Niels-Oliver Walkowski
Alexander Czmiel

Anschlussfähiges Wissen. Virtuelle Forschungsplattformen als Medium der Ressourcen- und Informationsorganisation/-integration am Beispiel des Wissensspeichers der BBAW

Vortrag, gehalten auf dem Workshop Virtuelle Forschungsplattformen an der Universität Trier

2009

Einleitung

Das Nachdenken über virtuelle Forschungsplattformen, oder auch virtuelle Forschungsumgebungen, virtuelle Forschungsinfrastrukturen wird in letzter Zeit durch eine Reihe von Veranstaltungen mit diesem Thema und Initiativen mit entsprechender Zielsetzung belebt. Die diskursive Aufmerksamkeit der eScience/Humanities Gemeinde scheint zum nächsten Sprung anzusetzen, der durch neue Triggerbegriffe vorbereitet wird. Am auffälligsten an diesen oben genannten Komposita ist die Entdeckung des Begriffs der Forschung für eine Terminologie der digitalen Wissenschaft. Wir reden nicht von Science als solche die um ein Präfix *e* erweitert wird oder von den Humanities die digital werden. Beide Begriffe, schon früh eingeführt zeugen noch von dem Wunsch überhaupt erstmal im virtuellen Raum präsent zu sein. Es sind Begriffe die ein soziales Feld und seine Existenzform beschreiben. Auch reden wir nicht von Objekten, wie bei dem Begriff *Digitale Edition* bzw. ihrem digitalen zu Hause, den Repositorien. Der Begriff Forschung meint die Tätigkeit der Entdeckung, Aneignung, Verarbeitung und Neubildung von Wissensbeständen selbst, er ist eine Handlung. Er setzt das Vorhandensein dieser Wissensbestände voraus. Genau an diesem Punkt macht die Konjunktur des Begriffs der *Forschungsplattform* auf den dahinter liegenden Prozess und auf ein bestimmtes Problem aufmerksam.

Noch vor kurzem war das Schlüsselwort *Content*. Der virtuelle Raum musste mit Inhalten erobert werden um ihn relevant werden zu lassen. Die Zeiten haben sich geändert. Wikisource[1] nennt 320 Einrichtungen die Digitaliserungsprojekte in deutscher Sprache unterhalten, das Directory of Open Acces Repositories[2] listet 1498 Einträge und wir reden hier nur von textuellen digitalen Wissensbeständen, die frei zugänglich sind. Je reichhaltiger die Fülle an wissenschaftlichen Angeboten im Internet wird desto stärker rücken Fragen nach Strukturierung, Organisation und Vermittlung in den Mittelpunkt. Aus der Systemtheorie wissen wir, dass Inhalte ihre Anschlussfähigkeit unter Beweis stellen müssen um zu gesellschaftlich wirksamen Wissen zu werden. Diese Anschlussfähigkeit ist bei überall verstreuten, häufig isoliert ihr Dasein fristenden Online-Ressourcen nicht gegeben, egal wie hoch ihre Qualität sein mag. An dieser Stelle bildet die Konzeption von Forschungsplattformen die Chance Anschlussfähigkeit herzustellen und Relevanz zu

1 http://de.wikisource.org (Stand: Oktober 2009)
2 http://www.opendoar.org/(Stand: Oktober 2009)

erzeugen, indem sie die Inhalte der Ressourcen semantisch und strukturell reorganisiert, auswertet, verarbeitet und rearrangiert.

Vor diesem Verständnis ist der Wissensspeicher der BBAW zu betrachten, dessen Ziel es sein soll, die zutiefst unterschiedlichen Ressourcen der Akademie zu integrieren und in einer verarbeiteten Form der wissenschaftlichen Auseinandersetzung anzubieten. Er stellt eine Weiterentwicklung der BBAW-Infrastruktur für die Präsentation der digitalen Wissensbestände dar und wird die Sichtbarkeit der Forschungstätigkeit der Akademie im Internet erhöhen. Grundlage für ein erfolgreiches Konzept ist die Integration des Wissensspeichers in die bestehende Infrastruktur. Er soll sich von Entwicklungen aus dem Bereich der Wissensorganisation und -präsentation im virtuellen Raum inspirieren lassen und diese im Sinne der besonderen Situation an der Akademie weiterentwickeln. Wir haben bisher gehört was technisch umsetzbar und was realisiert worden ist. Tatsächlich sind wir bisher wohl das unfertigste Projekt und noch so sehr am Anfang, das wir auch noch kein Akronym gefunden haben.

Welche Art an Informationen liegen an der Akademie vor?

