Dominique Ritze, Kai Eckert, Magnus Pfeffer

Forschungsdaten

In der Rolle als Informationsdienstleister wird der Umgang mit Forschungsdaten für Bibliotheken immer wichtiger. Forschungsdaten bilden in den meisten Disziplinen den Grundstein wissenschaftlichen Arbeitens und ihre Archivierung und Zugänglichmachung gewährleistet die Nachvollziehbarkeit und Reproduzierbarkeit von publizierten Ergebnissen. Eine besondere Rolle spielt dabei die Zitierbarkeit von Forschungsdaten und damit einhergehend die Verknüpfung von Veröffentlichungen und den ihnen zugrunde liegenden Forschungsdaten. Linked Open Data bietet eine Infrastruktur, um Forschungsdaten einheitlich zu beschreiben, eindeutig zu identifizieren und möglichst vielen Nutzern zur Verfügung stellen zu können. Die Verknüpfung von Forschungsdaten und Literatur eröffnet auch dem recherchierenden Benutzer neue Möglichkeiten: Die Originaldaten sind zugreifbar zur eingehenden Prüfung der publizierten Ergebnisse, darüber hinaus kann gezielt nach weiteren Publikationen gesucht werden, die auf den gleichen Forschungsdaten basieren. Aufbauend auf unseren Erfahrungen aus dem DFG-geförderten Projekt *InFoLiS* stellen wir die Herausforderungen sowie Möglichkeiten und Vorteile dar, die sich aus einer LOD-basierten Infrastruktur zum Beschreiben und Verlinken von Forschungsdaten ergeben.

Einleitung

„Daten sind die Währung der Wissenschaft, auch wenn Publikationen immer noch die Währung für die Festanstellung sind. Für die wissenschaftliche Produktivität, Kollaboration und Entdeckung neuer Ergebnisse ist es jedoch essentiell, Daten auszutauschen, zu kommunizieren, zu extrahieren, wiederzuverwenden und zu begutachten“ (Gold, 2007).[1] Obwohl durch den technologischen Fortschritt die Bereitstellung der Daten immer einfacher wird, ist es noch keine gebräuchliche Praxis, Daten zu veröffentlichen, die einer Publikation zugrunde liegen. Und wenn Daten veröffentlicht werden, dann ist nicht immer klar, welche Publikationen sich darauf beziehen. Auch die Referenzierung von Forschungsdaten innerhalb einer Publikation ist uneinheitlich, wenn überhaupt vorhanden, so dass es schwierig ist, die zugehörigen Daten zu finden. Ohne Zugang zu den

1 “[...]data is the currency of science, even if publications are still the currency of tenure. To be able to exchange data, communicate it, mine it, reuse it, and review it is essential to scientific productivity, collaboration, and to discovery itself.“

Forschungsdaten können weder die genannten Ergebnisse verifiziert oder reproduziert werden, noch die Daten mit anderen verglichen oder wiederverwendet werden.

Für die Organisation von Forschungsdaten und Publikationen bietet sich die dezentrale Architektur von Linked Open Data an. Ausgehend von einer detaillierten Beschreibung für die lokalen Bedürfnisse der Datenanwender lassen sich so Forschungsdaten einrichtungs- und bereichsübergreifend integrieren und vernetzen. Analog zur Verknüpfung von Publikationen untereinander basierend auf Zitationen können so die Forschungsdaten zitierbar gemacht werden und zu darauf aufbauenden Publikationen in Bezug gesetzt werden.

Im Folgenden wird zunächst auf die zyklische Natur wissenschaftlicher Forschung eingegangen und werden anschließend die an unterschiedlichen Punkten entstehenden Forschungsdaten klassifiziert und für den Kontext dieser Betrachtung genauer definiert. Die nächsten Abschnitte beschäftigen sich mit dem wichtigen Aspekt der dauerhaften Archivierung von Forschungsdaten und den Möglichkeiten, die eine auf Linked Open Data basierende Infrastruktur für die Referenzierung von Forschungsdaten bieten kann. Ein kurzer Überblick über das DFG-geförderte Projekt „InFoLiS“ zeigt exemplarisch, wie die genannten Möglichkeiten in der Praxis umgesetzt werden können. Eine Zusammenfassung mit einem Ausblick auf kommende Entwicklungen beschließt dieses Kapitel.

Forschungszyklus

Unabhängig vom Forschungsgebiet und dem konkreten Forschungsthema folgt die Forschung einem Ablauf, wie er in Abbildung 1 skizziert ist (angelehnt an (Baskerville & Wood-Harper, 1996)). Am Anfang steht eine *Recherche*, aus der sich entweder die Forschungsfrage ergibt oder die dazu dient, eine vorhandene Forschungsfrage zu untermauern. Aus den gefundenen Forschungsergebnissen kann eine konkrete Forschungsfrage formuliert werden. Es folgt die *Gestaltung einer Studie* bzw. eines Experiments deren *Durchführung* experimentelle Ergebnisse liefert, die weiter untersucht werden müssen. An dieser Stelle unterscheiden sich die einzelnen Disziplinen deutlich: Es kann sich dabei etwa um die Auswertung einer primären Textquelle in den Geisteswissenschaften handeln oder auch um die Durchführung einer Forschungsreise zum Nordpol mit entsprechenden Experimenten, die Daten liefern sollen. Unabhängig von der konkreten Forschungsaktivität folgt eine *Analyse der Ergebnisse*, die hoffentlich zur Beantwortung der ursprünglichen Forschungsfrage führt. Forschungsdaten fallen in

diesem Modell sowohl nach der Durchführung (als unbearbeitete Rohdaten) als auch nach der Analyse an (als aufbereitete Daten).

