JENSEITS DER DATEN

ÜBERLEGUNGEN ZU DATENZENTREN FÜR DIE GEISTESWISSENSCHAFTEN AM BEISPIEL DES KÖLNER 'DATA CENTER FOR THE HUMANITIES'

von Patrick Sahle und Simone Kronenwett

Zusammenfassung

„Auch in den Geisteswissenschaften werden Daten produziert, die dauerhaft gesichert und zugänglich gehalten werden müssen." Dieser Satz ist richtig, aber an einer Stelle problematisch: ‚Daten'. Aus der Sicht der Geisteswissenschaften ist unklar, ob der allgemeine, derzeit herrschende Datenbegriff die Situation in ihren Disziplinen wirklich treffend beschreibt und ob seine Konsequenzen die gleichen sind wie in anderen Feldern der Forschung. Der Beitrag geht von einer Spezifik geisteswissenschaftlicher Daten und dem dort vorhandenen Problem der schlechten Unterscheidbarkeit zwischen Primärdaten und Ergebnisdaten aus und beschreibt die Konsequenzen für den Aufbau eines geisteswissenschaftlichen Datenzentrums am Beispiel des im Dezember 2012 gegründeten ‚Data Centers for the Humanities' (DCH) an der Universität zu Köln. Zu klären ist dabei unter anderem, was Forschungsdatenmanagement für die beteiligten Forscher und Projekte bedeutet und wie die Dauerhaftigkeit eben nicht nur von ‚Daten', sondern auch von Forschung insgesamt sicher gestellt werden kann. Ausgehend von der Unterscheidung zwischen ‚Daten' und ‚Ressourcen' und der Frage, welche Leistungen von einem Datenzentrum eigentlich zu erwarten sind, wird der einerseits schichtweise, andererseits modulare Aufbau des DCH begründet. Die vielfältigen Aufgaben, die sich bei der Sicherung der Forschung ergeben, lassen sich über vier Paradigmen beschreiben, die einen begrifflichen Anschluss an die bestehenden Einrichtungen der Kulturerbesicherung ermöglichen. Ob dieser Anschluss nur metaphorisch ist, wenigstens eine didaktisch-erklärende Kraft hat oder sogar die Grundlage weiterer konzeptioneller Überlegungen sein kann, ist zu diskutieren.

Abstract

„Even in the humanities data are produced that must be permanently secured and kept accessible." This sentence might be true, yet, going into details the problem occurs with the term ‚data'. At the time being, from the perspective of the humanities it is not really clear what the term "data" actually means, how it is defined, and what belongs to it. This article reflects data specific in humanities research. Talking about research data in the humanities in general, the paper casts a light on the existing problem of separating so-called primary data from result data. Consequences for the setting up and development of a data center for the humanities are described by the example of the Cologne 'Data Center for the Humanities' (DCH). The meaning of research data management will be discussed especially with regard to the production of data and results as well as to the performances to be kept permanently secure and accessible. Based on the distinction between ‚data' and ‚resources' the design of the DCH is established in layers as well as in modules. The variety of tasks may be described by four paradigms borrowed from cultural heritage institutions.

Dieser Beitrag geht von einer Spezifik geisteswissenschaftlicher Daten aus. Im Fokus steht das in diesen Disziplinen vorhandene Problem einer schwierigen Trenn- und Unterscheidbarkeit von Primär- und Ergebnisdaten. Der Artikel beschreibt die sich daraus ergebenden Konsequenzen für den Aufbau eines geisteswissenschaftlichen Datenzentrums am Beispiel des im Dezember 2012 gegründeten ‚Data Center for the Humanities' (DCH) an der Universität zu Köln. Zu klären ist dabei unter anderem, was Forschungsdatenmanagement für die beteiligten Forscher und Projekte bedeutet und wie die Dauerhaftigkeit eben nicht nur von ‚Daten', sondern von Forschungsleistungen insgesamt sichergestellt werden kann. Ausgehend von der Unterscheidung zwischen ‚Daten' und ‚Ressourcen' und der Frage, welche Leistungen von einem Datenzentrum eigentlich zu erwarten sind, wird der einerseits schichtenweise, andererseits modulare Aufbau des DCH begründet. Die vielfältigen Aufgaben, die sich bei der Sicherung der Forschung ergeben, lassen sich mit vier Paradigmen beschreiben, die einen begrifflichen Anschluss an die bestehenden Einrichtungen zur Sicherung des kulturellen Erbes ermöglichen. Ob dieser Anschluss nur metaphorisch ist, wenigstens eine didaktisch-erklärende Kraft hat oder sogar die Grundlage weiterer konzeptioneller Überlegungen sein kann, ist jenseits dieses Beitrages zu diskutieren.

Ausgangslage

Wie in anderen Wissenschaften, so wird auch in der geisteswissenschaftlichen Forschung vornehmlich projektorientiert gearbeitet. Dies ist zunächst unabhängig davon, ob es um Vorhaben geht, die zum Beispiel durch eingeworbene Drittmittel gefördert werden, oder um Arbeiten, die im Rahmen der wissenschaftlichen Tätigkeit auf einer akademischen Stelle oder von Qualifikationsarbeiten (Promotion, Habilitation) durchgeführt werden. Nicht entscheidend ist außerdem, ob Projekte in kleineren oder größeren Teams oder als die für die Geisteswissenschaften typische Einzelforschung bearbeitet werden. Der Projektcharakter ergibt sich vielmehr aus der Orientierung an einem zu erreichenden Ziel beziehungsweise einem definierten Endprodukt auf der einen Seite und der zeitlichen Begrenzung durch einen Start- und Endpunkt auf der anderen Seite. Grundsätzlich gibt es auch nicht-projektbezogene Forschungsaktivitäten, die sich durch ihre unbestimmte Dauer zum Beispiel im Laufe eines kontinuierlichen institutionellen Rahmens, innerhalb eines dauerhaften Forschungsauftrages oder durch das Fehlen eines spezifizierten Zieles im Sinne eines abzuliefernden Produktes auszeichnen.

Will man von ‚Daten' sprechen, die im Forschungsprozess anfallen, dann liegt aufgrund der Produktorientierung auch in den Geisteswissenschaften zunächst eine theoretische Trennung zwischen Ergebnisdaten auf der einen und Ausgangsdaten oder ‚Primärdaten' auf der anderen Seite nahe. Wenn man geisteswissenschaftliche Forschung dauerhaft speichern will, wird man aber nicht umhin kommen, einen genaueren Blick auf die verschiedenen Datenarten und den Umgang mit ihnen zu werfen.[1]

1 Den Begriff der Daten in den Geisteswissenschaften und ihre Spezifik thematisiert auch Christof Schöch in seinem Blogpost "Big? Smart? Clean? Messy? Data in the Humanities", 29.07.2013, http://dragonfly.hypotheses.org/443. Alle URLs des Beitrags wurden zuletzt am 29.08.2013 überprüft.

Was also sind ‚Primärdaten' in den Geisteswissenschaften?[2] Es sind zunächst meist *keine* in der physischen Welt gemessenen Daten, keine Datenreihen, die Eigenschaften von abstrakt modellierten Untersuchungsgegenständen abbilden.[3] Es sind auch meistens keine systematischen Beschreibungen von Objekten, auch wenn beispiels-weise die Prinzipien der Erschließung und Katalogisierung eine große Tradition und Bedeutung in den Geisteswissenschaften haben. Eher selten sind es auch ‚empirische' Verfahren der Informationserhebung, wie sie in der Zeitgeschichte oder den Kulturwissenschaften unter anderem zum Beispiel bei Zeitzeugenbefragungen, Umfragen oder Interviews zum Einsatz kommen. Grundsätzlich unterscheiden sich die Verfahren der Informationserhebung stark von jenen in den Naturwissenschaften, während die Methoden etwa zum Beispiel der Sozialwissenschaften zwar zuweilen eingesetzt werden, aber ebenfalls nicht paradigmatisch sind.[4] Wenn man abstrakt von der Gewinnung von Ausgangsdaten sprechen will, sind vielmehr die folgenden beiden Verfahren typisch: Erstens wird man in den Geisteswissenschaften der „"Literatur" bzw. dem Text im allerweitesten Sinne selbst den Charakter von Primärdaten zusprechen müssen, da die häufig weit in die Vergangenheit zurückreichenden Texte in den meisten Fällen die wichtigste, manchmal sogar die ausschließliche Grundlage der Auseinandersetzung mit einer Forschungsfrage darstellen.[5]

Die Unterscheidung von Primärliteratur als Gegenstand der Forschung und Sekundärliteratur als Tradition der Auseinandersetzung mit diesem Gegenstand ist häufig unscharf und nicht konsequent durchzuhalten und soll deshalb hier auch nicht diskutiert werden. Die so genannte ‚Primärliteratur' ragt als ‚historisches Dokument' allerdings in den zweiten Bereich der Primärdaten hinein, der alle unveröffentlichten Quellen, Archivalien, Fundstücke et cetera umfasst. Hier ist insgesamt von *allen Überresten und Artefakten der menschlichen Kultur* zu sprechen, die entweder Gegenstand der geisteswissenschaftlichen Forschung sind oder Indizien zur Beantwortung von Forschungsfragen liefern können. Nur auf den ersten Blick paradox klingt dann die Erweiterung, dass in den Geisteswissenschaften auch die ‚Sekundärliteratur' zu den erweiterten Primärdaten zu rechnen ist.

Diesen ‚Ausgangsdaten' stehen traditionelle Formen der Ergebnissicherung und -präsentation gegenüber. Die Geisteswissenschaften produzieren fast ausschließlich Texte. Monografische Schriften und einzelne Aufsätze sind als Produkte der Forschung

2 Zur Definition von 'Geisteswissenschaften' siehe unter anderem Wissenschaftsrat: Empfehlungen zur Entwicklung und Förderung der Geisteswissenschaften in Deutschland, Köln 2006, S. 17, http://www.wissenschaftsrat.de/download/archiv/geisteswissenschaften.pdf.

