Humanities Data Centre

-

Angebote und Abläufe für ein geisteswissenschaftliches Forschungsdatenzentrum

Projektakronym:	HDC
Förderer:	Niedersächsisches Ministerium für Wissenschaft und Kultur/ Niedersächsisches Vorab
Fördernummer:	VWZN2941 Humanities Data Centre
Thema:	Forschungsdatenmanagement
Projektdauer:	01.05.2014 – 30.04.2016
Nummer des Berichts:	1
Titel des Berichts:	Humanities Data Centre – Angebote und Abläufe für ein geisteswissenschaftliches Forschungsdatenzentrum
Stand des Berichts:	Mai 2015
AP Nummer:	1/2
AP-Leitung:	BBAW/ MPIMMG
Autoren:	Andreas Aschenbrenner, Stefan Buddenbohm, Claudia Engelhardt, Ulrike Wuttke

Inhalt

1 Einleitung: Nachfrage und Angebot geisteswissenschaftlicher Forschungsdaten

Geisteswissenschaftliche Forschungsdaten sollen über die üblichen Lebenszyklen von Datenformaten und Software hinaus nutzbar sowie über die Anwesenheit der ursprünglichen Datenersteller hinaus interpretierbar und nachnutzbar sein. Dieser Bedarf kommt sowohl von Forschern aus Gründen der Guten Wissenschaftlichen Praxis[1], der Nachnutzbarkeit von Forschungsdaten und als Teil des wissenschaftlichen Diskurses, wie auch von Forschungsförderern aus Sicht der Effizienz von Projektfinanzierungen.

Ein geisteswissenschaftliches Forschungsdatenzentrum, das diesen Bedarf adressiert, muss unterschiedliche Akteure zusammenbringen: Forscher, Förderer, Rechenzentren und andere Institutionen (Forschungseinrichtungen, Fachgesellschaften, Standardisierungsinitiativen). In der Erhebung von Anforderungen und Rahmenbedingungen eines HDC treffen daher **unterschiedliche Perspektiven** aufeinander: konkret technisch vs. nachhaltig organisatorisch; von theoretisch machbar bis pragmatisch sinnvoll; von „unmittelbar“ bis „langfristig“.

Die langfristige Nachhaltigkeit und Nachnutzbarkeit von Forschungsdaten zu sichern, ist dabei nicht nur ein technisches, sondern vor allem ein organisatorisches Thema. Relevante Standards[2] konzentrieren sich daher nicht auf Software, sondern auf z.B. Abläufe und Dokumentation. Analog dazu sind die **Eckpfeiler des Angebots** eines HDC weniger Technologiespezifikationen, als vielmehr **definierte Abläufe und übergreifende Datenstrukturen**. Teil dieser Abläufe ist auch ein ständiger Blick nach Innen und Außen, um vorhandene Strukturen den sich ändernden, jeweils aktuellen Bedingungen anzupassen.

So wie die organisatorische Struktur eines HDC über längere Zeiträume kontinuierlich weiterentwickelt werden muss, muss sich auch sein **Angebot kontinuierlich in die Breite erweitern**. Methoden und Werkzeuge in den Geisteswissenschaften sind bei so unterschiedlichen Disziplinen wie Byzantinistik, medizinische Ethik und Musikwissenschaften sehr divers. Schon aus dem Wissenschaftsverständnis heraus widersprechen sie sich teilweise selbst innerhalb einer Disziplin und mit den Entwicklungen in den Digital Humanities verändern sie sich aktuell grundlegend weiter. Vor diesem Hintergrund können kaum alle Anforderungen aus allen Disziplinen erhoben und eingefroren, geschweige denn in ein umfassendes Technologiepaket umgesetzt werden. Stattdessen wird ein HDC seine Angebote auf Basis von konkreten (Nutzer-)Anforderungen aufbauen und diese „lokalen Durchstiche“ von der (Nutzer-)Schnittstelle bis in die Infrastruktur sukzessive erweitern müssen.

Auf der anderen Seite ist der Aufbau eines Forschungsdatenzentrums, das ständig lernt und seine Angebote erweitert, aus Sicht eines Rechenzentrums zwar technisch und organisatorisch vorstellbar, finanziell aber kaum realistisch umzusetzen. Rechenzentren funktionieren auf Basis von stabilen, generischen Diensten, die sie auf eine große Nutzergruppe skalieren können. Lokale Angebote für wenige Nutzer skalieren nicht und eignen sich kaum für ein nachhaltiges Geschäftsmodell. Die langfristige Finanzierbarkeit eines solchen HDCs wäre in ständiger Gefahr. Dieser Widerspruch[3]

[1] Gute Wissenschaftliche Praxis, DFG, http://www.dfg.de/foerderung/grundlagen_rahmenbedingungen/gwp/
[2] z.B. OAIS, TRAC/TDR; siehe dazu Kapitel 2.2 Projektumfeld: Standards und Richtlinien
[3] Diese Herausforderung zur „Skalierbarkeit der Angebote“ ist nicht spezifisch geisteswissenschaftlich, sondern betrifft alle Wissenschaftsdisziplinen.

zwischen der Notwendigkeit der kontinuierlichen Weiterentwicklung lokaler Angebote bei gleichzeitig (fehlender, bzw. schwieriger) **Skalierbarkeit der Angebote** ist eine der großen Herausforderungen beim Aufbau eines HDC.

Ein weiterer Aspekt ist der **Ausgleich zwischen Stabilität und Erneuerung**. Wie oben beschrieben muss ein HDC über längere Zeiträume lernen und seine Angebote anpassen. Gleichzeitig ist Stabilität der konkreten (technischen) Angebote aber wichtig[4]: für die technische Infrastruktur zum Aufbau stabiler technischer Dienste; für Nutzer eines HDC, um bereits zu Projektbeginn die Angebote eines HDC für ihr Forschungsdatenmanagement berücksichtigen und auch zu Projektende noch darauf vertrauen zu können. Stabilität und Innovation sind dabei zwar keine Gegensätze, ihr Verhältnis bedarf aber sorgfältiger Ausbalancierung.

Die Angebote eines HDC müssen daher einerseits so gezielt (spezifisch) wie möglich sein (um konkrete Anforderungen aus einem wissenschaftlichen Kontext zu adressieren), andererseits so generisch wie nötig (um ein nachhaltiges Geschäfts- und Organisationsmodell zu ermöglichen). Sie müssen möglichst stabil sein (für die technische Planbarkeit der Infrastruktur und das Vertrauen der Nutzer), gleichzeitig jedoch auch innovativ (um in den sich dynamisch entwickelnden Digital Humanities aktuell zu bleiben). Dieser Bericht sucht den Überschneidungen zwischen diesen Perspektiven, indem er neben der Erhebung konkreter Anforderungen auch Übertragbarkeit und Nachfrage thematisiert und eine Balance zwischen Stabilität und Innovation anstrebt. Die Anforderungen und Nutzungsszenarien in diesem Bericht bleiben auf einem abstrakten Niveau, so dass sie möglichst in unterschiedlichen Kontexten und über längere Zeit hinweg gültig bleiben. Auf Basis dieses Berichts sollen langfristige organisatorische Abläufe (TP1, Teilprojekt 1, ebenfalls in diesem Dokument), konkrete technische Angebote (HDC TP2) sowie Ansätze für das Geschäftsmodell (HDC TP3) abgeleitet werden können.

[4] Dieser Bericht konzentriert sich dabei mehr auf Anforderungen und Abläufe. Technische Strukturen und Spezifikationen werden im TP2 des HDC-Projektes erarbeitet und in einem parallelen Dokument veröffentlicht.

2 Projektumfeld und Methodik des HDC

Der Aufbau von nachhaltigen Infrastrukturen für Forschungsdatenmanagement und Gute Wissenschaftliche Praxis sind aktuelle Themen, die in der Forschungspolitik von großer Bedeutung sind.[5] Dementsprechend ist die Sichtbarkeit dieser Themen hoch und es gibt eine Reihe laufender Aktivitäten. Dieses Kapitel gibt einen Überblick über aktuelle Umfragen und Analysen (Kapitel 2.1), zu Standards und Richtlinien (Kapitel 2.2) sowie zu verwandten Initiativen (Kapitel 2.3).

Die Sichtung abgeschlossener und laufender Umfragen im relevanten Themenbereich hat die Anforderungsanalyse für diesen Bericht mit einer hinreichenden Menge an grundlegenden Daten unterlegt, sodass keine zusätzliche allgemeine Umfrage durchgeführt wurde. Stattdessen konzentrierte sich das Teilprojekt TP1 auf das Metastudium existierender Analysen von Nutzeranforderungen sowie vertiefende Expertengespräche mit Forschern, Rechenzentren und existierenden Infrastrukturinitiativen.[6] Dieses Vorgehen ermöglichte einen tiefen Einblick und schnellen Erkenntnisgewinn für das Projekt.

2.1 Erhebungen von Nutzeranforderungen an ein Forschungsdatenzentrum

In der wissenschaftlichen Community gibt es eine Vielzahl an Umfragen[7] und Analysen[8] zu den Anforderungen der Wissenschaftler an Angebote zum Forschungsdatenmanagement,

[5] Auswahl, der in diesem Kontext relevanten Empfehlungen und Richtlinien (politische Resolutionen, nationale Strategiebildung):
* UNESCO Guidelines for the preservation of digital heritage. (2003)
http://www.unesco.org/new/en/communication-and-information/resources/publications-and-communication-materials/publications/full-list/guidelines-for-the-preservation-of-digital-heritage/
* Council of the European Union: Council Conclusions on scientific information in the digital age: access, dissemination and preservation. (2007) https://ec.europa.eu/digital-agenda/sites/digital-agenda/files/council_conclusions_nov2007.pdf
* Kommission Zukunft der Informationsinfrastruktur: Gesamtkonzept für die Informationsinfrastruktur in Deutschland. (April 2011).
http://www.leibniz-gemeinschaft.de/fileadmin/user_upload/downloads/Infrastruktur/KII_Gesamtkonzept.pdf (S. B109)
* Wissenschaftsrat: Empfehlungen zu Forschungsinfrastrukturen in den Geistes- und Sozialwissenschaften. (Januar 2011)
http://www.wissenschaftsrat.de/download/archiv/10465-11.pdf
* Wissenschaftsrat: Empfehlungen zur Weiterentwicklung der wissenschaftlichen Informationsinfrastrukturen in Deutschland bis 2020. (April 2012) http://www.wissenschaftsrat.de/download/archiv/2359-12.pdf