Die Erkenntnisse aus der Analyse, ergänzt durch weitere Ergebnisse, werden zusammen mit einer Beschreibung des Vorgehens in einer *Publikation* festgehalten und somit anderen Wissenschaftlern zur Verfügung gestellt. Damit schließt sich der Kreis: die neue Publikation kann im Rahmen der Recherche anderen Wissenschaftlern als Ausgangsbasis dienen.

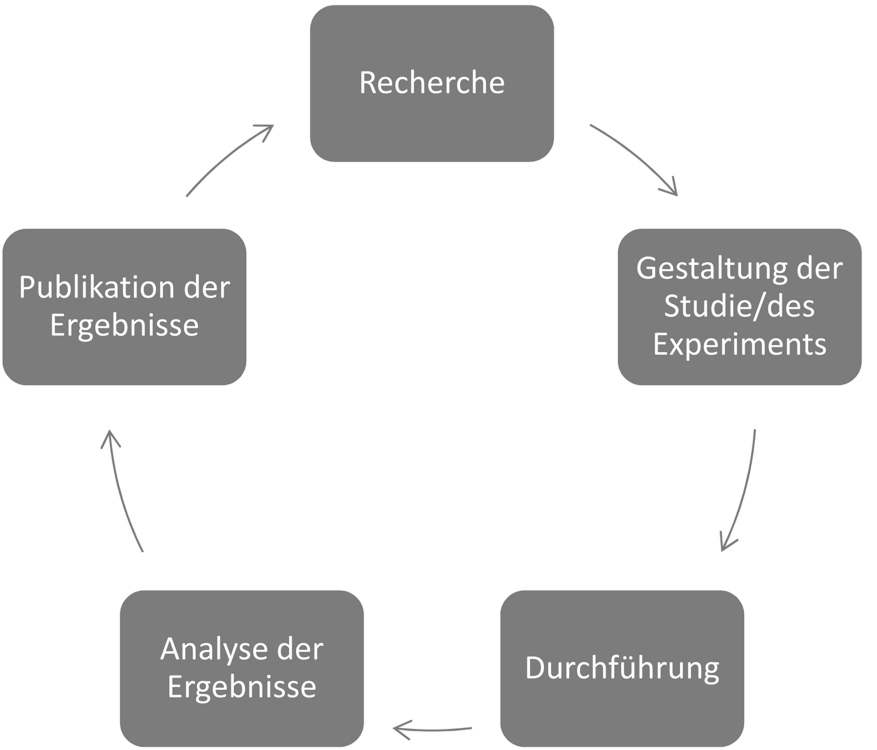

Abbildung 1: Forschungszyklus.

Auch wenn also die wesentlichen Erkenntnisse, die aus Sicht des durchführenden Wissenschaftlers aus den angefallenen Forschungsdaten gewonnen werden können, veröffentlicht werden, fehlt im Allgemeinen anderen Wissenschaftlern eine Möglichkeit, auf die ursprünglichen Daten (Rohdaten und aufbereitete Daten) zuzugreifen.

Forschungsdaten

Im Sinne einer klaren Abgrenzung für die unterschiedlichen denkbaren Arten von Forschungsdaten verwenden wir die Definition der Universität Edinburgh (Edinburgh University, 2011). Demnach sind Forschungsdaten alle Arten von Informationen, die gesammelt oder erstellt werden mit dem Zweck der Analyse, um daraus Forschungsergebnisse zu erhalten.[2]

Forschungsdaten werden dabei in fünf unterschiedliche Kategorien unterteilt:

Tabelle 1: Arten von Forschungsdaten.

Art	Charakteristik	Beispiel
Beobachtungen	Daten werden in Echtzeit erfasst meistens unersetzbar	Sensordaten Umfragedaten
Experimente	meist im Labor erstellt reproduzierbar aber teuer	Gensequenzen Chromatogramme
Simulationen	von Testmodellen generiert Model und die Metadaten wichtiger als Ausgabe	Klimamodelle Wirtschaftsmodelle
Abgeleitete Daten	aus anderen Daten abgeleitet oder kompiliert reproduzierbar	Textmining 3D-Modelle
Referenzen	Sammlung kleinerer Datensätze Meist publiziert	Gensequenzdatenbank Primäre Textquellen

Tabelle 1 zeigt die verschiedenen Arten von Forschungsdaten mit ihren spezifischen Charakteristiken. Primäre Kandidaten für eine dauerhafte Aufbewahrung, Dokumentation und Zugänglichmachung sind dabei die Daten aus nicht wiederholbaren Beobachtungen oder Experimenten, die oft nur unter hohem Aufwand reproduzierbar sind.

Zusätzlich liegen die Forschungsdaten auch noch in unzähligen Formaten vor. Beispiele dafür sind Textformate, numerische Formate, Multimedia, Modelle,

2 "Research data, unlike other types of information, is collected, observed, or created, for purposes of analysis to produce original research results."

Software- oder disziplinspezifische Formate. Daraus ergeben sich unterschiedliche Anforderungen an die Beschreibung und Speicherung der Daten. Eine Übersicht über die verschiedenen Arten von Forschungsdaten, ihr Umfang und die Formatvielfalt gibt Neuroth, Strathmann, Oßwald, Scheffel, Klump, & Ludwig, 2012.

Metadaten zu Forschungsdaten

Die Forschungsdaten selbst werden mit Hilfe von Metadaten beschrieben. Diese Angaben dienen neben der Beschreibung der Daten auch ihrer Interpretation. So kann zum Beispiel die Angabe des Befragungszeitraums einer Umfrage notwendig sein, um diese dem korrekten Kontext zuordnen zu können. Zusätzlich erschließen sie die Forschungsdaten für die Recherche, so dass andere Forscher sie in entsprechenden Informationsystemen auffinden können.