3 Ausnahmen finden sich zum Beispiel in der Phonetik oder bei der feldforschenden Linguistik.

4 Allerdings sei an dieser Stelle darauf hingewiesen, dass die Grenzen zwischen Geistes- und Naturwissenschaften in einigen Teilbereichen – gerade durch den Einsatz digitaler Methoden – teilweise zu "verschwimmen" scheinen, vgl. Wolfgang Pempe: Geisteswissenschaften, in: Heike Neuroth u.a. (Hrsg.): Langzeitarchivierung von Forschungsdaten. Eine Bestandsaufnahme, Boizenburg 2012, S. 138, http://nestor.sub.uni-goettingen.de/bestandsaufnahme/nestor_lza_forschungsdaten_bestandsaufnahme.pdf.

5 Siehe zu dieser Diskussion auch Jasmin Hügi, René Schneider: Digitale Forschungsinfrastrukturen in den Geistes- und Geschichtswissenschaften, Genf 2013, S. 17-21, http://doc.rero.ch/record/31535/files/Schneider_Digitale_Forschungsinfrastrukturen.pdf.

so vorrangig, dass alle anderen Formen in einer ersten Annäherung im Grunde zu vernachlässigen sind.[6]

Sie sind so paradigmatisch und dominant, dass neben diesen Primärdaten eine weitere Datenart nur selten diskutiert wird. Tatsächlich entstehen im Forschungsprozess nämlich regelmäßig auch intermediäre oder Arbeitsdaten. In den Projekten werden Materialien gesammelt und erschlossen, Aggregationsstufen erzeugt, Texte bewertet, kommentiert und annotiert, Verlinkungen hergestellt, Aufzeichnungen oder ganze Aktenreihen angelegt, Korrespondenzen geführt, der Arbeitsprozess dokumentiert oder vielfältige analytische oder narrative Zwischenstufen zum endgültigen Ergebnis erarbeitet. In der Literatur ist hier von einer Zwischenschicht zwischen ‚input' und ‚output' die Rede, die als ‚throughput' bezeichnet wird.[7] Auch diese Terminologie könnte allerdings eine scharfe Trennbarkeit der Schichten suggerieren, die bei einer digitalen Arbeitspraxis und Infrastruktur immer weniger gegeben ist. Hier dürfte vielmehr der ‚throughput' immer mehr mit einem ‚augmented and processed input' zusammenfallen.

Die verschiedenen Datenarten werden von der etablierten Infrastruktur der Informationsbewahrung und -bereitstellung auf unterschiedliche Weise erfasst. Die Primärdaten werden in Bibliotheken, Archiven und Museen gesammelt, bewahrt, erschlossen und zugänglich gemacht; prekär ist eher die Situation bei Materialien, die erst in der Forschung entstehen, wie die Sammlungen archäologischer Funde oder die Befragungen der zeitgeschichtlichen Forschung. Das Gros der Ausgangstexte, das in den Bibliotheken liegt, kann als gut erschlossen und dauerhaft gesichert bezeichnet werden, so dass hier – von außen betrachtet – kein grundsätzliches Problem zu bestehen scheint. Ähnlich sieht es für die Ergebnisdaten aus: Ein etabliertes System der Publikation über die Verlage und der Vorhaltung über die Bibliotheken sorgt dafür, dass Ergebnisdaten als Produkte auf den Markt der öffentlichen Kommunikation kommen und dort auf Dauer erreichbar und nutzbar bleiben. Nur im Bereich der ‚Zwischendaten' ist die Situation weniger klar. Im Regelfall werden die Materialien des Forschungsprozesses selbst nach seinem Abschluss oder nach dem Ableben der beteiligten Forscher[8] vernichtet.[9] In anderen Fällen gehen sie als Deposita oder Nachlässe an die Archive, häufig an Universitätsarchive, wo sie zwar materiell gesichert, oft aber nur oberflächlich erschlossen werden können und so zwar grundsätzlich, aber nicht unbedingt komfortabel oder leicht auffindbar und zugänglich sind.

[6] So werden Vorträge als verbreitete Produkte der Forschung traditionell entweder als Aufsätze veröffentlicht oder nicht eigens publiziert. Zu den Zwischenprodukten der Forschung weiter unten.
[7] So bei Hügi, Schneider: Forschungsinfrastrukturen, S. 20.
[8] Im Folgenden schließt die Verwendung des generischen Maskulinums immer auch Frauen mit ein.
[9] Die DFG weist allerdings im Rahmen ihrer "Vorschläge zur Sicherung guter wissenschaftlicher Praxis" auf eine zehnjährige Aufbewahrungspflicht von Primärdaten, die als Grundlage für Veröffentlichungen dienen, hin, vgl. DFG: Vorschläge zur Sicherung guter wissenschaftlicher Praxis: Empfehlungen der Kommission „Selbstkontrolle in der Wissenschaft", Denkschrift, Weinheim 1998, S. 12, http://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/download/empfehlung_wiss_praxis_0198.pdf.

Diese Charakteristika kennzeichnen die etablierte Struktur für die traditionelle Forschung. Angesichts der aktuellen Veränderungen, nicht zuletzt durch den Einsatz digitaler Medien und Werkzeuge, sieht die Situation in fast allen Aspekten der Datenerzeugung und Datennutzung im Forschungsprozess inzwischen und vor allem perspektivisch ganz anders aus. Primärdaten werden inzwischen von Bibliotheken und Archiven zunehmend digital zur Verfügung gestellt. Grundsätzlich können sie dabei immer noch in den Institutionen verbleiben und von der Forschung nur adressiert sowie gegebenenfalls in die eigenen Systeme der Ergebnispräsentation eingebunden werden. Tatsächlich werden Primärdaten aber oft in die Systeme des Forschungsprozesses selbst eingebunden, weil zum Beispiel ihre Bearbeitung, vertiefende Erschließung und Annotation die Grundlage der Forschung sind. Hinzu kommt, dass Primärdaten oft erst bei der wissenschaftlichen Arbeit selbst entstehen. Hier ist an eigens angefertigte Digitalisate ebenso zu denken wie auch an andere digitale ‚Aufnahmen' aus Disziplinen wie der Linguistik.

Sehr viel differenzierter wird zunehmend auch das Bild bei den Ergebnisdaten. Hier ist zunächst zu beobachten, wie sich die traditionellen Formen verändern und welche Rückwirkungen dies auf ihre Vorhaltung in den Informationsinfrastrukturen hat. Die Monografie und der wissenschaftliche Aufsatz scheinen auch in ihrer digitalen Gestalt klar umgrenzte, stabilisierte Produkte der Forschung zu sein, für die unter anderem mit den institutionellen Repositorien eine etablierte Speicher- und Publikationsstruktur zur Verfügung steht. Neue Herausforderungen entstehen dann, wenn sich diese narrativen Formen durch die Bedingungen unserer digitalen Umwelt zu verändern beginnen. Durch Multimedialisierung, die Verwendung von Links für interne und externe Bezüge, die Nutzung von Strukturen und Formaten jenseits des ‚Seitenparadigmas' und die prinzipielle Öffnung für fortlaufende Veränderungen, können die Abgrenzung und die Stabilität solcher ‚Publikationen' fragwürdig werden. Dies lässt sich vor allem bei den neueren Formen der wissenschaftlichen ‚Mitteilung' beobachten: Hypertexte, Wikis, Blogs und die teilweise sehr komplexen Informationsportale, die häufig gezielt und damit sehr individuell für einzelne Forschungsvorhaben entstehen, lassen sich kaum noch mit den Begriffen der traditionellen Publikationskultur beschreiben oder mit ihren Systemen verwalten. Hier sei nur auf die ungeklärten Fragen zur Zitierung und Referenzierung oder zur Kreditierung und Urheberschaft der ‚Beiträge' verwiesen, die sich als ‚Informationspartikel' in unterschiedlichen Granularitäten auf verschiedenen Ebenen einer wissenschaftlichen Ressource heute oft nur noch sehr schlecht identifizieren lassen. Aus methodischer Sicht lässt sich derzeit geradezu eine Trennlinie ausmachen, die zwei Arten von Ergebnisdaten zwei verschiedenen Infrastruktursparten zuweist: Stabilisierte, abgeschlossene Narrative in einfachen Strukturen und Formaten lassen sich auch in digitaler Form problemlos in den etablierten Systemen der Bibliotheken dauerhaft nachweisen und vorhalten. Offene, veränderliche Informationssysteme mit individuellen und komplexen Strukturen erweisen sich hier im Moment noch als große Herausforderung und können kaum dauerhaft gepflegt werden.

Hinter diesen Schwierigkeiten verbirgt sich eigentlich, wenn man die Situation unter der Perspektive der Forschungsdaten beschreiben will, ein anderes Phänomen. Der Forschungsprozess vollzieht sich heute – zumindest theoretisch – in digitalen Arbeitsumgebungen, in denen kollaborativ, inkrementell und kontinuierlich gearbeitet

wird. Dies untergräbt nicht nur die Vorstellung finaler, abgeschlossener Produkte, die sich bibliografisch klar fassen lassen würden. Es stellt vor allem die Abtrennbarkeit der Bereiche Primärdaten – Zwischendaten – Ergebnisdaten in Frage. Thesenhaft lautet die Zuspitzung: Durch die Digitalisierung des Forschungsprozesses verschmelzen die verschiedenen Arten von Forschungsdaten zu einem Kontinuum, das von den Ausgangsdaten bis zu den Narrativen der Ergebnisse der Forschung alle Schritte der Verarbeitung umfasst.

