Pflichtablieferung von Dissertationen mit Forschungsdaten an die DNB

Anlagerungsformen und Datenmodell

Dirk Weisbrod, Deutsche Nationalbibliothek (bis Jan. 2018, seitdem DIPF) Frankfurt am Main

Zusammenfassung:
Im Rahmen des DFG-Projektes „Elektronische Dissertationen Plus“ (eDissPlus) entwickeln die Humboldt-Universität zu Berlin (HU) und die Deutsche Nationalbibliothek (DNB) Lösungen für eine zeitgemäße Archivierung und Publikation von Forschungsdaten, die im Zusammenhang mit Promotionsvorhaben entstehen. Dabei müssen die unterschiedlichen Anlagerungsformen von Forschungsdaten an eine Dissertation berücksichtigt und in einem Datenmodell abgebildet sowie das von der DNB verwendete Metadatenschema XMetaDissPlus überarbeitet werden. Das ist notwendig, um die Relationen zwischen der Dissertation und den abgelieferten Forschungsdaten-Supplementen sowie den Daten, die auf externen Repositorien verbleiben sollen, nachzuweisen und im Katalog der DNB recherchierbar zu machen. Dieser Beitrag stellt das Datenmodell und die Änderungen im Metadatenschema vor.

Summary:
As a part of the DFG founded project “Electronic dissertations plus” (eDissPlus), the Humboldt University of Berlin (HU) and the German National Library (DNB) develop a prototype for an integrated system for archiving and publishing the research data generated or used by doctoral students as part of their dissertation project. In doing so, the different connections between thesis and research data must be considered and represented in a data model. Also, the metadata format XMetaDissPlus has to be revised according to the data model. This is necessary in order to trace the relationships between the thesis and the research data (which can either be submitted as a supplement to the document or be stored in an external repository) and make them searchable in DNB’s catalog. The paper describes the data model and the changes made in XMetaDissPlus.

Zitierfähiger Link (DOI): https://doi.org/10.5282/o-bib/2018H2S72-78
Autorenidentifikation: Weisbrod, Dirk: GND: 1079147012, ORCID: http://orcid.org/0000-0002-9455-4527
Schlagwörter: Forschungsdaten; Dissertationen; Anlagerungsformen; Datenmodell; Metadaten

1. Voraussetzungen für die Pflichtablieferung von Forschungsdaten an die DNB

Die Pflichtablieferungsverordnung (PflAV) der DNB definiert die Rahmenbedingungen, nach denen Forschungsdaten an die DNB abgeliefert werden können. Demnach sollen „Elemente, Software und Werkzeuge [gesammelt werden], die in physischer oder in elektronischer Form erkennbar zu den ablieferungspflichtigen Netzpublikationen gehören“ (§ 7, Absatz 2 PflAV), während selbständig veröffentlichte Primär-, Forschungs- und Rohdaten ausdrücklich nicht unter die Ablieferungspflicht

fallen (§ 9 PflAV).[1] Forschungsdaten, die erkennbar mit einer ablieferungspflichtigen Netzpublikation in Verbindung stehen, können somit von der DNB gesammelt, langzeitarchiviert und in ihrem Katalog nachgewiesen werden.[2] Um dieses Szenario beispielhaft durchzuspielen, erarbeitete die DNB im DFG-Projekt eDissPlus zusammen mit der Humboldt-Universität zu Berlin Lösungen für Forschungsdaten, die im Rahmen von Dissertationsprojekten an der HU generiert und zusammen mit der Dissertation abgeliefert werden. Ein erstes Ergebnis des Projektes ist die im vergangenen Jahr veröffentlichte Policy für dissertationsbezogene Forschungsdaten der DNB.[3] Die HU entwickelte zudem einen Ingest-Prozess für die Publikation von dissertationsbezogenen Forschungsdaten auf ihrem Publikationsserver. In einem weiteren Schritt musste der Pflichtablieferungsworkflow der DNB für die Ablieferung von Forschungsdaten angepasst werden. Es musste folglich ein Datenmodell entwickelt werden, dass den Bezug zwischen Dissertation und Forschungsdaten abbildet und eine angemessene Weiterbearbeitung der Forschungsdaten und der dazugehörigen Metadaten bei Pflichtablieferung und Langzeitarchivierung ermöglicht. Über diesen Teilaspekt von eDissPlus berichtet der folgende Artikel.