Durch die breite Spanne an denkbaren Datentypen sind die beschreibenden Metadatenelemente höchst unterschiedlich. Übergreifende Elemente beschränken sich auf allgemeine Angaben, wie zum Beispiel zu beteiligten Personen und Institutionen oder dem Erstellungsdatum, während ansonsten domänenspezifische Elemente erforderlich sind.

Grundsätzlich werden Metadaten in vier verschiedene Kategorien eingeteilt (Razum, 2011):

- Technische Metadaten
- Provenienz Metadaten
- Lizenz-Metadaten
- Deskriptive Metadaten

Technische Metadaten dienen dazu, die Forschungsdaten auf einer formalen Ebene zu beschreiben. Dazu gehören alle Angaben zu Dateiformat, -typ und -größe. Mithilfe dieser kann ein Datenpaket einer bestimmten Art von Forschungsdaten zugeordnet werden (zum Beispiel ein 3D-Modell, Text oder Video).

Provenienz Metadaten geben Aufschluss darüber, wie die Daten entstanden sind und bearbeitet wurden. Dazu gehören neben den Angaben zur Methodik auch die zu beteiligten Personen und Institutionen. Sie sollten eine erste Einschätzung zur Eignung und Vertrauenswürdigkeit der Daten erlauben und im besten Fall eine Rekonstruktion des Experiments für eine unabhängige Validierung ermöglichen.

Lizenz-Metadaten enthalten Angaben darüber, ob und in welchem Rahmen die Forschungsdaten von anderen Wissenschaftlern nachgenutzt werden dürfen. Jegliche zusätzliche Informationen, die spezifisch für die entsprechende Art von

Forschungsdaten sind, zum Beispiel bei einer Umfrage welche Personen befragt wurden und in welchem Zeitraum, werden als deskriptive Metadaten zusammengefasst. Entsprechend unterscheiden sich diese Metadaten sehr stark je nach Art der beschreibenden Forschungsdaten.

Die hohe Vielfalt der Forschungsdaten spiegelt sich also auch in den Metadaten wieder. Zur Integration verwendet man übergreifende Elemente, die sich auf allgemeine Angaben beschränken, wie zum Beispiel zu beteiligten Personen und Institutionen oder dem Erstellungsdatum, während ansonsten domänenspezifische Elemente erforderlich sind. Diese Heterogenität unterscheidet Metadaten zu Forschungsdaten deutlich von Metadaten in anderen Kontexten, wie zum Beispiel bibliografischen Beschreibungen, und ist eine besondere Herausforderung für die Speicherung und Recherche nach Forschungsdaten.

Identifikatoren für Forschungsdaten

Eine adäquate Referenzierung von Forschungsdaten in Publikationen wird immer stärker gefordert, z.B. durch die DataCite Initiative (Brase, 2009), damit unter anderem die Nachweisbarkeit sowie die Wiederverwendbarkeit ermöglicht werden. Um eine dauerhafte und eindeutige Identifizierung der Daten zu ermöglichen, kann es nicht ausreichend sein, diese einfach auf einem Webserver abzulegen und über einen URL zugreifbar zu machen. Persistente Identifikatoren sind ein Ansatz, der sich auch bei der Referenzierung von Online-Veröffentlichungen bewährt hat, sind. Diese sollten Teil der Metadaten sein und die Forschungsdaten eindeutig identifizieren sowie den Zugriff durch Dritte dauerhaft sichern. Es gibt zahlreiche technische Umsetzungen für solche Identifikatoren, die bekanntesten sind unter anderem DOI (Digital Object Identifier), URN (Uniform Resource Name) und PURL (Persistent Uniform Resource Locators) (Jensen, Katsanidou, & Zenk-Möltgen, 2011).

DOIs sind bei Online-Veröffentlichungen weit verbreitet und abstrahieren das digitale Objekt von seinem konkreten Speicherort in einem Netzwerk. Mit dem *DOI name* (dem eigentlichen Identifikator) können bei *DOI resolvern* Speicherort und Metadaten zu dem Objekt angefragt werden. Im Fall von Änderungen muss die neue Information nur in den Resolvern hinterlegt werden. Bereits jetzt ist es möglich, auch für Forschungsdaten DOIs zu beantragen. Dies kann in Deutsch-

land beispielsweise bei der Initiative DataCite[3] oder bei der Registrierungsagentur für sozialwissenschaftliche Daten da|ra[4] von GESIS[5] geschehen.

Veröffentlichung und Referenzierung von Forschungsdaten

Obwohl es bereits seit Jahren als gute wissenschaftliche Praxis beschrieben wird (Deutsche Forschungsgemeinschaft, 1998), werden Forschungsdaten derzeit nach wie vor selten veröffentlicht. Das hat mehrere Gründe (Weichselgartner, Günther, & Dehnhard, 2011):

Zeit- und Geldaufwand: Um Forschungsdaten zu veröffentlichen, müssen diese entsprechend aufgearbeitet werden. Dies beinhaltet eine aussagekräftige Beschreibung durch Metadaten, die Klärung von rechtlichen Gesichtspunkten und die dauerhafte Bereitstellung in einem geeigneten Repository. Für Wissenschaftler, die sich noch nie mit der Thematik beschäftigt haben, ist das insbesondere bei der ersten Veröffentlichung mit einem hohen Aufwand verbunden.