Implizit ist damit auch der Begriff der ‚Forschungsdaten' verändert, da eine Definition in der Abgrenzung zu den Ergebnisdaten heraus nicht mehr haltbar erscheint. Ihre Begründung erfährt die These vom Forschungsdatenkontinuum durch den Prozess der digitalen geisteswissenschaftlichen Forschung. Dieser ist dadurch geprägt, dass Primärdaten digitalisiert, in andere Repräsentationsformen überführt, immer weiter erschlossen, verknüpft, annotiert, kontextualisiert, extrahiert und analysiert werden, um so die Forschung im Sinne einer digitalen Transformation weiter gestalten zu können.

Die kritische Bearbeitung und die Herstellung angereicherter Formen von Repräsentationen ist selbst zentraler Bestandteil der wissenschaftlichen Arbeit und muss als eines ihrer Ergebnisse betrachtet werden. Damit bekommen viele Formen der Bearbeitung den Charakter von Zwischendaten, die als publizierte und nachnutzbare Leistungen zugleich Ergebnisdaten sind. Virtuelle Forschungsumgebungen und fragestellungsbedingte, tief erschlossene digitale Bibliotheken oder Editionen sind gerade durch ihre Zielstellung definiert, die ‚Ausgangsdaten', ihre kritische Erschließung und Analyse und die Ergebnisse der Forschung möglichst eng miteinander zu verbinden. Vor diesem Hintergrund erscheint es sinnvoll, den Begriff der ‚Forschungsdaten' noch einmal auf seine Operationalisierung für eine dauerhafte Sicherung von Forschungsleistungen hin zu prüfen.

Daten und Systeme

Wenn die Schichten, ‚Aggregierungsgrade' oder ‚Prozessstufen' von Forschungsdaten nicht ohne Weiteres scharf trennbar sind, und wenn die Daten in einer Abtrennung voneinander und von den Systemen ihrer Bearbeitung, Verbindung und Präsentation ihre Benutzbarkeit und damit ihren Sinn verlieren würden, dann muss möglicherweise eine andere Terminologie eingeführt werden, die sowohl den Charakter dieser Daten besser beschreibt, als auch eine Grundlage für die praktische Lösung der anstehenden Probleme liefern kann. Wenn man an der Vorstellung festhalten will, dass es einerseits ‚Daten' gibt und es andererseits darum geht, die ‚Ergebnisse der Forschung' zu sichern und nutzbar zu machen, dann kann man diesen Problemkreis möglicherweise auch als Spannungsverhältnis zwischen *Daten* und *Systemen* beschreiben.

Letztlich sind Daten die Grundlage aller wissenschaftlichen Erkenntnis. Und Daten sind archivierbar. Methoden und Systeme zur digitalen Langzeitarchivierung befinden sich in der Entwicklung.[10] Hier sind in den letzten Jahren große Forschungsleistungen

[10] An dieser Stelle sei stellvertretend das Projekt zum Aufbau eines Digitalen Archivs Nordrhein-Westfalen (DA NRW) genannt, vgl. Manfred Thaller (Hrsg.): Das Digitale Archiv NRW in der Praxis. Eine Softwarelösung zur digitalen Langzeitarchivierung (Kölner Beiträge zu einer geisteswissenschaftlichen Fachinformatik, Bd. 5), Hamburg 2013, http://www.danrw.de/.

erbracht worden, so dass davon auszugehen ist, dass in den kommenden Jahren zuverlässige Strukturen und Systeme verfügbar sein werden, um digitale Daten dauerhaft speichern, beschreiben, adressieren und wiederverwenden zu können. Das Angebot der Sicherung von Daten im engeren Sinne wird also zukünftig vorhanden sein. Davon unberührt und deshalb hier auch nicht zu diskutieren, ist das Problem der Nachfrage und der Bedürfnisse der *abliefernden* und der *nachnutzenden* Forscher. Für die erste Gruppe ist festzustellen, dass viele Forscher sich der Relevanz ihrer Primärdaten noch gar nicht bewusst sind oder nicht wollen, dass die Primärdaten sichtbar werden und nachgenutzt (und damit überprüft) werden können. Dies ist eine Frage der Wissenschaftsmentalität und Wissenschaftspolitik. Sie sollte bei Überlegungen zur Einrichtung von Datenzentren mit bedacht werden, kann aber von ihnen nicht gelöst werden.

Für die andere Seite, also die Nachnutzung der Daten, ist ebenfalls unklar, ob sich in den Geisteswissenschaften überhaupt eine Kultur der Zweitverwertung einmal erhobener und erschlossener Daten *auf der Ebene der Daten* und nicht auf der Ebene der Präsentations- und Publikationsformen etablieren wird. Der gegenwärtige Trend in den Digital Humanities zur Arbeit mit ‚big data' weist in diese Richtung. Trotzdem ist zu bedenken, dass die Auswahl und Aufbereitung von Daten in den Geisteswissenschaften häufig so sehr an eine bestimmte Fragestellung und eine bestimmte theoretische sowie methodische Perspektive gebunden sind, dass eine unmittelbare Übernahme fremder Grunddaten nicht ohne Weiteres möglich sein wird. Deshalb ist auch die genaue Dokumentation dieser Daten von hoher Bedeutung.

Jenseits der kulturellen Spezifika bleibt die Grundaufgabe klar: Forschungsdaten müssen in allgemeine Speicherinfrastrukturen übernommen, gesichert, beschrieben, nachgewiesen, auffindbar und zugänglich gemacht werden. Die Aufgabe eines Forschungsdatenmanagements im Rahmen von Datenzentren liegt darin, diese Sicherung und Zugänglichkeit durch gute Dokumentation und die Verwendung und Anwendung geeigneter Beschreibungsstandards und Formate zu unterstützen. Zu den Besonderheiten der Geisteswissenschaften gehört es dabei allerdings, dass (1.) die bestehenden Standards entweder sehr komplex (zum Beispiel TEI) oder in ihrer Anwendung unklar beziehungsweise sehr frei (zum Beispiel DC) sind, es (2.) eine große Vielfalt von Standards für die unterschiedlichen Wissensbereiche und Gegenstände gibt und es (3.) eher im Regel- als im Ausnahmefall eigene, stark idiosynkratische lokale Modelle für komplexe Wissensbestände gibt. Dies alles erhöht die Schwierigkeit der Nachnutzung von Primär- oder Zwischendaten. Es erhöht aber auch die Schwierigkeit der gemeinsamen Speicherung, Verwaltung und Bereitstellung der Daten in übergreifenden Datenzentren. Vor allem scheint es das grundsätzliche Paradigma von klar abgrenzbaren ‚Datensammlungen' oder gar einzelnen ‚Records' in Frage zu stellen. Für ein geisteswissenschaftliches Datenzentrum scheint die Chance, ‚Datensätze' aus verschiedenen Quellen und Projekten ohne große Informationsverluste unter einem gemeinsamen Schema verwalten und anbieten zu können, derzeit jedenfalls noch gering zu sein. Zweifellos muss im Forschungsdatenmanagement in diesem Bereich noch viel geleistet werden, um dafür zu sorgen, dass auch geisteswissenschaftliche Daten interoperabler oder wenigstens anschlussfähiger werden, um sie für andere Forschungsfragen oder in anderen Informationssystemen nachnutzbar zu machen. Diese Herausforderungen gilt es anzunehmen und durch die

Schaffung entsprechender Anreize Lösungswege zu finden, die zu einer maximalen Interoperabilität führen. Die bisherigen praktischen Erfahrungen deuten darauf hin, dass auch zwischen einer rein technischen beziehungsweise syntaktischen Interoperabilität und einer echten semantischen Interoperabilität zu unterscheiden ist. Für die letztere scheinen wiederum Absprachen, abgestimmte Praktiken und präzise Anwendungsrichtlinien entscheidend zu sein, die am ehesten in begrenzten Teil-Forschungsgemeinschaften etabliert werden können.

Wenn eine Nachnutzung von Grund- und Ausgangsdaten bislang eher schwierig und noch wenig verbreitet ist, dann kommt der Bereitstellung der wissenschaftlichen Erträge in den Systemen der Präsentation eine besondere Bedeutung zu. Offensichtlich liegt die am weitesten verbreitete Nutzung von Daten in den Geisteswissenschaften immer noch in ihrer browsenden, suchenden und dann lesenden Rezeption durch die Forschung. In vielen, vermutlich sogar in den meisten Fällen, entstehen in den Projekten eigene, maßgeschneiderte Präsentationssysteme. Nur diese scheinen das Potential der Inhalte (der Daten) in einer ersten Nutzung wirklich zielgenau auszuschöpfen. Um die Nutzbarkeit der Forschungsleistungen sicherzustellen, scheint es deshalb unumgänglich, solche Systeme dauerhaft zu pflegen und am Laufen zu halten.

Auf der Theorieebene ist damit klar, dass nicht nur zwischen der Aufbewahrung und der Nutzung stärker unterschieden werden muss.[11] Vielmehr muss bei der Nutzung auch differenziert werden, ob hier die Nutzung der reinen Daten oder von Systemen gemeint ist, die diese Daten präsentieren. Daten sind, wenn sie aufbewahrt werden, deshalb noch lange nicht unmittelbar oder gar einfach nutzbar. Die Nutzbarkeit setzt vielmehr, neben der Beschreibung nach möglichst allgemeinen, verbreiteten Standards und neben der Bereitstellung von Daten an Schnittstellen, oft auch die Bewahrung der präsentierenden Systeme und primären Nutzungsoberflächen voraus. Ganz besonders in den Geisteswissenschaften ist also auch von einer Diskussion nicht nur der Daten, sondern ebenso der Systeme auszugehen, wobei sich die Szenarien der Betreuung durchaus unterscheiden: Daten werden beschrieben, archiviert und zugänglich gemacht – Systeme werden gepflegt, aktualisiert und gewartet.

