Handbuch Forschungsdatenmanagement

Herausgegeben von
Stephan Büttner, Hans-Christoph Hobohm, Lars Müller

BOCK + HERCHEN Verlag
Bad Honnef
2011

ISBN 978-3-88347-283-6

BOCK+HERCHEN Verlag, Bad Honnef
Printed in Germany

1.2 Der Lebenszyklus von Forschungsdaten

Stefanie Rümpel
Fachhochschule Düsseldorf

Für Wissenschaftler sind Veröffentlichungen unentbehrlich und werden als „Währung“ angesehen. Die Forschungsdaten, auf denen die Publikation basiert, sind aber i.d.R. nicht enthalten.

> „Mit beginnender Analyse und Interpretation von Daten, werden unter Umständen nicht mehr alle Details eines Rohdatensatzes transportiert. […] [Beispielsweise] werden zusammengeführte Einzelmessungen unter Umständen nur noch als Mittelwert dargestellt, obwohl ursprünglich eine ganze Reihe von Forschungsdatensätzen erzeugt wurde […].“ (TIB Hannover, FIZ Chemie Berlin & Universität Paderborn, 2010, S. 26–27)

Doch gerade die Daten sind deutlich interessanter und relevanter für weitere Forschungsprozesse, um einen Mehrwert zu erreichen (Sietmann, 2009, S. 154).

> „ […] 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.“ (Gold, 2007)

Mit anderen Worten: Forschungsdaten nur als Grundlage für eine Publikation zu verwenden, missachtet deren Wert. Gegenwärtig wird im Forschungsprozess meist darauf verzichtet, auf bereits erhobene und gespeicherte Daten zurückzugreifen. Eher werden kostenintensive Messwiederholungen in Kauf genommen.

Die Technische Informationsbibliothek (TIB) formulierte, bezogen auf das Fachgebiet Chemie, die Missstände des Forschungsdatenmanagements deutlich.

> „Der bisherige Umgang mit Forschungsdaten in der Chemie beinhaltet keine allgemein anerkannten Standards hinsichtlich einer Nachnutzbarkeit oder langfristigen Verfügbarkeit. Überwiegend existiert keine Qualitätssicherung, keine gesicherte Langzeitarchivierung, kein gesicherter Nachweis sowie keine Erschließung der Forschungsdaten und somit keine Datensicherheit.“ (TIB Hannover, FIZ Chemie Berlin & Universität Paderborn, 2010, S. 5)

Diese Vorgehensweise ist vorherrschend, da Daten aus vielen Forschungsprozessen i.d.R. noch gar nicht dauerhaft gespeichert oder aufbereitet werden. Gründe sind sowohl auf Seiten der Wissenschaftler als auch der Institutionen zu finden, beispielsweise Unwissenheit über persistente und qualitative Verwaltung der Daten, Hemmnisse bezüglich der Datenspeicherung oder fehlende Transparenz der gespeicherten Daten in Repositorien.

Um die Distanz von Wissenschaftlern und allen Involvierten im Forschungsprozess gegenüber der Aufbereitung von Daten zu mindern, erscheint es wesentlich, das Bewusstsein der Wissenschaftler für die Notwendigkeit einer Nachvollziehbarkeit der Forschung zu fördern. Dabei muss beachtet werden, dass die dauerhafte Speicherung, Pflege und Bereitstellung von Forschungsdaten einen erheblichen Arbeitsaufwand erfordert.

Die Stimmen nach einem verantwortungsvollen und organisierten Umgang mit Forschungsdaten werden immer lauter:

> „Einhergehend mit der Bearbeitung von Forschungsdaten steigt die Gefahr von Fehlern und Fehlinterpretationen. Umso komplexer die Experimente, Datenstrukturen und Fragestellungen, desto relevanter wird die Verfügbarkeit von ursprünglichen Forschungsdaten, um Ergebnisse kritisch zu evaluieren. Daher ist der öffentliche Zugang zu den Forschungsdaten im wissenschaftlichen Erkenntnisgewinn eminent.“ (TIB Hannover, FIZ Chemie Berlin & Universität Paderborn, 2010, S. 27)

Der gegenwärtige Lebenszyklus von Forschungsdaten sieht jedoch anders aus.

