ZfdG

Zeitschrift für digitale Geisteswissenschaften

Artikel aus:
Zeitschrift für digitale Geisteswissenschaften

Titel:
Angebotsgenese für ein geisteswissenschaftliches Forschungsdatenzentrum

Autor/in:
Stefan Buddenbohm

Kontakt: buddenbohm@sub.uni-goettingen.de
Institution: Max-Planck-Institut zur Erforschung multireligiöser und multiethnischer Gesellschaften
GND: 1107805643 ORCID: 0000-0002-3469-6101

Autor/in:
Claudia Engelhardt

Kontakt: claudia.engelhardt@sub.uni-goettingen.de
Institution: Niedersächsische Staats- und Universitätsbibliothek Göttingen
GND: 1107806968 ORCID: 0000-0002-3391-7638

Autor/in:
Ulrike Wuttke

Kontakt: wuttke@akademienunion-berlin.de
Institution: Akademie der Wissenschaften zu Göttingen
GND: 1107808405 ORCID: 0000-0002-8217-4025

DOI des Artikels:
10.17175/2016_003

Nachweis im OPAC der Herzog August Bibliothek:
863827977

Erstveröffentlichung:

Letzte Überprüfung aller Verweise:
02.08.2016

GND-Verschlagwortung:
Langzeitarchivierung | Infrastruktur | Daten |

Zitierweise:
Stefan Buddenbohm, Claudia Engelhardt, Ulrike Wuttke: Angebotsgenese für ein geisteswissenschaftliches Forschungsdatenzentrum. In: Zeitschrift für digitale Geisteswissenschaften. 2016. PDF Format ohne Paginierung. Als text/html abrufbar unter DOI: 10.17175/2016_003.

Stefan Buddenbohm, Claudia Engelhardt, Ulrike Wuttke

Angebotsgenese für ein geisteswissenschaftliches Forschungsdatenzentrum

Abstracts

Dieser Beitrag widmet sich der Entwicklung einer Angebotsstruktur für ein geisteswissenschaftliches Forschungsdatenzentrum, d.h. der Frage, welche Angebote notwendig sind, um die Ergebnisse geisteswissenschaftlicher Forschung langfristig verfügbar zu halten und ihre Nachnutzung zu ermöglichen. Hierfür werden zunächst die grundlegenden Merkmale digitaler geisteswissenschaftlicher Forschungsdaten herausgearbeitet. Es wird aufgezeigt, dass neben dateibasierten Forschungsdaten – Texten, Bildern, Videos usw. – zunehmend komplexe Datenstrukturen in den Vordergrund treten und neue Herausforderungen an Einrichtungen der Forschungs(daten)infrastruktur stellen. Anschließend werden die Ergebnisse der in der Designphase des Humanities Data Centre (HDC) durchgeführten Testübernahmen für drei sogenannte Forschungsdatentypen – Datenbank, digitale Edition und interaktive Visualisierung – vorgestellt und die Schlussfolgerungen für das initiale, modular angelegte Angebot des HDC diskutiert, das in Abstimmung mit anderen Forschungsdatenzentren in Zukunft weiter ausgebaut werden kann. Der Beitrag spiegelt den Stand der Entwicklungen im HDC-Projekt vom Januar 2016. Informationen zu neuen Entwicklungen finden Sie auf der Webseite: http://humanities-data-centre.de.

This article describes the development of a service portfolio for a research data centre in the humanities; in particular, it explores which services are necessary to support the long-term preservation and re-use of the results of humanities research. First, the fundamental characteristics of digital research data in the humanities are examined. In addition to file-based research data – such as texts, images, videos, and so on – complex data structures become more prominent and pose new challenges for institutions working with research (data) infrastructures. Next, this paper describes the results of the test ingests using three types of research data – database, digital edition, and interactive visualisation – that were carried out during the design phase of the Humanities Data Centre (HDC), followed by a discussion of their implications for the HDC's initial modular service portfolio, which could be expanded in the future in coordination with other research data centres. This article reflects the developments in the HDC project as of January 2016. Information about new developments can be found on the website: http://humanities-data-centre.de/.

1. Einleitung

Wie muss das Angebot eines geisteswissenschaftlichen Forschungsdatenzentrums beschaffen sein, um Ergebnisse und Forschungsdaten geisteswissenschaftlicher Forschung langfristig verfügbar zu halten? Angesichts der Heterogenität geisteswissenschaftlicher Forschung, der Schwierigkeit, Forschungsdaten in den Geisteswissenschaften definitorisch zu erfassen sowie der unterschiedlichen Voraussetzungen, Anforderungen und Interessen der verschiedenen Akteure ist die Ausarbeitung eines Angebotsportfolios für ein geisteswissenschaftliches Forschungsdatenzentrum eine komplexe Aufgabe. Die größte Herausforderung besteht darin, einen Kompromiss zwischen den komplexen wissenschaftlichen Anforderungen an die nachhaltige Verfügbarkeit digitaler Daten und den begrenzten Ressourcen, sowohl seitens der zukünftig nutzenden Parteien als auch des

Infrastrukturanbieters, und fehlenden allgemeinen Standards für geisteswissenschaftliche Forschungsdatenzentren zu finden. Im vorliegenden Beitrag wird die Angebotsgenese des Projekts Humanities Data Centre (HDC) während der Designphase erörtert.

Ziel des vom Niedersächsischen Ministerium für Wissenschaft und Kultur (Niedersächsisches Vorab) geförderten HDC-Projekts ist es, ein Forschungsdatenzentrum für die Geisteswissenschaften aufzubauen. In der Designphase (Mai 2014–April 2016) werden zunächst die konzeptionellen Grundlagen für den Aufbau und Betrieb des Datenzentrums erarbeitet. Ausgangspunkt der Angebotsdefinition des HDC war neben der Aufarbeitung der Fachliteratur zu den Charakteristika geisteswissenschaftlicher Forschungsdaten, den Anforderungen der Geisteswissenschaftlerinnen und -wissenschaftler und konzeptionellen Studien ähnlich gelagerter Initiativen[1] der Blick auf die Forschungsvorhaben und Datenbestände im Konsortium.[2] Auf diese Weise konnten eine Reihe von Fällen identifiziert werden, die technologische und methodische Gemeinsamkeiten aufweisen und die zu sogenannten Forschungsdatentypen zusammengefasst werden können.

Im Folgenden werden zuerst die wesentlichen Charakteristika geisteswissenschaftlicher Forschungsdaten und die diesbezüglichen Anforderungen der Wissenschaftlerinnen und Wissenschaftler den Umsetzungsbedingungen eines Infrastrukturanbieters gegenübergestellt. Anschließend werden das Konzept der Forschungsdatentypen vor- und die Implikationen für die Angebotsgenese dargestellt. Folgende Fragen spielen dabei eine Rolle: Kann ein Forschungsdatenzentrum alle denkbaren Archivierungsfälle abdecken oder wie könnte eine sinnvolle Konzentration des Angebotes aussehen? Wie kann die möglichst einfache und vielfältige Nachnutzung geisteswissenschaftlicher Forschungsdaten ermöglicht werden? Inwiefern müssen sich unterschiedliche Anforderungen von Datengebern und -nutzern in der Angebotsstruktur widerspiegeln? Welche Schnittstellen sind mit Blick auf die Nachnutzung zu bedenken? Inwiefern ergeben sich Anforderungen zur Standardisierung bezüglich des Spektrums archivierungsfähiger Forschungsdaten? Insbesondere sollen anhand der Darstellung einiger im Rahmen der Designphase vorgenommener Testübernahmen die im Rückblick für die Angebotsentwicklung nützlichen Umwege verdeutlicht werden, die zeigen, dass die Angebotsentwicklung eines Datenzentrums kein linearer Prozess ist, sondern die Prämissen und Prototypen immer wieder auf die Tragfähigkeit in der Praxis überprüft werden müssen. Abschließend werden Schlussfolgerungen für ein umsetzbares Angebot abgeleitet und der derzeitige Stand (Januar 2016) des HDC-Angebotsportfolios vorgestellt.

2. Charakteristika geisteswissenschaftlicher Forschungsdaten

[1] Wir danken unseren Kolleginnen und Kollegen vom Konrad-Zuse-Zentrum für Informationstechnik Berlin (ZIB) und der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW) für Beiträge und Vorarbeiten zu diesem Artikel.

[2] Das HDC-Konsortium setzt sich aus folgenden Partnern zusammen: Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG), Akademie der Wissenschaften zu Göttingen (ADWG), Berlin-Brandenburgische Akademie der Wissenschaften (BBAW), Max-Planck-Institut zur Erforschung multireligiöser und multiethnischer Gesellschaften Göttingen (MPI MMG), Niedersächsische Staats- und Universitätsbibliothek Göttingen (SUB) sowie dem Konrad-Zuse-Zentrum für Informationstechnik Berlin (ZIB).

Geisteswissenschaften sind traditionell zu einem großen Teil Textwissenschaften, in dem Sinn, dass sowohl die Forschungsgrundlagen als auch die Publikationen von klassischen Textformaten geprägt sind. Mit dem Aufkommen neuer Medien und Methoden in der Forschung werden auch immer mehr Forschungsdaten und -ergebnisse in digitaler Form erzeugt und (nach)genutzt. Geisteswissenschaftliche Forschungsergebnisse bestehen zunehmend auch aus webbasierten Präsentationen verbundener Objektsammlungen[3], wie Datenbanken, digitalen Editionen oder interaktiven Datenvisualisierungen, bei denen oft ein großer Teil des Informationsgehalts beziehungsweise des wissenschaftlichen Mehrwerts im Design, das heißt in der Architektur der Softwareumgebung und der entsprechenden Aufbereitung der Daten, liegt.[4] In diesen Fällen, in denen Form und Inhalt nicht ohne Weiteres voneinander gelöst werden können, drohen bei einer Beschränkung der langfristigen Archivierung und Bereitstellung auf die Primärdaten, Informationsgehalt beziehungsweise Forschungsergebnisse und -leistungen verloren zu gehen, wie weiter unten anhand ausgewählter Beispiele ausgeführt wird.

Der Begriff Primärdaten ist an dieser Stelle bereits problematisch, da es auf dem Weg von der Forschungsfrage über die Datenerhebung und Auswertung hin zur Publikation zahlreiche Zwischenstufen geben kann, die für die Dokumentation und Nachnutzung relevant sein können, wie Sahle und Kronenwett zur schlechten Unterscheidbarkeit zwischen Primärdaten und Ergebnisdaten in den Geisteswissenschaften ausgeführt haben.[5] Dies rührt unter anderem daher, dass die Geisteswissenschaften eher eine interpretative als eine datengetriebene Wissenschaft sind, auch wenn es durchaus geisteswissenschaftliche Forschungszweige gibt, die mit quantitativen Daten arbeiten.[6] Diese definitorische Unsicherheit wirft nicht nur Probleme diesbezüglich auf, dass sich Geisteswissenschaftlerinnen und -wissenschaftler oftmals nicht darüber bewusst sind, dass sie überhaupt über Forschungsdaten beziehungsweise Primärdaten verfügen, die laut der DFG-Empfehlungen zur Sicherung guter wissenschaftlicher Praxis[7] bzw. der DFG-Leitlinien zum Umgang mit Forschungsdaten »in der eigenen Einrichtung oder in einer fachlich einschlägigen, überregionalen Infrastruktur für mindestens 10 Jahre archiviert werden«[8] sollen,[9] sondern auch bezüglich der Kriterien für die Auswahl und Bewertung von Forschungsdaten, einem bisher noch (zu) wenig diskutierten Problembereich. Mit anderen Worten, welche Daten können beziehungsweise müssen auf welche Art und Weise nach dem Abschluss geisteswissenschaftlicher Forschungsprojekte langzeitarchiviert werden? Diese Frage stellt sich insbesondere bei komplexen webbasierten Präsentationen verbundener Objekte, wie unten weiter ausgeführt wird.[10]