Problem und Lösung

Die Probleme projektgetriebener Forschung unter digitalen Bedingungen liegen auf der Hand. Daten, die nicht entweder als allgemeine Primärdaten in den Ausgangsinstitutionen wie Bibliotheken und Archiven verbleiben oder als abtrennbare Ergebnisdaten in die Infrastruktur der Publikation bei den Verlagen oder Bibliotheken eingehen, drohen verloren zu gehen. Dies betrifft zunächst die immer breiter

[11] Die grundsätzliche Unterscheidung betrifft eher die rechtlichen Bedingungen der Nachnutzung. So heißt es im DFG-Papier „Ergänzungen der Empfehlungen der Deutschen Forschungsgemeinschaft zur Sicherung guter wissenschaftlicher Praxis" vom Juli 2013: „Bei Primärdaten ist zwischen deren Nutzung und deren Aufbewahrung zu unterscheiden. Die Nutzung steht insbesondere dem/den Forscher(n) zu, die sie erheben. Im Rahmen eines laufenden Forschungsprojektes entscheiden auch die Nutzungsberechtigten (ggfs. nach Maßgabe datenschutzrechtlicher Bestimmungen), ob Dritte Zugang zu den Daten erhalten sollen. Sind an dem Vorgang der Datenerhebung mehrere Institutionen beteiligt, empfiehlt sich, die Frage vertraglich zu regeln.", DFG: Ergänzungen, S. 5.

werdende Schicht der angereicherten oder aufbereiteten Zwischendaten. Dies betrifft aber auch vor allem die komplexen Präsentationssysteme, welche mit dem Ende der Projektfinanzierung meist nicht mehr gewartet werden, vielleicht noch eine Weile weiterlaufen, dann aber irgendwann verwaisen, veralten, Softwareupdates nicht überleben und schließlich ganz abgeschaltet werden. Dass damit Investitionen in die Forschung vergeudet werden und wertvolle Ressourcen verloren gehen, ist offensichtlich.

Ebenso offensichtlich ist, dass diese Probleme nicht auf der Ebene des einzelnen Projekts oder des einzelnen Forschers, vielleicht noch nicht einmal auf der Ebene der einzelnen Forschungseinrichtung gelöst werden können. Dauerhafte Lösungen, die zudem spezielle technische und methodische Kompetenzen verlangen, können nur in einem institutionellen Rahmen geschaffen werden, in dem für eine kontinuierliche Beschäftigung mit dem Problem und für eine anhaltende Betreuung Sorge getragen wird. Die Forschung braucht deshalb Datenzentren, die jenseits der Projekte und Vorhaben eine anhaltende Pflege und Bereitstellung gewährleisten. Diese Datenzentren können zum Beispiel an bestehenden Institutionen wie Bibliotheken, Archiven oder Rechenzentren angesiedelt sein. Es muss aber geklärt werden, ob sie von ihren Anforderungen her in das bestehende Leistungsspektrum und das Kompetenzprofil dieser Einrichtungen passen. Vor allem muss die Frage beantwortet werden, ob Datenzentren eine fachspezifische Ausrichtung haben müssen, die den Rahmen der generischen Angebote der genannten Institutionsarten überschreitet. Wenn diese Frage nämlich mit ja zu beantworten ist, dann sind fachspezifische Datenzentren vielleicht besser an die bestehenden fachspezifischen Strukturen der Wissenschaft anzusiedeln.

Was wiederum diese Strukturen betrifft, so ist jeder Fall ein Sonderfall. An der Universität zu Köln besteht beispielsweise eine ungewöhnlich große Philosophische Fakultät, die alle geisteswissenschaftlichen Fächer umfasst. Es gibt hier eine sehr lange Tradition digitaler Forschung mit einer fast unüberschaubaren Zahl von Projekten und daraus entstandenen Angeboten und Portalen. Schon lange bestehen darüber hinaus zwei dedizierte Lehrstühle und mehrere Studiengänge für digitale Geisteswissenschaften. Etliche Projekte wurden im Bereich der digitalen Langzeitarchivierung durchgeführt oder laufen immer noch. Kölner Partner sind an den großen Infrastrukturprojekten DARIAH *und* CLARIN beteiligt. Seit 2009 gibt es mit dem Cologne Center for eHumanities (CCeH) außerdem ein etabliertes Zentrum, das viele Entwicklungen im Bereich der Digital Humanities zusammenführt, verknüpft und weiter ausbaut. Vor diesem Hintergrund hat die Philosophische Fakultät im Dezember 2012 die Gründung eines Datenzentrums für die Geisteswissenschaften beschlossen und für die erste Aufbauphase mit Stellen ausgestattet. Das ‚Data Center for the Humanites' (DCH) hat dabei den Auftrag, sich zunächst um die Bedürfnisse der Fakultät zu kümmern. Es soll seine Dienste dann aber ebenso anderen interessierten Einrichtungen in Nordrhein-Westfalen und darüber hinaus zur Verfügung stellen.

Das DCH hat dabei einen sehr klar gefassten definitorischen Rahmen. Seine *raison d'être* liegt in der Spezifik der Geisteswissenschaften, ihrer Gegenstände und Methoden. Auf der einen Seite wird davon ausgegangen, dass es die Geisteswissenschaften mit anderen Ausgangsdaten zu tun haben als andere Wissenschaftszweige, die noch am

ehesten benachbarten Sozialwissenschaften eingeschlossen. Sie verfügen zudem über andere Beschreibungsmodelle und -standards, andere Fragestellungen, andere Methoden und Praktiken. Eine Sicherung geisteswissenschaftlicher Forschungsdaten in fachfremden oder fachblinden, oder positiv gewendet: generischen Datenzentren scheint den besonderen Anforderungen dieses Forschungsfeldes nicht gerecht werden zu können. Hier würde ein tieferes Verständnis für die komplexen, häufig unscharfen, implizit stark vernetzten, jedenfalls regelmäßig durch ihre Historizität und Kontextualität entscheidend geprägten ‚Daten' jedenfalls ebenso wenig zu erwarten sein, wie eine breite Kompetenz für die vielfältigen Modelle, Formate und Standards. Jenseits des ‚records'-Paradigmas würden hier außerdem die individuell höchst unterschiedlichen, vielschichtigen ‚collections' und vor allem die geradezu individuellen, projektbezogenen Systeme der Bereitstellung und Präsentation einem Ansatz widersprechen, der aus Effizienzgründen eher nach übergreifenden, dann aber auch für die einzelnen Datenlieferanten verbindlichen Lösungen suchen würde. Wenn davon ausgegangen wird, dass für die Geisteswissenschaften eine besondere Definition von ‚Forschungsdaten' erforderlich ist und hier von anderen Wissenschaftszweigen abweichende Bedürfnisse bestehen, dann liegt die Bündelung von Forschungsdaten in einem spezialisierten Datenzentrum nahe. Dabei gibt es dann wiederum zwei mögliche Orientierungen: Entweder widmet sich ein solches Datenzentrum gezielt *einem* Fach oder einer eng verbundenen Fächergruppe, oder es versteht sich als lokaler oder regionaler Ansprechpartner für *alle* geisteswissenschaftlichen Fächer.

Das DCH hat den letzteren Weg eingeschlagen, um die Forscher vor Ort, aber auch regionale und überregionale Einrichtungen im Bereich des Forschungsdatenmanagements und der dauerhaften Sicherung und Bereitstellung ihrer Daten zu unterstützen. Die Übernahme von Daten aus inzwischen verwaisten oder gar abgestorbenen Projekten, möglicherweise verbunden mit der Revitalisierung oder der Neukonstruktion von Präsentationssystemen und Nutzungsoberflächen ist ein akut anstehendes Szenario, kann aber nur als der drittbester Weg bezeichnet werden, da hier unter Umständen bereits viel Energie in die Rekonstruktion der Intention und Semantik der Datenmodelle und -formate investiert werden muss – von der Neuentwicklung der Präsentationssysteme ganz zu schweigen. Die Übergabe von Daten und Systemen direkt zum Ende der Förderung ist dagegen die zweitbeste Lösung, da hier wenigstens eine nahtlose Fortführung und eine Abstimmung zwischen den Projektbeteiligten und dem Datenzentrum erfolgen kann. Der Königsweg und das beste Szenario sind dahingehend zweifellos die Begleitung des Forschungsprozesses in einem Projekt durch das DCH von Anfang an. Nur so kann sichergestellt werden, dass im Forschungsdatenmanagement und im Aufbau von Systemen der Ergebnispräsentation die Grundlagen für eine spätere dauerhafte Bereitstellung bereits gelegt werden. Bereits jetzt ist absehbar, dass jedes Vorhaben, egal ob abgeschlossen, noch laufend oder jetzt erst startend, eigene Anforderungen haben wird und dass das Datenzentrum für jedes Projekt ein eigenes Angebot der Begleitung und Betreuung wird machen müssen. Nicht nur dafür ist ein enger Kontakt zwischen dem DCH und den Fachwissenschaftlern unerlässlich.

In diesem Kontext ist es nicht unüblich, dass im Bereich des Forschungsdatenmanagements und der Verstetigung von Forschungsleistungen heute noch aktiv Überzeugungsarbeit geleistet werden muss. Vielfach sind die damit

zusammenhängenden Probleme den Fachforschern nämlich noch nicht im wünschenswerten Maße bewusst, so dass hier nicht nur Aufklärung nötig ist, sondern sogar Anreize zur Bereitstellung von Daten und zum nachhaltigen Aufbau von Systemen geschaffen werden müssen. Dabei sollte ein besonderes Augenmerk auf die Zitierbarkeit gelegt werden, die für die Wissenschaftscommunity von unmittelbar einsichtiger Bedeutung ist. Letztlich besteht das Ziel in der *"freiwillige[n] Übergabe der Daten durch den Wissenschaftler"*, wie es beispielsweise die DFG in den *Empfehlungen zur gesicherten Aufbewahrung und Bereitstellung digitaler Forschungsprimärdaten* formuliert.[12] Ungeachtet der Postulat der Forschungsförderungsorganisationen müsste es ohnehin im Interesse der *abliefernden* Forscher sein, dass die Archivierung, Vorhaltung, Zitierfähigkeit, Kreditierung und dauerhafte Sichtbarkeit ihrer Forschungsleistungen durch entsprechende Datenzentren unterstützt und gewährleistet wird.