2. Von den Anlagerungsformen zum Datenmodell

Die erkennbare Zugehörigkeit von Forschungsdaten zu einer Netzpublikation ergibt sich unter anderem aus der Anlagerungsform und damit aus der formalen Beziehung zwischen Forschungsdaten und Publikation bei der Ablieferung. Dieser Aspekt war für die Entwicklung des Datenmodells besonders wichtig, da die Verarbeitungsmöglichkeiten der Daten von deren Ablieferung abhängen. Die Data Publication Pyramid (Abbildung 1), die im Rahmen des EU-Projektes Opportunities für Data Exchange (ODE) unter Teilnahme der DNB erarbeitet wurde, visualisiert diese Anlagerungsformen, indem sie Forschungsdaten in Bezug auf ihre Aufbereitung im Forschungs- und Publikationskontext in 5 Typen einteilt.[4]

Für eDissPlus war allerdings nur die Spitze der Pyramide relevant, da reine Datenpublikationen (Typ 4) sowie nicht publizierte Daten (Typ 5) keinen erkennbaren Konnex zu einer ablieferungspflichtigen Netzpublikation besitzen und deswegen von der DNB nicht berücksichtigt werden dürfen. Außerdem konnten bei der Entwicklung des Datenmodells die vollständig in eine Dissertationsschrift integrierten Daten vernachlässigt werden (Typ 1). Das sind zum Beispiel Grafiken oder Tabellen, die in ein PDF eingebettet oder in einer Printversion abgedruckt werden. Forschungsdaten, die in dieser Form vorliegen, sammelt die DNB schon immer – sozusagen automatisch als Bestandteil der abgabepflichtigen Hochschulschrift. Diese Daten werden nicht gesondert verzeichnet. Relevant für

1 Vgl. „Verordnung über die Pflichtablieferung von Medienwerken an die Deutsche Nationalbibliothek," zuletzt geprüft am 01.01.2018, http://www.gesetze-im-internet.de/pflav/index.html.

2 Vgl. hierzu Dirk Weisbrod, *Policy der Deutschen Nationalbibliothek für dissertationsbezogene Forschungsdaten*, Version 1.1 (Frankfurt, M.: Deutsche Nationalbibliothek, 2017), urn:nbn:de:101-2017092701, 3: „Die DNB fordert die Ablieferung von Forschungsdaten nicht aktiv ein, sondern überlässt die Definition entsprechender Rahmenbedingungen der jeweiligen Hochschule."

3 Vgl. Weisbrod, *Policy der Deutschen Nationalbibliothek*.

4 Vgl. Susan Reilly et al., *Opportunities of Data Exchange: Report on Integration of Data and Publications* (2011), 5–6, zuletzt geprüft am 01.01.2018, http://hdl.handle.net/10013/epic.40198.

die Lösung der oben erläuterten Fragestellung sind hingegen die Anlagerungsformen „Supplemente“ (Typ 2) und „referenzierte Daten“ (Typ 3).

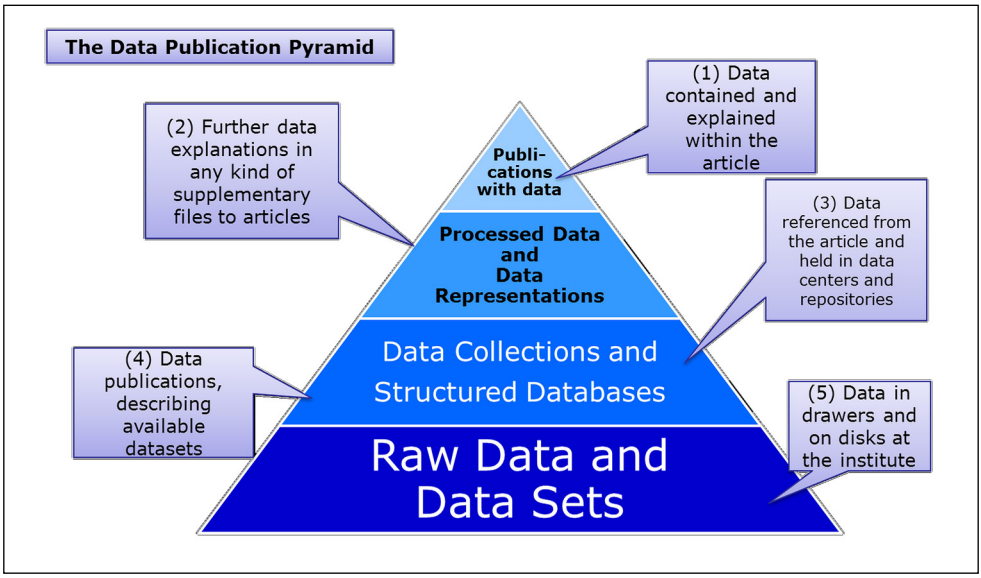

Abb. 1: Data Publication Pyramid[5]