Nachteil im wissenschaftlichen Wettbewerb: Stehen Daten anderen Wissenschaftlern zur freien Verfügung, so können sie diese entsprechend für ihre eigene Forschung verwenden. Die Ergebnisse können publiziert werden und möglicherweise dem ursprünglichen Ersteller der Forschungsdaten zuvorkommen.

Reputationsverlust: Durch die Veröffentlichung der Daten können Schwächen in der zugehörigen Veröffentlichung sichtbar werden, die ansonsten nicht aufgefallen wären. Obwohl eine solche inhaltliche Auseinandersetzung Kern des wissenschaftlichen Diskurses ist, kann dies dem Ruf der betroffenen Wissenschaftler schaden.

Vorteile werden gesehen bei:

Zugang zu Ressourcen: Es bestehen Förderprogramme für das Veröffentlichen von Forschungsdaten. Da die Daten in der Regel bereits erfasst wurden, können Wissenschaftler auf diese Weise zusätzliche Ressourcen für eine „Zweitverwertung" der geleisteten Arbeit erhalten.

Reputationsgewinn und erhöhte Sichtbarkeit: Mit der Veröffentlichung hochwertiger Daten, die durch Dritte nachgenutzt werden, ist ein Reputationsgewinn verbunden, da auf die Nutzung der Daten in Veröffentlichungen hingewiesen wird. Dadurch erhöht sich automatisch auch die Sichtbarkeit in der Fachcommunity, was insbesondere für Wissenschaftler am Anfang ihrer Karriere relevant sein kann.

3 http://www.datacite.org/
4 http://www.gesis.org/dara/
5 GESIS - Leibniz-Institut für Sozialwissenschaften: http://www.gesis.org

Kooperationsmöglichkeiten: Die Veröffentlichung von Daten kann durch die Nachnutzung Kooperationen zwischen Wissenschaftlern befördern und zu neuen Ansätze und Ideen für die eigene Arbeit führen.

Insbesondere der mögliche Zugewinn an Reputation könnte für Wissenschaftler ausschlaggebend sein, den initialen Aufwand für die Veröffentlichung eigener Daten in Kauf zu nehmen. Dazu bedarf es aber einer Kultur der aktiven Nutzung und sauberen Referenzierung von Daten, die von Dritten bereitgestellt werden. In Fachdisziplinen wie der Teilchenphysik, die auf die gemeinsame Nutzung von Daten aus extrem aufwendigen Experimenten angewiesen sind, ist dies bereits heute der Fall. In anderen Disziplinen vollzieht sich der Wandel langsamer und in der individuellen Abwägung werden die Vorteile für die wissenschaftliche Community (Vermeidung von Doppelarbeit, Nachvollziehbarkeit von Ergebnissen) die Nachteile für den Einzelnen nicht ausgleichen können.

Neben der Möglichkeit der Förderung von Datenveröffentlichungen können auch die Herausgeber von wissenschaftlichen Zeitschriften und Konferenzbänden direkt dazu beitragen, dass mehr Daten veröffentlicht werden: Autoren werden verpflichtet, ihre Daten im Fall einer Veröffentlichung bereitzustellen. Diese Möglichkeit wird allerdings noch nicht übergreifend genutzt, wie eine Untersuchung im Rahmen des EDaWaX-Projekts[6] mit wirtschaftswissenschaftlichen Zeitschriften zeigt: Von den 141 untersuchten Zeitschriften verfügen nur 20% über Richtlinien, die das Einreichen der Forschungsdaten zu einem gewissen Zeitpunkt einfordern. Oftmals werden zudem nur die Datensätze selbst gefordert, aber zum Beispiel nicht zwingend die Programme, die zum Erstellen oder Ausführen benötigt werden (Vlaeminck, 2012). Ohne die dazugehörigen Programme kann jedoch eine Rekonstruktion eines Experiments nicht gewährleistet werden.

LOD als Infrastruktur

Um LOD als Infrastruktur für Forschungsdaten verwenden zu können, müssen die LOD-Prinzipien (Bizer, Heath, & Berners-Lee, 2009) entsprechend verfolgt werden. Um Forschungsdaten per URI identifizierbar zu machen, wird zuerst pro Ressource ein eindeutiger Identifier benötigt. Beispielsweise kann dafür ein Schlüssel aus der Datenbank genommen werden, falls die Daten bereits in einer Datenbank gespeichert sind. Basierend auf diesem Identifier kann eine HTTP URI pro Ressource erstellt werden. Normalerweise besteht der URI aus einem festen Teil, der für alle Daten gleich ist und dem Identifier.

6 http://www.edawax.de

Ein Beispiel wäre http://link.bib.uni-mannheim.de/primo/ als fester Teil und MAN_ALEPH001437414 als interner Identifier. Zusammen würde http://link.bib.uni-mannheim.de/primo/MAN_ALEPH001437414 einen URI ergeben, die die entsprechende Ressource referenziert. Beim Erstellen der URIs sollte darauf geachtet werden, dass man selbst im Besitz dieser URIs ist und diese nicht bereits anderweitig belegt sind. Ist die Ressource bereits mit einer DOI versehen, so kann dieser Identifier entsprechend verwendet werden.

Damit die Daten in den standardisierten Formaten vorliegen, müssen gegebenenfalls einige Transformationen ausgeführt werden. Zuerst muss man sich für ein Vokabular entscheiden, in dem die Forschungsdaten beschrieben werden.

Inzwischen sind bereits einige Vokabularien vorhanden, die wiederverwendet werden. Diese können sehr generell oder spezialisiert sind. Beispielweise können mit Hilfe des Vokabulars Dublin Core[7] die Kernmetadaten ausgedrückt werden, die alle Ressourcen teilen, zum Beispiel einen Titel und einen Autor.