[6] Eine chronologische Liste der HDC-Aktivitäten wird im Projektwiki bzw. im Projektabschlussbericht geführt. Als Auswahl aus den verschiedenen Gesprächen sind folgende zu nennen.
Nutzer: Teilnahme an Konferenzen wie DHd Passau und Graz;
Organisation einer fokussierten Session im Rahmen des THATCamp in Göttingen (2014).
Infrastrukturen: Workshop im August 2014 mit Peter Wittenburg (EUDAT, CLARIN, RDA); enge Abstimmung in DARIAH.
Rechenzentren: Neben den am HDC beteiligten Rechenzentren GWDG und ZIB sind auch das KIT Karlsruhe, RZ Garching und FZ Jülich im Rahmen von DARIAH-DE in laufende Diskussionen eingebunden.
Universitäten: Gespräche mit den Stabsstellen Forschungsdaten an der Universität Göttingen, HU Berlin, FU Berlin, FH Potsdam und anderen.
Verwandte Initiativen: enge Gespräche vor allem mit DCH Köln, IANUS am DAI Berlin, GAMS in Graz sowie internationalen Partnern in DARIAH-EU, z.B. DANS in den Niederlanden. (siehe Kapitel 2.3)
Durchführung von (Kreativ-)Workshops mit Vertretern aller Perspektiven in Berlin (30.8.2014) und Göttingen (17.10.2014)
Teilnahme an Arbeitsgruppen: u.a. DHd-Arbeitsgruppe "Aufbau von Datenzentren" (siehe Kapitel 2.3)

[7] Umfragen zu Nutzeranforderungen an Forschungsdatenarchive und -dienste, Shortlist:
→ abgeschlossene Umfragen, Ergebnisse einsehbar
* Bedarfsanalyse für ein Angebot „Digitale Langzeitarchivierung“ in den Geisteswissenschaften (2008)
http://www.sagw.ch/dms/sagw/laufende_projekte/infrastrukturinitiative/Fragebogen_dt.pdf
http://www.sagw.ch/dms/sagw/laufende_projekte/DDZ/Beilagen/Auswertung
* Bamboo Planning Project (2008-2010)
https://wikihub.berkeley.edu/display/pbamboo/Bamboo+Planning+Project+-+April+2008+to+September+2010

Langzeitarchivierung oder Infrastrukturen in den Geisteswissenschaften und speziell in den Digital Humanities. Im Rahmen der Angebotsdefinition hat das HDC diese Umfragen analysiert und mit laufenden Vorhaben Kontakt aufgenommen, um Einsicht in aktuelle Zwischenergebnisse zu erhalten. Wiederkehrende oder besonders herausstechende Punkte dabei waren:

- **Digitale Forschungsdaten sind weit verbreitet** und mehr als digitale Repräsentationen von analogen Materialien: Ein Großteil der Wissenschaftler (90% in der SAGW-Analyse von 2008[9]) erzeugt digitale Forschungsdaten. Obwohl Dokument-Formate vorherrschen, arbeiten auch über 70% bereits mit Datenbanken. Hingegen werden nur für knapp die Hälfte der Forschungsdaten auch Metadaten erfasst.
- **Nachnutzbarkeit** wird von den Wissenschaftlern gewünscht, beunruhigt sie jedoch gleichzeitig: Ein Großteil der Wissenschaftler (91% in der PARSE-Analyse von 2009) findet Nachnutzbarkeit von Forschungsdaten wichtig, allerdings vor allem im Hinblick auf die Daten anderer, denn nur 25% stellen ihre eigenen Daten zur Verfügung. Dies hat teilweise wahrscheinlich technisch-organisatorische Gründe, aber auch Angst vor „Missbrauch“ und rechtliche Gründe spielen dabei eine Rolle (laut PARSE).
- Publikation und **Visibilität ihrer Forschungsergebnisse** wünscht sich eine überwiegende Mehrheit der Wissenschaftler von einer digitalen Forschungsinfrastruktur in den Geisteswissenschaften (über 80% in der DARIAH-DE Umfrage, 2014). Als weitere wichtige

https://wikihub.berkeley.edu/display/pbamboo/Project+Bamboo+Literature+Review
https://wikihub.berkeley.edu/display/pbamboo/Project+Bamboo+Scholarly+Practice+Report
* PARSE.insight: Insight into digital preservation of research output in Europe. Survey Report. (2009).
http://www.parse-insight.eu/downloads/PARSE-Insight_D3-4_SurveyReport_final_hq.pdf
* Empirische Untersuchung zu digitalen Arbeitspraktiken in den Geisteswissenschaften an der Universität Regensburg (März-Mai 2013)
http://www.ur.de/sprache-literatur-kultur/medieninformatik/medien/materialien/infografik-umfrage-dh-ur.pdf
* Meghan Banach Bergin. 2013. "Sabbatical Report: Summary of Survey Results on Digital Preservation Practices at 148 Institutions" The Selected Works of Meghan Banach Bergin. http://works.bepress.com/meghan_banach/7
→ laufende Umfragen (das HDC konnte Vorabergebnisse und Rohdaten einsehen)
* DARIAH-EU: digitale Ressourcen, Methoden und Werkzeuge
http://surveys.dcu.gr/index.php/934532/lang-de
* DARIAH-DE: 1. Werkzeuge für die Lizenzerstellung. 2. Anforderungen an die DARIAH-DE Service Unit (DEISU). 3. Erfolgskriterien. 4. Gebrauch von digitalen Werkzeugen und Diensten während des geisteswissenschaftlichen Forschungsprozesses. (https://survey.gwdg.de/index.php/934152/lang-de)
* Europeana Cloud: Researcher Needs. 2014
http://www.pro.europeana.eu/web/europeana-cloud/work-package-1-researcher-needs
* IAG „Zukunft des wissenschaftlichen Publikationssystems“
http://www.publikationssystem.de/newsletter/online-konsultation-publikationssystem
* Union der Akademien / ALLEA: Bestandsaufnahme und Analyse geistes- und sozialwissenschaftlicher Grundlagenforschung an den europäischen Wissenschaftsakademien. (BMBF-Projekt)
http://www.akademienunion.de/BMBF_Projekt/
[8] (Beispiele für Analysen von Nutzeranforderungen für Forschungsdatenarchive und -Infrastrukturen)
* RePAH: A User Requirements Analysis for Portals in the Arts and Humanities (Final Report) (2006)
http://repah.dmu.ac.uk/report/pdfs/RePAHReport-Complete.pdf
* Warwick, C., M. Terras, P. Huntington, N. Pappa. “If You Build It Will They Come? The LAIRAH Study: Quantifying the Use of Online Resources in the Arts and Humanities through Statistical Analysis of User Log Data.” Literary and Linguistic Computing 23, no. 1 (2008): 85–102. http://discovery.ucl.ac.uk/176758/
* Martin Feijen, What Researchers Want, SURFfoundation (2011).
http://www.surf.nl/binaries/content/assets/surf/en/knowledgebase/2011/What_researchers_want.pdf
[9] Bedarfsanalyse für ein Angebot 'Digitale Langzeitarchivierung' in den Geisteswissenschaften (2008)
http://www.sagw.ch/dms/sagw/laufende_projekte/infrastrukturinitiative/Fragebogen_dt.pdf
http://www.sagw.ch/dms/sagw/laufende_projekte/DDZ/Beilagen/Auswertung

Mehrwerte werden z.B. spezialisierte Funktionen zur Datenanalyse, Annotation und Kollaboration genannt.

- Wissenschaftler brauchen **lokale und spezifische Lösungen** sowie **Beratung** direkt bezogen auf ihr Forschungsumfeld (SURF, 2011).

Aus diesen Erkenntnissen hat Knowledge Exchange – ein Zusammenschluss europäischer Forschungsförderer und Infrastruktureinrichtungen bzw. -initiativen mit dem Ziel, die digitale Infrastruktur zu verbessern – eine Diskussion zur Etablierung einer "Kultur des Datenaustausches" initiiert.[10] Teil davon sind Bestrebungen, eine Art „Impact Factor“, wie er für „traditionelle“ Publikationen bekannt ist, auch für Forschungsdaten zu etablieren, um so den Reputationsgewinn durch Datenpublikationen und mithin die Motivation von Forschenden, Daten zu veröffentlichen, zu steigern.[11]

Obwohl einige dieser Umfragen bereits älter sind, bestätigen aktuelle Erhebungen deren Ergebnisse bzw. unterstreichen den sich verstärkenden Trend zu digitalen Forschungsdaten und Werkzeugen. Die Erkenntnisse beziehen zwar auch die Natur- und Wirtschaftswissenschaften mit ein, doch auch hier zeigen die aktuellen Umfragen in den Geisteswissenschaften (DARIAH, ALLEA, Europeana), dass sie weitgehend auf die Geisteswissenschaften übertragbar sind.

2.2 Standards und Richtlinien

Beim Aufbau von geisteswissenschaftlichen Forschungsdatenzentren müssen unterschiedliche äußere Einflüsse beachtet werden. Dabei sind diese Richtlinien nicht (nur) als Einschränkungen, sondern vielmehr als Hilfestellungen zu sehen, um in einem komplexen Umfeld navigieren zu können.

- **Gute Wissenschaftliche Praxis** von Forschungsförderern[12] sowie **Richtlinien** von u.a. Kompetenzzentren, Hochschulen und Fachgesellschaften.[13]
- **Lizenzen** und datenspezifische rechtliche Vorgaben.

[10] Knowledge Exchange Workshop: Making Data Count. (April 2013, Berlin).
http://www.dfg.de/dfg_magazin/aus_der_wissenschaft/archiv/knowledge_exchange_workshop_2013/%5C

[11] Knowledge Exchange: The Value of Research Data Metrics for datasets from a cultural and technical point of view. (2013)
http://www.knowledge-exchange.info/datametrics

[12] Die relevantesten Richtlinien von Forschungsförderern in Deutschland sind diejenigen der DFG und der Europäischen Kommission:
* DFG (2009): Empfehlungen zur gesicherten Aufbewahrung und Bereitstellung digitaler Forschungsprimärdaten.
http://www.dfg.de/download/pdf/foerderung/programme/lis/ua_inf_empfehlungen_200901.pdf
* Vgl. das relevante Kapitel zu Datenmanagement in der DFG-Antragsstruktur, sowie
http://www.forschungsdaten.org/index.php/F%C3%B6rderorganisationen
* European Commission: Guidelines on Data Management in Horizon 2020. Version 1.0, 11.December 2013.
http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

[13] Beispiele für Empfehlungen zum Forschungsdatenmanagement sind:
* Jens Ludwig, Harry Enke (Hrsg.): WissGrid, Leitfaden zum Forschungsdaten-Management - Handreichungen aus dem WissGrid-Projekt (2012). http://www.wissgrid.de/publikationen/Leitfaden_Data-Management-WissGrid.pdf
* Büttner, Stephan; Hobohm, Hans-Christoph und Müller, Lars: Handbuch Forschungsdatenmanagement. Bad Honnef: Bock + Herchen, 2011. http://opus.kobv.de/fhpotsdam/volltexte/2011/241/pdf/HandbuchForschungsdatenmanagement.pdf
* Ratgeber zum Forschungsdatenmanagement in den Wirtschafts- und Sozialwissenschaften - "Auffinden-Zitieren-Dokumentieren". Leibniz Informationszentrum Wirtschaft, GESIS und dem RatSWD.
http://auffinden-zitieren-dokumentieren.de/download/

- Standards zur **Struktur von digitalen Datenrepositorien**, aus organisatorischer Sicht: z.B. die ISO-Standards OAIS[14] und TRAC/TDR[15].
- **Metadaten**: Für digitale Archivierung zu beachten ist PREMIS[16]. Funktions- und disziplinspezifische Metadaten wie CIDOC-CRM[17] oder TEI[18] müssen fallweise herangezogen werden.
- **Technische Standards** (z.B. für Persistent Identifier, Single-Sign-On) werden in diesem Dokument nicht betrachtet.