Forschungsdaten-Supplemente konnten schon in der Vergangenheit, etwa auf einer CD-ROM oder einem anderen Datenträger, einer gedruckten Dissertation beigefügt werden. Im Falle einer Netzpublikation kann man sich vorstellen, dass Promovierende Forschungsdaten-Supplemente zusammen mit der Dissertation auf den Publikationsserver einer Hochschule hochladen und von dort die Dissertation und die zugehörigen Supplemente an die DNB abgeliefert werden. Bislang wurden Supplemente allerdings nicht im Katalog der DNB nachgewiesen. Mit eDissPlus soll sich das nun ändern.

Referenzierte Daten sind Forschungsdaten, die zwar in einer Dissertation verwendet, jedoch nicht mit dieser abgeliefert werden. So kann es vorkommen, dass Promovierende ein renommiertes Fachrepositorium oder ein institutseigenes Datenarchiv nutzen und bei der Ablieferung lediglich den Persistenten Identifikator (PID) der Daten (z.B. DOI, URN oder handle) angeben. Der Integrationsgrad zwischen Publikation und Daten ist hier geringer als bei Supplementen, da die Forschungsdaten zunächst nicht physisch an die DNB abgeliefert werden. Man kann aber davon ausgehen, dass auch referenzierte Daten, ebenso wie die Supplemente, wesentlich zum Verständnis und zur Nachprüfbarkeit der in der Dissertation präsentierten Forschungsergebnisse beitragen. Im Rahmen von eDissPlus wurde deshalb entschieden, dass die PIDs der Daten wenigstens in den Metadaten der abzuliefernden Dissertation verzeichnet und damit dauerhaft referenziert werden. Für beide Anlagerungsformen musste nunmehr ein Datenmodell für die Pflichtablieferung gefunden werden.

5 Vgl. Reilly et al, *Opportunities of Data Exchange*, 6.

3. Forschungsdaten-Supplemente

Ausgehend vom existierenden Workflow bot sich für die Projektbeteiligten zunächst eine einfache Variante für die Ablieferung von Supplementen an: Dissertationsschrift und Forschungsdaten in einem Container. Da die DNB für die Ablieferung von Dissertationen in der Regel eine OAI-PMH-Schnittstelle nutzt, versehen die Abliefernden oder ein entsprechend beauftragter Dienstleister (im Fall von Dissertationen also die Hochschulbibliotheken) einen solchen Container mit einer Transfer-URL und stellen diese an der Schnittstelle zu Verfügung. Die DNB harvestet dann den Container und speist diesen in ihren Pflichtablieferungs-Workflow ein. Dieses Vorgehen ist aber sehr unbefriedigend, da nunmehr der gesamte Container, also Dissertation und Forschungsdaten gemeinsam, durch die mitgelieferten Metadaten beschrieben werden. Das hat gleich zwei Nachteile. Erstens können dann die Forschungsdaten aufgrund fehlender Metadaten nicht mehr eigenständig im Katalog der DNB verzeichnet werden; andererseits können Dissertation und Daten im gesamten Pflichtablieferungs- und Langzeitarchivierungs-Workflow nicht mehr ohne weiteres voneinander unterschieden werden, ein forschungsdatenspezifisches Preservation Planning ist dann nicht mehr möglich.