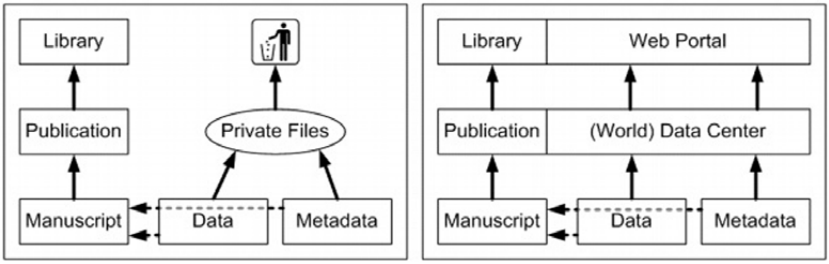

Abb. 1: *(links) Schematische Darstellung des wissenschaftlichen Informationsflusses in der Forschung (= bekannter Weg), (rechts) Veränderter Umgang mit Daten* (Klump et al., 2006, S. 80 nach Helly, Staudigel & Koppers, 2003, S. 2)

> „[...] Forschungsdaten sind die Grundlage jeglicher wissenschaftlicher Arbeit. Ausgehend vom Experiment durchlaufen Forschungsdaten viele, dem Wissenschaftler bekannte Stadien, die letztendlich als Erkenntnisgewinn in einer wissenschaftlichen Publikation münden. Danach verliert sich der bis dahin so klare Weg der Forschungsdaten, was deren Dokumentation, langfristige Speicherung oder Nachnutzbarkeit für andere Wissenschaftler betrifft.“ (TIB Hannover, FIZ Chemie Berlin & Universität Paderborn, 2010, S. 7)

Der Anspruch ist, die Daten aus den „Papierkörben“ der Forscher heraus in das Licht der Öffentlichkeit zu bringen. Sicher, die Speicherung von vielen einzelnen Daten ist arbeitsintensiver als die Speicherung einer einzelnen Text-Publi-

kation. Sie besitzen außerdem eine enorme Heterogenität und Komplexität, wodurch sie zu „[…] eine[r] wertvolle[n], jedoch schwierig zu handhabende[n] Ressource […]“ (NESTOR, 2009, S. 1) werden. Erforderlich ist also eine konsequent qualitative und persistente Verwaltung.

Hilfe bei der komplexen Verwaltung von Forschungsdaten gibt deren Lebenszyklus. Dieses wird im Folgenden mit Hilfe von zwei Modellen gezeigt. Beide beschreiben die verschiedenen Lebensphasen von Forschungsdaten, betrachten dies jedoch aus verschiedenen Blickwinkeln. Im ersten Modell werden die theoretischen Anforderungen an den Umgang mit Daten aufgeführt, im anderen die notwendigen technischen Bedingungen im Laufe des Lebenszyklus benannt.

1.2.1 Curation-Lifecycle-Model

Um einen Mehrwert von Forschungsdaten zu erhalten, ist eine adäquate Verwaltung notwendig. Ihr Lebenszyklus erstreckt sich über verschiedene Phasen, die von der Entstehung in wissenschaftlichen Arbeitsprozessen bis zur nachnutzbaren Archivierung reichen. Die Anforderungen an das Management von Forschungsdaten gehen weit über die Langzeitarchivierung hinaus (NESTOR, 2009, S. 1–2). Alle Tätigkeiten des Forschungsdatenmanagements werden durch das „*Curation Lifecycle Model*“, erstellt vom *Digital Curation Centre* (DCC), identifiziert (DCC, 2010).

> „It is important to note that the model is an ideal. In reality, users of the model may enter at any stage of the lifecycle depending on their current area of need.“ (DCC, 2010)

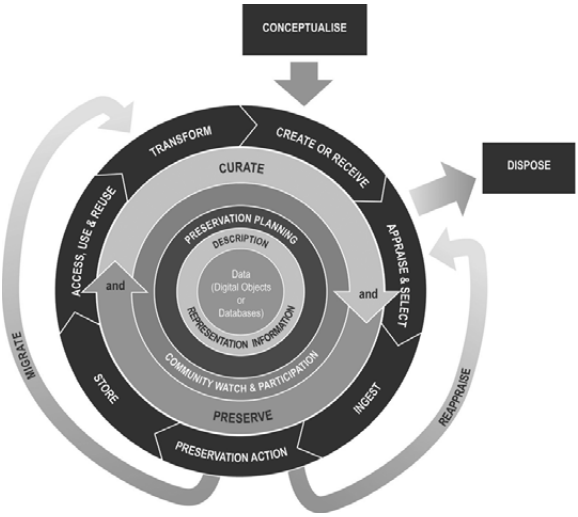

Abb. 2: *Curation Lifecycle Model* (DCC, 2010)