Zu den spürbarsten Erfolgen von Richtlinien gehören ein gemeinsames Verständnis der Kriterien, welche beim Aufbau und Betrieb eines Forschungsdatenzentrums beachtet werden sollten, und referenzierbare Terminologien. Sie bringen Struktur in die heterogenen Bereiche Forschungsdaten und digitale Langzeitarchivierung, in denen es oft zu Missverständnissen durch unterschiedliche Terminologien und Anforderungen/Perspektiven kommt. In Kapitel 3.1 arbeiten wir daher ebenso ein Verständnis von Forschungsdaten und Archivierung aus Sicht eines HDC heraus.

Relevant in diesem Bereich ist das Ebenen-Konzept der Langzeitarchivierung, das durch das WissGrid-Projekt definiert wurde.[19] Die von WissGrid genannten Ebenen Bitstream-Preservation, Content Preservation und Data Curation (siehe Abbildung 1) werden im HDC-Antrag übersetzt als:

- **Bitstream Preservation**: Daten werden Bit für Bit bewahrt. Dies betrifft primär die Datenintegrität (z.B. Kopien sichern die Daten gegen Beschädigung und Verlust). Aufgrund der Halbwertszeit der physischen Speichermedien schließt Bitstream Preservation den regelmäßigen Transfer (z.B. alle 3 Jahre) auf aktuelle Speichermedien mit ein (Media Refreshing). Die Aufbewahrungsdauer ist üblicherweise definiert durch die Vorgaben im Projektantrag bzw. Fördervertrag und durch die Gute Wissenschaftliche Praxis (in geisteswissenschaftlichen Projekten meist 10 Jahre).[20]
- **technische Nachnutzbarkeit**: Datenformate werden für Software interpretierbar gehalten, so dass die Daten in menschenlesbarer Form zugänglich bleiben. Das bedeutet primär die regelmäßige Konvertierung der Daten in aktuelle Formate.
- **intellektuelle Nachnutzbarkeit**: Nutzer können die Daten verstehen. Damit ist nicht eine umfassende Dokumentation gemeint, so dass jeder die Daten verstehen kann, sondern, dass die Datenersteller den Kontext der Datenerstellung und die Datenstrukturen so dokumentieren, dass außenstehende Experten (in den meisten Fällen werden dies Wissenschaftler derselben Disziplin oder angrenzender Disziplinen sein) sich die Bedeutung der Daten erschließen können.

[14] Reference Model for an Open Archival Information System (OAIS) - ISO 14721:2003. CCSDS - Consultative Committee for Space Data Systems (2003). http://nost.gsfc.nasa.gov/isoas/
[15] Trustworthy Digital Repository Checklist (TDR). ISO 16363 (2012). http://www.crl.edu/archiving-preservation/digital-archives/metrics-assessing-and-certifying-0
[16] PREMIS preservation metadata. http://www.loc.gov/standards/premis/
[17] CIDOC-CRM. ISO 21127:2006. http://www.cidoc-crm.org/
[18] TEI Text Encoding Initiative. http://www.tei-c.org/
[19] WissGrid: Generische Langzeitarchivierungsarchitektur für D-Grid. Version - 14, Januar 2010. S. 48ff. http://www.wissgrid.de/publikationen/deliverables/wp3/WissGrid-D3.1-LZA-Architektur-v1.1.pdf
[20] Im HDC-Projektantrag wird für "einfache" Gute Wissenschaftliche Praxis noch eine geringe Zugriffswahrscheinlichkeit angenommen und damit die Ablage auf ein Tape als Speichermedium suggeriert. Da Speichermedien und ergo Speicherkosten primär das Geschäftsmodell betreffen, wird diese Unterscheidung an dieser Stelle nicht übernommen.

Die Analyse von unterschiedlichen Arten von Forschungsdaten in Kapitel 4 nutzt diese drei Ebenen zur Strukturierung der Anforderungen an das Forschungsdatenmanagement.

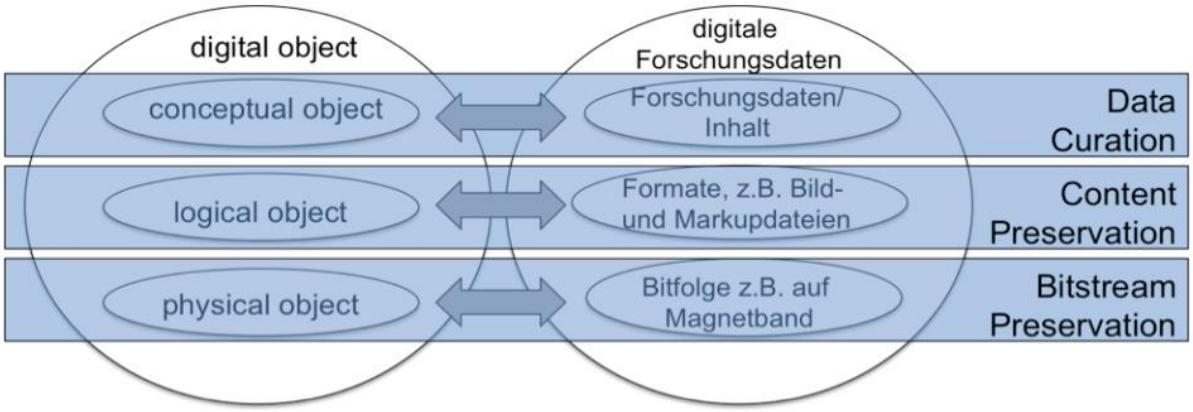


Abbildung 1[21]: Die WissGrid-Ebenen der Langzeitarchivierung behandeln unterschiedliche Aspekte digitaler Objekte bzw. Forschungsdaten.

2.3 Verwandte Initiativen

Seit einigen Jahren arbeitet eine Vielzahl von Initiativen am Aufbau von Diensten zum Forschungsdatenmanagement. Dabei hat sich der Fokus von technischen Themen inzwischen zu organisatorischen, rechtlichen und finanziellen Rahmenbedingungen verschoben, was durchaus als ein Zeichen der zunehmenden Reife der Aktivitäten gewertet werden kann. Obwohl nicht alle Initiativen genannt werden können, werden im Folgenden die für das HDC relevantesten Initiativen kurz beschrieben:

- **internationale Initiativen zum Forschungsdatenmanagement**: Digitale Infrastrukturen lancieren inzwischen vermehrt Empfehlungen zu Datenmanagement (z.B. das GRDI-Projekt[22]). Spezialisierte Organisationen im Forschungsdatenmanagement sind z.B. CODATA[23] und das ICSU World Data System[24], prominente Mitgliederorganisationen in den Naturwissenschaften. Die durch sie vernetzten Datenzentren sind lokal aufgebaut und finanziert. Die ICSU World Data Center zeigen, wie einzelne Zentren sich zu globalen Autoritäten entwickeln können. Die Zentren sind jeweils disziplinspezifisch, genießen Vertrauen in ihren Disziplinen und sind fester Bestandteil der disziplinären Forschungsprozesse.
  Die Research Data Alliance (RDA)[25], eine jüngere Initiative im Forschungsdatenmanagement, ist stärker auf Forschungsdaten und kollaborative Prozesse orientiert und hat starken politischen Rückhalt seitens Förderinstitutionen in Australien, den USA und Europa. Die RDA zeigt, wie brennend das Thema aktuell diskutiert wird, bleibt aber (derzeit) zu abstrakt für einzelne Wissenschaftler.

[21] Quelle: WissGrid 2010, S. 6.
[22] GRDI2020 - A Vision for Global Research Data Infrastructures. http://www.grdi2020.eu/
Fred Friend, Jean-Claude Guédon, Herbert Van de Sompel: Beyond Sharing and Re-using: Toward Global Data Networking. (October 2011). http://www.grdi2020.eu/pdf/Toward_Global_Data_Networking.pdf
[23] CODATA (Committee on Data for Science and Technology). http://www.codata.org/
[24] ICSU (International Council of Science) World Data System. http://www.icsu-wds.org/
[25] RDA - Research Data Alliance. https://rd-alliance.org/

- Aufbau von **Datenmanagement-Diensten in Europa**: Projekte wie EUDAT[26] bauen mit lokalen Rechenzentren ein Netzwerk aus Datenmanagement-Angeboten auf. EGI[27], eine der Vorläuferinitiativen, –die sich auf eine bestimmte Technologie (Grid) konzentriert hat – blieb zu weit von den Wissenschaftlern entfernt und ist heute kaum noch präsent.[28]
- **nationale Dateninfrastruktur-Initiativen:** Im Rahmen von D-Grid[29], dem nationalen Programm des BMBF für eine digitale Infrastruktur, errichtete das Projekt F&L-Grid (2007-2010)[30] einen Dienst für Backup und Archivierung von wissenschaftlichen Daten. Durch die Ansiedelung des Dienstes am Deutsches Forschungsnetz (DFN) hatte F&L-Grid von Beginn an eine Nachhaltigkeitsstrategie. Gleichzeitig wurde der Dienst von den Technologiepartnern nach dem Prinzip „Build it and they will come“ gebaut. Unseres Wissens bleibt das Nutzerinteresse allerdings bisher verhalten und den Webseiten des DFN ist nicht zu entnehmen, ob das Angebot überhaupt noch besteht. Aktuell errichtet das DFG-Projekt RADAR (2013-2016)[31] ein generisches Repositorium für Forschungsdaten. Das Projekt ist von der Grundstruktur her dem HDC ähnlich, daher werden die Ergebnisse für das HDC interessant sein. Außerdem ist RADAR mit Partnern aus der Chemie und Biochemie näher an den Nutzern als andere „generische“ Infrastrukturen.
- **Datenmanagement an Universitäten:** Zum Nachweis ihres wissenschaftlichen Outputs und um ihren Mitarbeitern Forschungsdatenmanagement nach Guter Wissenschaftlicher Praxis zu erleichtern, erarbeiten immer mehr Hochschulen Strategien und Richtlinien zum Forschungsdatenmanagement und richten z.T. auch Archive für Forschungsdaten ein.[32] Aufgrund der Breite der Aktivitäten sind diese Archive meist sehr generisch und ohne disziplinspezifische Angebote für die Zugänglichkeit und Nachnutzbarkeit der Daten. Es wird spannend, in der nahen Zukunft zu beobachten, ob und wie diese Archive mit Publikationsrepositorien, internationalen Infrastrukturen und disziplinspezifischen Angeboten verwoben werden können.
- **Internationale Infrastrukturen in den Geisteswissenschaften**: Die aktuell vermutlich größten Initiativen in Europa mit substanzieller politischer Unterstützung durch direkte Mitwirkung von nationalen Ministerien[33] sind DARIAH[34] und CLARIN[35]. Eine vergleichbar große Initiative in den