---

[3] Als Beispiel sei hier die komplexe Webapplikation *Republic of Letters* (http://republicofletters.stanford.edu/) genannt.
[4] Vgl. Pempe 2012, S. 141.
[5] Vgl. Sahle / Kronenwett 2013, S. 80f.
[6] Vgl. Borgman 2007, S. 213.
[7] Vgl. DFG 1998/2013, S. 21f.
[8] DFG 2015, S. 1.
[9] So berichten z.B. Kindling et al. 2013, S. 51, im Zusammenhang mit einer an der Humboldt Universität zu Berlin durchgeführten Umfrage zum Umgang mit digitalen Forschungsdaten: »Schon die Verwendung der verschiedenen Bezeichnungen ›Forschungsdaten‹ (in der Umfrage) und ›Primärdaten‹ (in den Grundsätzen) sorgte für einige Verwirrung. Durch die mit der Bezeichnung ›Primärdaten‹ assoziierte empirische (quantitative) Arbeitsweise beurteilten manche Teilnehmer die Grundsätze als nicht anwendbar für geisteswissenschaftliche Fachbereiche.« Ähnliches stellt Schöch 2013 fest: »Most of my colleagues in literary and cultural studies would not necessarily speak of their objects of study as ›data‹.«
[10] Hügi / Schneider 2013, S. i sprechen deshalb von »Forschungsprodukten« die »sowohl Datensätze aus Datenbanken, wie auch Primär- und Sekundärquellen, Digitalisate oder Hilfsmittel umschließen.«

Zunächst ist festzuhalten, dass alle kulturellen Artefakte und Phänomene, von Papyri, über Handschriften, Büchern, Gemälden, Filmen, Musik, Gebäuden bis zu ganzen Bibliotheken oder Gesellschaften zum Gegenstand geisteswissenschaftlicher Forschung und – übersetzt in eine digitale Form – zu relevanten Forschungsdaten werden können.[11] Dabei kann es sich um sehr unterschiedliche Arten digitaler Daten handeln, die sowohl unterschiedliche Strategien der Langzeitarchivierung (LZA) erfordern als auch unterschiedliche Nachnutzungsszenarien ermöglichen. Der Konzeption des HDC liegt die folgende, im Rahmen von DARIAH-DE entwickelte Definition digitaler geistes- und kulturwissenschaftlicher Forschungsdaten zu Grunde:

> »Unter digitalen geistes- und kulturwissenschaftlichen Forschungsdaten werden innerhalb von DARIAH-DE all jene Quellen/Materialien und Ergebnisse verstanden, die im Kontext einer geistes- und kulturwissenschaftlichen Forschungsfrage gesammelt, erzeugt, beschrieben und/oder ausgewertet werden und in maschinenlesbarer Form zum Zwecke der Archivierung, Zitierbarkeit und zur weiteren Verarbeitung aufbewahrt werden können.«[12]

Digitale geisteswissenschaftliche Daten sind sehr heterogen. Die Spannbreite reicht von Einzelobjekten bis zu höchst komplex miteinander verknüpften Objekten, von Digitalisaten und Textdaten (mit oder ohne Tiefenerschließung mittels inhaltlichem Markup) beliebiger Quellen (Blumenbach-online, Wolfenbütteler Digitale Bibliothek), Fotografien von Kunstobjekten oder Architektur mit erschließenden Angaben (Deutsche Inschriften Online), Multimediadaten (GlobalDiverCities), 3D-Scans (Blumenbach-online) bis zu Personen- und Ortsdatenbanken (Berliner Klassik) – um nur einige zu nennen –, die in einer Vielfalt von Formaten auftreten können, darunter TIFF, JPEG, PNG, PDF, DOC, XML, RTF, MP3, MP4 oder MySQL.[13] All diese digitalen Objekte können wiederum mit verschiedenen Metadaten versehen werden, wobei auch hier in den Geisteswissenschaften eine hohe Diversität herrscht.[14]

Wenn wir die traditionelle geisteswissenschaftliche Publikationskultur und die damit vorliegenden Ergebnisdaten betrachten, ist festzuhalten, dass diese von klassischen Textformaten wie Monographien und Einzelschriften dominiert wird, zumindest wenn die im Forschungsprozess entstehenden intermediären oder Arbeitsdaten (der sog. *throughput*) vernachlässigt werden.[15] Zwar sind noch nicht alle Fragen der Langzeitarchivierung

[11] Vgl. Sahle / Kronenwett 2013, S. 78; Schöch 2013; Borgmann 2015, S. 167.
[12] https://de.dariah.eu/forschungsdaten.
[13] Vgl. u.a. Pempe 2012, S. 142f.; Borgman 2015, S. 167. Eine von IANUS in den gesamten Altertumswissenschaften durchgeführte Stakeholderanalyse ergab bspw. 462 in der Community genutzte Dateiformate (vgl. Heinrich / Schäfer 2015, Folie 17). Auch wenn die Formate in den ›klassischen‹ textbasierten Geisteswissenschaften sicherlich in einigen Punkten abweichen, ist prinzipiell für ein geisteswissenschaftliches Forschungsdatenzentrum, das alle Fälle abdecken möchte, eine ähnlich diverse Ausgangsituation zu erwarten.
[14] Auf Grund der Heterogenität geisteswissenschaftlicher Daten ist auch kein einheitliches Metadatenschema in Sicht. Inzwischen hat sich Dublin Core als ein weitverbreiteter Standard etabliert. Dublin Core (DC) ist jedoch sehr variabel und für individuelle Konstellationen anpassbar und konzentriert sich zudem vorwiegend auf deskriptive Metadaten; er muss also um Aspekte der Langzeitarchivierung (LZA), wie zum Beispiel im Fall des TextGrid-Metadaten-Schemas, erweitert werden (vgl. Pempe 2012, S. 151f). DC ist auch als Mindestanforderung (zum Beispiel für die Beschreibung von Primärdaten, DFG 2009, S. 3) im geisteswissenschaftlichen Kontext in Bezug auf Nachnutzung nur selten ausreichend. Auch andere, speziell geisteswissenschaftliche Metadatenstandards wie TEI (für die Repräsentation von Texten in digitaler/maschinenlesbarer Form, momentan Version P5) oder MEI (für die Repräsentation von Musik in digitaler/maschinenlesbarer Form, momentan Version 2.1.1) sind sehr komplex und individuell modifizierbar.
[15] Vgl. Sahle / Kronenwett 2013, S. 78–79.

bis ins Detail geklärt, dennoch stellt sich zumindest von technischer Seite in diesem Fall sowohl auf der Seite der Primärdaten als auch auf der Seite der Ergebnis- bzw. Publikationsdaten die Situation weniger komplex dar[16], als im Fall der sich durch den zunehmenden Einsatz digitaler Medien und Methoden und der Multimedialisierung der Publikationen grundlegend veränderten Forschungs- und Publikationskultur in Richtung komplexer webbasierter Präsentationen verbundener Objekte.[17] Die Einbindung von Primärdaten in den Forschungsprozess, z.B. mittels vertiefender Erschließung und Annotationen, und das Überschreiten der Grenzen klassischer Publikationen durch komplexe webbasierte Informationsportale stellt die zur Verfügung stehenden Speicher- und Publikationsinfrastrukturen in Frage[18] und erfordert eine Erweiterung der Angebote, um den aus der Wissenschaft erwachsenden Herausforderungen – wie der Heterogenität der Forschungsdaten, der langfristigen Bewahrung ihrer intellektuellen Nachnutzbarkeit sowie Zitations-, Referenzierungs- und Urheberrechtsfragen – sinnvoll zu entsprechen.

Durch die Einbindung von Primärdaten in den Forschungsprozess ergeben sich weitere Problematiken, mit denen sich ein Forschungsdatenzentrum auseinanderzusetzen hat. Inzwischen gibt es verschiedene Projekte im Bereich der Digital Humanities, die auf der Sekundärnutzung von Daten aufbauen, die von Datenzentren oder Bibliotheken (und ähnlichen Gedächtnisinstitutionen) bereitgestellt und vorgehalten werden. Zwar könnte man durchaus behaupten, dass »bei Sekundärnutzung keine zusätzliche Archivierung mehr nötig [ist], weil diese Datenbestände von den liefernden Datenzentren bereits langfristig archiviert werden« und »die für die Forschungsfrage verwendeten Datensätze [...] lediglich in der Publikation dokumentiert [werden]«[19], dies funktioniert jedoch bei webbasierten Anwendungen, die auf Sekundärnutzungen aufbauen, nur begrenzt. Ein gutes Beispiel sind hier die im Rahmen des GlobalDivercities-Projekts entstandenen interaktiven Datenvisualisierungen zu Migrationsbewegungen zwischen Staaten. Ein konkretes Anschauungsbeispiel ist hier mit Global Migration Flows gegeben. In diesem Fall wurden die der Datenvisualisierung zugrunde liegenden Daten nicht im Projekt selbst erhoben, sondern basieren auf frei zugänglichen Daten der United Nations Population Division, d.h. die wissenschaftliche Leistung besteht nicht in der Datenerhebung, sondern in ihrer Auswahl und der Entwicklung und Programmierung der Datenvisualisierung. Müsste man in diesem Fall zusätzlich die der Visualisierung zugrunde liegenden Daten sichern und zugänglich machen? Da die Datenzentren auf die Lösung der wissenschaftsrechtlichen Details nur bedingt Einfluss haben, sei hier nur die Anmerkung gestattet, dass es an der Zeit ist, die rechtlichen Aspekte dahingehend zu klären, dass der Zugang und die nachhaltige Bereitstellung der einem Forschungsprojekt bzw. dessen Ergebnissen zu Grunde liegenden Daten möglichst einfach realisierbar ist.[20]

[16] Einfacher in dem Sinn, dass es sich hier um digitale Einzelobjekte handelt, die durch die objektbasierte Ablage im Rahmen bereits etablierter Infrastrukturen wie Museen, Archiven und Bibliotheken bedient werden können, man denke unter anderem an auf PDF-Formate spezialisierte Publikationsrepositorien.
[17] Vgl. Sahle / Kronenwett 2013, S. 79.
[18] Vgl. Sahle / Kronenwett 2013, S. 80.
[19] Beide Zitate aus Kindling et al. 2014, S. 9f.
[20] Beispielhaft sei hier das Eintreten von LIBER für eine Anpassung der bestehenden Gesetzgebungen zur Erleichterung von Text- und Datamining genannt. Vgl. u.a. http://libereurope.eu/text-data-mining/.

Es zeichnet sich ab, dass Datenzentren für die Gewährleistung der langfristigen Verfügbarkeit komplexer webbasierter Informationsportale, aber auch anderer digitaler Ergebnisse geisteswissenschaftlicher Forschung in Zukunft eine tragende Rolle zukommen wird. Sie ergänzen diesbezüglich schon bestehende Angebote etablierter Infrastruktureinrichtungen wie Rechenzentren und Bibliotheken in Bezug auf das Grundszenario der Aufbewahrung der Forschungsdaten für mindestens zehn Jahre und kooperieren mit diesen zur Gewährleistung dieses Ziels.[21] Denn die Anforderung ›langfristig verfügbar‹ (ganz im Sinne der verstärkten Forderung eines möglichst offenen Zugangs zu Forschungsdaten) übersteigt die Anforderung ›langfristig archiviert‹ und macht neue Ansätze, auch organisatorischer Art, notwendig, um zu für die Infrastrukturanbieter und die Wissenschaftlerinnen und Wissenschaftler gleichermaßen befriedigenden Lösungen zu gelangen, die auch bisher unvorhergesehene bzw. bisher wenig verbreitete Nachnutzungsszenarien erlauben.

