anspruch hier ist natürlich, einmal zu jeder Person in Wikipedia auch eine PND-ID anbieten zu können, dies wird die automatische Erstellung von individualisierten PND-Datensätzen (Tp) erfordern.

Mit dem Projekt PeEnDe (http://lists.wikimedia.org/pipermail/wikide-l/2009-November/022201.html) wird die automatische Erstellung von PND-Datensätzen aus strukturierten Wikipedia-Daten demonstriert. Die PeEnDe-Datenbank mit 400.000 Einträgen ist von Anfang an wie die sonstigen Wikipedia-Inhalte und im Gegensatz zur PND von jedermann auch zu kommerziellen Zwecken frei nutzbar.

Mit PND BEACON entwickelten Wikipedianer im Februar 2010 ein minimalistisches Dateiformat, das die Propagierung von personenbezogenen und mit der PND erschlossenen Netzressourcen erleichtert. Eine valide BEACON-Datei enthält eine Liste der PND-IDs, die im lokalen Projekt

verwendet werden, außerdem eine URL mit Platzhalter für die jeweilige PND. Durch das Einfügen einer konkreten PND in den Platzhalter wird die jeweilige Fundstelle für die Inhalte zu einer gegebenen Person abgeleitet. Durch die einfache Propagierung der sehr kleinen BEACON-Dateien wird es möglich, passgenaue Schnittstellen zwischen an sich völlig unterschiedlichen Internetprojekten ohne weiteren manuellen Pflegeaufwand zu schaffen. Jede Seite pflegt ihre BEACON-Liste , ggf. einen Redirectdienst zur Umleitung von PND-gestützer URL zum eigentlichen Permalink und propagiert die URL neuer Versionen dieser Datei, für den Rest sind die jeweiligen Nachnutzer zuständig.
Die Wikipedia-Personensuche (z..B. http://toolserver.org/~apper/pd/person/Karl_Marx) hat eine Liste von BEACON-Dateien integriert und zeigt jeweils zu konkreten Personen nur jene Seiten, die auch zur jeweiligen Person Inhalte vorzuhalten angeben. Bibliotheken können als optionalen Parameter der BEACON-Datei die Zahl der Fundstücke mitgeben. In der praktischen Arbeit helfen Aggregatoren wie z.B. http://beacon.findbuch.de/. Zusammen mit dem von Jakob Voss programmierten SeeAlso ist es eine Sache von zwei Zeilen JavaScript, in eine PND-erschlossene Webseite eine BEACON-Linkliste einzubauen.

Im Zuge einer Kooperation zwischen dem Bildarchiv des Deutschen Bundesarchivs und Wikimedia Deutschland arbeiteten Freiwillige innerhalb weniger Monate über 50.000 Datensätze durch, um, wenn möglich, einen Link zwischen der BArch-internen Personendatei, der PND und den Artikeln der deutschsprachigen Wikipedia zu setzen. Im Gegenzug gab das Bundesarchiv 100.000 Bilder aus seinem Bestand unter einer für Wikipedia nutzbaren CC-Lizenz heraus.

Seit Oktober 2011 läuft ein Experiment mit 32.000 Testartikeln aus den Landesdiensten der Deutschen Presseagentur dpa aus dem Monat Mai 2010 zur Erforschung von Werkzeugen zum Taggen von Personennamen in dpa-Meldungen.

Eine frühe Demo ist unter http://toolserver.org/~apper/dpa öffentlich.

Unter http://toolserver.org/~magnus/granDPA/index.php?id=14244 ist ein Serviervorschlag für mit Normdaten angereicherten dpa-Datensätzen publiziert.