Im Gegensatz dazu gibt es auch Vokabulare, die auf Forschungsdaten aus bestimmten Bereichen spezialisiert sind, wie die DDI[8] für Forschungsdaten aus den Sozialwissenschaften.

Tabelle 2: Metadaten.

Metadaten Feld	Inhalt
dc.contributor.creator	GESIS
dc.identifier.uri	doi:10.4232/1.1000
dc.subject.keyword	Gesellschaft
dc.title	Allgemeine Bevölkerungsumfrage der Sozialwissenschaften ALLBUS 1980
ddi.questionScheme.questionitem.literalText.text	Wie beurteilen Sie heute Ihre eigene wirtschaftliche Lage?

Beispiele dafür sind in **Tabelle 2** aufgezeigt. Mit Dublin Core lassen sich grundlegende Informationen festhalten wie der Ersteller, der Identifier, Schlagwörter oder auch der Titel. Allerdings kann man nur mit spezielleren Vokabularien wie DDI die Forschungsdaten genauer beschreiben. Wie in der Tabelle gezeigt, kann man entsprechend zu der Umfrage ALLBUS auch direkt die Fragen festhalten, in diesem Fall, wie die Befragten ihre eigene wirtschaftliche Lage bewerten.

7 http://dublincore.org/
8 http://www.ddialliance.org/

Zusätzlich gibt es bereits Ontologien, die zur Verknüpfung der Forschungsdaten mit anderen Datensätzen verwendet werden können, wie die CiTO[9] Citation Typing Ontologie. Diese bietet ein reichhaltiges Vokabular, um die Beziehungen zwischen Dokumenten über das reine „A zitiert B“ zu qualifizieren. So kann zum Beispiel ausgedrückt werden, dass der Autor eines Dokuments dem Inhalt eines anderen widerspricht (Propertiy `cito:disagreesWith`). Um die Beziehungen zwischen Dokumenten und Daten auszudrücken, gibt es in der Ontologie zwei zueinander inverse Property-Paare: `cito:citesAsDataSource` und `cito:isCitedAsDataSourceBy` sowie `cito: providesDataFor` und `cito:usesDataFrom`.

Falls keins der zur Verfügung stehenden Vokabularien eine passende Beschreibung enthält, ist es jeder Zeit möglich, selbst entsprechende Spezifikation zu erstellen. Werden die bereits vorhandenen Vokabularien genutzt, so können Verknüpfungen zu anderen Daten einfacher gefunden werden.

Sobald das Vokabular definiert ist, kann man die Forschungsdaten damit beschreiben. Liegen die Daten bereits in einem anderen Format vor, zum Beispiel in einer Datenbank, so kann man diese mit Hilfe von Transformationen in das gewünschte Format überführen. Dieser Schritt kann abhängig vom aktuell verwendeten Format einen gewissen Aufwand erfordern. Speziell wenn dieses Format in keiner Weise standardisiert ist, kann es aufwendig sein, die enthaltenen Informationen zu extrahieren.

Anschließend können die Daten entsprechend unter dem jeweils dafür erstellten URI verfügbar gemacht werden. Dabei sollte darauf geachtet werden, dass die Informationen über die Ressource auch menschen- und nicht nur maschinenlesbar sind. Wenn ein Benutzer an diesen Informationen interessiert ist, dann sollte es zum Beispiel nicht nötig sein, RDF zu kennen und zu wissen, wie man damit umgeht.

Als Letztes können noch die Verknüpfungen zu anderen Datensätzen eingefügt werden. Diese Verknüpfungen können sehr unterschiedlich sein. Beispielsweise sind Verknüpfungen zu Publikationen, aber auch zu weiteren Informationen über die Stadt, in der die Forschungsdaten erhoben wurden, denkbar. All diese Verknüpfungen helfen zum einen, die Daten entsprechend im Web zu finden, zum anderen dienen sie dem Benutzer dazu, mehr Informationen zu erhalten. Sind diese Verknüpfungen nicht angegeben, so muss sich der Benutzer gegebenenfalls diese Informationen selbst in mühsamer Arbeit zusammensuchen.

9 http://purl.org/spar/cito/ (Namensraum cito:)

Forschungsdaten in Recherchesystemen

Wie im Forschungszyklus beschrieben, beginnt jede Forschung mit einer Recherche vorhandener Publikationen. Jedoch können nicht nur die Publikationen, sondern auch die Forschungsdaten selbst für den Wissenschaftler in dieser Phase hilfreich sein. Gibt es beispielsweise schon Studien oder Umfragen zu einem Thema, so könnte der Forscher diese je nach Verfügbarkeit selbst nutzen, zum Beispiel um einen bestimmten Aspekt zu analysieren. Meist sind die Forschungsdaten sowie schon ihre Metadaten und Beschreibungen in einer Vielzahl von verschiedenen Systemen unterschiedlicher Institutionen gespeichert. Entsprechend schwierig ist es daher für Wissenschaftler, die passenden Daten zu finden. Deshalb kann es hilfreich sein, die Forschungsdaten direkt in ein Recherchesystem, zum Beispiel einer Bibliothek, zu integrieren, in dem der Forscher sowieso nach existierenden Publikationen recherchiert.

Um Forschungsdaten über Systemgrenzen hinweg zu finden, ist eine Möglichkeit, eine Metasuche anzubieten, die wiederum in den einzelnen Systemen nach entsprechenden Daten sucht und die Ergebnisse zusammenfasst. Allerdings müsste man dafür zuerst die Metasuche entwickeln und hat ein zusätzliches System, in dem nur Forschungsdaten gefunden werden können. Möchte man die Forschungsdaten in bereits existierende Systeme, z.B. in Bibliothekskataloge, integrieren, so gibt es generell zwei Möglichkeiten: server- und clientseitige Anreicherung (Ritze & Eckert, 2012).