Strategie und Aufbau des DCH

Das DCH wurde Ende 2012 an der Philosophischen Fakultät der Universität zu Köln gegründet, um den Belangen, Bedürfnissen und Besonderheiten der geisteswissenschaftlichen Fachforschung hinsichtlich eines professionellen Forschungsdatenmanagements und der dauerhaften Sicherung von Forschungsleistungen gerecht zu werden. In der Aufbauphase des Datenzentrums wird eine doppelte Strategie verfolgt, bei der sich ein bottom-up-Ansatz und ein top-down-Ansatz gegenseitig ergänzen. Im top-down-Ansatz liegt der Fokus auf der Konzeption, im bottom-up-Ansatz auf der Realisation eines Modells, das auf mehreren Schichten verschiedene modulare Komponenten verbindet. Archivierung, Bereitstellung, Adressierbarkeit, Präsentation und die Nutzung von Diensten und Werkzeugen bauen in diesem Modell aufeinander auf. Sie sind aber auch unabhängig voneinander nutzbar, so dass für jedes Projekt ein individuelles Leistungsprofil angeboten werden kann. Dadurch soll insgesamt eine (1) langfristige Sicherung, (2) dauerhafte Bereitstellung, (3) allgemeine Zugänglichkeit und (4) erhöhte Sichtbarkeit der Forschungsdaten gewährleistet werden, welche zugleich eine bessere Vernetzung der Projekte und Daten bedeutet und die Grundlage für eine zukünftige Nutzung in Forschung und Lehre bildet (vgl. Abb. 1).

[12] Vgl. Marleen Burger u.a. Forschungsdatenmanagement an Hochschulen. Internationaler Überblick und Aspekte eines Konzepts für die Humboldt-Universität zu Berlin, Version 1.1., 03.06.2013, Berlin 2013, S. 17, https://edoc.hu-berlin.de/docviews/abstract.php?lang=ger&id=40138 sowie DFG: Empfehlungen zur gesicherten Aufbewahrung und Bereitstellung digitaler Forschungsprimärdaten, Januar 2009, Bonn 2009, S. 2, http://www.dfg.de/download/pdf/foerderung/programme/lis/ua_inf_empfehlungen_200901.pdf.

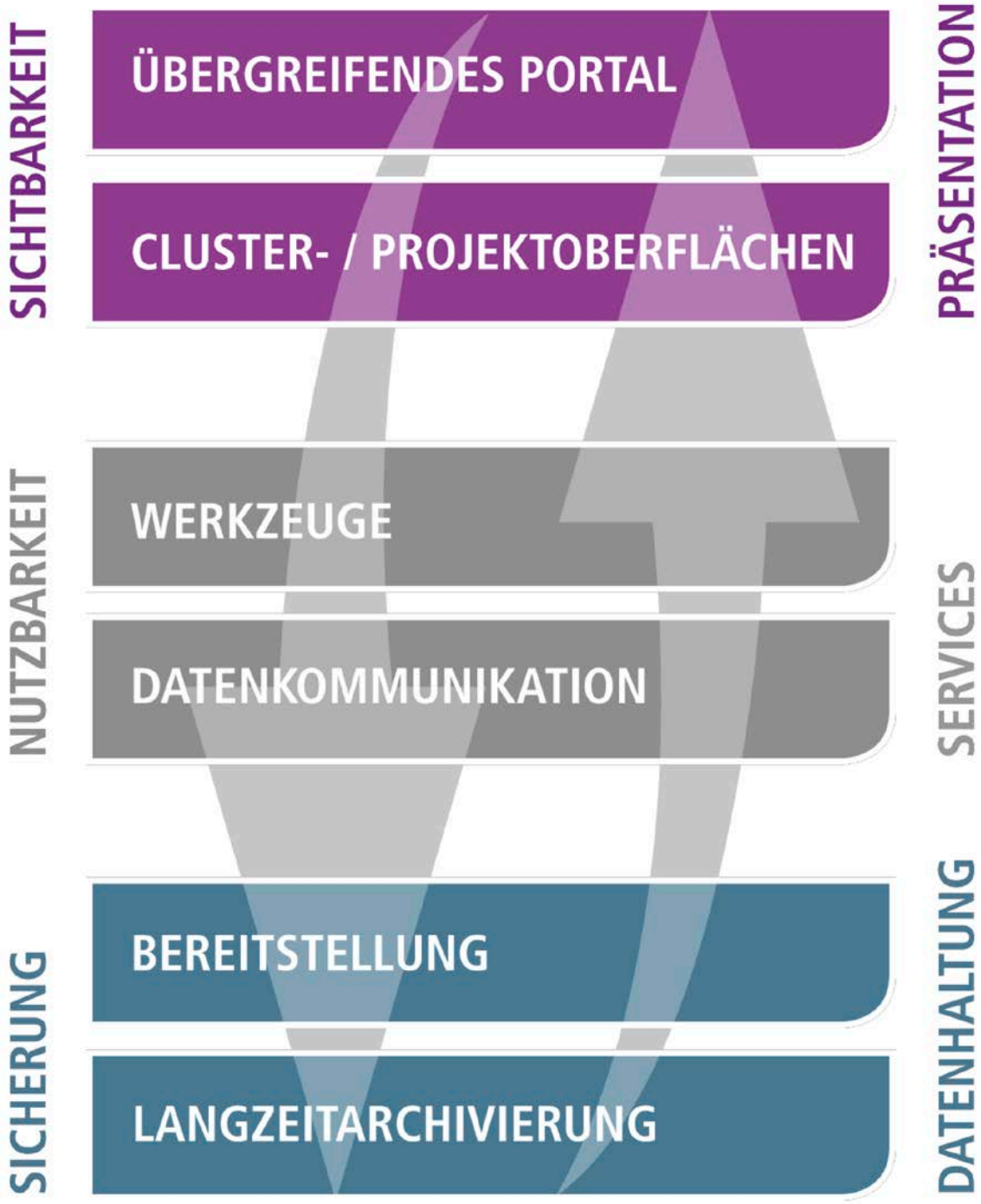

Abbildung 1: Schichtenmodell des Data Center for the Humanities (DCH).

Konkret lassen sich die einzelnen Schichten wie folgt beschreiben:

Schicht 1: Datensicherung und Datenvorhaltung.

Die langfristige Sicherung und Vorhaltung der digitalen Daten basiert auf den Archivierungs- und Speichersystemen der Universität zu Köln, welche vom Regionalen Rechenzentrum der Universität (RRZK) gemäß dem jeweiligen Nutzerbedarf betrieben werden. Dadurch wird einerseits eine Langzeitarchivierung der Forschungsdaten

ermöglicht und andererseits der andauernde direkte Zugriff auf die gespeicherten Daten gewährleistet, da sie tatsächlich permanent vorgehalten werden. In diesem Kontext sollen auch Dienste von institutionalisierten Digitalen Archiven eingebunden werden. Für ein zukünftiges Digitales Archiv NRW (DA NRW) läuft derzeit ein Pilotprojekt, das die technischen Lösungen für eine solche Einrichtung entwickelt.[13] Sowohl für die Archivierung als auch für die permanente Bereitstellung ist die Beschreibung der Daten über Metadaten von entscheidender Bedeutung. Aufgrund der Heterogenität geisteswissenschaftlicher Forschungsdaten ist zudem der Dokumentation von Projekten und Sammlungen sowie der Beschreibung von Daten beziehungsweise Datensammlungen durch geeignete Metadatenschemata eine besondere Beachtung zu widmen. Das DCH wird dabei versuchen, sich an der Entwicklung geeigneter Standards und Praktiken zum Beispiel im Rahmen von 'Project Description Languages' oder 'Collection Level Description Languages' zu beteiligen. Hinsichtlich der dauerhaften Referenzierung und Zitation von Daten spielen Persistent Identifier (PI) und PID-Systeme eine zentrale Rolle. Hier liegt die Präferenz im Moment beim DOI-System, wobei nicht ausgeschlossen ist, dass die angestrebte Konvergenz zu den parallelen Infrastrukturentwicklungen auf nationaler und internationaler Ebene zur Einbindung weiterer PID-Systeme führen wird. Die Spezifik geisteswissenschaftlicher Daten und die Tradition ihrer Zitierweisen legen es außerdem nahe, zu prüfen, ob für bestimmte Ressourcen nicht eher – oder zusätzlich – sprechende PURL-Systematiken verwendet werden sollten, um ihrer hohen Komplexität und Granularität gerecht zu werden.

Schicht 2: Schnittstellen, Dienste, Werkzeuge.

Im Unterschied zu anderen Initiativen (wie unter anderem IANUS) ist das DHC nicht fachspezifisch, sondern innerhalb der geisteswissenschaftlichen Disziplinen fachübergreifend ausgerichtet. Um den unterschiedlichen Anforderungen aus den verschiedenen Forschungsprojekten gerecht zu werden, werden deshalb intern die verschiedenen Bedürfnisse aus den Fachwissenschaften in drei 'Clustern' für 'Sprache', 'Dokumente' und 'Objekte' organisiert. In diesen fachgruppen- beziehungsweise materialorientierten Clustern sollen Projekte und Datensammlungen aus verschiedenen geisteswissenschaftlichen Bereichen zusammengefasst und gemeinsam zugänglich gemacht werden. Dabei müssen Sprachressourcen, Dokumentdaten und Objektdaten in der gleichen Weise abrufbar sein, sie haben aber teilweise ihre eigenen Standards und Werkzeuge und erfordern daher spezifische Fachkompetenzen, um beispielsweise im Bereich des Forschungsdatenmanagements optimal betreut werden zu können.