Die Berlin-Brandenburgische Akademie der Wissenschaften beschäftigt sich seit über 300 Jahren mit der Generierung von Wissen. Daran hat sich bis heute nichts geändert. Die Art, oder aus informationstechnischer Sichtweise, die Struktur dieser Informationen unterscheidet sich allerdings von Vorhaben zu Vorhaben, in der Natur der Dinge begründet, beträchtlich. So gibt es Daten, die eine klare Strukturierung aufweisen, z.B. das Wörterbuch von Jacob und Wilhelm Grimm, Bibliographien oder Prosopographien, aber auch das neue, von der DFG geförderte Projekt, Personendaten-Repositorium, welches eine Infrastruktur schafft um biografische Daten zu historischen Personen des 19. Jh. zu sammeln und zu vernetzen. Dagegen gibt es aber auch Wissen in semistrukturierter Form, wie z.B. bei historisch-kritischen Editionen und Dokumentationen. Vor allem Editionen, die sich mit der Textgenese befassen, wie z.B. die Leibniz-Edition an der BBAW, sind ein gutes Beispiel dafür.

Des weiteren unterscheidet sich der Entstehungskontext digitaler wissenschaftlicher Ressourcen. Gedrucktes Wissen, das vor der Einführung des Computers auf traditionelle Art und Weise erzeugt wurde, und seit einigen Jahren retrodigitalisiert wird unterscheidet sich an Struktur und elektronisch erschlossener Informationstiefe deutlich von

sogenannten „born digital“-Arbeiten, dem aktuellen Normalzustand. Letztere bieten, wenn solide konzipiert, Zugriff auf eine Informationsmenge und -tiefe, die weit über die der retrodigitalisierten Bestände hinausgeht. So kann man „born digital“-Editionen mit allen relevanten inhaltlichen Informationen anreichern, die man aber nicht alle gedruckt werden müssen; was oft aus Platzgründen auch gar nicht möglich ist. Bei einer retrodigitalisierten Edition hat man hauptsächlich formale Informationen, wie Angaben zu Seiten und Zeilen, digital vorliegen, die zwar die Zitierfähigkeit nach der gedruckten Ausgabe gewährleisten, eine Vernetzung mit reinen Onlineressourcen allerdings erschweren, da die inhaltlichen Anknüpfungspunkte meist manuell nachgetragen werden müssen.

Auch aus technischer Sicht ist die Heterogenität der verschiedenen Ressourcen der Akademie eine Herausforderung. Die verwendeten Formate gehen von einfachen Textdateien über Word, Tustep, LaTeX und PDF bis hin zu TEI-konformen XML. Dazu kommen verschiedenste relationale Datenbanken, sowie große Mengen an Digitalisaten und anderen Binärdaten. Und nicht zu vergessen die Publikationen der Akademiemitglieder und -mitarbeiter auf dem eDoc-Server der Akademiebibliothek[3].

Die Integration all dieser verschiedenen Ressourcen in ein System erleichtert nicht nur die Zugänglichkeit, sie erzeugt auch erheblichen inhaltlichen Mehrwert. Die Schnittmengen der Akademienvorhaben sind häufig offensichtlich. Viele Inhalte zeigen chronologische, geografische oder inhaltliche Überschneidungen. Aber viele dieser Wissensbestände stecken ebenfalls voller impliziter Referenzen aufeinander, Informationen die es gilt explizit zu machen. So wäre es zum Beispiel für Altphilologen interessant zu erfahren in welchen neuzeitlichen Texten oder Werken ihre Autoren zitiert oder selbst als Autor geführt werden. Zudem ist man in der Lage auf diese Art und Weise neues Wissen zu generieren:

Suche als Prozess

Wie kann nun diese Reorganisation und Verarbeitung der Ressourcen aussehen, die den inhaltlichen Mehrwert erzeugen soll? Der Zugriff auf die Ressourcen wird üblicherweise über einen Suchprozess realisiert. Er ist die heikle Schnittstelle zwischen dem Wissen, welches der Benutzer mitbringt und dem Wissen, welches in den Ressourcen liegt. Heikel deshalb, weil beide Horizonte, also das Verständnis für ein Thema und die Form innerhalb dessen ein Thema rezipiert wird von Benutzer und Ressource unterschiedliche sind. Um so verwunderlicher ist es also, dass bei den meisten Suchen nur unzureichend zwischen

3 http://edoc.bbaw.de/

Suchendem und den Inhalten vermittelt wird. Das Phänomen, dass man eigentlich schon wissen muss, was man erwarten kann um eine erfolgreiche Suche durchzuführen ist vielen vertraut. Noch schwieriger wird es, wenn man eigentlich gar nicht so genau weiß, was man suchen kann und noch kein ausgereiftes Vorwissen für ein Thema mitbringt. Gerade an dieser Stelle können Forschungsplattformen Anschlussfähigkeit erzeugen indem sie die Inhalte, die sie beherbergen, verarbeiten und die Suche als eine Sphäre der Auseinandersetzung mit den Inhalten nutzen.