Serverseitige Anreicherung

Bei einer serverseitigen Anreicherung werden die Beschreibungen der Forschungsdaten direkt in das Recherchesystems, genauer gesagt die Datenbank des Systems, geladen. Im Normalfall reicht es dabei nicht aus, die Daten in ein Format zu überführen, das vom Recherchesystem verarbeitet werden kann. Oftmals gibt es in einem System, das auf Publikationen spezialisiert ist, keine entsprechenden Felder, mit denen die Daten adäquat ausgedrückt werden können. Es gibt zwar einige Gemeinsamkeiten, wie Urheber, Titel und Jahr, aber auch spezifische Informationen, zum Beispiel den Erhebungszeitraum einer Studie. Speziell wenn man an Modelle aus Simulationen denkt, wird es recht schwer, diese Daten ohne Erweiterung im Recherchesystem abzubilden. Dementsprechend muss man das System teilweise sehr anpassen, um alle Informationen laden zu können. Allerdings sind die Forschungsdaten nach dem Laden Bestandteil des Katalogs und können analog zu allen anderen Daten verwendet werden.

Ein besonderes Problem der serverseitigen Anreicherung ist, dass die erforderlichen Schritte nicht von einem auf ein anderes System übertragen werden können. Je nach Komplexität kann es sogar durchaus sein, dass eine solche Anpassung noch nicht einmal 1:1 auf eine andere Instanz des gleichen Systems übertragen werden kann, da durch die grundsätzlich vorhandenen Anpassungsmöglichkeiten der Systeme einrichtungsspezifische Unterschiede bestehen können. Ein weiterer Nachteil kann auch die notwendige Datenreplikation sein, speziell die daraus resultierende Frage nach der Aktualität. Entsprechend müsste ein manueller Updatemechanismus eingebunden werden. Grundlegende Änderungen zum Beispiel am Format oder den anzuzeigenden Daten erzeugen einen relativ hohen Wartungsaufwand. Nicht zuletzt müssen ggf. auch Lizenzvereinbarungen beachtet werden, wenn externe Daten auf diese Weise dupliziert werden.

Auf der anderen Seite ist es vorteilhaft, dass die externen Informationen direkt in die Datenbank geladen werden. So verhalten sie sich wie jegliche andere Art von Daten im System und können dadurch auch recherchierbar gemacht werden. Auch andere Funktionen wie die Facettierung können genutzt werden. Eine serverseitige Anreicherung ist generell erforderlich, wenn Forschungsdaten direkt im System wie alle anderen Ressourcen gesucht werden sollen. Am einfachsten wäre eine generelle Integrierbarkeit jeglicher Art von LOD in das System. Dies ist allerdings bei den aktuellen Systemen, zum Beispiel Primo von Ex Libris[10] , soweit wir wissen, nicht möglich. Sonst wäre es gegebenenfalls möglich, die Daten ohne jegliche Transformation direkt zu integrieren.

Clientseitige Anreicherung

Unter einer clientseitigen Anreicherung verstehen wir, dass Daten erst zur Laufzeit im Client eingebunden werden, im Fall von Primo also zum Beispiel per JavaScript im Browser des Benutzers. Dies könnten zum Beispiels Links zu Forschungsdaten sein, die für den Benutzer entsprechend der aktuellen Suche relevant sein könnten. Eine Transformation der Daten in das interne Format des Recherchesystems ist deshalb nicht nötig.

Natürlich bedeutet eine Anreicherung der Daten in der Präsentationsschicht, dass die zusätzlichen Daten nicht direkt für die Recherche genutzt werden können. Allerdings lässt sich damit dennoch die Recherche unterstützen.

Die clientseitige Anreicherung bietet den Vorteil, dass zusätzliche Informationen auf einfache Weise zur Verfügung gestellt werden können, auch ohne direkte Unterstützung des Recherchesystems. Bei geschickter Implementierung kann die

10 http://www.exlibrisgroup.com/

systemabhängige Einbindung in den Client von der eigentlichen Umsetzung der Anreicherung getrennt werden, so dass eine weitgehende Systemunabhängigkeit erreicht wird. Durch die direkte Einbindung der externen Datenquellen stehen auch stets die aktuellsten Daten zur Verfügung.

Für die Frage, ob eine client- oder serverseitige Anreicherung verwendet wird, kann auch die gewünschte Präsentation für den Benutzer eine Rolle spielen. Viele clientseitige Anreicherungen verzögern die Anzeige. Zwar können die Daten asynchron im Hintergrund geladen werden, doch die Verzögerung, mit der die zusätzlichen Daten dann auf der Seite eingeblendet werden, ist für den Benutzer wahrnehmbar und unter Umständen nicht gewünscht.

Ähnlich wie bei der serverseitigen Anreicherung gibt es auch für diesen Ansatz Anwendungsfälle. Ein sehr typisches Beispiel ist das Einbinden von Wikipedia-Artikeln zu Publikationen oder Autoren. Die Informationen werden normalerweise nicht für die Recherche verwendet, sondern lediglich dem Benutzer bei Bedarf zur Verfügung gestellt, um dem Benutzer zu helfen.

Bei welchem Anwendungsfall welche Art der Anreicherung am besten verwendet wird, hängt demnach von einigen Faktoren ab. Diese sollten von der jeweiligen Institution, die eine Anreicherung plant, im Voraus betrachtet und untersucht werden.