Auf der Ebene der fach- beziehungsweise materialspezifischen Services wird deshalb dafür gesorgt, dass alle vorliegenden Daten abgerufen werden können und auch für automatisierte Abfragen zugänglich sind. Für den Datenzugriff müssen (soweit rechtlich möglich) definierte Schnittstellen mit definierten Protokollen vorhanden sein, wie allgemeine Harvesting-Schnittstellen wie OAI-PMH. Daneben sind Dienste vorgesehen, die den Zugriff über spezialisierte REST- oder SOAP-Schnittstellen auf Daten von einzelnen Projekten, Projektgruppen oder Teilarchiven ermöglichen.

[13] Vgl. DA NRW (Digitales Archiv Nordrhein-Westfahlen), http://www.danrw.de/.

Das DCH versteht sich nicht nur als reiner Datenspeicher, sondern es wird auch Werkzeuge für die Verarbeitung und Analyse von Daten hosten. Dazu gehören neben bereits etablierten Browser-Webdiensten wie dem DFG-Viewer weitere Services, welche in oder für verschiedene geisteswissenschaftliche Forschungsprojekte entwickelt werden. Auch diese müssen gesichert, wenn möglich standardisiert und damit für andere Projekte zur Nachnutzung zur Verfügung gestellt werden. Mit dem *Language Archive Cologne* (LAC) und dem CLARIN-Kurationsprojekt *Cologne Language Archive Services* (CLASS) wird derzeit eine Arbeitsumgebung realisiert, die webbasierte Dienste komfortabel an Sprachressourcen anbindet. CLASS implementiert hierzu die Poio-API, eine Umsetzung des ISO-Standards 24612, zur generischen und formatübergreifenden Interaktion mit linguistischen Annotationen. Diese (neu) entwickelten Tools unterstützen das Browsing beziehungsweise die Suche und die Auswahl von Daten und Sammlungen, stellen Konversionsroutinen zur Verfügung oder erlauben die weitere Annotation von Daten, was vor allem für die Nachnutzung in anderen Forschungsarbeiten von Interesse sein kann. Das Beispiel des CLASS-Projektes zeigt für den Umgang mit Daten aus Spracharchiven, wie im Zuge der verschiedenen Projektarbeiten zum Teil Werkzeuge entstehen, die dann auch in das Datenzentrum eingebunden und mit denen die Daten direkt verarbeitet werden können.

Schicht 3: Präsentation und Oberflächen.

Gemäß dem oben beschriebenen Ansatz, was in den Geisteswissenschaften als Forschungsdaten zu sichern ist, besteht der besondere Anspruch des DCH darin, die Präsentations- und Nutzungsoberflächen einzelner Projekte und Projektverbünde soweit zu bewahren und dauerhaft zu pflegen, wie es den Bedürfnissen der jeweiligen Fachforschung und den verfügbaren Ressourcen entspricht. Dies kann auch die dauerhafte Bereitstellung von Ergebnispräsentationen beinhalten, so wie sie aus den einzelnen Projekten hervorgehen. Zu diesen ‚lebenden Systemen', welche dauerhaft gepflegt, aktualisiert und gewartet werden müssen, gehören Webapplikationen oder Sammlungspräsentationen, um nur einige Beispiele zu nennen. Dadurch soll schließlich auch erreicht werden, dass solche digitalen Leistungen zitierfähig bleiben und die Sichtbarkeit und Anerkennung der dahinter stehenden wissenschaftlichen Leistungen gesichert werden.

Die Präsentation von Daten und Forschungsergebnissen erfolgt insgesamt auf drei verschiedenen Wegen:

1. Auf der Ebene (einzelner) Projekte wird – soweit erforderlich und möglich – für eine stabile und kontinuierliche Erreichbarkeit der jeweils eingesetzten (Backend-)Systeme und Oberflächen gesorgt. Dazu werden inhaltliche und zeitliche Garantien nur soweit übernommen, wie sie sich aus expliziten Betreuungsvereinbarungen beziehungsweise Service-Level-Agreements (mit entsprechender Ressourcenbereitstellung) ergeben. Auf dieser Ebene ist dafür zu sorgen, dass jedes System nur so speziell oder individuell wie nötig und so generisch wie möglich ist, um den Wartungsaufwand zu minimieren. Es muss aber auch klar sein, dass es vorkommen *kann*, dass Präsentationssysteme (zum

Beispiel durch Updates der beteiligten Softwarekomponenten) zusammenbrechen und nicht mehr im Rahmen der verfügbaren Ressourcen revitalisiert werden können. In diesem Fall werden die Projekte auf die zuvor beschriebene Schicht der reinen Datenbereitstellung durch Schnittstellen zurückgeführt.

2. Im Gegensatz zur Bewahrung der Nutzungsoberflächen auf der Ebene des Einzelprojekts richtet sich die Präsentation der Fachbereichs-Cluster vornehmlich an die einzelnen Fachgruppen. Wenn sich die Geisteswissenschaften einerseits von anderen Wissenschaftszweigen abgrenzen lassen, sie aber andererseits auch über *gemeinsame* Fragestellungen, Ausgangsmaterialien und Methoden verfügen, dann lassen sich diese in einem zweiten Schritt erneut differenzieren und Binnengruppen identifizieren. Im DCH setzen auf die infrastrukturelle Basis fachgruppen- beziehungsweise materialorientierte Säulen auf, die Projekte und Sammlungen aus verschiedenen Bereichen zusammenfassen. Auch auf der Ebene der Präsentation sind die drei Cluster ‚Sprache' (zum Beispiel Linguistik, auch feldforschende Linguistik und multimodale Sprachdaten), ‚Dokumente' (zum Beispiel Literaturwissenschaften, Geschichte, Philosophie und ähnliche) und ‚Objekte' (zum Beispiel Archäologie, Kunstgeschichte) herauszubilden, [14] die inhaltlich, organisatorisch und in der Präsentation zu unterscheiden sind, weil diese Bereiche zwar über gemeinsame Datentypen (Text, Bild, Metadaten) verfügen, aber sich teilweise durch spezielle Formate, Standards und Services unterscheiden. In diesen Clustern werden die einzelnen Projekte angebunden, sichtbar gemacht und soweit wie möglich aufeinander abgestimmt. Hier werden außerdem die erforderliche Fachkompetenz zur inhaltlichen Betreuung sowie die notwendige technische Kompetenz zur kontinuierlichen Pflege der laufenden Präsentationssysteme gebündelt.

3. Schließlich wird der zentrale Nachweis über alle Inhalte, Projekte und Datensammlungen des DCH durch ein zentrales Portal, das Gesamtportal des Datenzentrums erfolgen. Dadurch werden insgesamt eine übergreifende Suche und ein Browsing nach Projekten, Sammlungen oder Daten ermöglicht.

[14] An dieser Stelle muss darauf hingewiesen werden, dass zu beiden Clustern 'Sprache' und 'Dokumente' natürlich auch 'Texte' gehören; allerdings liegen diesen jeweiligen Texten meist unterschiedliche Daten, Datenmodelle, Datenformate, Standards oder Methoden, sowie je nach Fragestellung und Disziplin eine andere theoretische Fundierung und Ausrichtung zugrunde, so dass es einerseits sinnvoll ist, dies durch Cluster zu trennen, andererseits lassen sich Überschneidungen im Bereich der Korpora nicht vermeiden.

Um die aufgezeigten Ziele zu erreichen, ist die Zusammenarbeit des DCH mit verschiedenen Akteuren und ihren jeweiligen Kompetenzen grundlegend. Im Zentrum stehen dabei zunächst die Fachwissenschaftler und ihre Forschungsprojekte, deren Inhalte (Forschungsdaten und Publikationen) nachhaltig gesichert werden müssen. Auf der technischen Ebene ist das RRZK vor allem für die Bereitstellung und Pflege einer Server-Infrastruktur unverzichtbar. Die Universitäts- und Stadtbibliothek (USB) Köln übernimmt jene Teilaufgaben des Datenzentrums, die in ihren eigenen Aufgaben- und Kompetenzbereich fallen. Dazu gehört die Aufnahme von traditionellen Ergebnisdaten – also Texten – in ihrem institutionellen Repositorium, das Hosting flach erschlossener digitaler Sammlungen, sowie die Mitarbeit bei der Beschreibung von Projekten, Sammlungen und Objekten auf der Ebene der Metadaten. Das CCeH (Cologne Center for eHumanities) als fakultätsweite Einrichtung der digitalen Forschung übernimmt die Gesamtorganisation des Datenzentrums sowie die Koordination für das Zusammenspiel der einzelnen Komponenten und deren Aufbau. Dazu gehören die Einrichtung der fachgruppenspezifischen Cluster, die Betreuung der Inhalte in diesen Teilzentren sowie die Entwicklung von Services und Präsentationsoberflächen. Dabei werden die Präsentationssysteme gehostet und dauerhaft gepflegt, Services zur Datenkommunikation aufgebaut und Tools zur Arbeit mit den Daten bereitgestellt.