Aus diesem Grund wird die Suche beim Wissensspeicher in einen mehrstufigen Prozess aufgelöst. Den Einstieg bildet ein flexibel gestaltetes traditionelles Suchformular. Der nächste Schritt beinhaltet eine Oberfläche, auf der unterschiedliche inhaltliche Kontexte auf der Basis der vorangegangenen Suche und der im System befindlichen Ressourcen gebildet werden. Diese Kontexte werden auf der Basis von automatisierten semantischen Interpretationsmethoden wie dem TextMining, der Auswertung vorangegangener Suchprozesse und -statistiken, auf der Basis von Feedback zu den Ressourcen aber auch klassisch durch Metadaten erzeugt. Es werden semantische Cluster gebildet, die nicht nur zu einem Abgleich zwischen Suchendem und den Ressourcen führt, sondern selbst schon Wissen vermitteln. So werden z.B. relevante Schlüsselbegriffe im Kontext der Suche oder die Rezeptionsgeschichte eines Themas transparent. Repräsentiert werden so nicht nur die Ressourcen, sondern das in den Ressourcen gelagerte Wissen selbst. Dieser Schritt kann beliebig oft wiederholt, die Suche so modifiziert und die Auseinandersetzung vorangetrieben werden, bis der Übergang in eine Ergebnisliste und damit die Rückführung auf konkrete Ressourcen gewünscht ist. Ebenso kann eine Ressource als Wiedereinstieg in den eben beschriebenen Prozess dienen. Der Wissensspeicher versucht so die Suche als Teil der Auseinandersetzung mit den Inhalten zu begreifen.

Zusätzlich zur Aufgabe der Wissensorganisation soll der Wissensspeicher auch zum Voranbringen anderer Fragestellungen, wie etwa die der Langzeitarchivierung, genutzt werden. Je nach Implementierung auf der Ebene des Portals bietet der Wissensspeicher die Möglichkeit in einer sinnvollen Weise die Standardisierung bei der Erstellung digitaler Ressourcen an der Akademie zu unterstützen und als Werkzeug der Interaktion und Kommunikation zwischen den Vorhaben zu dienen.

Für die Organisation eines Wissensspeichers aus technischer Sicht existieren

verschiedene konzeptuelle Herangehensweisen. Man kann die unterschiedlichen digitalen Ressourcen und bestehende isolierte Ressourcensyteme, wie Repositorien und Datenbanken, unter einem „Dach“ bündeln. Durch vielfältige Schnittstellen lassen sich darüber hinaus auch Inhalte maschinell erfassen und weiterverarbeiten, die bisher noch nicht über solche Schnittstellen verfügten. Diese Integration kann entweder über ein zentrales Repositorium erfolgen, in dem alle Ressourcen komplett erfasst und vernetzt sind oder man belässt die eigenständigen Datenbanken und Ressourcensysteme als solche und baut eine zentrale Datenbank auf, welche lediglich Metainformationen zu den einzelnen Ressourcen enthält. So wäre der Wissensspeicher als verteiltes System, also Cloud oder Grid realisiert.
Ein weiterer wichtiger Punkt ist die breite Verfügbarkeit des Wissens. Neben einem Portal, das wie eben beschrieben, die Suche für den menschlichen Nutzer in den Ressourcen der Akademie ermöglicht, sind insbesondere technische Zugriffsmöglichkeiten nötig, die nicht nur eine Verknüpfung der digitalen Ressourcen der Akademie erlauben, sondern diese auch z.B. in Form von Webservices für andere Systeme als Open-Access zur Verfügung stellen. Die dazu verwendeten Protokolle und Technologien sollen offen, gut dokumentiert und im besten Fall normiert (wie z.B. die PND) sein. So kann die Vielfalt und Akzeptanz der angebotenen Schnittstellen relevant für die Verbreitung des Wissens sein.