Im Kapitel „Datenanreicherung auf LOD-Basis“ wird der Aspekt der Kataloganreicherung noch genauer besprochen, allerdings ohne Fokus auf Forschungsdaten.

Praktische Anwendung im InFoLiS-Projekt

Das Projekt „Integration von Forschungsdaten und Literatur in den Sozialwissenschaften“, kurz *InFoLiS*,[11] ist ein DFG-gefördertes Projekt dreier Kooperationspartner. Dazu gehören die GESIS – Leibniz-Institut für Sozialwissenschaften, die Universitätsbibliothek Mannheim und der Lehrstuhl für Künstliche Intelligenz an der Universität Mannheim.

Obwohl die Wichtigkeit von Forschungsdaten für die Forschung und die wissenschaftliche Informationsversorgung bekannt ist, ist der Umgang mit Forschungsdaten oftmals nicht adäquat. Zum einen gibt es sehr viele verschiedene Systeme, in denen jeweils einzelne Forschungsdaten gespeichert sind, und zum anderen sind die Forschungsdaten weder untereinander noch mit den resultie-

11 http://www.gesis.org/en/research/external-funding-rojects/projektuebersicht-drittmittel/infolis/

renden Publikationen verknüpft. Dies stellt besonders bei den empirisch ausgerichteten Wissenschaften wie den Sozialwissenschaften eine unvorteilhafte Situation dar. Sind Wissenschaftler daran interessiert, Forschungsdaten zu finden, auf denen Publikationen basieren, müssen einige Schritte ausgeführt werden.

Zuerst muss der Wissenschaftler manuell in den Publikationen nach Referenzen zu Forschungsdaten suchen. Anders als bei Zitierungen gibt es normalerweise keine Liste mit verwendeten Daten. Zusätzlich werden Forschungsdaten nicht standardisiert referenziert. Sie können unter anderem im Text selbst, in Fußnoten oder auch Bildunterschriften genannt sein. Außerdem werden häufig synonyme Namen oder Abkürzungen zum Referenzieren verwendet, wie zum Beispiel „ALLBUS“ für die „Allgemeine Bevölkerungsumfrage der Sozialwissenschaften“. Deshalb kann es notwendig sein, dass der Wissenschaftler die Publikation Satz für Satz durchlesen muss, um die entsprechenden Referenzen zu finden. Im Anschluss muss der Wissenschaftler das richtige System finden, in dem die entsprechenden Forschungsdaten gespeichert sind. Durch die vielen unterschiedlichen Namen für Forschungsdaten und den zahlreichen Systemen stellt dies neben dem Finden der Referenzen selbst ein weiteres Problem dar.

Im InFoLiS-Projekt haben sich daher die Kooperationspartner zum Ziel gesetzt, zum einen Referenzen zu Forschungsdaten automatisch zu erkennen und so Forschungsdaten (zumindest ihre Metadaten) zu einer Publikation über das Recherchesystem der Universitätsbibliothek Mannheim direkt zugänglich zu machen sowie umgekehrt Publikationen (zumindest ihre Metadaten) über das System der GESIS. Unabhängig vom genutzten System soll es möglich sein, Forschungsdaten sowie Publikationen gleichzeitig zu suchen und die entsprechenden Verknüpfungen angezeigt zu bekommen. Somit kann die Arbeit der Wissenschaftler enorm erleichtert und zusätzlich die Transparenz der Forschung erhöht werden.

Um Referenzen von Forschungsdaten in Publikationen zu finden, haben wir einen entsprechenden Algorithmus entwickelt. Dieses sogenannte Bootstrapping Verfahren versucht Muster zu erkennen, die typisch für die Referenzierung von Forschungsdaten, hier insbesondere Studien, sind. Mit diesen Mustern können neue Studien gefunden werden, die gegebenenfalls wieder neue Muster liefern (Boland, Ritze, Eckert, & Mathiak, 2012). Dabei verwenden wir insbesondere die Volltexte der Publikationen, da Referenzen andernorts, zum Beispiel in Titeln oder Abstracts, kaum zu finden sind. Das reine Suchen nach Studiennamen schlägt zudem meist fehl, da es zum einen keine vollständige Liste mit allen verfügbaren Studien gibt und zum anderen oftmals Abkürzungen oder alternative Namen verwendet werden.

Nachdem wir diesen Algorithmus auf unsere Daten angewendet haben, sollen die gefundenen Verknüpfungen als LOD veröffentlicht werden. Somit

kann jedes System diese Verknüpfungen in das eigene System einfügen. Zusätzlich stehen sie auch allen anderen frei zur Verfügung und können entsprechend nachgenutzt werden.

Die Integration der Daten selbst ist zurzeit noch nicht abgeschlossen. Allerdings haben wir sehr schnell festgestellt, dass die Forschungsdaten aus da|ra[12] entsprechend nicht ohne weiteres in das Bibliotheksystem, in diesem Fall Primo, eingebunden werden können. Eine serverseitige Anreicherung würde demnach sehr viel Aufwand bedeuten, da alle Daten zuerst transformiert werden müssten. Zudem ist das System auf Publikationen ausgelegt und spezielle Informationen zu den Studien können nicht adäquat abgebildet werden. Deswegen arbeiten wir daran, LOD als Infrastruktur zu verwenden.

Schlussfolgerungen

Der Forschungszyklus sollte sich erweitern so wie in **Abbildung 2** dargestellt. Nach der Durchführung des Experiments/der Studie etc. sollten die Forschungsdaten möglichst direkt veröffentlicht werden. Damit können zum Beispiel auch Gutachter von Publikationen direkt überprüfen, ob die Ergebnisse korrekt sind und richtige Schlüsse aus den Daten gezogen wurden. Neben den Publikationen können die Forschungsdaten ebenso für andere Wissenschaftler während der Recherche verwendet werden.