[26] EUDAT - European Data Infrastructure. www.eudat.eu
[27] EGI - European Grid Infrastructure. http://www.egi.eu/
[28] Obwohl es vorstellbar ist, dass in lokalen und nationalen Rechenzentren vielleicht EGI-Dienste genutzt werden.
[29] D-Grid. http://d-grid-ggmbh.de/
[30] F&L-Grid, BMBF-Projekt (2007-2010). http://d-grid-ggmbh.de/index.php?id=54&L=
[31] RADAR - Research Data Repositorium. DFG-Projekt, 2013-2016. http://www.radar-projekt.org/
[32] In Deutschland gibt es unter anderem folgende Aktivitäten an Universitäten:
* Forschungsdatenmanagement an der Universität Bielefeld, https://data.uni-bielefeld.de/
Grundsätze zu Forschungsdaten an der Universität Bielefeld (19. Juli 2011). https://data.uni-bielefeld.de/policy
Resolution zum Forschungsdatenmanagement (12. November 2013). https://data.uni-bielefeld.de/de/resolution
* Forschungsdaten-Leitlinie der Universität Göttingen (einschl. UMG) vom 01. Juli 2014. http://www.uni-goettingen.de/de/01-juli-2014-forschungsdaten-leitlinie-der-universitaet-goettingen-einschl-umg/488918.html
* Forschungsdaten-Policy der HU Berlin: Grundsätze zum Umgang mit Forschungsdaten an der Humboldt-Universität zu Berlin (8.7.2014). https://www.cms.hu-berlin.de/ueberblick/projekte/dataman/policy
Simukovic, E., Kindling, M., & Schirmbacher, P. (2014). Unveiling Research Data Stocks: A Case of Humboldt-Universität zu Berlin. In iConference 2014 Proceedings (p. 742 - 748). http://hdl.handle.net/2142/47259
* TU Berlin - Servicezentrum Forschungsdaten und -publikationen. http://www.szf.tu-berlin.de/
http://www.fu-berlin.de/sites/open_access/Veranstaltungen/oa_berlin/poster/Open-Access-an-der-TU-Berlin_Dagmar-Schobert_TU-Berlin.pdf?1412673857
[33] ESFRI - European Strategy Forum for Research Infrastructure. Europäischer Rat.
http://ec.europa.eu/research/infrastructures/index_en.cfm?pg=esfri
[34] DARIAH - Digital Research Infrastructure for the Arts and Humanities. www.dariah.eu
[35] CLARIN - Common Language Resources and Technology Infrastructure. www.clarin.eu

USA, Bamboo[36], ist an der Skepsis und fehlenden Mitwirkung der Wissenschaftler gescheitert. DARIAH und CLARIN sind hier besser aufgestellt, da sie durch ihre Partnerstruktur sowohl große internationale Projekte als auch kleine lokale Projekte aus allen Teildisziplinen einbinden. Im Hinblick auf Datenmanagement bauen DARIAH und CLARIN auf die Mitwirkung von lokalen Initiativen – so betreibt z.B. der HDC-Partner BBAW einen der Daten-Knoten in CLARIN. Insofern verhält sich das HDC-Projekt komplementär zu europäischen Infrastrukturen wie DARIAH und CLARIN.

- **Disziplinäre und sammlungsspezifische Forschungsdatenzentren**: Es gibt eine Vielzahl[37] von Forschungsdatenzentren bezogen auf spezielle Datensammlungen, die meist aus wissenschaftlichen Projekten oder an Institutionen gewachsen sind und (teilweise) verstetigt wurden. Diese Forschungsdatenzentren sind aber meist nur einem eingeschränkten Nutzerkreis zugänglich bzw. verfolgen strenge Selektionskriterien für die Aufnahme von Daten. Projekte wie IANUS[38] (Archäologie und Altertumswissenschaften) haben dabei durchaus das Potenzial, in ihrer Disziplin zu autoritativen Zentren zu werden und die Kriterien eines World Data Center zu erfüllen. Dazu müssen sie nachhaltige Strukturen aufbauen, sich der gesamten Community öffnen[39] und umgekehrt in ihrer Community weithin akzeptiert sein.

Trotz der Vielzahl an bereits existierenden Forschungsdatenmanagement-Initiativen gibt es nur wenige Initiativen mit der Ausrichtung des HDC, basierend auf einem nachhaltigen Geschäftsmodell und offen für alle Arten von geisteswissenschaftlichen Forschungsdaten.

Eines der Vorbilder ist DANS[40] in den Niederlanden, das sowohl ein Portal zur Archivierung und Nachnutzung von Forschungsdaten anbietet, als auch direkt Datenmanagement-Aufgaben im Rahmen von Forschungsprojekten übernimmt. Als nationales Angebot in den Niederlanden ist DANS stark diversifiziert und kämpft an vielen Fronten gleichzeitig – vor allem die direkte Beratung skaliert hierbei nicht beliebig.

Im deutschsprachigen Raum ist GAMS[41] vom Zentrum für Informationsmodellierung der Universität Graz zu nennen, das sukzessive existierende Datensammlungen an der Universität in das GAMS-Repositorium integriert. Dabei werden die Datenstrukturen und Metadaten in GAMS neu modelliert und die Forschungsdaten über GAMS online angeboten.

Das DCH der Universität Köln[42] geht darüber hinaus, indem es zusätzlich zu einem übergreifenden Repositorium zur Datenablage auch davon ausgeht, dass teilweise spezifische Präsentationssysteme und Werkzeuge notwendig sein werden, um unterschiedliche Typen von Forschungsdaten adäquat zugänglich machen zu können.[43] Es arbeitet dazu eng mit dem Rechenzentrum der Kölner Universität zusammen, ist aber noch eine vergleichsweise junge Initiative (gegründet 2012).

[36] Project Bamboo (2008-2012). www.projectbamboo.org
[37] siehe re3data.org - Registry of Research Data Repositories
[38] IANUS - Forschungsdatenzentrum Archäologie und Altertumswissenschaften. http://www.ianus-fdz.de/
[39] Speziell in der Archäologie bedeutet das u.a. die Beantwortung von rechtlichen Fragestellungen, die einen Austausch von Forschungsdaten teilweise über Ländergrenzen hinweg verbieten oder an bestimmte Lizenzen gebunden sind.
[40] DANS - Data Archiving and Networked Services. http://www.dans.knaw.nl/en
[41] GAMS - Geisteswissenschaftliches Asset Management System. http://gams.uni-graz.at/
[42] Data Center for the Humanities - Universität zu Köln. http://dch.phil-fak.uni-koeln.de/startseite.html?&L=0
[43] 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. http://libreas.eu/ausgabe23/09sahle/

Alle genannten Einrichtungen – DANS, GAMS und DCH – ...

- ... stehen in direktem Kontakt mit den Erzeugern der Forschungsdaten, und zwar vom Zeitpunkt der Erstellung der Forschungsdaten an bis zur Übergabe der Daten an das Archiv sowie auch in gemeinsamen Forschungsprojekten zur Evaluierung der Nutzeranforderungen und der Erweiterung des Angebots.
- ... haben das Vertrauen ihrer Nutzer, weil ihre Angebote für diese einen Mehrwert bringen und weil sie zudem ein nachhaltiges Geschäftsmodell haben.
- ... orientieren sich in ihren Zielen und ihrer Kultur an den Wissenschaften. Sie verfolgen ständige Innovation, wurzeln dabei aber in einer stabilen technischen Infrastruktur (ein lokales Rechenzentrum bzw. DANS in der nationalen Infrastruktur).

Das DCH, GAMS und weitere Initiativen kommen in der DHd-Arbeitsgruppe[44] „Aufbau von Datenzentren“ zusammen, um gemeinsame Ausrichtungen, Terminologien und Standards zu diskutieren – womit diese ein besonders relevantes Forum für das HDC darstellt.

[44] DHd - Digital Humanities im deutschsprachigen Raum. Arbeitsgruppen. http://www.dig-hum.de/dhd-ags

3 Ziele eines HDC

Ein HDC sollte es ermöglichen, dass geisteswissenschaftliche Forschungsdaten über die üblichen Lebenszyklen von Datenformaten und Software hinaus nutzbar sowie über die Anwesenheit der ursprünglichen Datenersteller hinaus interpretierbar bleiben. Nach der Übergabe der Forschungsdaten eines Projekts an ein HDC sollten die Forschungsdaten zumindest für den vom Projektförderer geforderten Zeitraum und idealerweise so lange wie möglich darüber hinaus verfügbar und nachnutzbar bleiben. Die Ziele eines HDC sind wissenschaftsgetrieben, d.h. im Vordergrund steht die Ermöglichung der Nachnutzung von Forschungsdaten durch die Wissenschaft.

Zur Unterstützung der Interpretation dieses übergreifenden Ziels betrachten die folgenden Kapitel das zugrunde liegende Verständnis von Forschungsdaten, die Interessen der Akteure und die Erfolgsfaktoren für die Archivierung und Nachnutzung geisteswissenschaftlicher Forschungsdaten in detaillierter Weise.

3.1 Begriffsklärung: Forschungsdaten

Forschungsdaten können alle (digitalen) Artefakte sein, die während des Forschungsprozesses entstehen. Sie können in unterschiedlichen Datenformaten und Aggregationsstufen vorliegen. Artefakte können neben Daten auch Funktionen auf den Daten bzw. komplette Anwendungen oder Benutzeroberflächen sein.[45]

Forschungsdaten, die an ein Datenzentrum übergeben werden, haben einen **langfristigen Wert** – entweder als Ergebnisse der Forschung oder als wesentliche Zwischenschritte, die den Forschungsprozess dokumentieren.[46] „Langfristig“ bedeutet dabei zumindest für die Dauer der relevanten Guten Wissenschaftlichen Praxis[47], potenziell natürlich unbefristet.