Die Abbildung zeigt ein sich aus mehreren Ebenen zusammensetzendes Kreismodell. Mittig sind die Tätigkeiten angeordnet, welche Daten während des gesamten Lebenszyklus begleiten: „*Data Preservation*“ (Datenerhaltung) und „*Data Curation*“ (Datenpflege). Beide ergänzen sich und bilden die Kernprozesse der *Digital Curation*. Diese Arbeiten müssen im gesamten Lebenszyklus von Forschungsdaten erfolgen. *Preservation* bezieht sich auf die Bewahrung der Daten im Sinne der digitalen Langzeitarchivierung. Um Daten nutzbar zu gestalten und zu behalten, wird eine Pflege notwendig, subsumiert unter dem Begriff „*Data Curation*“.

Die sequenziellen Tätigkeiten sind im äußeren Kreis dargestellt. Mit der *Konzeption* des Forschungsvorhabens erfolgt der Einstieg in den Kreislauf. Bereits vor der eigentlichen Forschungstätigkeit sind Überlegungen bezüglich der anfallenden Daten in dem Modell integriert. Die Deutsche Forschungsgemeinschaft (DFG) fordert beispielsweise seit kurzem die Berücksichtigung des Forschungsdatenmanagements bereits bei Beantragung von Forschungsvorhaben. Vor dem Start der Forschung müssen nun alle relevanten Fragen bezüglich des Umgangs mit Forschungsdaten beantwortet werden (Winkler-Nees, 2010, S. 23).

Der nächste Schritt der Datenverwaltung ist die *Datenerstellung* und die *Datenübernahme*. Der Punkt „*Datenübernahme*“ macht klar, dass es sich um einen Zyklus handelt bzw. handeln kann. In dieser „Lebensphase“ kann nach erhobenen Forschungsdaten recherchiert und diese im eigenen Forschungsprozess übernommen werden. Die Daten sind aber nur dann wiederverwendbar, wenn deren Anreicherung mit Informationen so umfassend ist, dass sie transparent werden. Wo, wann und wie sind die Daten erhoben worden? In vielen wissenschaftlichen Fächern ist es gegenwärtig jedoch nicht möglich nach passenden Forschungsdaten zu recherchieren, weil es keine ausreichenden Übersichten über die vorliegenden Daten gibt. Der Schwerpunkt bei der Verwaltung liegt gegenwärtig noch auf der Verwaltung von neu erhobenen Daten.

Nicht alle Daten, die erhoben wurden, müssen gespeichert werden.

> „Derzeit werden in den meisten Institutionen alle Primärdaten so lange gespeichert, bis diese irgendwann schleichend verloren gehen.“ (Severiens & Hilf, 2006, S. 29)

Es muss eine *Bewertung* erfolgen, welche Daten speicherwürdig sind. Daran schließt sich die *Selektion* jener Forschungsdaten an, die letztendlich gespeichert werden. Daten, die bei der Prüfung nicht als speicherwürdig erachtet wurden, können im Sinne der Richtlinien bzw. rechtlichen Anforderungen aussortiert werden. Die DCC bezeichnet diesen Vorgang als „*Dispose*“.

> „[…] im Laborbetrieb [wird] eine große Menge an Forschungsdaten produziert, die eher in den Bereich der Qualitätskontrolle von laufenden Prozessen fallen und nicht relevant für Publikationen sind. Für solche

> Forschungsdaten ist eine Speicherung in institutionellen Repositorien vorstellbar. Erst bei der Zusammenfassung von wissenschaftlichen Ergebnissen und deren Aufbereitung für eine Veröffentlichung werden Forschungsdatensätze für die Untermauerung wissenschaftlicher Erkenntnisse und Thesen herangezogen. Solche Forschungsdaten sind von Relevanz für die langfristige Speicherung und öffentliche Zugänglichkeit.“ (TIB Hannover, FIZ Chemie Berlin & Universität Paderborn, 2010, S. 30)

Mit der Speicherung werden Maßnahmen zur *Preservation* notwendig.

> „Preservation actions should ensure that data remains authentic, reliable and usable while maintaining its integrity. Actions include data cleaning, validation, assigning preservation metadata, assigning representation information and ensuring acceptable data structures or file formats.“ (DCC, 2010)

Nach der Durchführung der *Preservation* schließt sich die „Langzeitspeicherung“ an. Dabei muss der *Zugriff* auf die Daten bzw. auch die *Benutzung* oder die sich daraus resultierende *Wiederverwendung* gewährleistet sein.