In eDissPlus wurde stattdessen ein Datenmodell erarbeitet, das Dissertation und Forschungsdaten schon bei der Ablieferung deutlich voneinander trennt. Jede Entität erhält vom Abliefernden einen eigenen Metadatensatz im DNB-eigenen Format für Hochschulschriften XMetaDissPlus. Zugleich wird in diesem Modell auch der Bezug zwischen Dissertation und Daten abgebildet, denn der Nachweis dieses Bezugs ist ja Voraussetzung dafür, dass die Daten überhaupt von der DNB gesammelt werden dürfen. Somit können zukünftig Forschungsdaten mit eigenem Metadaten-Satz an der OAI-PMH-Schnittstelle bereitgestellt werden. Dabei handelt es sich um einen im tar- oder zip-Format gepackten Ordner, in dem die abgelieferten Supplemente beliebig angeordnet werden können. Auch die Ablieferung mehrerer solcher Forschungsdaten-Ordner ist möglich, wenn Promovierende oder die abliefernde Hochschule das für sinnvoll halten und damit zum Beispiel Teilprojekte der Dissertation oder wichtige Einzeldaten sichtbar gemacht werden sollen. Jeder Ordner muss durch einen Metadatensatz beschrieben werden, ebenso wie die Dissertationsschrift, die bisher schon separat abgeliefert wird.

Hierfür wurde von der DNB ein Metadaten-Kernset für Forschungsdaten definiert, in dem die obligatorischen und fakultativen Metadaten definiert sind. Es lehnt sich an das DNB-Kernset für Hochschulschriften an und wurde um forschungsdatenspezifische Elemente wie zum Beispiel Relationstypen, Datentypen oder Lizenzinformationen ergänzt.[6] Ebenso wurde das Kernset für Hochschulschriften erweitert, um Relationen zu Daten von Seiten der Dissertationen abbilden zu können. Die Verweisung erfolgt somit bidirektional.

Die Kernsets dienten zudem als Vorlage für die Erweiterung von XMetaDissPlus, dem DNB-eigenen Metadatenschema für Hochschulschriften. Die Ablieferer sind verpflichtet die Publikationen an der OAI-Schnittstelle mit XMetaDissPlus-Metadaten bereitzustellen, folglich müssen die im Kernset

6 Vgl. Deutsche Nationalbibliothek, *Lieferung von Metadaten für monografische Netzpublikationen an die Deutsche Nationalbibliothek*, Version 2.0 (Frankfurt, M.: Deutsche Nationalbibliothek, 2014), 6, urn:nbn:de:101-2014071100.

definierten Elemente und Attribute dort auch implementiert werden. Besonderes Augenmerk wurde dabei auf die Abbildung der Relation zwischen Dissertation und Forschungsdaten gelegt, was bisher in XMetaDissPlus nicht möglich war. Die Projektpartner orientierten sich dabei am Metadatenschema von DataCite, das genau diese Aufgabe durch den Einsatz von Relationstypen bereits vorbildlich gelöst hatte.[7] Diese Relationstypen wurden nunmehr in XMetaDissPlus implementiert und zugleich das Element <relatedIdentifier> mit den entsprechenden Attributen definiert. Die Relation zu einer Dissertation sieht in einem Forschungsdatensatz dann in etwa so aus:

```
<ddb:relatedIdentifier ddb:relatedIdentifierType="URN"
      ddb:relationType="IsPartOf">
      urn:beispiel-urn-derDissertation
</ddb:relatedIdentifier>
```

Vice versa wird diese Relation auch in den Metadaten zu einer Dissertation verzeichnet, dann aber mit dem Relationstyp „HasPart" und unter Angabe der Forschungsdaten-PID. Die Relationstypen „HasPart" und „IsPartOf" sind hierbei verbindlich vorgeschrieben. In der Regel soll als <relatedIdentifierType> eine URN angegeben werden, da in der DNB alle ablieferungspflichtigen Objekte eine URN erhalten. Alternativ ist aber auch die Angabe von DOIs, handles oder anderer PID-Formaten möglich.

Die Relation Dissertation - Forschungsdaten ist eine 1-n-Relation, d.h. einem Dissertationsdatensatz können beliebig viele Forschungsdaten zugeordnet werden. Umgekehrt ist nur eine 1-1-Beziehung möglich. Damit trug man der gesetzlichen Vorgabe Rechnung, dass nur Daten, die Teil einer Veröffentlichung sind, gesammelt werden dürfen und somit jedes Forschungsdatenpaket in Abhängigkeit zu einer bestimmten Dissertation stehen muss.