3. Anforderungen an ein Forschungsdatenzentrum

3.1 Aus Sicht der Wissenschaft

In der Anforderungsanalyse (zum methodischen Vorgehen vgl. Abschnitt 4.1) zu Beginn der Designphase wurden aus Sicht des HDC Nachhaltigkeit, Präsentation und Integration als maßgebliche Dimensionen identifiziert, aus denen sich weitere Anforderungen ableiten lassen und die den Forschungsdatenlebenszyklus[22] sowie die verschiedenen Aggregationsstufen von Daten[23] vollständig abdecken. Diese Dimensionen adressieren sowohl Datengeber bzw. Datenproduzenten, als auch Datennutzer. Die Anforderungen der Datennutzer übersteigen diejenigen der Datenproduzenten in dem Sinne, dass die öffentliche Bereitstellung von Forschungsdaten für eine möglichst unbeschränkte, disziplinübergreifende Nachnutzung auf Seiten der Datenproduzenten einen erheblichen Mehraufwand erforderlich macht. Dies bezieht sich vor allem auf die Einhaltung von Standards bei der Erzeugung und der Aufbereitung der Daten sowie die ausführliche Dokumentation aller nötigen Kontextinformationen, um die Daten verstehen und einordnen zu können. Datenzentren sollten sich angesichts dessen nicht auf die Rolle eines Anbieters technischer Infrastruktur zurückziehen, sondern Forschende im Sinne eines Kompetenzzentrums beim Datenmanagement beraten und unterstützen – von der Auskunft zu konkreten, einzelnen

[21] Depping 2014, S. 84f. unterscheidet zwei Grundszenarien: 1) Minimalziel: Aufbewahrung der Forschungsdaten für zehn Jahre, 2) komplexe Aufgabe: Überprüfung und Nachnutzung durch andere Wissenschaftler und Wissenschaftlerinnen. Ersteres verortet er als Aufgabe bei Hochschulbibliotheken und Rechenzentren, Letzteres bei überregionalen fachlichen Datenrepositorien, wobei den Bibliotheken eine Beratungs- und Maklerfunktion zukommen könnte. Diese Beratungs- und Maklerfunktion wird in der Angebotskonzeption des HDC im Rahmen der HDC-Datenkuratorinnen und -kuratoren, die an wissenschaftlichen Bibliotheken und Einrichtungen angesiedelt sein können, aufgegriffen.
[22] Nach dem auf dem vereinfachten Datenlebenszyklus des UK Data Archive beruhenden Datenlebenszyklus von IANUS handelt es sich um die Phasen Erstellung, Verarbeitung, Analyse, Archivierung, Zugang und Nachnutzung. Vgl. http://www.ianus-fdz.de/it-empfehlungen/lebenszyklus.
[23] Bspw. Rohdaten, prozessierte und aufbereitete Daten, Tabellen, Abbildungen und Publikationen.

Fragen über die Bereitstellung von Informations- und Schulungsmaterialien bis hin zur Beteiligung an Projekten im Sinne des Embedded Data Managements.[24]

Insgesamt bleibt festzuhalten, dass die Geisteswissenschaftlerinnen und -wissenschaftler von Anfang an möglichst eng in die Erarbeitung sowohl der technischen als auch der forschungsbegleitenden Beratungs- und Schulungsangebote des Forschungsdatenzentrums einzubeziehen sind, da allein die Bereitstellung eines Angebots nicht ausreicht, um die Akzeptanz durch die potenziellen Nutzerinnen und Nutzer zu erzeugen.[25]

3.1.1 Nachhaltigkeit

Die Wissenschaftlerinnen und Wissenschaftler erwarten, dass die Langzeitarchivierung und die langfristige Bereitstellung der Forschungsdaten, die sie dem Datenzentrum übergeben, sichergestellt sind. Zum einen spielen diesbezüglich Anforderungen von Forschungsförderern und wissenschaftlichen Institutionen eine Rolle,[26] zum anderen ist die zum Beispiel von der DFG geforderte Mindestaufbewahrungsdauer von zehn Jahren angesichts der oft sehr langen Relevanz geisteswissenschaftlicher Forschungsergebnisse eine nur sehr kurze Zeitspanne.[27] So ergab eine Umfrage an den europäischen Akademien bezüglich des letzten Punktes, dass sich Forschende stark um die Nachhaltigkeit digitaler Daten bzw. Archive sorgen.[28] Die Langzeitarchivierung und nachhaltige Bereitstellung geisteswissenschaftlicher Forschungsdaten ist nicht nur ein technologisches Problem, sondern angesichts sich momentan gerade erst im Entstehen befindlicher kooperativer Strukturen und noch weitgehend fehlender bewährter Geschäfts- und Finanzierungsmodelle[29] sowie fehlender oder zumindest unzulänglicher Anreiz- und Unterstützungssysteme für Forschende, ihre Daten für die Nachnutzung angemessen aufzubereiten, auch ein organisatorisches Problem. Dennoch sind die technologischen Herausforderungen nicht zu unterschätzen. Hierzu gehören nicht nur Bitstream Preservation, sondern auch die Erhaltung der Nutzungs- und Präsentationsumgebungen. In diesen Bereich spielt auch die Berücksichtigung der

[24] Orientiert am Konzept der Embedded Librarianship sieht das Embedded Data Management die Unterstützung von Wissenschaftlerinnen und Wissenschaftlern durch Datenmanagementexpertinnen und -experten vor, die als integraler Teil eines Projektteams, einer Arbeitsgruppe o.ä. (›eingebettet‹ in die Zielgruppe) agieren und somit gegenüber der Wissenschaft weniger in der Rolle als Dienstleister, denn vielmehr als Partner auf Augenhöhe auftreten (vgl. Cremer et al. 2015, S. 15ff).

[25] Vgl. z.B. Nelson 2009, S. 160. Auch Hügi / Scheider 2013, S. iii betonen, dass für die Wissenschaftlerinnen und Wissenschaftler persönliche Anreize zur Benutzung der Angebote des Datenzentrums geschaffen werden müssen und die bereitgestellte Infrastruktur sehr gut den Bedürfnissen ihres Publikums entsprechen muss.

[26] Grundlegend ist hier DFG 1998/2013, aber auch institutionelle Richtlinien wie Göttingen 2014. Die Anforderungen an das Datenmanagement sind erst kürzlich in DFG 2015 präzisiert worden: Die Daten müssen »in der eigenen Einrichtung oder in einer fachlich einschlägigen, überregionalen Infrastruktur für mindestens 10 Jahre archiviert werden«.

[27] Vgl. Borgman 2007, S. 214: »Literature in the humanities goes out of print long before it goes out of date«, und Borgmann 2007, S. 217: »Both the literature and the data of humanities are long lived. [...] Determining which sources are worth digitizing and preserving may be the most difficult to accomplish in the humanities.«

[28] Vgl. Leathem / Adrian 2015, S. 112: »Für die oftmals auf Jahrzehnte angelegten Langzeitprojekte der befragten Organisationen sind die Zeiträume, in denen sich DFW [Digitale Forschungswerkzeuge, die Autoren] verändern, zu kurz. Während sich historische Quellen über Jahrhunderte in Papierform konservieren ließen, kann zu diesem Zeitpunkt noch niemand mit Sicherheit sagen, wie lang digital gespeicherte Informationen zugänglich bleiben werden.«

[29] Das prominenteste Beispiel dafür, welch fatale Konsequenzen dies haben kann, ist sicherlich die Abwicklung des AHDS (Arts and Humanities Data Service, UK) im Jahr 2008 aufgrund fehlender finanzieller Mittel. Vgl. http://www.ahds.ac.uk/.

Datennutzung in einer zum Übergabezeitpunkt nicht intendierten Weise eine wichtige Rolle. Die Attraktivität des Bestands eines Forschungsdatenzentrums hängt nicht zuletzt entscheidend davon ab, welche Nachnutzungsszenarien die Forschungsdaten ermöglichen. Neben dem Kurationsaufwand sind hiermit selbstverständlich auch rechtliche und finanzielle Fragestellungen verbunden.

3.1.2 Präsentation

Wie bereits angerissen, bedeutet die Sicherung im Sinne einer Langzeitarchivierung nicht unbedingt, dass die Daten auch unmittelbar zur Nachnutzung bereitstehen. Die Dimension Präsentation bezieht sich insbesondere auf den die objektbasierte Ablage von Daten übersteigenden Aspekt des Erhalts von komplexen Datenobjekten und Präsentationssystemen. Ein weiterer Aspekt in dieser Hinsicht ist auch die generelle Sichtbarkeit der Daten, die a) auffindbar [30] sein sollten – wobei wiederum die detaillierte inhaltliche Erschließung mit Metadaten eine Rolle spielt – und b) eindeutig referenzierbar sein sollten, was z.B. mittels persistenter Identifikatoren (PIDs) gewährleistet werden kann. Auch wenn bislang die Zitation von Forschungsdaten in den Geisteswissenschaften im Vergleich zu den Naturwissenschaften weniger etabliert ist,[31] spielt die Erhöhung der Sichtbarkeit der Forschungsleistungen und digitalen Ergebnisse, auch einer breiteren Öffentlichkeit gegenüber, für Forschende wie Trägerinstitutionen auch in den Geisteswissenschaften eine nicht zu vernachlässigende Rolle.

Als einschränkender Aspekt bei der Präsentation von Forschungsdaten muss an dieser Stelle die Zugänglichkeit erwähnt werden. Auch wenn – entsprechend dem Open Access-Prinzip – alle Forschungsdaten grundsätzlich frei und ohne Beschränkungen zugänglich sein sollten,[32] wird dies nicht in allen Fällen möglich sein, insbesondere aufgrund urheber- oder verwertungsrechtlicher Einschränkungen oder aus Datenschutzgründen. Diese Gründe müssen jedoch einer Archivierung der Daten in einem Forschungsdatenzentrum nicht grundsätzlich entgegenstehen, da über Access Policies verschiedene Zugangsstufen definiert werden können und mittels entsprechender technischer Lösungen (bspw. einer AAI[33]) der Zugang kontrolliert und gesteuert werden kann. Diese Form der Zugangskontrolle ist auf eindeutige Informationen in den Metadaten der fraglichen Ressourcen angewiesen, die im System eine Freigabe des Zugangs in automatisierter Form erlauben.

3.1.3 Integration

[30] Bspw. Suche im Katalog des Datenzentrums, Suche im gemeinsamen Katalog aller Datenbanken bzw. Integration in bibliothekarische Nachweissysteme. So sind etwa alle digitalen Editionen der WDB (Wolfenbütteler Digitalen Bibliothek) über den Opac nachgewiesen.
[31] Vgl. Hügi / Schneider 2013, S. 36.
[32] Vgl. Informationsplattform Open-Access.net.
[33] Eine AAI (Authentication and Authorization Infrastructure) erlaubt es, die Identität einer Person zu verifizieren (Authentication / Authentifizierung), die Zugriffsberechtigung auf die angefragte Ressource zu prüfen und den Zugriff entsprechend zu erlauben bzw. zu verweigern (Authorization / Autorisierung); vgl. Advancing Technologies and Federating Communities 2012, S. 31.

Unter die Dimension Integration fallen der Metadaten- und Datenaustausch mit anderen Forschungsdatenzentren, die Anbindung von virtuellen Forschungsumgebungen oder die Möglichkeit, verschiedene Datensammlungen zusammenzuführen, um mit ihnen neue Forschungsfragen zu beantworten. Insbesondere für die Nachnutzung von Forschungsdaten durch andere Akteure als die ursprünglichen Datenerzeugerinnen und -erzeuger sind möglichst standardisierte Metadaten und umfangreiche Kontextinformationen sowie geeignete Schnittstellen unabdingbar.[34]

Hier stellt sich auch die Frage nach der Verteilung der Verantwortlichkeiten für die verschiedenen mit der Datenkuration und -bereitstellung verbundenen Aufgaben. Anders als bei den Dimensionen Nachhaltigkeit und Präsentation verschiebt sich bei der Dimension Integration der Fokus vom ursprünglich abliefernden Forschungsprojekt hin zu zukünftigen Vorhaben, die die Bestände eines Forschungsdatenzentrums für eigene Forschungsfragen nutzen wollen. Das abliefernde Forschungsprojekt wird sich auf die für es wichtigen Aspekte bei der Archivierung konzentrieren (bspw. Dokumentationspflicht gegenüber dem Förderer, Referenzierbarkeit von Forschungsdaten), während die Ermöglichung von Nachnutzungsszenarien für Dritte in den meisten Fällen eine untergeordnete Rolle spielen wird, zumal dafür innerhalb des Forschungsprojekts meist auch keine Ressourcen vorgesehen sind.