Ein Forschungsdatenzentrum trägt aktiv dazu bei, dass Forschungsdaten weiter **in den Wissenschaftszyklus eingebunden** sind. Das bedeutet, dass sie zugänglich[48], referenzierbar und

[45] Bspw. um Datenbanken für Zugänglichkeit zu öffnen, oder im Fall von interaktiven Anwendungen. In der Definition der Significant Properties betrifft dies die Dimensionen „Rendering“ und „Behaviour“.

[46] Hierbei ist zu beachten: Der "langfristige Wert" von Forschungsdaten schließt nicht unbedingt die Bewertung ihrer "Qualität" mit ein. Wahrnehmung von Qualität ist oft abhängig von Forschungsmethode, wissenschaftlicher Schule und aktuellen (technischen) Möglichkeiten. Auch z.B. unvollständige Zwischenschritte können aber für die Dokumentation eines wissenschaftlichen Prozesses essenziell sein.

[47] Umgekehrt bedeutet das: Löschung der Forschungsdaten nach einer bestimmten Aufbewahrungsdauer – z.B. den in den Richtlinien der Guten Wissenschaftlichen Praxis vorgeschriebenen 10 Jahren – ist denkbar und eventuell nicht ungewöhnlich.

[48] Zugänglichkeit bedeutet hier, dass der Inhalt der Forschungsdaten möglichst direkt einsehbar ist. Je nach Significant Properties ist die Zugänglichkeit von Fall zu Fall zu bewerten: So kann bspw. eine statische Video-Sequenz einer interaktiven Visualisierung zusammen mit ausführlicher Dokumentation aus fachwissenschaftlicher Sicht für den Zweck der Aufbewahrung hinreichend zugänglich sein; nicht zugänglich hingegen ist der Abzug (Dump) einer SQL-Datenbank, der ohne die Installation einer (speziellen) Datenbanksoftware nicht zu entschlüsseln ist.

vernetzt[49], suchbar/auffindbar[50] sowie nachnutzbar[51] sind. Über längere Zeiträume werden in allen diesen Aspekten Migrationsschritte als Maßnahmen der digitalen Erhaltung[52] notwendig sein.

Solange Forschungsdaten im Datenzentrum registriert sind, sind sie **eingefroren**. Speziell bei interaktiven Anwendungen können temporäre oder administrative Daten zwar in der Anwendung abgespeichert werden, aber sie sollen nicht durch neue Inhalte ergänzt und verändert werden. Dies ist sowohl aus fachwissenschaftlicher Sicht (Zitierbarkeit), als auch aus Gründen der technischen Aufbewahrung (z.B. Sicherung der Datenintegrität durch redundante Kopien) gefordert.[53]

Für ihre Interpretierbarkeit müssen Forschungsdaten mit **Metadaten und Dokumentation** verknüpft sein. Aus fachwissenschaftlicher Sicht sind das u.a. Informationen über Forschungsfrage, -methode und Entstehungskontext.

Im Allgemeinen wird ein Forschungsdatenzentrum **keinen Einfluss auf (wissenschaftliche) Inhalte** nehmen, aber eng mit Wissenschaftlern zusammenarbeiten bzw. sich auf externe Bewertungen verlassen. Wo doch Einfluss auf Inhalte genommen wird, sollte dies transparent dokumentiert und begründet werden.[54] Detailliertere Empfehlungen zu Erstellung und Qualität von Forschungsdaten sollten daher auch vorzugsweise nicht durch ein HDC, sondern direkt durch Institutionen, disziplinspezifische Fachgesellschaften oder andere wissenschaftliche Autoritäten gegeben werden.[55]

3.2 Selbstverständnis des Forschungsdatenzentrums

Die übergreifende Ausrichtung eines HDC wurde in der Einleitung und in Kapitel 3.1 bereits angesprochen und wird in den folgenden Kapiteln noch detaillierter beschrieben. Aufgrund der Bandbreite der Geisteswissenschaften und der Anforderungen bzw. Nachfrage aus den unterschiedlichen Disziplinen können sich HDCs aber in Bezug auf ihren Entstehungskontext unterscheiden: hinsichtlich der Zielgruppe (disziplinspezifisch, institutionell, regional oder offen), auf Basis ihrer Angebote und der Abdeckung unterschiedlicher Forschungsdatentypen sowie bezüglich ihres aktiven Beitrags zur geisteswissenschaftlichen Forschung. Zum besseren Verständnis, was ein HDC ist bzw. was ein HDC nicht ist, wird in diesem Abschnitt daher die Abgrenzung zu verwandten Einrichtungen herausgearbeitet.

Ein HDC ist kein bloßes **Archiv** in dem Sinn, dass Daten lediglich als Backup „im Keller eines Rechenzentrums“ liegen. Eine Backup-Funktion kann Teil der technischen Nachhaltigkeit sein.

[49] Referenzierbar z.B. für Zitate in wissenschaftlichen Publikationen; vernetzt z.B. mit (externen) Normdatenkatalogen.
[50] Auffindbarkeit kann aufgrund der Beschaffenheit der Daten auch spezialisierte Suchfunktionen erfordern; eine reine Suche in den Metadaten oder auch eine Google-ähnliche Volltextsuche kann unzureichend sein.
[51] Z.B. in aktuellen Virtuellen Forschungsumgebungen außerhalb des Forschungsdatenzentrums
[52] nestor Handbuch - Eine kleine Enzyklopädie der digitalen Langzeitarchivierung. Version 2.3 - 2010. Kapitel 8: Digitale Erhaltungsstrategien. http://nestor.sub.uni-goettingen.de/handbuch/
[53] Zur Illustration des Begriffs Einfrieren und seiner Notwendigkeit ein einfaches Beispiel: eine Anwendung, die über eine Schnittstelle Daten in Echtzeit aggregiert und auswertet, bspw. Twitter oder Börsenkurse, wird zum Übergabezeitpunkt angehalten/ eingefroren.
[54] Gemeinsam mit Wissenschaftlern ist unter anderem zu bewerten: (a) welche Forschungsdaten aufgenommen werden sollen/können, (b) ob Metadaten und Dokumentation ausreichend sind, (c) ob die Significant Properties adäquat definiert sind – sowie zu einem späteren Zeitpunkt (d) ob Migrationsschritte trotz eines dabei eventuell eintretenden Informationsverlustes durchgeführt werden können und (e) ob die Forschungsdaten nach der Aufbewahrungsdauer gelöscht werden können oder ob es eine Folgefinanzierung gibt. Siehe dazu Kapitel 7 (Abläufe).
[55] Siehe dazu auch Kapitel 2.2 (Projektumfeld. Standards und Richtlinien).

Entscheidend aber ist, dass darüber hinaus ein HDC die Forschungsdaten weiterhin in den Wissenschaftszyklus einbindet, so dass sie auffindbar, zugänglich und nachnutzbar sind. Üblicherweise verstehen sich aktuelle Projekte der Langzeitarchivierung bzw. Langzeitverfügbarkeit in diesem Sinn und mangels einer autoritativen Definition der Begriffe, werden sie im Zusammenhang mit dem Aufbau eines HDC auch in dieser Weise benutzt.

Ein HDC ist eng mit der Forschung verknüpft und wird bei wesentlichen Entscheidungen über Forschungsdaten durch Geisteswissenschaftler beraten (z.B. bei Migration, bei Löschung). Ein HDC ist jedoch **kein DH-Zentrum**, da es sich üblicherweise nicht (über die Aufgaben eines HDC hinaus, zu denen durchaus auch die Schulung gehört) in der DH-Lehre engagiert, DH-Forschungsprojekte anstrengt oder vermittelt oder Virtuelle Forschungsumgebungen aufbaut und hostet.

Ein HDC bietet einen Dienst für Forschende in den Geisteswissenschaften und erstellt **keine eigenen Forschungsdaten**. In diesem Sinne betreibt ein HDC keine eigene DH-Forschung, obwohl es gemeinsam mit Forschern aktiv seine Angebote weiterentwickelt (auch im Rahmen von Forschungsprojekten).

Natürlich kann z.B. eine Universität ein HDC zusammen mit einem DH-Zentrum unter einem Dach einrichten und die Aufgaben der beiden stark miteinander verknüpfen. Es ist jedoch nicht a priori sicher, ob sich tatsächlich Synergien und personelle Überlappungen zwischen den unterschiedlichen Aufgaben und Verantwortungen ergeben.