Probleme und Grenzen des Machbaren bei der Realisierung des Wissenspeichers ergeben sich vor allem auf Grund der angesprochenen großen Vielfalt der in der BBAW vorhandenen Ressourcen, die sowohl technischer als auch inhaltlicher Natur ist. Aus beiden Bereichen sei kurz ein Beispiel genannt. Da die Forschung an der BBAW Forschungsgegenstände aus unterschiedlichen Regionen und Epochen umfasst, haben wir es häufig mit Ressourcen in unterschiedlicher Sprache zu tun. Insbesondere tritt diese Vielfältigkeit vor Augen wenn man bedenkt, das zum Beispiel die Turfanforschung Ressourcen in den Sprachen Sogdisch, Sakisch und Parthisch veröffentlicht. Das Erzeugen ressourcenübergreifender Kontexte durch automatisierte Prozesse wie des TextMining wird hier unmöglich. Ein anderes Problem sind die angesprochenen unterschiedlichen Systeme und Formate die an der BBAW existieren. Dies wirft die Frage auf wie all diese unterschiedlichen Quellen in ein System integriert werden können und tatsächlich können einige das nicht. Insofern geht der Aufbau des Wissensspeichers an der BBAW einher mit einer Propagierung der Benutzung offener Standards und

exportfähiger Systeme. Dies ist keine Selbstverständlichkeit, da viele Wissenschaftler z.B. exotische Datenbanksysteme nutzen, die in ihrem Fachkreis eine gewisse Verbreitung besitzen, aus technischer Sicht und der Zukunftsfähigkeit wegen jedoch nicht zu empfehlen sind.

In einem ersten Schritt, in dem wir uns aktuell befinden, werden alle elektronischen Ressourcen der BBAW erfasst und mit Metadaten versehen. Für diese Erschließung haben wir auf das *Metadata Object Description Scheme* (MODS) zurückgegriffen. Die Entscheidung für MODS beruht auf seiner ausgesprochenen Flexibilität in der Beschreibung Ressourcen unterschiedlichen Typs, sowie in der Fähigkeit Beziehungen zwischen Objekten ausdrücken zu können. Grundsätzlich ist der Ansatz in den MODS Metadaten, die Ressourcen technisch so zu beschreiben, dass das System durch diese Informationen in die Lage versetzt wird den vollen Zugriff auf die Ressourcen zu bekommen. Da wir es nicht nur mit Ressourcen sondern auch mit Ressourcesystemen zu tun haben nutzen wir das <relatedItem>-Element um diese Beziehung darzustellen:

<relatedItem type=“host“ xlink=“12082009-23-WSP-BBAW“ />

Über das <extension> Element ermöglicht MODS die Integration eines eigenen Vokabulars in die MODS Datei. Wir benutzen das Element um technische Spezifizierungen der Ressourcen vorzunehmen, auf Grund deren das System in der Zukunft den Zugriff auf die Inhalte erhalten soll.

```
<extension>
    <wsp:systemType>exist</wsp:systemType>
    <wsp:accesMode>xquery</wsp:accesMode>
    <wsp:rootPath>xmldb://telota/ressources/</wsp:rootPath>
</extension>
```

Der Zugriff auf diese inhaltlichen Metadaten erfolgt über eine zentrale Suche, die auf der neugestalteten Homepage der BBAW angeboten werden soll. Dadurch wird ein erster, einfacher Zugang zu allen Ressourcen eingerichtet werden. Die Suche in den Metadaten ermöglicht eine Einordnung der Suchergebnisse in die verschiedenen Fachgebiete und ein gezieltes Weiterleiten in die relevanten Forschungsmaterialien.

Die Einrichtung von Schnittstellen und Services wird auch externen Open-Access-Anbietern den Zugriff auf die Metadaten und die Verarbeitung der Ressourcen ermöglichen. Erste Ergebnisse werden im 1. Quartal 2010 auf der Website der BBAW zu sehen sein. Ziel ist es den Wissensspeicher, auch über die technischen Schnittstellen, mittelfristig in die internationale Forschungslandschaft zu integrieren. Denkbar wäre ebenfalls eine Anbindung an kommerzielle oder durch sogenanntes Crowdsourcing erstellte Ressourcen: Google-Maps, Google-Books, Wikipedia.

Die Webservices können die Wissenschaftler bei der Integration ihrer Ressourcen in das System unterstützen, die Standardisierung von Arbeitsprozessen bei der Erstellung digitaler Inhalte sowie die Kommunikation unter den Benutzern fördern oder auch für eine effektive Einbettung der Ressourcen in den Arbeitsalltag von Wissenschaftlern sorgen.

Das auf der gemeinsamen Verwaltung verschiedener, zuvor isolierterRessourcensysteme aufbauende verknüpfen der Ressourcen und der Organisation des in ihm enthaltenen Wissens, wie in diesem Vortrag vorgestellt, bildet die Grundlage für die nächste Projektphase, welche nach Möglichkeit im nächsten Jahr beginnen soll.