Für die konkrete Umsetzung der skizzierten Pläne ist die genannte Doppelstrategie aus bottom-up- und top-down-Ansatz maßgeblich. Im bottom-up-Ansatz werden derzeit erste konkrete Bausteine für das zu errichtende Gesamtsystem entwickelt, da es gilt, die anstehenden Probleme *jetzt* zu lösen und nicht weiter in die Zukunft zu verschieben. In diesen Ansatz fließen die theoretischen Erfahrungen und praktischen Bausteine verschiedener aktueller Projekte (wie zum Beispiel CLASS oder LAC der Sprachwissenschaften) mit in die Entwicklung und den Aufbau des Datenzentrums ein. Gleichzeitig wird im Rahmen des komplementären top-down-Ansatzes an einem umfassenden Entwicklungsplan für den nachhaltigen Aufbau des DCH gearbeitet, der sich auf nationaler wie internationaler Ebene in die laufende Infrastrukturentwicklung von Datenzentren für die Geisteswissenschaften und ihre nationale Koordination einpasst. Unter der Einbeziehung der aktuellen Entwicklungen werden hier institutionelle und strukturelle Fragen ebenso wie methodische und technische Probleme thematisiert und Lösungen dazu erarbeitet. Vor diesem Hintergrund gilt es nicht nur, die Abstimmung mit den oben genannten Partnern des DCH zu konkretisieren und zu kodifizieren, sondern auch für die nachhaltige institutionelle Absicherung zu sorgen. Denn nur durch das eindeutige Bekenntnis zur dauerhaften Übernahme der Verantwortung durch Fakultät und Universität wird es möglich, dass das DCH als zuverlässiger und beständiger Teil der Infrastruktur der Langzeitarchivierung und des Forschungsdatenmanagements auftritt und sich als Akteur und innovativer Partner in der Wissenschaftsgemeinschaft positionieren kann. Dies ist auch die Voraussetzung für die angestrebte Vernetzung und Abstimmung mit anderen Einrichtungen auf nationaler wie internationaler Ebene (wie DARIAH oder an anderen Standorten geplanten geisteswissenschaftlichen Datenzentren), bei der alle Beteiligten voneinander profitieren. Durch die bestehenden und angestrebten Kooperationen und Partnerschaften würde das DCH jedenfalls ideal in die sich im Auf- und Ausbau befindlichen Forschungs- und Informationsinfrastrukturen für die

Geisteswissenschaften eingebettet.[15] Die Zusammenarbeit mit den weiteren Partnern vor Ort ermöglicht es, dass sich das DCH einerseits auf seine Stärke – nämlich die Spezifik geisteswissenschaftlicher Daten – konzentrieren und andererseits für infrastrukturelle Basisdienstleistungen auf die Expertise und langjährige Erfahrung seiner etablierten Partner zurückgreifen kann. Zusammengefasst wird durch die gemeinsamen Anstrengungen versucht, der Problematik einer dauerhaften Bereitstellung von geisteswissenschaftlichen Forschungsdaten aktiv zu begegnen.

Hinsichtlich der konkreten konzeptionellen Umsetzung des DCH stehen vier große Arbeitsbereiche im Vordergrund: (1.) ‚Methodik', (2.) ‚Services', (3.) ‚Technik' und (4.) ‚Organisation'. Die (1.) Methodik umfasst die ganze Spannbreite des klassischen Forschungsdatenmanagements. Sie erstreckt sich von der Datenaufnahme, bei der es um Formate, Standards, Metadaten, Interoperabilität auf technischer sowie semantischer Ebene geht, über die Projektbeschreibung beziehungsweise eine zu definierende *Project Description Language*, um die einzelnen Projekte verwalten zu können, bis hin zu einer anhaltenden Technology Watch, Tool Watch und Scene Watch, um jeweilige neue Entwicklungen zu evaluieren und gegebenenfalls in das Datenzentrum mit einzubinden. Zu den (2.) Services zählt hauptsächlich das Hosting von Daten und Projekten, welches Server-, Speicher- und Archivierungsdienste beinhaltet. Neben dem Monitoring der digitalen Inhalte und der Evaluierung bestehender Dienste und Schnittstellen für die Migration werden beispielsweise auch Authentifizierungs- und PID-Dienste zur Verfügung gestellt und weitere, bereits entwickelte Services anderer Projekte wie zum Beispiel kollaborative Arbeitsumgebungen getestet und integriert. Der (3.) technische Arbeitsbereich umfasst grob die Wartung und technische Betreuung der einzelnen oben beschriebenen Schichten. Neben der Langzeitarchivierung und Datenhaltung gehören dazu vor allem die Service-Ebenen sowie die Pflege der Präsentationschicht(en). Schließlich umfasst der (4.) Aspekt der Organisation sowohl den internen institutionellen Aufbau des DCH als auch die entsprechenden externen Kooperationen mit den jeweiligen für das Forschungsdatenmanagement essentiellen Einrichtungen (CCeH, USB, RRZK et cetera). Da diese Bereiche relativ umfangreich sind, werden an dieser Stelle nur die wichtigsten Stichpunkte genannt: Geschäftsmodell, Zertifizierung, Rechtemanagement, Marketing und Öffentlichkeitsarbeit, Positionierung in der nationalen und internationalen Community, Beratung, Begleitung und Qualifizierung von Wissenschaftlern in ihren Forschungsvorhaben und -projekten, Cluster-Struktur, Use Cases, Best-Practice-Beispiele.

Vier Paradigmen

Zeitgemäße Datenzentren für die Geisteswissenschaften stehen vor der Aufgabe, die Leistungen der Forschung auf Dauer zugänglich und nutzbar zu halten. Die Komplexität dieser Herausforderung ergibt sich hauptsächlich daraus, dass sie sich

[15] Eine Übersicht über die aktuellen Entwicklungen im Bereich der Forschungsinfrastrukturen für die Geistes- und Sozialwissenschaften bietet eine BMBF-Broschüre, welche im Februar 2013 veröffentlicht wurde und die explizit auf einen ‚Nachholbedarf in den Geistes- und Sozialwissenschaften' hinweist, vgl. BMBF (Hrsg.): Forschungsinfrastrukturen für die Geistes- und Sozialwissenschaften, Stand: Februar 2013, Mühlheim an der Ruhr 2013, S. 3f., http://www.bmbf.de/pub/forschungsinfrastrukturen_geistes_und_sozialwissenschaften.pdf.

nicht in der Archivierung klar abgrenzbarer Daten, Datensätze oder Datensammlungen erschöpft. Würde der Auftrag nämlich so eng definiert, wären die Daten zwar *theoretisch* gesichert und zugänglich gemacht. Angesichts der Arbeitspraxis der Geisteswissenschaften und der Bedeutung spezieller Systeme der Präsentation würden sie aber tatsächlich kaum genutzt werden können. Ein zeitgemäßes Datenzentrum für die Geisteswissenschaften sollte deshalb insgesamt vier Paradigmen verfolgen, die zugleich eine plakative Beschreibung seiner Aufgaben und Leistungen liefern können. Dabei sind so weit wie möglich andere bereits bestehende Einrichtungen als Partner einzubinden. Alle Leistungen und Funktionen, für die es bereits zuständige Einrichtungen und funktionierende Lösungen gibt, sollen im Rahmen des Datenzentrums selbstredend *nicht* neu geschaffen, sondern in das Leistungsspektrum integriert werden.

Erstes Paradigma: Archiv.

Ein Datenzentrum erfüllt eine Archivfunktion. Daten werden übernommen, gegebenenfalls kassiert, physisch gesichert, beschrieben, katalogisiert, wenn nötig in andere Formate konvertiert und für den Abruf bereitgestellt. Im Vordergrund steht hier aber zum einen die physische Sicherung und zum anderen die grundsätzliche Auffindbarkeit und Abrufbarkeit. Dies kann bedeuten, dass Daten nicht direkt abgerufen werden, sondern erst auf Anforderung aus den Sicherungssystemen ausgelesen werden. Im Idealfall erfolgt die Datenarchivierung in Zusammenarbeit mit generischen digitalen Archiven.

Zweites Paradigma: Bibliothek.

Ein Datenzentrum erfüllt eine Bibliotheksfunktion. Dieses Paradigma umfasst eine tiefer gehende, auch fachwissenschaftlich kompetente Erschließung und Verzeichnung. Daten und Sammlungen werden hier mit persistenten Adressen auf verschiedenen Granularitätsebenen versehen und stehen für einen permanenten und unmittelbaren Abruf über standardisierte Schnittstellen zur Verfügung. Ein zentraler Katalog erlaubt die übergreifende Recherche in allen Beständen. Übersichten über die Projekte und Sammlungen erlauben außerdem eine stöbernde Orientierung über den Gesamtbestand und damit über den möglichen Suchraum. Das Datenzentrum als Bibliothek implementiert außerdem ein formalisiertes Rechtemanagement, das für jede Sammlung die Zugänglichkeit für bestimmte Gruppen oder individuelle Nutzer regelt. Die Bibliotheksfunktion reicht von der generischen Server-Infrastruktur, die zum Beispiel von einem Rechenzentrum betrieben wird, über die eingespielten Leistungen einer wissenschaftlichen Bibliothek bis hin zu neu zu schaffenden Angeboten und Routinen des Datenzentrums.

Drittes Paradigma: Museum.