Dissertationsbezogene Forschungsdaten können somit zukünftig über die von vielen Hochschulen für Dissertationen schon genutzte OAI-PMH-Schnittstelle an die DNB abgeliefert werden – unter der Voraussetzung, dass in den Metadaten die Relation zu einer Dissertation eingetragen ist. Die Metadaten werden in das Katalogsystem der DNB importiert, die Forschungsdaten-Supplemente strikt von der Dissertation getrennt weiterverarbeitet und der Langzeitarchivierung zugeführt. Das alles läuft vollautomatisch ab. Die Supplemente sind nach abgeschlossenem Import als eigenständige Einträge im Katalog recherchierbar; zugleich wird über den Relationstyp im Katalogeintrag die Abhängigkeit zu einer Dissertation nachgewiesen und die Supplement-Funktion der Daten für diese wissenschaftliche Arbeit sichtbar gemacht.

7 DataCite ist ein internationales Konsortium, das sich unter anderem für die einheitliche Registrierung von Forschungsdaten einsetzt und dabei DOI als Objekt-Identifikator verwendet. Für die Registrierung der Daten wurde von DataCite zudem ein Metadatenschema entwickelt. Vgl. DataCite Meatadata Working Group., *DataCite Metadata Schema Documentation for the Publication and Citation of Research Data*, Version 4.0 (2016), 18–20, http://doi.org/10.5438/0012.

4. Referenzierte Daten

Über dieses Modell lassen sich auch referenzierte Daten ohne Probleme nachweisen. Hierzu wird in den Metadaten der Dissertation eine PID als <relatedIdentifier> eingetragen, die Daten auf externen Repositorien referenziert. Die Daten können nun über diese PID aus dem DNB-Katalog heraus aufgerufen werden. Da ein Dissertationsdatensatz n-Relationen zu Forschungsdaten enthalten darf, können auf diesem Wege beliebig viele Forschungsdaten zu einer Dissertation nachgewiesen werden – ganz gleich ob diese als Supplemente bei der DNB oder auf externen Repositorien vorliegen.

Auf Supplemente wendet die DNB ihren kompletten LZA-Workflow an, d.h. sie führt ein Preservation Planning durch und archiviert die Daten in ihrem Langzeitarchiv.[8] Bei referenzierten Daten ist jedoch zunächst das jeweilige externe Repositorium für diese Aufgabe zuständig. Da einige Repositorien sich auf die von der DFG vorgeschlagene 10-jährige Aufbewahrungsfrist[9] für Forschungsdaten beschränken und keine Langzeitarchivierung durchführen,[10] entwickelte die DNB im Rahmen von eDissPlus ein Konzept, das vorsieht, Daten auf externen Repositorien zu harvesten und sie in den eigenen Langzeitarchivierungs-Workflow einzuspeisen.

5. Zusammenfassung

Das Datenmodell von eDissPlus ermöglicht es der DNB, dissertationsbezogene Forschungsdaten in Zukunft als eigenständige Einträge in ihrem Katalog vorzuhalten und sie damit recherchierbar zu machen.[11] Zugleich wird über die Metadaten die Relation zur betreffenden Dissertation abgebildet. Benutzer und Benutzerinnen der DNB können damit zukünftig im Rahmen der gesetzlichen Regelungen im Lesesaal oder online auch Forschungsdaten einsehen. Über die Metadaten der Dissertation wird ihnen auch der Zugriff auf externe Daten ermöglicht, soweit deren PIDs bei Abgabe der Dissertation von den Promovierenden angegeben wurden. Alle Daten, die an die DNB abgeliefert wurden, werden nach Möglichkeit langzeitarchiviert.[12]

Damit leistet die DNB einen Beitrag zur aktuellen Diskussion zu/über Forschungsdaten, die einen Höhepunkt in den 2016 veröffentlichen Empfehlungen des Rates für Informationsinfrastruktur (RFII) fand.[13] Entsprechend ihrer gesetzlichen Vorgaben sammelt die DNB allerdings nicht alle

8 Vgl. Deutsche Nationalbibliothek, *Langzeitarchivierungs-Policy der Deutschen Nationalbibliothek*, Version 1.0 (Frankfurt, M. : Deutsche Nationalbibliothek, 2013), urn:nbn:de:101-2013021901.

9 Vgl. Deutsche Forschungsgemeinschaft, *Vorschläge zur Sicherung guter wissenschaftlicher Praxis: Empfehlungen der Kommission „Selbstkontrolle in der Wissenschaft", ergänzte Auflage* (Weinheim: Wiley-VHC., 2013), 21, https://doi.org/10.1002/9783527679188.oth1.