Des Weiteren müssen bereits vor oder spätestens bei Übergabe der Forschungsdaten wichtige rechtliche Fragen geklärt werden, u.a. wem die Daten gehören und wer sie in welcher Form nutzen darf.[35] Diesbezüglich bedarf es keiner besonderen technischen Lösungen (abgesehen von einer AAI), allerdings ist eine eindeutige (maschinenlesbare) Kennzeichnung, z.B. mittels der Metadaten, notwendig. Hierfür steht zum Beispiel im geisteswissenschaftlich relevanten Metadatenstandard Dublin Core das Feld dc:rights[36] zur Verfügung.

3.2 Weitere Anforderungen an ein Forschungsdatenzentrum

Neben den oben genannten wissenschaftlichen Anforderungen sind einige Teilaspekte der Anforderungen zu benennen, die im Interesse von Gedächtnisinstitutionen stehen, wie bspw. die Erstellung von Digitalisaten seltener Objekte zur Schonung der physischen Originale, die dann natürlich auch dementsprechend nachgewiesen und präsentiert werden müssen.[37] Zudem kann es im Interesse von Gedächtnisinstitutionen sein, Digitalisate oder genuin digitale Objekte nachhaltig aufzubewahren, die sie z.B. aus urheberrechtlichen Gründen (noch) nicht einer breiteren wissenschaftlichen Öffentlichkeit zur Verfügung stellen können. Dies kann zum

[34] Die Ermöglichung auch ursprünglich nicht vorgesehener Nutzungsszenarien kristallisierte sich u.a. als eine wichtige Anforderung auf dem HDC-Workshop am 16.09.2015 an der Universität Hamburg (im Rahmen der FORGE 2015) heraus.
[35] Eine Einführung in juristische Fragen speziell im Kontext der digitalen Geisteswissenschaften und die Problematik von Lizenzen bieten Klimpel / Weitzmann 2015. Darüber hinaus sei hier auf Beer et al. 2014 verwiesen.
[36] Vgl. http://dublincore.org/documents/usageguide/elements.shtml#rights.
[37] Ein Digitalisat ist aus wissenschaftlicher Sicht natürlich kein Surrogat für das Original (vgl. Schöch 2013, Borgmann 2015, S. 216f), auf der anderen Seite ermöglichen gut tiefenerschlossene Digitalisierungsprojekte oftmals ungeahnte Forschungsmöglichkeiten, nicht nur weil die Wissenschaftlerinnen und Wissenschaftler nicht mehr persönlich in jede Bibliothek oder Archiv reisen müssen.

einen der digitalen Bestandserhaltung[38] dienen oder sogar als ›Reservekopie‹ für den Fall des Verlusts des physischen Objekts. Exemplarisch seien hier der Einsturz des Historischen Archivs der Stadt Köln[39] und der Brand in der Herzogin-Anna-Amalia-Bibliothek in Weimar[40] genannt. In beiden Fällen sind viele Bücher und Dokumente unwiederbringlich verloren gegangen, die noch zur Verfügung stehen würden, wenn sie vorher digitalisiert geworden wären.

3.3 Umsetzungsbedingungen für ein Forschungsdatenzentrum aus Sicht eines Infrastrukturanbieters

Im Gegensatz zu einem Publikationsrepositorium[41], das sich sowohl bei den Formaten als auch bei den Metadaten nur auf eine vergleichsweise geringe Bandbreite an Formaten und Technologien einzustellen braucht, steht ein Forschungsdatenzentrum für geisteswissenschaftliche Daten vor einer größeren Herausforderung. Hier geht es – wie oben bereits ausführlich beschrieben (vgl. hierzu Abschnitt 2) – um die Archivierung und Erhaltung einer Vielfalt von Formaten sowie komplexer Datentypen. Die Formatvielfalt ist nicht zuletzt eine Folge der Diversität der in der geisteswissenschaftlichen Forschung eingesetzten digitalen Werkzeuge, unter denen sich momentan noch viele individuell an die konkrete Forschungsfrage angepasste Eigenentwicklungen befinden.

Ein wesentliches Kennzeichen komplexer Datentypen ist, dass oft nur ein kleiner Kern standardisierbar ist, nämlich deskriptive, technische und administrative Metadaten. Hieraus folgt für die Gewährleistung der Langzeitarchivierung und Nachnutzbarkeit dieser Daten, dass in der Regel mehr oder weniger maßgeschneiderte Lösungen erforderlich sind. So muss zunächst die Struktur der komplexen Datenobjekte ermittelt werden, d.h. es muss untersucht werden, aus welchen einzelnen Daten und Anwendungen sie sich zusammensetzen. Dann müssen passende Archivierungsstrategien entwickelt, getestet und ggf. revidiert werden. Die Entwicklung und Durchführung derartiger individueller Lösungen erfordert einen hohen Ressourceneinsatz, dessen Umfang von Fall zu Fall erheblich variieren kann, was wiederum die Bestimmung der Kosten für die Langzeitarchivierung und -bereitstellung schwierig macht. Die geringe Standardisierung der Daten stellt auch für die Nachnutzung eine Herausforderung dar.

Die Rechercheoberfläche eines Forschungsdatenzentrums sollte zum einen möglichst intuitiv bedienbar sein, zum anderen aber ausreichend relevante Treffer und Möglichkeiten zur Ergebnisverfeinerung bieten. Hier wirft jedoch bereits der beschreibende Kern von Metadaten für die Entwicklung einer Suche über einen Bestand von Forschungsdaten die Frage auf, wie granular die beschreibenden Metadaten sein müssen bzw. sein können,

---

[38] Vgl. zum Beispiel im Fall des Deutschen Literaturarchivs Marbach: Digitale Bestandserhaltung.
[39] Vgl. Einsturz des Historischen Archivs Köln.
[40] Vgl. Brand der Anna Amalia Bibliothek 2004.
[41] Disziplinunabhängige Standards und Empfehlungen gibt zum Beispiel das DINI-Zertifikat für Open-Access-Repositorien und Publikationsdienste (DINI AG Elektronisches Publizieren et al. 2013) vor. Vgl. zu den Metadaten- und Datenformaten insbesondere die Angaben in Kapitel 2.6 »Erschließung und Schnittstellen«(DINI Arbeitsgruppe Elektronisches Publizieren 2013, S. 25ff.) sowie Kapitel 2.8 »Langzeitverfügbarkeit« (DINI Arbeitsgruppe Elektronisches Publizieren 2013, S. 31).

denn Metadaten werden keine Suche in Volltexten ersetzen können. Eine weitere Frage in diesem Zusammenhang ist, wie die Interoperabilität der Metadatenschemata verschiedener Forschungsdatenzentren und Gedächtnisinstitutionen gewährleistet werden kann.[42]

Während Kosten aus Sicht der Wissenschaft idealerweise nur eine untergeordnete Rolle spielen, sind Infrastrukturanbieter dazu angehalten, ihre Angebote so kosteneffizient wie möglich bereitzustellen. Die Kosteneffizienz von Angeboten zur Erhaltung und Bereitstellung von Forschungsdaten wird für einen Infrastrukturanbieter vor allem durch die Faktoren Standardisierbarkeit und Skalierbarkeit beeinflusst:

- **Standardisierbarkeit der Daten und Dienste**: Diese ist vor allem mit Blick auf die Interoperabilität mit anderen Diensten und Anbietern sowie im Interesse effizienter Kurationsabläufe innerhalb des Forschungsdatenzentrums wünschenswert. Je standardisierter Daten übernommen werden können, desto weniger Aufwand entsteht für die Kuration und desto einfacher ist die Umsetzung von Mehrwertdiensten, bspw. ein übergreifendes Suchportal basierend auf dem Austausch von Metadaten per OAI-PMH.
- **Skalierbarkeit der Dienste**: Je standardisierter die Daten sind, desto besser skalieren die Dienste für ihre Archivierung und Bereitstellung. Die mit zunehmender Heterogenität der Daten und Abläufe abnehmende Skalierbarkeit der Dienste liegt nicht unbedingt im Interesse eines Infrastrukturanbieters.

Wie oben bereits angedeutet, führt jedoch die Heterogenität geisteswissenschaftlicher Forschungsdaten dazu, dass viele Dienste zu ihrer Archivierung und Bereitstellung nicht skalieren. Skalierbare Dienste sind noch am ehesten für einfache Objektklassen vorstellbar, da hier der Aufwand für den Ingest, d.h. die Überführung ins Datenzentrum, und die Bereitstellung nach einer gewissen Zeit auf Grund von Erfahrungswerten gut abgeschätzt und kalkuliert werden kann. Da es sich bei der Übernahme komplexer Datenobjekte jedoch voraussichtlich meist um Einzelfälle handeln wird, ergeben sich als Konsequenz schwer kalkulierbare Personalressourcen bzw. Kosten.

3.4 Zwischenresümee

Das Angebotsportfolio eines Forschungsdatenzentrums muss einen Kompromiss zwischen den Anforderungen der Wissenschaft und den Anforderungen bzw. Möglichkeiten des Infrastrukturanbieters finden. Eine naheliegende Option bildet ein Angebotsportfolio, das neben generischen, skalierenden Diensten für einfache Objektmodelle auch spezialisierte Angebote für die Erfassung, Archivierung, Präsentation und wissenschaftliche Nachnutzung komplexer Datenstrukturen bereitstellt und auf eine kooperative Ergänzung mit den Angeboten anderer Forschungsdatenzentren bzw. Arbeitsteilung ausgerichtet ist. Da sich angesichts der ohnehin unumgänglichen engen Zusammenarbeit mit den Wissenschaftlerinnen und Wissenschaftlern die Angebotsspezialisierung der verschiedenen

[42] Beispielhaft für die Verknüpfung bzw. Gewährleistung der Interoperabilität verschiedener Datenbestände im Bereich Digital Humanities sei hier die parallele Entwicklung des DARIAH-DE und des TextGrid Repositories genannt (vgl. Blümm et al. 2015, besonders S. 309ff).

Forschungsdatenzentren auf bestimmte Bereiche von Forschungsdaten anbietet, könnte so eine vermeintliche Schwierigkeit zu einer Stärke werden.

4. Forschungsdatentypen als Instrument bei der Angebotsgenese

4.1 Vorgehensweise

Als Hilfsmittel bei der Angebotsgenese[43] und um einen sinnvollen Kompromiss zwischen den Anforderungen der Wissenschaft und der Infrastrukturanbieter zu finden, wurden in der HDC-Designphase sogenannte Forschungsdatentypen definiert, für die auf der Grundlage der im Konsortium vorhandenen Forschungsvorhaben und Erfahrungen entlang der weiter oben beschriebenen wissenschaftlichen Anforderungsdimensionen konkrete Angebote konzipiert wurden.

Hierfür wurden zunächst die heterogenen Datenbestände der an der HDC-Designphase beteiligten Konsortialpartner hinsichtlich möglicher Gemeinsamkeiten untersucht und zu Gruppen von Forschungsdatentypen zusammengeführt. Ziel war es, ausgehend von der Vielzahl der möglichen Datenmodelle einen gangbaren Weg zur Einengung des Feldes zu finden, der gleichzeitig zu der oben angesprochenen Arbeitsteilung zwischen verschiedenen Forschungsdatenzentren führen könnte. Unter einem Forschungsdatentypen wird im HDC-Projektkontext die idealtypische Repräsentation einer Gruppe von Forschungsdaten verstanden, die – bspw. hinsichtlich der Erhebungs-, Analyse- und Darstellungsweisen – technologische, methodische und informationswissenschaftliche Gemeinsamkeiten aufweisen.[44] Innerhalb des Konsortiums wurden folgende grundlegenden Forschungsdatentypen identifiziert:

- Datenbanken,
- digitale Editionen,
- Bildformate,
- Videoaufzeichnungen (von Interviews),
- Anwendungen zur interaktiven Visualisierung von Daten.