Die Komplexität des Datenmanagements wird offensichtlich. Gegenwärtig existieren Bereiche der Wissenschaft, in denen die Datenverwaltung schon recht gut funktioniert. In anderen wiederum trifft man beim Thema „Datenverwaltung“ auf Diskrepanzen als auch auf Vorbehalte.

Auch im folgenden zweiten Modell werden diese Komplexität sowie die Notwendigkeit von Fachpersonal für die Umsetzung des Forschungsdatenmanagements deutlich.

1.2.2 Data Curation Continuum

Treloar und Harboe-Ree (2008) veranschaulichten in ihrem Model „*Data Curation Continuum*“, das an der *Monash University* in Australien entwickelt wurde, die unterschiedlichen Phasen im Lebenszyklus von Forschungsdaten. Der Forschungsprozess wurde in diesem Modell in drei Domänen unterteilt. Es illustriert, dass jeder Bereich unterschiedliche, teilweise gegensätzliche Ansprüche besitzt. Teils werden sogar verschiedene Technologien erforderlich.

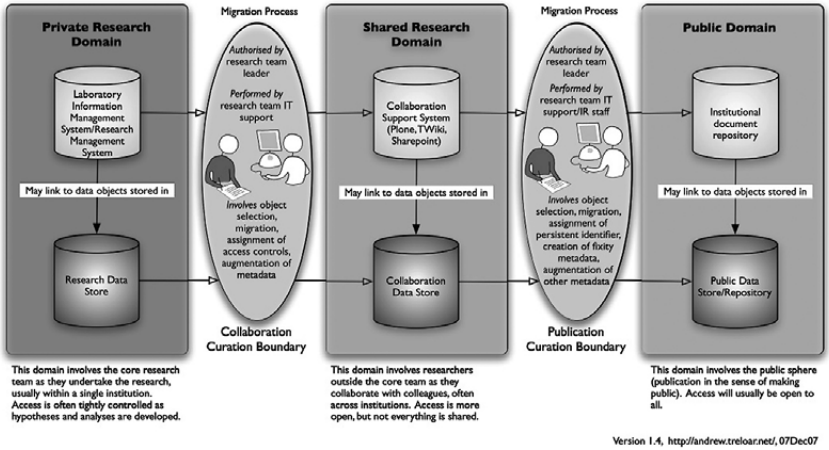

Abb. 3: *Data Curation Continuum* (Treloar & Harboe-Ree, 2008, S. 6)

Daten werden in einer Forschergruppe, der sog. „*Private Research Domain*“ erzeugt. Für die Arbeit in dieser Phase werden „*research management*“ Systeme benötigt, die einen Überblick über die gesamten Datenbestände geben. Ab der Entstehungsphase müssen Sie mit einem hohen Zugriffsschutz und Metadaten versehen werden. Metadaten ergeben sich einerseits durch die gerätespezifische Generierung, andererseits vergibt der Wissenschaftler zusätzlich Metadaten, um eine persönliche Verwaltung zu erhalten (TIB Hannover, FIZ Chemie Berlin & Universität Paderborn, 2010, S. 32). Die Datenspeicherung erfolgt in einem „*Research Data Store*“. Sind die Forscher bereit, Teilergebnisse ihrer Forschung anderen für Analysen zugänglich zu machen, erfolgt eine Migration in die sog. „*Shared Research Domain*“. Dies ergibt sich beispielsweise, wenn dem Vorgesetzen oder kooperierenden Wissenschaftlern die bisherige Forschung präsentiert wird (TIB Hannover, FIZ Chemie Berlin & Universität Paderborn, 2010, S. 32). Für diesen Austausch, auch bezeichnet als „*Data Sharing*“, werden Systeme notwendig, die eine kollektive Arbeitsweise unterstützten, wie Plone oder TWiki. Die Datenobjekte selbst befinden sich in Repositorien. Somit kann eine starke Strukturierung der Datensammlungen erfolgen und ausgefeilte Zugriffsrechte formuliert werden. Mit dem Abschluss der Forschungstätigkeiten erfolgt die Migration zur sog. „*Public Domain*“. Die fertigen Forschungsergebnisse (beispielsweise Publikationen) werden in die institutionellen Repositorien migriert, ein bekannter Prozess. Zusätzlich muss eine Verlinkung auf die mit *Digital Object Identifiers* (DOI) und Metadaten versehene Datenobjekte erfolgen, die sich in einem öffentlichen Forschungsdaten-*Repository* befinden. (Treloar & Harboe-Ree, 2008, S. 5–7)

Elementar ist die Anreicherung der Daten mit vollständigen Metadaten, damit eine Recherche, Identifizierung und Wiederverwendbarkeit eindeutig gewährleistet wird (TIB Hannover, FIZ Chemie Berlin & Universität Paderborn, 2010, S. 33).