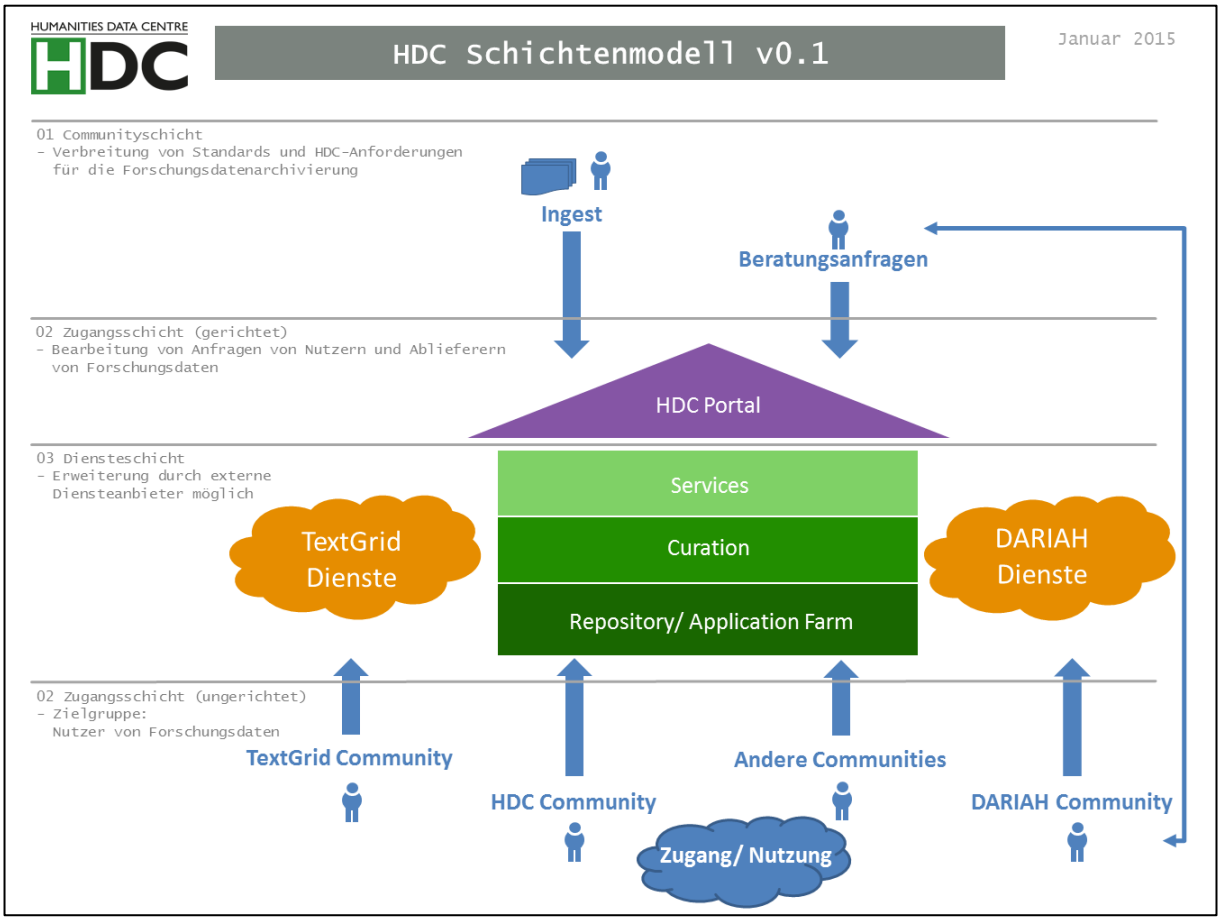


Abbildung 2: Die Interaktionskategorien eines HDC, bspw. das Abliefern von Forschungsdaten, die Abwicklung von Beratungsanfragen aus Forschungsprojekten oder die Nutzung von externen Diensten über das HDC als Hub.

3.3 Akteure

a. **Forschungsförderer**:

Ein großer Teil wissenschaftlicher Forschungsprojekte ist abhängig von der Finanzierung durch Forschungsförderer. In den Geisteswissenschaften dürfte der Prozentsatz an Projekten, auf die dies zutrifft, besonders hoch sein, da hier der Anteil an Auftragsforschung, bspw. für Industrie oder Wirtschaft, gering ist. Als Geldgeber haben Forschungsförderer ein Interesse daran, dass die von ihnen bereitgestellten Mittel auf sinnvolle, effiziente und nachhaltige Weise genutzt werden. Dazu gehört, dass Projektergebnisse zur Sicherung der **Qualität von Forschung** auch nach Projektende nachvollziehbar und verifizierbar bleiben sollen. Für die **Effizienz und Nachhaltigkeit der Forschungsförderung** sollen Projektergebnisse außerdem in einer Form langfristig erhalten bleiben, in der sie durch Folgeprojekte und durch die interessierte Öffentlichkeit nachgenutzt werden können.

Forschungsförderer fordern vermehrt die Aufbewahrung und öffentliche Zugänglichkeit von (mit öffentlichen Geldern finanzierten) Forschungsergebnissen über einen disziplinbezogenen Mindestzeitraum, üblicherweise 10 Jahre.[56] Zunehmend sollen dazu als Teil von Projektanträgen auch Datenmanagementpläne eingereicht werden, deren Reichweite aber durch die Projektdauer begrenzt ist. Eine „Datenpauschale“ (Einmalzahlung für die adäquate Aufbewahrung von Daten über den disziplinbezogenen Mindestzeitraum hinweg) ist bei unterschiedlichen Förderern in der Diskussion.

b. **Forschende in den Geisteswissenschaften**: von universitären und außeruniversitären Forschern über institutionsübergreifende Forschungsprojekte bis hin zu institutionsungebundenen Einzelprojekten (citizen science)

Forscher suchen Lösungen zur **vertrauenswürdigen Aufbewahrung** ihrer Projektergebnisse nach den Vorgaben von Projektförderern und den Regeln Guter Wissenschaftlicher Praxis, die ihnen selbst möglichst **wenig Mehraufwand und keine Mehrkosten** abverlangen.

Forscher wünschen sich **Sichtbarkeit ihrer Forschungsergebnisse** in Form von Daten und Software, ergänzend zu gängigen wissenschaftlichen Publikationen wie Zeitschriftenartikeln und Buchkapiteln. Obwohl die zitierbare „Publikation von Forschungsdaten“ (noch) nicht Teil der Wissenschaftstradition ist, wird der Bedarf danach zunehmend sichtbarer.[57] In diesem Zusammenhang stellt sich auch die Frage, ob sich ein „Impact Factor“ für Forschungsdaten etablieren kann oder ob alternativ Paper- oder Buchpublikationen mit vernetzten Forschungsdaten eine relevante Erhöhung ihres Impact Factors bekommen können.

Der **Wissenschaftliche Diskurs** auf Basis der Forschungsdaten – von der Diskussion von Zwischenschritten bis hin zur Verifikation von Forschungsergebnissen – hat positive Effekte für die

[56] European Commission: Guidelines on Data Management in Horizon 2020. Version 1.0, 11.December 2013. http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf
DFG (2009): Empfehlungen zur gesicherten Aufbewahrung und Bereitstellung digitaler Forschungsprimärdaten. http://www.dfg.de/download/pdf/foerderung/programme/lis/ua_inf_empfehlungen_200901.pdf
Vergleiche auch das relevante Kapitel zu Datenmanagement in der DFG Antragsstruktur, sowie http://www.forschungsdaten.org/index.php/F%C3%B6rderorganisationen

[57] Siehe z.B. Rettet die Wissenschaft: Die Folgekosten können hoch sein. Interview mit Jürgen Zöllner. In: Die Zeit, 01/2014 (5. Januar 2014). http://www.zeit.de/2014/01/wissenschaft-zoellner-forschung-transparenz/komplettansicht

Forschung. Er dient der Qualitätssicherung, der Vernetzung mit verwandten Projekten sowie der Generierung von Folgefragen und Folgeprojekten. Gleichzeitig muss es einen „privaten Raum“ geben und der Forscher muss die **Kontrolle** über Zeitpunkt und Form der Publikation behalten.[58]

Forscher suchen relevante Forschungsdaten und (Analyse-)Werkzeuge **zur Nachnutzung**. Die Nachnutzung kann mehrere positive Effekte haben:

(I) Obwohl die Forschenden auch wenn nachnutzbare Forschungsdaten vorhanden sind noch viel Energie in die Erschließung dieser Quellen investieren werden müssen (z.B. um deren Passgenauigkeit auf ihre Forschungsfrage sowie die Qualität der Daten zu prüfen und sie ggf. nachzubearbeiten), können sie einige Schritte abkürzen (z.B. Digitalisierung, Transkription) und damit **wertvolle Projektzeit sparen**.

(II) Vorhandene Forschungsdaten, idealerweise in Verbindung mit Analysewerkzeugen, ermöglichen bei bestimmten Forschungsfragen die Durchführung von begrenzten Prototypen. Solche Demonstratoren können mit vergleichsweise geringem Aufwand Annahmen festigen und Trends darstellen, die zwar für die abschließende Beantwortung der Forschungsfrage unzureichend sind, aber das **Projektrisiko senken** können.

(III) Die Vernetzung und Analyse einer großen Menge an Forschungsdaten ermöglicht die Bearbeitung **neuer Forschungsfragen** (im Geiste der Big Data-Bewegung), die ein Projekt alleine aufgrund des Aufwands der Quellenerschließung nicht übernehmen könnte.

c. **Institutionen:** Universitäten, außeruniversitäre Forschungseinrichtungen sowie disziplinspezifische **Fachgesellschaften**

Die verschiedenen wissenschaftlichen Institutionen bündeln eine große Vielfalt von Forschungsaktivitäten und beginnen zunehmend, analog zum inzwischen verbreiteten Aufbau von Publikationsrepositorien, selbst Forschungsdatenzentren aufzubauen. Ihre Motive sind dabei vielfältig und bewegen sich parallel zu den Interessen der Forscher.

Ein Forschungsdatenzentrum ist ein Mittel, um den wissenschaftlichen Output einer Einrichtung nachzuweisen und sichtbar zu machen. Ziel ist dabei, die **Attraktivität** der Forschungseinrichtung für mögliche Mitarbeiter, Partner und Drittmittelförderer zu erhöhen.

Zusammen mit der Sichtbarkeit von Forschungsergebnissen ist es für den Ruf der Einrichtung essenziell, dass die Forschung entlang der Leitlinien **Guter Wissenschaftlicher Praxis** durchgeführt und dokumentiert wird. Obwohl ein Forschungsdatenzentrum nicht die Qualität der Forschungsinhalte sichern kann oder will, ist es doch ein mögliches Werkzeug, um Forscher bei der Einhaltung von Qualitätskriterien in Bezug auf das Datenmanagement zu unterstützen und die Daten zur Überprüfung vorzuhalten.

Dabei soll ein Forschungsdatenzentrum der **Effizienz** dienen und den Forschern die mit dem Datenmanagement verbundene Arbeit soweit wie möglich abnehmen, indem es die mit dem Datenmanagement verbundenen (administrativen) Aufgaben und Kompetenzen zentral bündelt.

[58] Andrew Treloar, Cathrine Harboe-Ree: Data management and the curation continuum: How the Monash experience is informing repository relationships. VALA 2008. http://www.valaconf.org.au/vala2008/papers2008/111_Treloar_Final.pdf

Dabei geht man davon aus, dass sich sowohl durch die Beauftragung von Datenkuratoren mit spezialisierten Kompetenzen, als auch durch die Errichtung von skalierbaren Datenmanagement-Diensten der Gesamtaufwand über alle Projekte hinweg substanziell verringert.

d. **Rechenzentrum / Datenzentrum:** als Heimateinrichtung oder Infrastrukturanbieter für ein HDC

Viele Angebote eines HDC betreffen den Aufbau einer stabilen technischen Infrastruktur und der dafür notwendigen Organisationsstruktur. Wenn die Ressourcen, Kompetenzen und Strukturen eines Rechenzentrums für das HDC nicht neu aufgebaut werden sollen, liegt die Ansiedlung (von Teilen) eines HDC an einem Rechenzentrum nahe.

Ein Rechenzentrum sucht generell kontinuierlich nach **neuen Zielgruppen** und Angeboten und kann durch ein HDC-Angebot im Wettbewerb mit anderen Rechenzentren hervorstechen.

Dabei achtet ein Rechenzentrum auf die Einbettung in vorhandene Ressourcen und Strukturen sowie auf die Skalierbarkeit der Dienste, um ein möglichst **nachhaltiges Geschäftsmodell** für das HDC aufzubauen.