10 Lt. DNB Policy werden derzeit das Data Seal of Approval (DSA) und das nestor-Siegel als vertrauenswürdige Zertifizierungen für Repositorien angesehen. Vgl. Weisbrod, *Policy der Deutschen Nationalbibliothek*, *4*.

11 eDissPlus beschäftigt sich nur mit Netzpublikationen. Printpublikationen mit Forschungsdaten wurden nicht bearbeitet.

12 „Je nach Wissenschaftsdisziplin kann es sich bei Forschungsdaten um Messdaten, Beobachtungsdaten, Umfrageergebnisse, oder andere Arten von Daten handeln. Aufgrund der Vielfalt und Komplexität der Datenformate kann die Langzeitarchivierung im Sinne einer unbeschränkten Nutzbarkeit über lange Zeiträume nicht für alle Forschungsdaten von der DNB gewährleistet werden." Vgl. Weisbrod, *Policy der Deutschen Nationalbibliothek*, *5*.

13 Vgl. Rat für Informationsinfrastrukturen, *Leistung aus Vielfalt: Empfehlungen zu Strukturen, Prozessen und Finanzierung des Forschungsdatenmanagements in Deutschland* (Göttingen, 2017), urn:nbn:de:101:1-201606229098.

Forschungsdaten, sondern nur Supplemente und referenzierte Daten, die ablieferungspflichtigen Netzpublikationen zugeordnet werden können. Damit trägt sie im Sinne der guten wissenschaftlichen Praxis zur Nachprüfbarkeit der in den abgelieferten wissenschaftlichen Arbeiten niedergelegten Forschungsergebnisse bei und ermöglicht deren Nachnutzung. Außerdem wird die DNB die weitere Entwicklung beobachten und bei Bedarf die Lösung von eDissPlus weiterentwickeln. Auf die Möglichkeit, Daten auf Repositorien, die keine langfristige Archivierung anstreben zu identifizieren und zu harvesten, wurde weiter oben schon eingegangen. Zudem wird genau zu beobachten sein, wie die Empfehlungen des Rates für Informationsinfrastrukturen umgesetzt werden und welche Rolle die DNB in einer noch zu schaffenden nationalen Infrastruktur für Forschungsdaten spielen könnte.

Literaturverzeichnis

- Bundesministerium für Justiz und für Verbraucherschutz (BMJV). „Verordnung über die Pflichtablieferung von Medienwerken an die Deutsche Nationalbibliothek“. Zuletzt geprüft am 01.01.2018. http://www.gesetze-im-internet.de/pflav/index.html.

- DataCite Metadata Working Group. *DataCite Metadata Schema Documentation for the Publication and Citation of Research Data*. Version 4.0. 2016. http://doi.org/10.5438/0012.

- Deutsche Forschungsgemeinschaft. *Vorschläge zur Sicherung guter wissenschaftlicher Praxis: Empfehlungen der Kommission „Selbstkontrolle in der Wissenschaft“*. Ergänzte Auflage. Weinheim: Wiley-VHC., 2013. https://doi.org/10.1002/9783527679188.oth1.

- Deutsche Nationalbibliothek. *Langzeitarchivierungs-Policy der Deutschen Nationalbibliothek*. Version 1.0. Frankfurt, M.: Deutsche Nationalbibliothek, 2013. urn:nbn:de:101-2013021901.

- Deutsche Nationalbibliothek. *Lieferung von Metadaten für monografische Netzpublikationen an die Deutsche Nationalbibliothek*. Version 2.0. Frankfurt, M.: Deutsche Nationalbibliothek, 2014. urn:nbn:de:101-2014071100.

- Rat für Informationsinfrastrukturen. *Leistung aus Vielfalt: Empfehlungen zu Strukturen, Prozessen und Finanzierung des Forschungsdatenmanagements in Deutschland.* Göttingen, 2017. urn:nbn:de:101:1-201606229098.

- Reilly, Susan, Wouter Schallier, Sabine Schrimpf, Eefke Smit und Max Wilkinson. *Opportunities of Data Exchange: Report on Integration of Data and Publications*. 2011. Zuletzt geprüft am 01.01.2018. http://hdl.handle.net/10013/epic.40198.

- Weisbrod, Dirk. *Policy der Deutschen Nationalbibliothek für dissertationsbezogene Forschungsdaten*. Version 1.1. Frankfurt, M.: Deutsche Nationalbibliothek, 2017. urn:nbn:de:101-2017092701.