Die stark miteinander verwobenen Ausgangs-, Zwischen- und Ergebnisdaten der geisteswissenschaftlichen Forschung müssen häufig auch jenseits ihrer grundsätzlichen Abrufbarkeit in ihren jeweiligen Systemen der Präsentation zugänglich und nutzbar bleiben. Wie in einem Museum muss ein geisteswissenschaftliches Datenzentrum die Projekte und Sammlungen vorstellen, ‚präsentieren' und so einzeln, in Gruppen oder in gezielten Zusammenstellungen

sichtbar und nutzbar machen. Die Daten bleiben nicht nur abstrakt hinter ihren beschreibenden Metadaten, über die sie abgerufen werden können. Die Benutzung sollte nach Möglichkeit keine intellektuelle Rekonstruktion der Entstehungszusammenhänge und der verwendeten Datenmodelle zur Voraussetzung der Nachnutzung machen. Vielmehr müssen mindestens jene Ressourcen, die von anhaltendem Interesse sind, auch anhaltend, gewissermaßen als Dauerausstellung präsentiert werden. Dabei impliziert das Bild von der Museumsfunktion auch, dass Ausstellungen veralten und – wenn das Interesse dies rechtfertigt und die Ressourcen es erlauben – von Zeit zu Zeit aktualisiert werden (müssen).[16] Der Begriff des ‚data curation' ist für das Forschungsdatenmanagement und die Langzeitarchivierung bereits eingeführt, um die Pflege von Forschungsdaten zu beschreiben. Das ‚Kuratieren' von Sammlungen und Projektergebnissen ist aber ebenso unter der Perspektive der Museumsfunktion eines Datenzentrums sinnvoll, bei der aus den Daten von Zeit zu Zeit neue ‚Ausstellungen' organisiert werden. Die Museumsfunktion scheint derzeit nur im Rahmen neu aufzubauender Kompetenzen und Leistungsangebote eines Datenzentrums realisierbar zu sein.

Viertes Paradigma: Werkstatt.

Daten und Ressourcen befinden sich – so behauptet zumindest die Theoriebildung – in stetiger Bearbeitung, Anreicherung, Erschließung und Weiterentwicklung. Digitale Bibliotheken und Informationsportale entwickeln sich zu virtuellen Forschungsumgebungen weiter, die auch Werkzeuge zur wissenschaftlichen Arbeit integrieren. Die Systeme der Bereitstellung von Forschungsdaten sollen neue Ergebnisse direkt einbinden. Das Datenzentrum pflegt solche Systeme und soll daneben auch andere Werkzeuge bereitstellen, die eine Bearbeitung von Daten unterstützen. Wenn die Oberflächen der Benutzung aber auch die Umgebungen zur aktiven Auseinandersetzung mit den Daten werden, dann wird ein geisteswissenschaftliches Datenzentrum zur Werkstatt, in der die Arbeit kontinuierlich fortgesetzt werden kann. Hier ist allerdings darauf zu achten, dass solche Werkzeuge möglichst in Abstimmung und im Austausch mit anderen Datenzentren entwickelt und allgemein nutzbar gemacht werden – schließlich sind sie nicht durch die lokal verfügbaren Daten bestimmt, sondern durch Materialgruppen, Datenmodelle und Formate, die sich in den Datenzentren wiederholen dürften.

Fazit

Die Überlegungen zu Datenzentren in den Geisteswissenschaften am Beispiel des Data Center for the Humanities (DCH) haben gezeigt, dass aufgrund der Unterschiede zwischen den Natur- und Geisteswissenschaften die eher von den Naturwissenschaften ausgehenden Methoden und Praktiken des Forschungsdatenmanagements und der Forschungsinfrastrukturen nicht einfach auf die Geisteswissenschaften übertragen werden können. Aufgrund der besonderen Charakteristika geisteswissenschaftlicher

[16] Das Konzept der Museumsfunktion im Gegensatz zur Archivfunktion bei digitalen Daten ist für den Bereich der digitalen Editionen bereits von Edward Vanhoutte etabliert worden, vgl. Edvard [so der Vorname dort fälschlich] Vanhoutte: Where is the editor? Resistance in the creation of an electronic critical edition, in: Human IT, 1 (1999), http://etjanst.hb.se/bhs/ith/1-99/ev.htm.

Forschungsprozesse und Forschungsdaten müssen hier spezifische Lösungen entwickelt werden, die sich aber möglichst konvergent zu den generischen Lösungen oder den Lösungen auf benachbarten Feldern verhalten sollten.

Die Initiative der Philosophischen Fakultät der Universität zu Köln zur Gründung eines geisteswissenschaftlichen Datenzentrums versucht, diesen Herausforderungen aktiv zu begegnen. Die Sicherung und Zugänglichkeit geisteswissenschaftlicher Forschungsdaten und Forschungsprozesse im Rahmen eines professionellen Forschungsdatenmanagements durch das DCH soll einerseits die Überprüfbarkeit der Ergebnisse im Sinne der guten wissenschaftlichen Praxis verbessern und andererseits die Grundlage für weitere zukünftige Forschungsarbeiten schaffen. Es ist zu hoffen, dass es der Kölner Universität, die sich ihrer Verantwortung für ihre Forschungsaktivitäten auf eine so konstruktive Weise gestellt hat, gelingt, die weitere Institutionalisierung des Datenzentrums zu sichern.

Literaturverzeichnis

Blanke, Tobias, Hedges, Mark: Scholarly Primitives. Building Institutional Infrastructure for Humanities e-Science. In: F*uture Generation Computer Systems, 29/2 (2013),* 654-661. DOI: 10.1016/j.future.2011.06.006, URL: http://linkinghub.elsevier.com/retrieve/pii/S0167739X11001178.

Borgman, Christine L.: Scholarship in the Digital Age. Information, Infrastructure, and the Internet, Cambridge/MA 2007.

Bundesministerium für Bildung und Forschung (BMBF) (Hrsg.): Forschungsinfrastrukturen für die Geistes- und Sozialwissenschaften, Stand: Februar 2013, Mühlheim an der Ruhr 2013, URL: http://www.bmbf.de/pub/forschungsinfrastrukturen_geistes_und_sozialwissenschaften.pdf.

Burger, Marleen u.a.: Forschungsdatenmanagement an Hochschulen. Internationaler Überblick und Aspekte eines Konzepts für die Humboldt-Universität zu Berlin, Version 1.1., 03.06.2013, Berlin 2013, URL: https://edoc.hu-berlin.de/docviews/abstract.php?lang=ger&id=40138.

Burrows, Toby: Sharing Humanities Data for e-Research. Conceptual and Technical Issues, in: Sustainable data from digital research. Humanities perspectives on digital scholarship. Proceedings of the conference held at the University of Melbourne, 12-14th December 2011, Sydney 2011, URL: http://ses.library.usyd.edu.au/handle/2123/7938.

Deutsche Forschungsgemeinschaft (DFG): Vorschläge zur Sicherung guter wissenschaftlicher Praxis: Empfehlungen der Kommission „Selbstkontrolle in der Wissenschaft“, Denkschrift, Weinheim 1998, URL: http://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/download/empfehlung_wiss_praxis_0198.pdf.

Dies.: Empfehlungen zur gesicherten Aufbewahrung und Bereitstellung digitaler Forschungsprimärdaten, Januar 2009, Bonn 2009, URL: http://www.dfg.de/download/pdf/foerderung/programme/lis/ua_inf_empfehlungen_200901.pdf.

Dies.: Die digitale Transformation weiter gestalten - Der Beitrag der deutschen Forschungsgemeinschaft zu einer innovativen Informationsinfrastruktur (AWBI-Positionspapier), 03.07.2012, Bonn 2012, URL: http://www.dfg.de/download/pdf/foerderung/programme/lis/positionspapier_digitale_transformation.pdf.

Dies.: Ergänzungen der Empfehlung der deutschen Forschungsgemeinschaft zur Sicherung guter wissenschaftlicher Praxis, Juli 2013, Bonn 2013, URL: http://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/download/empfehlung_wiss_praxis_0198_ergaenzungen.pdf.

European Science Foundation: Research Infrastructures in the Digital Humanities. Science Policy Briefing 42, Straßburg 2011, URL:

http://www.esf.org/fileadmin/Public_documents/Publications/spb42_RI_DigitalHumanities.pdf.

Higgins, Sarah: The DCC Curation Lifecycle Model, in: International Journal of Digital Curation, 3/1 (2008), S. 134-140, DOI: 10.2218/ijdc.v3i1.48, URL: http://www.ijdc.net/index.php/ijdc/article/view/69.

Hochschulrektorenkonferenz (HRK): Gute wissenschaftliche Praxis an deutschen Hochschulen, Empfehlungen der 14. Mitgliederversammlung der HRK am 14. Mai 2013 in Nürnberg, Bonn 2013, URL: http://www.hrk.de/uploads/tx_szconvention/Empfehlung_GutewissenschaftlichePraxis_14052013_02.pdf.

Hügi, Jasmin, Schneider, René: Digitale Forschungsinfrastrukturen in den Geistes- und Geschichtswissenschaften, Genf 2013, URL: http://doc.rero.ch/record/31535/files/Schneider_Digitale_Forschungsinfrastrukturen.pdf.

Molloy, Laura: Oh, the Humanities! A Discussion About Research Data Management for the Arts and Humanities Disciplines, in: JISC MRD - Evidence Gathering 2011, URL: http://mrdevidence.jiscinvolve.org/wp/2011/12/16/oh-the-humanities-a-discussion-about-research-data-management-for-the-arts-and-humanities-disciplines/.

Pempe, Wolfgang: Geisteswissenschaften, in: Heike Neuroth u.a. (Hrsg.): Langzeitarchivierung von Forschungsdaten. Eine Bestandsaufnahme, Boizenburg 2012, S. 138-160, URL: http://nestor.sub.uni-goettingen.de/bestandsaufnahme/index.php.

Thaller, Manfred (Hrsg.): Das Digitale Archiv NRW in der Praxis. Eine Softwarelösung zur digitalen Langzeitarchivierung (Kölner Beiträge zu einer geisteswissenschaftlichen Fachinformatik, Bd. 5), Hamburg 2013, URL: http://www.danrw.de/.

Winkler-Nees, Stefan: Der Umgang mit Forschungsdaten in Wissenschaft und Lehre, Vortrag, 12.03.2013, Bad Honnef 2013, URL: http://www.dfg.de/download/pdf/dfg_magazin/wissenschaftliche_karriere/heisenberg_treffen_2010/forschungsdaten.pdf.