Sie wurden im Hinblick auf die folgenden Dimensionen untersucht, um daraus konkrete Anforderungen an die zu entwickelnde Infrastruktur abzuleiten:

- technische Ebene (bspw. Format, Anwendung),
- informationswissenschaftliche Ebene (Metadaten, Interoperabilität),

[43] Vgl. hierzu auch den ausführlichen Projektbericht Aschenbrenner et al. 2015.
[44] Dessen ungeachtet kann es innerhalb eines Forschungsdatentyps eine teilweise sehr stark ausgeprägte Fallspezifik geben, die zu einem späteren Zeitpunkt anhand von Praxiserfahrungen ggf. zur Bildung von Unterkategorien führen könnte.

- fachwissenschaftliche Ebene (significant properties, wissenschaftlicher Gehalt).[45]

Im Folgenden werden drei Forschungsdatentypen anhand von konkreten Beispielen aus dem Konsortium genauer beschrieben.

4.2 Forschungsdatentyp: Datenbank

Der Begriff ›Datenbank‹[46] verweist im Kontext des HDC-Projekts nicht auf eine bestimmte wissenschaftliche Methode, sondern dient als abstrakte, zusammenfassende Bezeichnung für »(geordnete) Listen von (semi-)strukturierten Einträgen«[47]. Eine Datenbank kann das primäre Ergebnis eines Forschungsprojektes sein, sie kann aber auch ein Werkzeug sein, das z.B. lediglich als Hilfsmittel zur Texterschließung in einem Editionsvorhaben aufgebaut wird. In beiden Fällen ist sie jedoch ein wesentliches (Zwischen-)Ergebnis der Forschung. Der Forschungsdatentyp Datenbank wird am Beispiel der Berliner Klassik näher erläutert (vgl. Abbildung 1).

Das Akademievorhaben *Berliner Klassik* (BBAW, Laufzeit 2000–2013) untersuchte die Kulturblüte in Berlin zwischen 1786 und 1815. Hierfür wurden künstlerische, wissenschaftliche und gewerbliche Leistungen erfasst. Neben schriftlichen Publikationen zählen zu den Forschungsergebnissen fünf Datenbanken, die sich der Darstellung der Vernetzung der Berliner Gesellschaft widmen. Die Datenbanken der Berliner Klassik dienten zunächst als Arbeitsinstrument, dann als Präsentationsinstrument. Vier der Datenbanken (die Personendatenbank, die Literaturdatenbank, die Nationaltheaterdatenbank und die Geselligkeitsdatenbank) sind momentan über ein gemeinsames Portal ansprechbar. Bei den erfassten Inhalten handelt es sich vor allem um Namen, Orte, bibliographische Angaben und Datumsangaben Durch die Verknüpfung der einzelnen Datenbanken haben sich zusätzliche Suchfunktionen ergeben, die bei Verlust der Verknüpfungen verloren zu gehen drohen.

[45] Für eine detaillierte Beschreibung der konzeptionellen Grundlagen und Vorarbeiten sei auf Aschenbrenner et al. 2015, S.7ff. verwiesen. Das Konzept der Forschungsdatentypen diente im HDC-Kontext vor allem der pragmatischen Annäherung an die Entwicklung der Angebote des Forschungsdatenzentrums. Es wäre jedoch interessant, es in Zukunft theoretisch und praktisch eingehender zu untersuchen, insbesondere mit Blick auf gemeinsame Erfahrungen und die Arbeitsteilung mit anderen Forschungsdatenzentren.

[46] Vgl. hierzu auch Aschenbrenner et al. 2015, S. 29ff.

[47] Aschenbrenner et al. 2015, S. 29.

Abb. 1: Vereinfachte Darstellung des Forschungsdatentyps Datenbank anhand der Berliner Klassik. Zitiert aus: Andreas Aschenbrenner et al.: Humanities Data Centre – Angebote und Abläufe für ein geisteswissenschaftliches Forschungsdatenzentrum. HDC-Projektbericht Nr. 1. O.O. 2015, S. 30. [online]

Da der Forschungsdatentyp Datenbank nur ungenau eingegrenzt werden kann und nicht alle projektinternen Ad-hoc-Lösungen in die LZA überführt werden müssen, ist eine genaue Abschätzung der Nachfrage momentan schwer möglich. In Zukunft erscheint es, wie auch bei den folgenden Forschungsdatentypen, wichtig, dass der Nachhaltigkeitsaspekt von Anfang an in die Planung und Umsetzung von Datenbanken durch die Projekte einbezogen wird, wobei der Beratung durch Datenzentren wie dem HDC eine zentrale Rolle zukommen kann (vgl. Abschnitt 6: generische Anwendungssysteme).

4.3 Forschungsdatentyp: Digitale Edition

Genauso wenig wie ›die‹ Datenbank, gibt es auch nicht ›die‹ digitale Edition.[48] Zum einen kann der Begriff digitale Edition unterschiedlich definiert werden,[49] zum anderen können bei der konkreten Umsetzung unterschiedliche Methoden zum Einsatz kommen. Auch wenn digitale Editionen zu Präsentationszwecken in Datenbanken abgelegt werden können, unterscheiden sie sich doch zumeist vom Forschungsdatentyp Datenbank dadurch, dass in ihnen literarische und historische Quellen vollständig erschlossen und in eine geeignete Darstellungsform überführt werden. Komplexe digitale Editionen, die unterschiedliche Repräsentationsebenen wie Bild, diplomatische Transkription, edierter Text, Erläuterungen und Sekundärtexte gleichzeitig anzeigen können, lassen sich nur unter Informationsverlust, d.h. Verlust eines mehr oder weniger großen Teils der Forschungsleistung, in eine lineare Darstellung oder in eine für den Druck geeignete Form überführen. Während sich für digitale Editionen auf der Ebene der Primärdaten Metadatenstandards und Standardformate wie XML

[48] Vgl. hierzu auch Aschenbrenner et al. 2015, S. 25ff.
[49] So führt z.B. Sahle 1998 aus: »Wie schon angedeutet, plädiere ich für eine Unterscheidung von digitalen Texten und digitalen Editionen. Digitalisierte Texte aus Editionen gehören zur ersten Gruppe, solange sie eine primär lineare Form beibehalten. Ohne Zweifel weisen auch digitalisierte Texte gedruckter Editionen erhebliche Vorteile gegenüber ihren Vorlagen auf.«

TEI, XML MEI, TUSTEP oder TIFF etabliert haben, gestaltet sich die Umsetzung der interaktiven Darstellungsform recht unterschiedlich.[50]

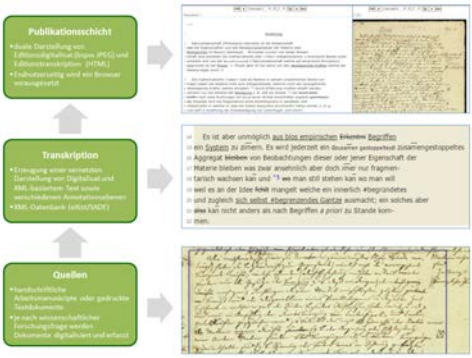

Abb. 2: Vereinfachte Darstellung des Forschungsdatentyps digitale Edition anhand der Kant-Edition. Zitiert aus: Andreas Aschenbrenner et al.: Humanities Data Centre – Angebote und Abläufe für ein geisteswissenschaftliches Forschungsdatenzentrum. HDC-Projektbericht Nr. 1. O.O. 2015, S. 28. [online]

Als Beispiel soll hier die Online-Edition des *Opus Postumum* Immanuel Kants (BBAW, Laufzeit seit 2001, vgl. Abbildung 2) dienen, die eine Neuedition des Kant'schen Nachlasswerkes anstrebt. Die technische Struktur gliedert sich folgendermaßen:

1. Primärdatenschicht/Quellen: Die Primärdaten umfassen Texte inkl. Annotationen (in XML/TEI), damit verknüpfte Digitalisate (bspw. Bildformate) sowie Apparate und Register (in XML).
2. Transkriptionsschicht: In einer XML-Datenbank, basierend auf eXist, werden die Transkriptionen des Originals mit den entsprechenden Digitalisaten vernetzt.
3. Präsentationsschicht: Über eine webbasierte Publikationsschicht mittels XML-Datenbank und XSLT-Skripten, d.h. in einem herkömmlichen Browser, wird die Transkription in verschiedenen Darstellungsformen als HTML ausgegeben. Dabei wird »die Rohtranskription in wechselseitiger abschnittsweiser Verbindung mit den digitalisierten Faksimiles in der diplomatischen Abfolge des Manuskripts«[51] präsentiert.

4.4 Forschungsdatentyp: Datenvisualisierung

Im Kontext der HDC-Angebotsgenese wird der Forschungsdatentyp Datenvisualisierung[52] als eine auf Datenbanken beruhende, für das menschlichen Auge ansprechende, interaktive Präsentation von Forschungsdaten definiert. Im Mittelpunkt steht in diesem Fall das die

[50] Vgl. z.B. Sahle 1998 zu einigen Möglichkeiten der interaktiven Darstellungsformen digitaler Editionen.
[51] Online-Edition des Opus Postumum.
[52] Vgl. hierzu auch Aschenbrenner et al. 2015, S. 31ff.

Präsentation und Interaktion ermöglichende Werkzeug, nicht die der Visualisierung zugrunde liegenden Daten.[53]

Als konkretes Beispiel für eine interaktive Datenvisualisierung soll hier auf die am MPI MMG [54]entwickelte Anwendung *Global Migration Flows* [55] näher eingegangen werden, die Migrationsbewegungen zum Thema hat und die Zusammenstellung individueller Datensets zu Migrationsbewegungen zwischen 1970 und 2011 innerhalb eines herkömmlichen Browsers ermöglicht (vgl. Abbildung 3). Hierdurch können auf einfache und anschauliche Art und Weise Wanderungsgewinne oder -verluste veranschaulicht werden.

Da Datenvisualisierungen stark an ihre Präsentationsumgebung gebunden sind, müssen Langzeitarchivierungslösungen ihrer besonderen Mehrschichtigkeit Rechnung tragen, um ihren Mehrwert zu erhalten. In der Regel bestehen auf einer Browseranwendung beruhende Datenvisualisierungen aus mindestens drei Schichten:

1. Den aufbereiteten und normalisiert in einer Datenbank bereitgestellten Daten. Die Aufbereitung richtet sich nach dem Zweck der Datenvisualisierung und kann unterschiedlich komplex sein – das Spektrum reicht von einfachen Excel-Tabellen bis hin zu komplexen Datenbanken.
2. Einer Prozessierungsschicht (Middleware), die die normalisierten Daten übernimmt und an die Client-Anwendung des Nutzers oder der Nutzerin weitergibt.
3. Eine darauf aufsetzenden Präsentationsschicht als Nutzerschnittstelle. [56]

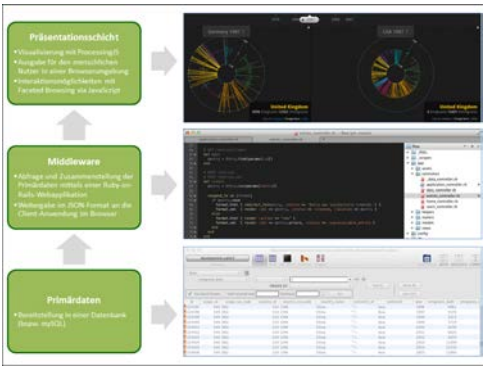

Abb. 3: Vereinfachte Darstellung des Forschungsdatentyps Datenvisualisierung am Beispiel der Global Migration Flows. Zitiert aus: Andreas Aschenbrenner et al.: Humanities Data Centre – Angebote und Abläufe für ein geisteswissenschaftliches Forschungsdatenzentrum. HDC-Projektbericht Nr. 1. O.O. 2015, S. 34. [online]

[53] Streng genommen wäre zur Erfüllung der Anforderung der Sicherung der Primärdaten nur die Archivierung selbiger notwendig. Eine derart minimalistische Sicht ist nicht ungewöhnlich, in diesem Fall steht jedoch der Aspekt der Nachnutzung der interaktiven Anwendung durch den Erhalt der Anwendung selbst oder zumindest der eine spätere Rekonstruktion ermöglichenden Parameter im Vordergrund.
[54] Der Internetauftritt des MPI MMG ist erreichbar unter http://www.mmg.mpg.de.
[55] Die Anwendung ist erreichbar unter http://www.mmg.mpg.de/data-visualization/.
[56] Die für die Entwicklung und Programmierung der Prozessierungs- wie auch der in der nächsten Ebene angesprochenen Präsentationsschicht kann aus ressourcentechnischen Gründen nicht in jedem Forschungsprojekt selbst zur Verfügung stehen. Dies macht die Nachnutzung derartiger Anwendungen besonders attraktiv.