3.4 Erfolgsfaktoren

Neben den Grundfunktionen eines HDC lassen sich aus den Beschreibungen der Akteure einige potenzielle Mehrwerte ableiten, die zum Erfolg eines HDC beitragen können[59]. Diese sind idealtypisch formuliert:

- wenig Mehraufwand (Zeit) und keine Mehrkosten (finanzielle Ressourcen) für die Nutzer, was – natürlich zusammen mit dem Mehrwert – den Nutzern gut kommuniziert werden muss
- ein Angebot wertvoller Daten, so dass das Datenzentrum von Beginn an eine hohe Visibilität der Inhalte leistet
- Vernetzung mit Infrastrukturen, Katalogen und Virtuellen Forschungsumgebungen, sodass Forscher mit einer Registrierung im HDC gleichzeitig eine hohe Verbreitung, Visibilität und Nachnutzbarkeit ihrer Forschungsdaten erreichen können (z.B. Nachweis der Inhalte in der DARIAH Collection Registry, einfache Erstellung von Enhanced Publications[60] in Publikationsrepositorien)

Dazu zählen weiterhin "weiche" und nur mittelbar erzeugbare Faktoren, die nichtsdestotrotz entscheidend zum Erfolg eines HDC beitragen:

- ein Umfeld, das die Publikation von Forschungsdaten fordert und fördert[61]
- das Vertrauen der Nutzer in die Nachhaltigkeit des HDC

[59] Nach Kano werden hier also die Begeisterungsmerkmale aufgezählt, nicht die Basismerkmale (z.B. Datenspeicher) und nicht die Leistungsmerkmale (z.B. Metadatensuche, langfristige Finanzierung). http://de.wikipedia.org/wiki/Kano-Modell

[60] OpenAire: What is an Enhanced Publication? (Website, Updated on 09 September 2013) https://www.openaire.eu/en/component/content/article/76-highlights/344-a-short-introduction-to-enhanced-publications

[61] Vgl. z.B. die in Kapitel 2.1 beschriebenen Aktivitäten der DFG in Knowledge-Exchange: Knowledge Exchange Workshop: Making Data Count. (April 2013, Berlin). http://www.dfg.de/dfg_magazin/aus_der_wissenschaft/archiv/knowledge_exchange_workshop_2013/%5C

- kontinuierliche Innovation (parallel zur Entwicklung in den Geisteswissenschaften und Digital Humanities)
- Standardisierungsdruck, sowohl durch Vorgaben von Förderern (weniger technische Vorgaben, als vielmehr Anforderungen für Prozesse und Garantien, bei gleichzeitiger Anerkennung von HDCs zur Durchführung dieser Prozesse und Garantien), als auch durch sanfte Anreize für HDC-Nutzer (durch Zusatzangebote für z.B. Werkzeuge und Interoperabilität)

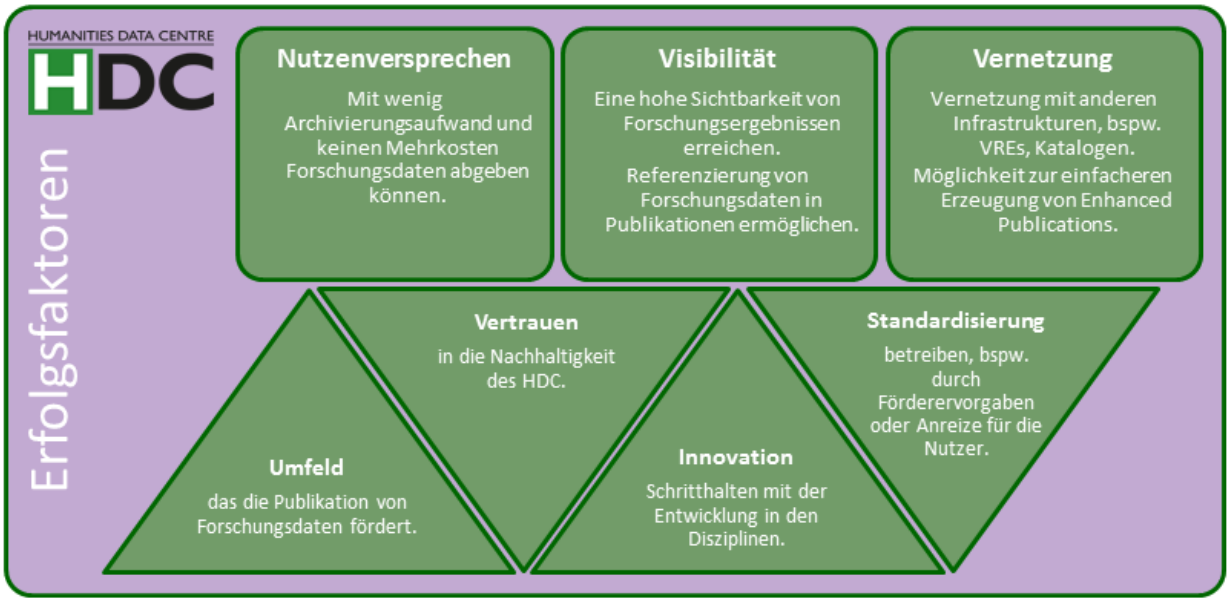


Abbildung 3: Nutzerrelevante Erfolgsfaktoren eines HDC

Die HDC-Projektpartner bringen dabei einige wertvolle Erfolgsfaktoren bereits mit – im Sinne des Lean Canvas[62] handelt es sich dabei um herausstechende *Advantages* des Konsortiums:

1. Die Partner besitzen umfangreiche Sammlungen an wertvollen Daten, die als „Erstausstattung“ des HDC bereits hohe Sichtbarkeit bedeuten können.
2. Die Partner haben in anderen Projekten (z.B. DARIAH) bereits gemeinsam Grundstrukturen einer Dateninfrastruktur aufgebaut und besitzen daher eine technische Basis und das grundlegende Vertrauen von Nutzern.
3. Die Partner sind institutionell divers und bilden bereits selbst ein Netzwerk aus institutionellen Datenkuratoren, die als Vorbild für weitere Institutionen gelten können.

[62] Ash Maurya: Why Lean Canvas vs. Business Model Canvas? Blog, 27 February 2012. http://practicetrumpstheory.com/why-lean-canvas/

4 Vom Datenmodell zu Forschungsdatentypen

Für Publikationsrepositorien der Open Access-Bewegung galt ursprünglich meist ein einfaches Datenmodell, in dem eine Datei einmalig in das Repository geladen und mit Metadaten verknüpft wurde. Die intellektuelle Nachhaltigkeit dieser Daten beruht dabei vor allem auf dem Daten- und Metadatenformat. Obwohl wissenschaftliche Forschungsdatenzentren (mitunter auch) Erben dieser Community sind, **unterscheiden sich Forschungsdaten** doch in vielen Fällen in wesentlichen Punkten von Publikationen:

a) Forschungsdaten sind meist Sammlungen von Objekten unterschiedlicher Formate. Forschungsdatenzentren müssen zusätzlich zu den Daten selbst auch die Beziehungen der Objekte zueinander abbilden.
b) Forschungsdaten müssen nicht per se dateibasiert sein. Ein wachsender Anteil geisteswissenschaftlicher Daten basiert z.B. auf Datenbanken oder interaktiven Visualisierungen. Deren Abbildung auf Dateien geht meist mit Informationsverlust einher.[63]
c) Forschungsdaten sind oft abhängig von ihrer Umgebung (s.o.): inhaltlich im Hinblick auf die verwendeten Terminologien und Strukturen sowie funktional, z.B. durch die Nutzung von externen Diensten.
d) Die signifikanten Eigenschaften (*significant properties*) von Forschungsdaten sind individuell abhängig von Forschungsfrage, -methode und -projekt. Nur ein administrativer Kern wird über alle Sammlungsstücke eines Forschungsdatenzentrums hinweg standardisierbar sein.

HUMANITIES DATA CENTRE HDC **Prämissen für das Datenmodell**

	Publikationen	Forschungsdaten
Repräsentation	Text/ verknüpfte Objekte Dateibasiert	Objektsammlung Datei/ nicht Dateibasiert (bspw. Datenbanken/ Visualisierungen)
Inhalte	Inhalte statisch, objektbasiert und eindimensional abbildbar	Inhalte können interaktiv und mehrdimensional sein, Objektbeziehungen können relevant sein
Formate	Formate für Inhalte und Metadaten gut standardisierbar	Formate stark divers nur ein administrativer Kern (Metadaten) ist übergreifend gut standardisierbar
Signifikante Eigenschaften	signifikante Eigenschaften eingrenzbar (Text, Erscheinung des Textes, Kontext)	signifikante Eigenschaften stark abhängig von Forschungsfrage, -methode und –projekt
Migration	Migration i.d.R. mit überschaubarem Informationsverlust und Aufwand möglich	Migration nur mit schwer kalkulierbarem Informationsverlust und bisweilen erheblichem Aufwand möglich

Abbildung 4: Vom normalisierten, zentralen Datenmodell zu adaptierbaren, vernetzten Forschungsdatentypen. Anhand eines Vergleiches von Merkmalen von textbasierten Publikationen und Forschungsdaten wird deutlich, welche Herausforderungen sich bei der Konzeption eines Datenmodells für Forschungsdaten stellen.

Gleichzeitig definiert Kapitel 3.1 als **wissenschaftliche Anforderungen**, dass Forschungsdaten im Wissenschaftszyklus eingebunden, wissenschaftlich zitierbar, zugänglich und möglichst direkt einsehbar sein sollen. Diese Nachnutzbarkeit und Zugänglichkeit ist gerade bei Forschungsdaten

[63] Mit anderen Worten: die Aspekte *Appearence* und *Behaviour* (siehe Kap. 4.1.1) der signifikanten Eigenschaften gehen verloren; die Daten sind nicht mehr unmittelbar so nutzbar, wie es die Ersteller ursprünglich geplant hatten.

durch die oben genannten Aspekte i.d.R. nicht durch einen standardisierten Ablageprozess zu erreichen.[64]

So wie viele geisteswissenschaftliche Daten nicht direkt in gängigen dateiorientierten Repositorien abgebildet und durch großflächig automatisierbare Prozesse gesammelt werden können, so unterscheidet sich auch ein HDC in seiner Struktur und in seinen Angeboten von den gängigen Publikationsrepositorien. Analog geht ein HDC auch **über typische Angebote von *generischen* Forschungsdatenzentren** hinaus: Die WissGrid Service-Level[65] für generische Forschungsdatenzentren gehen ursprünglich von dateibasierten Forschungsdaten aus. Die erst auf der Ebene „intellektuelle Nachnutzbarkeit" eingeschriebenen Aspekte zur Interpretierbarkeit der Forschungsdaten und des Kontextes sind aus wissenschaftlicher Sicht eine Kernanforderung, ohne die selbst – auch an sich langfristig– archivierte Daten gegebenenfalls nutzlos sind.