Es ist möglich, die Gesamtheit des Forschungsprozesses durch die Nutzung eines *Repository* umzusetzen. Gegenwärtig gibt es jedoch unterschiedliche Anforderungen in den verschiedenen wissenschaftlichen Domänen, wodurch die technische Umsetzung in der Praxis komplex ist (Treloar & Harboe-Ree, 2008, S. 7). Bisher werden grundsätzlich institutionelle Repositorien verwendet, die für die Verwaltung und den Nachweis von Dokumenten konzipiert wurden. Diese ebenfalls für das Forschungsdatenmanagements zu verwenden, ist wegen ihrer Inflexibilität schwierig. Für Speicherung und *Curation* von Forschungsdaten muss eine Lösung existieren, die beispielsweise eine variable Vergabe von Metadaten erlaubt, wie es von der jeweiligen Disziplin gefordert wird, um zukünftig die Daten auch nachnutzen zu können. Die meisten Softwarelösungen für Repositorien (OPUS, *eprints* etc.) unterstützen dies noch nicht. Fedora macht hier eine Ausnahme. Mit dem *Open Source Projekt **F**lexible **E**xtensible **D**igital **O**bject **R**epository **A**rchitecture Commons* (FEDORA), entwickelt an der *Cornell University*, steht ein fertiges System zur Verfügung, das beliebige digitale Objekte (Daten, Textdateien, Metadaten, Bilder, Videos, Webseiten etc.) verwalten kann. (siehe Beitrag von Aschenbrenner & Neuroth, Kap. 2.5)

Neben institutionellen und disziplinären Repositorien werden Forschungsdatenspeicher notwendig, die allesamt jedoch miteinander verknüpft arbeiten sollten.

1.2.3 Fazit

Die Darlegung des Lebenszyklus von Forschungsdaten anhand der beiden Modelle verdeutlicht einerseits die Komplexität andererseits aber auch die theoretische Machbarkeit der Speicherung von Forschungsdaten. Wie bereits geschildert, existieren Einrichtungen, in denen das Forschungsdatenmanagement erfolgreich praktiziert wird. Beispielsweise wird Pangaea vom Alfred-Wegener-Institut für Polar- und Meeresforschung (AWI) gemeinsam mit dem Zentrum für Marine Umweltwissenschaften (MARUM) gehostet.

> „The information system PANGAEA is operated as an Open Access library aimed at archiving, publishing and distributing georeferenced data from earth system research. The system guarantees long-term availability of its content through a commitment of the operating institutions.“ (AWI & Center for Marine Environmental Sciences)

Die Diskussionen zu „Forschungsdaten“ und insbesondere deren Management wird weiter dadurch erschwert, da es eine hohe Domänenspezifik gibt, die eine Übertragung auf andere Wissenschaftsdisziplinen nicht per se zulässt.

> „Vielmehr muss es Ziel sein, in Zusammenarbeit mit den Fachgesellschaften disziplinspezifische Ansätze zu entwickeln, die dann prototypisch realisiert werden können. Dabei wird es Disziplinen geben, die wie die Geowissenschaften eine zentrale Datenzentrenstruktur benötigen, aber auch Disziplinen, die unter Verwendung allgemeingültiger Standards individuelle Lösungen in Form von verteilten Repositorien betreiben.“ (TIB Hannover, FIZ Chemie Berlin & Universität Paderborn, 2010, S. 14)

Neben der Entwicklung der notwendigen Techniken und Systeme zur Datenspeicherung gibt es noch einige grundsätzliche Fragen, deren Beantwortung derzeitig noch nicht erfolgte. Sietmann zählt dazu die folgenden:

> „So wirft die Transformation der gesamten Prozesskette von der Erzeugung über die Speicherung bis zur Bewahrung und Pflege von Forschungsdaten, die sich in dem Begriff „*Open Data*“ verdichtet, Fragen über Fragen auf. Wer standardisiert die Metadaten? Wer setzt die „*Data Policies*“? Wie erzeugt man die Anreize, dass Forscher ihre Daten und Programme verfügbar machen? Wer trägt die Aufwendungen, dass sie verfügbar bleiben?“ (Sietmann, 2009, S. 160)

Sicherlich könnten durch den Austausch von Erfahrungen, die Einen von den Anderen lernen. Ein Problem, das alle Disziplinen und Beteiligte betrifft, ist die Frage nach der personellen Umsetzung des Forschungsdatenmanagements. Wissenschaftler werden i.d.R. die Daten nicht verwalten. Wer dabei welche Unterstützung gibt, bzw. ob eine mehr oder weniger vollständige Ausgliederung dieser Arbeit möglich ist, bleibt Gegenstand der Diskussion.

Literaturhinweise

AWI (Alfred Wegener Institute for Polar and Marine Research) & Center for Marine Environmental Sciences. *PANGAEA.* Data Publisher for Earth & Environmental Science. Online: http://www.pangaea.de/about/ [Zugriff am 17.07.2011].

DCC (Digital Curation Centre), 2010. *DCC Curation Lifecycle Model*. Online: http://www.dcc.ac.uk/resources/curation-lifecycle-model [Zugriff am 01.06.2011].

Gold, A., 2007. Cyberinfrastructure, Data, and Libraries, Part I. A Cyberinfrastructure Primer for Librarians. *D-Lib Magazine*, 13(9/10). Online: http://www.dlib.org/dlib/september07/gold/09gold-pt1.html [Zugriff am 25.04.2011].

Helly, J. Staudigel, H. & Koppers, A, 2003. ‚Scalable models of data sharing in Earth sciences'. *Geochemistry, geophysics, geosystems.* G 3, an electronic journal of the earth sciences, 4(1), S. 1–14. Online: http:// www.beamreach.org/research/data_sharing_model_GC2002.pdf [Zugriff am 17.07.2011].

Klump, J. et. al., 2006. Data Publication in the Open Access Initiative. *Data Science Journal*, 5(15 June 2006), S. 79–83. Online: http:// www.mad.zmaw.de/fileadmin/extern/Publications/datapublication.pdf [Zugriff am 17.07.2011].

NESTOR – Kompetenznetzwerk Langzeitarchivierung, 2009. *Digitale Forschungsdaten bewahren und nutzen – für die Wissenschaft und für die Zukunft*. NESTOR Arbeitsgruppe Grid /e-sience und Langzeitarchivierung. (NESTOR-Bericht) Online: http://nbn-resolving.de/nbn:de:0008-2009071031. [Zugriff am 17.07.2011].

Severiens, T. & Hilf, E.R., 2006. *Langzeitarchivierung von Rohdaten – Studie zum Stand vorhandener Forschungsdaten und Rohdaten aus wissenschaftlichen Tätigkeiten: Erfordernisse und Eignung zur Archivierung bzw. Zurverfügungstellung in Deutschland (Primärdaten)*. Online: http:// nbn-resolving.de/urn:nbn:de:0008-20051114018.

Sietmann, R., 2009. Rip. Mix. Publish. Der Wissenschaft steht ein radikaler Wandel im Umgang mit Forschungsdaten bevor. *c't*, (14), S. 154–161.

TIB (Technische Informationsbibliothek) Hannover, FIZ Chemie Berlin & Universität Paderborn, 2010. *Konzeptstudie „Vernetzte Primärdaten-Infrastruktur für den Wissenschaftler-Arbeitsplatz in der Chemie“*. Online:

http://www.tib-hannover.de/fileadmin/projekte/primaer-chemie/Konzeptstudie_Forschungsdaten_Chemie.pdf [Zugriff am 25.04. 2011].

Treloar, A. & Harboe-Ree, C., 2008. *Data management and the curation continuum. How the Monash experience is informing repository relationship*. Online: http://www.valaconf.org.au/vala2008/papers2008/111_Treloar_Final.pdf [Zugriff am 15.05.2011].

Winkler-Nees, S., 2010. *Der Umgang mit Forschungsdaten in Wissenschaft und Lehre*. Bad Honnef. Online: http://www.dfg.de/download/pdf/dfg_magazin/wissenschaftliche_karriere/heisenberg_treffen_2010/forschungsdaten.pdf [Zugriff am 01.06.2011].