Aus dieser Mehrschichtigkeit ergeben sich verschiedene Konsequenzen. Im Unterschied zu Texten, Bildern oder Videos spielten digitale Datenvisualisierungen in bibliothekarischen oder archivarischen Kontexten bisher so gut wie keine Rolle. Aufgrund ihrer spezifischen Eigenschaften können sie auch nicht einfach in ein dateibasiertes Repositorium überführt werden, ohne an Informationswert zu verlieren, da sie meist nicht datei-, sondern datenbankbasiert sind und Beziehungen von Objekten zueinander darstellen.

Das Thema wissenschaftliche Datenvisualisierung gewinnt vor dem Hintergrund der Digitalisierung der Forschung und der einfacheren Zugänglichkeit (im Sinne von Bedienbarkeit und Standardisierung) von Werkzeugen zur Aufbereitung und Darstellung von digitalen Forschungsdaten sowie der immer besseren Verfügbarkeit (großer Mengen) digitaler Daten zunehmend an Bedeutung. Durch den Einsatz neuer Technologien und der Möglichkeit zur Erschließung und Vernetzung verteilter Datenbestände werden attraktive Darstellungen von Zusammenhängen – gerade auch für eine interessierte Öffentlichkeit – möglich. Visualisierungen können textgebundene Publikationen anreichern und qualitativ aufwerten und entsprechen so einer Nachfrage von Wissenschaftlerinnen und Wissenschaftlern und Fördereinrichtungen. Derzeit ist für die Umsetzung von Visualisierungen noch sehr spezielle IT-Expertise notwendig, aber es ist wahrscheinlich, dass es in Zukunft Werkzeuge geben wird, mit denen auch informationstechnische Laien relativ einfach Visualisierungen erzeugen können.[57] Somit müssen Forschungsdatenzentren in der Lage sein Langzeitarchivierungslösungen für diesen Forschungsdatentyp anzubieten und ggf. über generische Anwendungssysteme für die Wissenschaftlerinnen und Wissenschaftler zur Nutzung zur Verfügung zu stellen.

5. Ergebnisse der Testübernahmen

5.1 Teilergebnisse: Berliner Klassik

Am Beispiel der Berliner Klassik wurde im Projektverlauf der Versuch unternommen, eine bestehende Anwendung technologisch auf neue Füße zu stellen und dabei die signifikanten Eigenschaften zu erhalten. Hierfür wurden die Daten aus der über die Jahre hinweg etwas wildwüchsig geratenen Struktur einer PostgreSQL-Datenbank testweise in eine dokumentenbasierte Datenbank (Mongo-DB, JSON) überführt und das Zope-basierte Frontend der Anwendung auf Basis eines neuen Webframeworks (Express.JS) reimplementiert.

Ziel dieses Experiments war eine Bestimmung des Migrations- und Reimplementierungsaufwands in der Hoffnung, perspektivisch den längerfristigen Betrieb einer ganzen Reihe einfacherer Datenbankanwendungen auf einem einheitlichen Software-Stack sicher betreiben zu können und die Auslotung von Möglichkeiten für einen generischen Viewer, also einer vereinfachten, generischen Präsentationsumgebung für Datenbanken in entsprechend standardisierter, normalisierter Form (vgl. Abschnitt 5.2).

[57] Schon heute gibt es relativ einfache geisteswissenschaftliche Visualisierungsmöglichkeiten wie zum Beispiel das auf der DARIAH-DE Webseite zur Verfügung gestellte Textvisualisierungstool *Digivoy* (https://de.dariah.eu/digivoy).

Obwohl die Überführung manuell durchgeführt wurde und sich komplex gestaltete, hielt sich der Aufwand mit ein bis zwei Personenmonaten dennoch in Grenzen. Wenngleich der Prototyp individuell auf die Anforderungen der *Berliner Klassik* zugeschnitten war, deutet die einfache Übertragbarkeit von einer Datenbank auf eine andere prinzipiell darauf hin, dass generische Datenmodelle und Systemstrukturen entwickelt werden können. Im Bereich der geisteswissenschaftlichen Datenbanken zeichnet sich derzeit noch keine *Best Practice* in der Community ab.

Aus informationswissenschaftlicher Sicht galt es vor allem, bei der Überführung die Verknüpfungen zwischen den Datensätzen und bestimmte Suchfunktionalitäten zu erhalten, die Art der Darstellung im Webinterface spielte eine nachgeordnete Rolle. Die Bewertung des ersten Prototyps durch Fachwissenschaftlerinnen und Fachwissenschaftler fiel jedoch gemischt aus. Dies lag vor allem daran, dass der ursprünglichen Darstellung einschließlich aller Suchfunktionen und Details ein hoher Wert beigemessen wurde, diese aber bei Überführung in einen generischen Viewer nicht vollständig erhalten werden konnten. Das Problembewusstsein für das Thema Langzeitverfügbarkeit der gesamten Anwendung stand hinter diesen Gesichtspunkten zurück. Der Sinn des Migrationsversuchs verstand sich daher nicht von selbst. Erst eine intensive und individuelle Auseinandersetzung im Gespräch konnte eine gewisse Akzeptanz für das Ergebnis erzielen. Erforderliche Nachbesserungen des Prototypen stellten in der Regel weniger eine technische als vielmehr eine kommunikative Herausforderung dar, weil es oft um Änderungen der Darstellung im Detail ging, deren Bedeutung für Fachfremde nicht unmittelbar ersichtlich ist. Selbst wenn das globale Layout der Darstellung nahezu beliebig verändert werden kann, kommt bestimmten Teilaspekten wie Gliederung oder Interpunktion unter Umständen dennoch eine große Bedeutung zu.

Insgesamt zeichnete sich ab, dass eine solche Migrationsstrategie schon bei Anwendungen geringerer Komplexität einen hohen Aufwand seitens des Datenzentrums erforderlich macht, um zu einem für die Wissenschaftlerinnen und Wissenschaftler akzeptablen Ergebnis zu führen. Die Abwägung zwischen Anpassungsaufwand und Zusatznutzen führte zu einer Entscheidung zugunsten einer Konservierung der unveränderten Anwendung in Form einer Virtualisierung (auf einem abgesicherten Server) als Bestandteil des Angebotsportfolios. Diese sogenannte Anwendungskonservierung wird ohne tiefergehende technische Eingriffe solange betrieben, wie in der aktuellen Umgebung die Funktionalität der einzelnen Komponenten gewährleistet ist – auch unter Berücksichtigung von Sicherheitsaspekten.

5.2 Teilergebnisse: Opus Postumum (Kant-Edition)

Das Opus Postumum Immanuel Kants stand dem HDC exemplarisch für den Forschungsdatentyp digitale Edition zur eingehenden Analyse zur Verfügung. Darüber hinaus treiben mit BBAW und SUB zwei der HDC-Konsortialpartner größere digitale Editionsprojekte voran, so dass sich gute Möglichkeiten für einen Austausch ergaben. Abstrakt bekannt, aber erst durch konkrete Beispiele wirklich erfassbar, war das folgende Dilemma: Mit der in XML-TEI ausgezeichneten Transkription liegen einerseits hochgradig strukturierte Daten vor, die einen enormen Aufwand bei der Erfassung bedeuten und prinzipiell vielfältige Möglichkeiten der

technischen Verarbeitung bieten. Andererseits ist TEI ein Standard der in hohem Maße darauf ausgelegt ist, die Auszeichnung an die individuellen Erfordernisse eines Projektes anpassen zu können. Schon auf Ebene der verwendeten Auszeichnungskonventionen ergeben sich somit Inkompatibilitäten zwischen verschiedenen Editionsprojekten.

Ein Ziel der Arbeit am Prototypen Kant-Edition war die Klärung der Umsetzungsmöglichkeiten für sogenannte generische Viewer (nicht nur für Editionsprojekte). Dahinter steht die Idee, verschiedene Editionen nach der Überführung in ein Forschungsdatenzentrum in einer generischen Umgebung anzeigen zu können. Der *DFG-Viewer* [58] zur Anzeige von Digitalisaten kann als bereits umgesetztes Instrument gut zur Illustration der grundsätzlichen Möglichkeiten herangezogen werden. Ebenso wie bei der *Berliner Klassik* stellte sich aber heraus, dass mittelfristig wohl nur eine Anwendungskonservierung eine breitere Akzeptanz bei den Nutzerinnen und Nutzern erlangen kann, wenn es um die Langzeitverfügbarkeit einer digitalen Edition als Webanwendung geht. Die Idee des generischen Viewers wird weiterverfolgt, aber aufgrund der begrenzten Projektressourcen vorerst zurückgestellt.

Die zugrunde liegenden Daten, im Falle der Kant-Edition also TEI-kodierte Transkriptionen und Annotationen sowie die Digitalisate des Manuskripts, können aber auch über die Lebensdauer einer Anwendungskonservierung hinaus einen Wert an sich darstellen. Eine objektbasierte Archivierung dieser Daten wird daher seitens des HDC im Rahmen des Angebots Repositorium angestrebt und an geeigneten Verfahren gearbeitet. Insbesondere sollen die Möglichkeiten von TEI, aber etwa auch des im LZA-Kontext gebräuchlichen Standards METS[59] ausgereizt werden, um die abschnittsweise Verknüpfung zwischen Transkription und Faksimile abzubilden. Im Rahmen eines Beratungsangebots des HDC böte sich perspektivisch die Gelegenheit, darauf hinzuwirken, dass bei künftigen Editionsprojekten etwaige Verknüpfungen möglichst konsequent innerhalb der Daten abgebildet und nicht erst über die Anwendungslogik hergestellt werden.

5.3 Teilergebnisse: Global Migration Flows

In der HDC-Designphase wurde die Überführung der oben beschriebenen Anwendung *Global Migration Flows* in eine virtuelle Maschine getestet. Hierfür wurde die Server-Client-Struktur auf Seiten des Datengebers vorbereitet, d.h. um überflüssige und unerwünschte Bestandteile bereinigt und um eine für die Überführung ausreichende Dokumentation der technischen Spezifikationen ergänzt. Danach wurde die Umgebung geklont und auf eine den Spezifikationen entsprechende virtuelle Maschine überführt.

Im Rahmen dieser Testüberführung zeigte sich, dass sich aus der Mehrschichtigkeit von Datenvisualisierungen für ein Datenzentrum zwei Herausforderungen ergeben. Zum einen besteht die Notwendigkeit zur Überführung der vollständigen Server-Client-Struktur. Dies

[58] Vgl. hierzu http://dfg-viewer.de/ueber-das-projekt/.
[59] Vgl. hierzu http://www.loc.gov/standards/mets/.

schließt auch Abhängigkeiten von externen Anwendungen wie etwa Geoinformationssystemen (GIS), Normdateien (wie bspw. die GND) oder Schriftarten ein. Der hierfür in Betracht zu ziehende Arbeitsaufwand ist stark davon abhängig, wie aufwendig sich die Bereinigung beziehungsweise Anpassung der Server-Client-Struktur im Einzelfall gestaltet. Prinzipiell handelt es sich hierbei um eine Aufgabe im Verantwortungsbereich des Datengebers, nicht des Infrastrukturanbieters, der idealerweise im Vorfeld die für die Überführung notwendigen Anpassungen vornehmen müsste, um den Ingest-Spezifikationen zu entsprechen. Zum anderen ist es unklar, für wie lange die Bereitstellung seitens des Datenzentrums garantiert werden kann, wobei vor allem die Bereitstellung der konservierten Anwendung nach außen kritisch ist. Früher oder später kommen diesbezüglich aufgrund veraltender Komponenten IT-spezifische Sicherheitsfragen zum Tragen. Hier ist dann ggf. auch eine Authentifizierungsschicht notwendig.