12 http://www.gesis.org/dara/

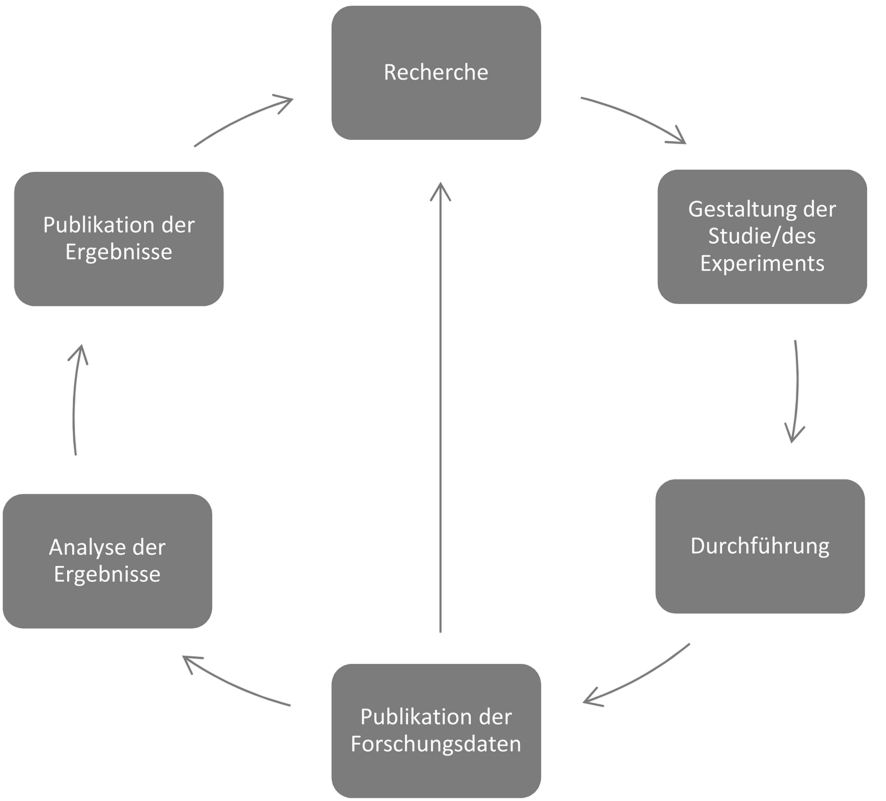

Abbildung 2: Verbesserter Forschungszyklus mit Forschungsdaten.

Literaturverzeichnis

Baskerville, R., & Wood-Harper, A. T. (1996). A Critical Perspective on Action Research as a Method for Information Systems Research. *Journal of Information Technology 11 Nr. 3*, S. 235-246.

Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems (IJSWIS)*, S. 1-22.

Boland, K., Ritze, D., Eckert, K., & Mathiak, B. (2012). Identifying References to Datasets in Publications. *Theory and Practice of Digital Libraries*, S. 150-161. Berlin Heidelberg: Springer.

Brase, J. (2009). DataCite - a global registration agency for research data. *Eleventh Interlending and Document Suppy Conference*, S. 257-261. New York: IEEE.

Deutsche Forschungsgemeinschaft. (1998). *Sicherung Guter Wissenschaftlicher Praxis*. Weinheim: Wiley VCH Verlag.

Edinburgh University. (2011). *Edinburgh University Data Library Research Data Management Handbook*. Abgerufen am 13. Mai 2013 von http://www.docs.is.ed.ac.uk/docs/data-library/EUDL_RDM_Handbook.pdf

Gold, A. (2007). Cyberinfrastructure, Data, and Libraries, Part 1: A Cyberinfrastructure Primer for Librarians. *D-Lib Magazine 13 Nr. 9*. Abgerufen am 13. Mai 2013 von http://www.dlib.org/dlib/september07/gold/09gold-pt1.html

Jensen, U., Katsanidou, A., & Zenk-Möltgen, W. (2011). Metadaten und Standards. In S. Büttner, H.-C. Hobohm, & L. Müller, *Handbuch Forschungsdatenmanagement*, S. 83-100. Bad Honnef: BOCK + HERCHEN Verlag.

Neuroth, H., Strathmann, S., Oßwald, A., Scheffel, R., Klump, J., & Ludwig, J. (2012). *Langzeitarchivierung von Forschungsdaten. Eine Bestandsaufnahme.* Boizenburg: Verlag Werner Hülsbusch.

Razum, M. (2011). Systeme und Systemarchitekturen für das Datenmanagement. In S. Büttner, H.-C. Hobohm, & L. Müller, *Handbuch Forschungsdatenmanagement*, S. 123-138. Bad Honnef: BOCK + HERCHEN Verlag.

Ritze, D., & Eckert, K. (2012). Data Enrichment in Discovery Systems using Linked Data. *GfKl 2012*. to be published.

Vlaeminck, S. (2012). *Wirtschaftswissenschaftliche Forschungsergebnisse replizierbar machen-das Projekt EDaWaX*. Abgerufen am 13. Mai 2013 von urn:nbn:de:0290-opus-12794

Weichselgartner, E., Günther, A., & Dehnhard, I. (2011). Archivierung von Forschungsdaten. In S. Büttner, H.-C. Hobohm, & L. Müller, *Handbuch Forschugnsdatenmanagement*, S. 191-202. Bad Honnef: BOCK + HERCHEN Verlag.