5.4 Schlussfolgerungen

Aus den exemplarischen Testübernahmen während der HDC-Designphase lassen sich folgende Schlussfolgerungen für die Angebotsgestaltung geisteswissenschaftlicher Forschungsdatenzentren ziehen:

Angesichts der zunehmend komplexer werdenden Datenstrukturen in den Geisteswissenschaften sind seitens eines Forschungsdatenzentrums Angebote gefragt, die über objektbasierte Lösungen, wie bspw. ein Repository, hinausgehen. Lösungen für komplexere Archivierungsfälle gehen meist mit einem größeren Aufwand für die Infrastrukturanbieter, etwa bezüglich des Ingests und der Kuration, einher, ebenso aber auch im Bereich der IT-Sicherheit, wenn es bspw. um die Konservierung von Umgebungen zu einem bestimmten Zeitpunkt geht.

Für die Entwicklung generischer Viewer (z.B. für digitale Editionen oder Datenbanken), die den wissenschaftlichen (Detail-)Anforderungen der datengebenden und der datennutzenden Wissenschaftlerinnen und Wissenschaftler entsprechen, sind eine enge Zusammenarbeit mit der Community und ein erheblich höherer Mitteleinsatz als ursprünglich geschätzt notwendig.

Forschungsdatentypenspezifische Angebote müssen durch zielgerichtete, individuelle Beratungs- und Unterstützungsangebote hinsichtlich einer möglichst effizienten Überführung der Daten in ein Forschungsdatenzentrum abgerundet werden, um die Wissenschaftlerinnen und Wissenschaftler auf diesem Gebiet zu entlasten.

6. Initiales Angebot des Humanities Data Centre

Unter Berücksichtigung der eingangs beschriebenen wissenschaftlichen Nutzungsszenarien Nachhaltigkeit, Präsentation und Integration (vgl. Abschnitt 3.1) und der Herausforderungen der Langzeitarchivierung komplexer Datenstrukturen, wurde in der HDC-Designphase ein initiales Angebotsportfolio entwickelt (vgl. dazu auch Abbildung 4). Dieses

besteht aus technischen Angeboten und Angeboten zur Beratung und Unterstützung von Wissenschaftlerinnen und Wissenschaftlern, die ineinander übergreifen und sich gegenseitig ergänzen.

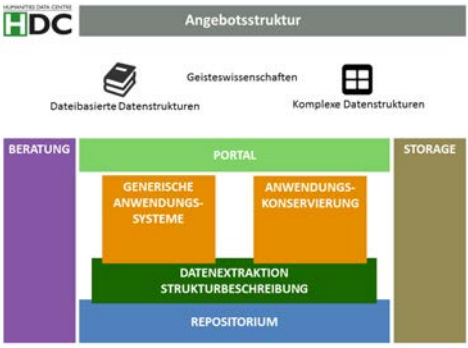

Abb. 4: Vorläufiges Angebotsportfolio des Humanities Data Centre, Stand: Dezember 2015, eigene Grafik.

Technische Angebote:

1. **Anwendungskonservierung**: Mit Blick auf bereits vorhandene LZA-Angebote (bspw. Publikationsrepositorien) und damit einhergehende Standards sowie den Bedarf für die LZA dateibasierter geisteswissenschaftlicher Forschungsdaten und die damit verbundene zu erwartende hohe Akzeptanz bei den datengebenden Wissenschaftlerinnen und Wissenschaftlern zählt ein Repositorium sicherlich zu den Kernbestandteilen eines jeden Forschungsdatenzentrums. Über die Angebotskomponente Repositorium wird die langfristige Verfügbarkeit von Forschungsdaten auf Dateiebene sowie die Referenzierbarkeit mittels persistenter Identifikatoren wie DOIs oder ePIC-Handles sichergestellt. Für das Repositorium ist ein abgestuftes Zugangsmodell vorgesehen, das es erlaubt, auch Forschungsdaten im Repositorium zu archivieren, die nur eingeschränkt oder gar nicht zugänglich sein sollen. Hier ist zunächst die Einbindung des DARIAH-DE-Repositoriums vorgesehen. Daneben werden weitere Repositorien aufgebaut bzw. integriert, sofern dies für bestimmte Objekttypen notwendig ist.
2. **Anwendungskonservierung**: Über die Anwendungskonservierung können komplexe Datenstrukturen und Anwendungen wie zum Beispiel Datenvisualisierungen in ihrem Übergabezustand vorgehalten werden. Sie werden in ihrem Übergabestand quasi ›eingefroren‹, d.h. in ihrer ursprünglichen Struktur (bspw. Client-Server, dazugehörige Anwendungen und Bibliotheken) in das Forschungsdatenzentrum übernommen. Im Vordergrund der Anwendungskonservierung steht die Präsentation und Nachvollziehbarkeit von Forschungsergebnissen und -methoden, weniger die direkte Nachnutzung der Forschungsdaten. Forschungsdaten werden mittels der Anwendungskonservierung unter anderem aufgrund von Sicherheitserwägungen (veraltende Software mit zunehmendem Risiko von Sicherheitslücken) nur für einen

begrenzten Zeitraum ab Übergabe an das Forschungsdatenzentrum präsentiert werden können.
3. **Generische Anwendungssysteme**: Während die Anwendungskonservierung erst nach Projektende ansetzt und die Präsentation von Daten und Ergebnissen abgeschlossener Projekte im Fokus hat, werden im Rahmen generischer Anwendungssysteme standardisierte Umgebungen und Werkzeuge mitgeliefert, die direkt von Wissenschaftlerinnen und Wissenschaftlern eingesetzt werden können – und zwar von Beginn eines Vorhabens an, sodass nach Projektende die Überführung der produzierten Daten und Anwendungen ins Repositorium und/oder die Anwendungskonservierung mit dem geringstmöglichen Aufwand vonstattengehen kann und die Wahrscheinlichkeit, dass sie langfristig und ohne signifikante Verluste erhalten werden können, steigt.

Die technischen Hauptangebote Repositorium, Anwendungskonservierung und generische Anwendungssysteme werden durch die Komponenten Datenextraktion/Strukturbeschreibung, Storage und Portal unterstützt:

1. **Datenextraktion / Strukturbeschreibung**, d.h. die Sicherung der Inhalte aus den Forschungsdaten ohne Format- oder Anwendungsbeschränkungen, soweit dies unter den gegebenen Ressourcen und Kompetenzen möglich ist. Dies kann auf einer sehr basalen Ebene eine flat file-Struktur, bspw. plain text, sein. Dies soll insbesondere die Zusammenführung verschiedener Datenbestände sowie ursprünglich nicht vorgesehene Nutzungsszenarien ermöglichen. Absehbar ist, dass diese Form der Datenaufbereitung mit einem Kurationsaufwand einhergehen wird, über dessen Erbringung fallweise zu entscheiden sein wird.
2. **Storage- bzw. Speicherschicht**: Hierbei handelt es sich um die basale technische Infrastruktur zur Gewährleistung der Bitstream Preservation der übergebenen Forschungsdaten sowohl im Rahmen des HDC-Repositoriums, als auch der Anwendungskonservierung und der Generischen Anwendungssysteme. Perspektivisch ist vorgesehen, aus dieser konventionellen Bitstream Preservation ein vollwertiges Angebot zur Langzeitarchivierung zu entwickeln.
3. **Portal**: Das HDC-Portal bietet den Nutzerinnen und Nutzern einen Zugang bzw. Informationen zu den technischen Angeboten und den Beratungs- und Unterstützungsangeboten und dient als Anlaufstelle für den Zugang zu Schnittstellen für den Austausch von Metadaten oder die Vernetzung von Diensten.

Beratungs- und Unterstützungsangebote:

Neben den bisher beschriebenen technischen Angeboten bildet der für das HDC äußerst wichtige Aspekt der Beratung die zweite Säule des Serviceportfolios. Sie bringt die verschiedenen Akteure zusammen und unterstützt die Wissenschaftlerinnen und Wissenschaftler bezüglich der Aufgaben des Forschungsdatenmanagements sowie der Nutzung der technischen Angebote. Forschungsdatenmanagement und Langzeitarchivierung sind – anders als vielleicht das Publikationsmanagement – Aufgaben, bei denen Self Service-

Angebote, zumindest zum gegenwärtigen Zeitpunkt, nur bedingt Erfolg versprechend sind. Auch wenn die Beratung im Vorfeld von Forschungsprojekten recht arbeitsintensiv ist, kann sie idealerweise im weiteren Verlauf des Projekts beziehungsweise am Ende der Projektlaufzeit Probleme und Kosten bezüglich der Langzeitarchivierung minimieren oder gänzlich neue Möglichkeiten der Präsentation und Nachnutzung ermöglichen. Daher sollte bei der Angebotsgestaltung eines Forschungsdatenzentrums nicht der Fehler begangen werden, sich allein auf die Entwicklung technischer Dienste oder den Aufbau von Infrastrukturen zu konzentrieren.

Das HDC sieht als Antwort auf den Beratungsbedarf den Aufbau eines Netzwerks von Datenkuratorinnen und Datenkuratoren vor, die am Datenzentrum selbst, aber auch an assoziierten wissenschaftlichen Einrichtungen verankert sind, und dort die Wissenschaftlerinnen und Wissenschaftler direkt unterstützen. Zu den Kernaufgaben der Datenkuratorinnen und -kuratoren werden die persönliche Beratung der Wissenschaftlerinnen und Wissenschaftler zu allen Aspekten Forschungsdatenmanagements (u.a. zu Datenmanagementplänen, fachspezifischen Standards, empfehlenswerten Technologien und rechtlichen Fragen) gehören. Diese findet im Idealfall projektbegleitend über die gesamte Laufzeit eines Forschungsvorhabens statt, d.h. von der Beratung in der Planungs- bzw. Antragsphase bis zur Betreuung der Überführung von Forschungsdaten und -ergebnissen in das Datenzentrum bei Projektende. Die Angebote zur Beratung und Unterstützung werden durch solche zur Schulung ergänzt, die aus Online-Angeboten (bspw. Materialien zur Nutzung der HDC-Angebote und Tutorials zu Datenmanagementthemen) und Workshops bestehen. Während die Beratung sich vornehmlich an Datengeberinnen und -geber wendet und einzelfallbezogen ist, richten sich die Schulungen an ein – sowohl hinsichtlich des Spektrums als auch zahlenmäßig – breiteres Publikum, das auch diejenigen einbezieht, die Daten nachnutzen.

Das initiale Angebotsportfolio des HDC stellt das Ergebnis eines iterativen Prozesses während der Designphase dar. Die Ergebnisse der Pilotübernahmen haben die auf Grundlage der Anforderungen aus der Wissenschaft starke Ausrichtung der Angebote auf komplexe Forschungsdatentypen bestätigt. Sie haben jedoch auch zur Rückstellung der Entwicklung generischer Viewer zum momentanen Zeitpunkt geführt, was vor allem mit Blick auf die vorhandenen Ressourcen wie auch die vergleichsweise geringe Akzeptanz bei den Wissenschaftlerinnen und Wissenschaftlern erfolgte. Zudem haben sie verdeutlicht, wie wichtig neben der Bereitstellung technischer Angebote eine intensive Beratung und Unterstützung der Wissenschaftlerinnen und Wissenschaftler ist, sodass diese Angebote im überarbeiteten Portfolio noch stärker gewichtet wurden.

Diese zunächst begrenzten initialen Angebote sollen unter Berücksichtigung der Schwerpunktsetzungen anderer geisteswissenschaftlicher Forschungsdatenzentren und der Rückkopplung aus den Fachwissenschaften insbesondere über das Netzwerk der Datenkuratorinnen und -kuratoren im Rahmen der dem HDC zur Verfügung stehenden Ressourcen ausgebaut werden, z.B. um Angebote für weitere Forschungsdatentypen ergänzt werden, wenn die technischen Voraussetzungen und Ressourcen gegeben sind. Dieses

modulare Vorgehen bietet den Vorteil, von Anfang an ein funktionales Angebot bereitstellen zu können.

7. Fazit

Eingangs wurde die Frage nach der Angebotsstruktur für ein geisteswissenschaftliches Forschungsdatenzentrum gestellt. Welche Angebote sind notwendig, um die Ergebnisse geisteswissenschaftlicher Forschung langfristig verfügbar zu halten und ihre Nachnutzung zu ermöglichen?

Es bleibt festzuhalten, dass das konkrete Angebot eines Forschungsdatenzentrums immer einen Ausgleich zwischen den Anforderungen der Wissenschaftlerinnen und Wissenschaftler und den Umsetzungsbedingungen der Infrastrukturentwickler und -betreiber wird finden müssen. Auf der Seite der Wissenschaft verlangt die Vielfalt an Forschungsdatentypen bzw. an Formaten, Inhalten, Methoden, Standards und Technologien sowie die selbst innerhalb eines Forschungsdatentyps durch die jeweils individuellen Projektkonstellationen sich ergebende hohe Varianz in der konkreten Umsetzung in hohem Maße nach aufwendigen Einzelfalllösungen. Demgegenüber haben Infrastrukturanbieter aufgrund der Notwendigkeit eines effizienten Ressourceneinsatzes sowie technologischer Beschränkungen und Praktikabilitätserwägungen ein starkes Interesse an standardisierten Lösungen. Der Ausgleich zwischen diesen gegensätzlichen Anforderungen wurde im initialen HDC-Angebotsportfolio durch die Konzentration auf eine Reihe von Forschungsdatentypen gepaart mit umfangreichen Angeboten zur intensiven Beratung und Unterstützung der Wissenschaftlerinnen und Wissenschaftler geschaffen.

In Anbetracht des breiten Spektrums von Forschungsdatentypen in den Geisteswissenschaften, von dem die in diesem Beitrag besprochenen Beispiele nur einen kleinen Ausschnitt andeuten, ist es ohnehin unwahrscheinlich, dass in Zukunft ein Datenzentrum allein die Lösungen für ihre Gesamtheit bereitstellen kann. Dies wird nur in Zusammenarbeit mit anderen geisteswissenschaftlichen Datenzentren gelingen, weshalb die Entwicklung einer kooperativen arbeitsteiligen Angebotsstruktur geboten scheint, im Rahmen derer die einzelnen, jeweils auf einen bestimmten Ausschnitt des Spektrums geisteswissenschaftlicher Daten spezialisierten Datenzentren in ihrer Gesamtheit Lösungen für möglichst viele geisteswissenschaftliche Forschungsdaten bzw. Forschungsdatentypen bereitstellen können.

Bibliographische Angaben

Advancing Technologies and Federating Communities. A Study on Authentication and Authorisation Platforms for Scientific Resources in Europe. Final Report. A study prepared for the European Commission, DG Communications Networks, Content & Technology. Hg. von der Europäischen Kommission 2012. [online]

Andreas Aschenbrenner / Frank Dickmann / Harry Enke / Bernadette Fritzsch / Michael Lautenschlager / Benjamin Löhnhardt / Jens Ludwig / Torsten Rathmann / Angelika Reiser / Florian Schintke / Jens Stegmann / Stefan Strathmann: Generische Langzeitarchivierungsarchitektur für D-Grid. Arbeitspaket 3. Hg. vom WissGrid-Projekt. 14. Januar 2010. [online]

Andreas Aschenbrenner / Stefan Buddenbohm / Claudia Engelhardt / Ulrike Wuttke: Humanities Data Centre – Angebote und Abläufe für ein geisteswissenschaftliches Forschungsdatenzentrum. HDC-Projektbericht Nr. 1. Hg. vom Humanities Data Centre. Mai 2015. [online]

Nikolaos Beer / Kristin Herold / Wibke Kolbmann / Thomas Kollatz / Matteo Romanello / Sebastian Rose, Niels-Oliver Walkowski / Felix Falko Schäfer / Maurice Heinrich: Datenlizenzen für geisteswissenschaftliche Forschungsdaten - Rechtliche Bedingungen und Handlungsbedarf. DARIAH-DE Working Papers 6. 2014. Hg. von DARIAH-DE, Niedersächsische Staats- und Universitätsbibliothek. 2014. [online] [Nachweis im GBV]

Mirjam Blümm / Stefan E. Funk / Sibylle Söring: Die Infrastruktur-Angebote von DARIAH-DE und TextGrid. In: Information. Wissenschaft & Praxis 66 (2015), H.5–6, S. 304–312. [online] [Nachweis im GBV]

Fabian Cremer / Claudia Engelhardt / Heike Neuroth: Embedded Data Manager - Integriertes Forschungsdatenmanagement: Praxis, Perspektiven und Potentiale. In: Bibliothek – Forschung und Praxis 39 (2015), H. 1, S. 13–31. [online] [Nachweis im GBV]

Christine L. Borgman: Scholarship in the Digital Age: Information, Infrastructure, and the Internet. Cambridge, London 2007. [Nachweis im GBV]

Christine L. Borgman: Big Data, Little Data, No Data: Scholarship in the networked world. Cambridge, London 2015. [Nachweis im GBV]

DINI-Zertifikat für Open-Access-Repositorien und -Publikationsdienste, Arbeitsgruppe »Elektronisches Publizieren«, Version 4.0. Hg. von der DINI-Arbeitsgruppe Elektronisches Publizieren. Oktober 2013. [online] [Nachweis im GBV]

Ortwin Dally et al.: IANUS. Die Konzeption eines nationalen Forschungsdatenzentrums für die Archäologie und die Altertumswissenschaften. In: Archäologie und Informationssysteme. Vom Umgang mit archäologischen Fachdaten in Denkmalpflege und Forschung. Hg. von Stefan Winghart. Hameln 2013, S. 118–127. (= Arbeitshefte zur Denkmalpflege in Niedersachsen 42). [online] [Nachweis im GBV]

Ralf Depping: Publikationsservices im Dienstleistungsportfolio von Hochschulbibliotheken. Eine (Neu-)Verortung in der wissenschaftlichen Publikationskette. In: o-bib 1 (2014), H. 1, S. 71–91. PDF. [online] [Nachweis im GBV]

Empfehlungen zur gesicherten Aufbewahrung und Bereitstellung digitaler Forschungsdaten. In: DFG.de. Januar 2009. [online]

Forschungsdaten-Leitlinie der Universität Göttingen (einschl. UMG). Hg. von Georg-August-Universität Göttingen. 1. Juli 2014. [online]

Guidelines on Data Management in Horizon 2020. Version 2.0. Hg. von der Europäischen Kommission, Directorate – General for Research & Innovation. 30. Oktober 2015. [online]

Handbuch Forschungsdatenmanagement. Hg. von Stephan Büttner / Hans-Christoph Hobohm / Lars Müller. Bad Honnef 2011. [online] [Nachweis im GBV]

Maurice Heinrich / Felix F. Schäfer: IANUS als fachspezifisches Forschungsdatenzentrum für die Altertumswissenschaften in Deutschland (Folien Hauptvortrag FORGE 2015, Hamburg). 17. September 2015. [online]

Jasmin Hügi / René Schneider: Digitale Forschungsinfrastrukturen in den Geistes- und Geschichtswissenschaften. Genf 2013. [online]

Maxi Kindling / Peter Schirmbacher / Elena Simukovic: Forschungsdatenmanagement an Hochschulen: das Beispiel der Humboldt-Universität zu Berlin. In: LIBREAS. Library Ideas 23 (2013), S. 43–63. [online] [Nachweis im GBV]

Maxi Kindling / Peter Schirmbacher / Elena Simukovic / Alexander Struck / Raphael Thiele: Was sind Ihre Forschungsdaten? Interviews mit Wissenschaftlern der Humboldt-Universität zu Berlin. Bericht, Version 1.0. Berlin 2014. [online]

Paul Klimpel / John H. Weitzmann: Forschen in der digitalen Welt. Juristische Handreichung für die Geisteswissenschaften. DARIAH-DE Working Papers 12. 2015. [online] [Nachweis im GBV]

Camilla Leathem / Dominik Adrian: Bestandsaufnahme und Analyse geistes- und sozialwissenschaftlicher Grundlagenforschung an den europäischen Wissenschaftsakademien und ähnlichen Forschungseinrichtungen. Berlin 2015. [online]

Langzeitarchivierung von Forschungsdaten. Eine Bestandsaufnahme. Hg. von Heike Neuroth, Stefan Strathmann, Achim Oßwald, Regine Scheffel, Jens Klump, Jens Ludwig. Boizenburg 2012. [online] [Nachweis im GBV] [Nachweis im GBV]

Leitfaden zum Forschungsdaten-Management. Handreichungen aus dem WissGrid-Projekt. Hg. von Jens Ludwig / Harry Enke. Glückstadt 2013. [online] [Nachweis im GBV] [Nachweis im GBV]

Leitlinien zum Umgang mit Forschungsdaten. In: DFG.de. 30. September 2015. [online]

Bryn Nelson: Data sharing: Empty archives. In: Nature 461 (2009), S. 160–163. [online] [Nachweis im GBV]

Wolfgang Pempe: Geisteswissenschaften. In: Langzeitarchivierung von Forschungsdaten. Eine Bestandsaufnahme. Hg. von Heike Neuroth, Stefan Strathmann, Achim Oßwald, Regine Scheffel, Jens Klump, Jens Ludwig. Boizenburg 2012, S. 137–159. [online] [Nachweis im GBV] [Nachweis im GBV]

Patrick Sahle: Digitale Editionen, Publikation auf der Webseite des Autors. Februar 1998. [online]

Patrick Sahle / Simone Kronenwett: Jenseits der Daten. Überlegungen zu Datenzentren für die Geisteswissenschaften am Beispiel des Kölner 'Data Center for the Humanities'. In: LIBREAS. Library Ideas, 23 (2013), S. 76–96. [online] [Nachweis im GBV]

Christoph Schöch: Big? Smart? Clean? Messy? Data in the Humanities. In: Journal of Digital Humanities 2 (2013), H. 3, S. 2–14. DOI dx.doi.org/10.5281/zenodo.8432 [Nachweis im GBV]

Vorschläge zur Sicherung guter wissenschaftlicher Praxis. Denkschrift. Ergänzte Auflage. Hg. von Deutsche Forschungsgemeinschaft. Weinheim, 1998/2013. [online] [Nachweis im GBV] [Nachweis im GBV]

Abbildungslegenden und -nachweise

Abb. 1: Vereinfachte Darstellung des Forschungsdatentyps Datenbank anhand der Berliner Klassik. Zitiert aus: Andreas Aschenbrenner et al.: Humanities Data Centre – Angebote und Abläufe für ein geisteswissenschaftliches Forschungsdatenzentrum. HDC-Projektbericht Nr. 1. O.O. 2015, S. 30. [online]

Abb. 2: Vereinfachte Darstellung des Forschungsdatentyps digitale Edition anhand der Kant-Edition. Zitiert aus: Andreas Aschenbrenner et al.: Humanities Data Centre – Angebote und Abläufe für ein geisteswissenschaftliches Forschungsdatenzentrum. HDC-Projektbericht Nr. 1. O.O. 2015, S. 28. [online]

Abb. 3: Vereinfachte Darstellung des Forschungsdatentyps Datenvisualisierung am Beispiel der Global Migration Flows. Zitiert aus: Andreas Aschenbrenner et al.: Humanities Data Centre – Angebote und Abläufe für ein geisteswissenschaftliches Forschungsdatenzentrum. HDC-Projektbericht Nr. 1. O.O. 2015, S. 34. [online]

Abb. 4: Vorläufiges Angebotsportfolio des Humanities Data Centre, Stand: Dezember 2015, eigene Grafik.