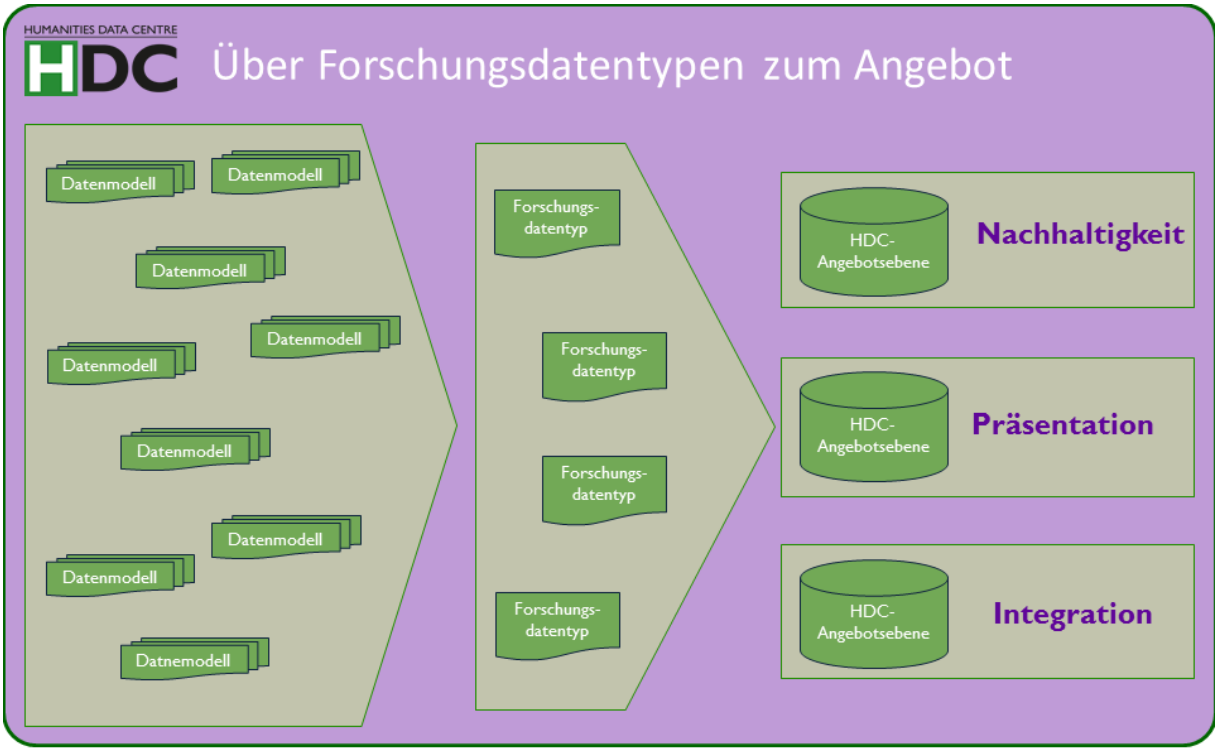


Abbildung 5: Dargestellt ist das Verfahren der Angebotsgestaltung für ein HDC. Aus der Vielzahl der gebräuchlichen Forschungsdatenmodelle werden über Auswahl und Zusammenfassung verschiedene Forschungsdatentypen definiert (z.B. Editionen, Visualisierungen, Datenbanken). Auf Grundlage dieser Forschungsdatentypen, die auf die spezifischen Kompetenzen der HDC-Partner ausgerichtet sind, werden konkrete Angebote konzipiert. Bei einem HDC teilen sich diese auf drei grundlegende Serviceebenen auf.

Angebote für ein HDC orientieren sich daher weg von einem normalisierten, dateiorientierten Datenmodell (mit globalen Standards) hin zu einem Modell, das **wissenschaftliche Abläufe und Forschungsmethoden** (mit individuell zugeschnittenen Angeboten) einbezieht. Obwohl dies für Wissenschaftler offensichtlich ist, ist es im Hinblick auf die **Skalierbarkeit** eines HDC problematisch. Eine individuelle Betreuung von Projekten und die Anpassung der technischen Dienste an spezifische Anforderungen (von wissenschaftlichen Disziplinen oder Methoden) sind zwar rein technisch und

[64] Während z.B. eine PDF-Datei automatisiert in einem Publikationsrepository registriert und anschließend in der aktuellen Anwendungsumgebung des Nutzers dargestellt werden kann, ist dies bei Datenbanken oder bei interaktiven Webanwendungen nicht der Fall.
[65] Vgl. Kapitel 2.2.

organisatorisch machbar, aber auf Grund der damit verbundenen Kosten nur schwer in ein nachhaltiges Geschäftsmodell integrierbar.

Als Abstufung zur völlig individuellen Behandlung von Forschungsprojekten entwickelt ein HDC spezialisierte Angebote für Cluster von Forschungsdaten (**Forschungsdatentypen**, FT). Dieses Vorgehen baut auf der Erfahrung auf[66], dass **wiederkehrende wissenschaftliche Methoden und Projektkontexte zu wiederkehrenden Mustern in den Forschungsdaten führen**. Beispielhaft dargestellt: Ähnlich dem DFG-Viewer[67], der eine Art Lingua Franca für Digitalisierungsprojekte darstellt, könnten *-Viewer für weitere Forschungsdatentypen geschaffen werden. Beispiele und eine Initialmenge an Forschungsdatentypen finden sich in den folgenden Unterkapiteln. Weitere Ableitungen aus dieser Herangehensweise sind:

- Die Angebote eines geisteswissenschaftlichen Forschungsdatenzentrums sind forschungsdatentypenspezifisch, das heißt für jeden Forschungsdatentyp müssen individuell unterschiedliche Dienste und Abläufe entwickelt werden. Geisteswissenschaftliche Forschungsdatenzentren können sich durch die Spezialisierung auf unterschiedliche Forschungsdatentypen unterscheiden.
- Daraus folgt, dass es nicht ein geisteswissenschaftliches Forschungsdatenzentrum geben sollte, sondern eine Reihe von Forschungsdatenzentren mit unterschiedlichen Spezialisierungen über ein generisches Grundangebot hinaus. Auf einem bestimmten Forschungsdatentyp basierende Angebote müssen nicht über die Forschungdatenzentren übergreifend standardisiert werden, allerdings wird vermutlich Konvergenz auf Grund von Angebot und Nachfrage entstehen. Mit anderen Worten: Es wird nicht „ein“ oder „das perfekte“ Angebot für einen Forschungsdatentyp geben, sondern die unterschiedlichen geisteswissenschaftlichen Datenzentren werden auch individuelle Angebote haben.
- Obwohl sich die forschungsdatentypischen Angebote gemeinsam mit den Forschungsmethoden entwickeln müssen, ist davon auszugehen, dass sie sich weniger schnell entwickeln werden als zum Beispiel spezielle Datenformate oder Software. Sie bieten daher einen hinreichend stabilen Container für das Management von Migrationszyklen.

4.1 Struktur der Beschreibung von Forschungsdatentypen

Die unten beschriebene Struktur dient als Inspiration und Anleitung für die Beschreibung der Forschungsdatentypen. Zielgruppe für die Beschreibungen sind primär Datenkuratoren und Verantwortliche für technische Infrastruktur (z.B. in einem Rechenzentrum). Die Fachlichkeit kann dabei verkürzt dargestellt werden, solange Einflüsse auf Technik, Abläufe und Geschäftsmodelle in der Beschreibung angedeutet sind.

[66] Diese basiert unter anderem auf den Erfahrungen aus dem nunmehr 15-jährigen Betrieb von TELOTA (BBAW) sowie aus der Entwicklung der Kern-Codierung in TextGrid.
http://www.textgrid.de/fileadmin/TextGrid/reports/TextGrid_Kerncodierung_070615.pdf
Vgl. hierzu auch die Herangehensweise des Kölner Data Center for the Humanities (DCH):
Patrick Sahle & Simone Kronenwett: Jenseits der Daten: Überlegungen zu Datenzentren für die Geisteswissenschaften am Beispiel des Kölner ‘Data Center for the Humanities’. LIBREAS #23. http://libreas.eu/ausgabe23/09sahle/
[67] DFG-Viewer. http://dfg-viewer.de/

4.1.1 Significant Properties / wissenschaftliche Anforderungen

Die in diesem Absatz aufgeführten Eigenschaften von Forschungsdatentypen stehen im Mittelpunkt des Interesses bei der Aufbewahrung über lange Zeiträume hinweg und bei Migrationsschritten. Bei Verlust einer dieser Eigenschaften würde der Forschungsdatentyp an Information und Wert verlieren. Dabei ist ein gewisser Informationsverlust (aus ökonomischen oder technischen Gründen) nicht zu verhindern bzw. unter wissenschaftlichen Gesichtspunkten möglicherweise auch akzeptabel. Die hier vorliegende, strukturierte Liste hilft dabei, die Risiken zu analysieren und bewusst in die Entscheidungsfindung einzubeziehen.

Eine gängige Strukturierung von Significant Properties ist diejenige in die vom Project INSPECT[68] herausgearbeiteten Bereiche: Content, Appearance, Behaviour, Context und Structure.[69] Zusätzlich zu diesen Eigenschaften betrachten wir an dieser Stelle die wichtigsten Anforderungen von Geisteswissenschaftlern an Datenzentren: die Nachnutzbarkeit und Sichtbarkeit der Daten.[70] Ausgehend von diesen Punkten dient die folgende Struktur als Werkzeug für strukturierte Beratungen mit Geisteswissenschaftlern, auf deren Basis das Angebot des HDC dargestellt, die individuellen Anforderungen für die Langzeitverfügbarkeit dokumentiert und ein Vertrag zwischen Wissenschaftler und HDC geschlossen werden können.

Checkliste Significant Properties und wissenschaftliche Anforderungen

Ablage der Daten (Bit Preservation) (> Content, Structure)

a) Identifikation der relevanten Daten sowie Beschreibung der Datenstruktur bzw. Überführung in eine projekt- und systemlandschaftsunabhängige Struktur zur Ablage, als Basis für Bit-Preservation nach gängigen Ansätzen der Langzeitarchivierung (LZA)

Interpretierbarkeit (intellektuelle Nachnutzbarkeit) (> Context)

b) Beschreibung des Projektkontextes, des wissenschaftlichen Ausgangspunktes und der wissenschaftlichen Methodik[71]

c) Identifikation und Entschärfung von inhaltlichen Abhängigkeiten, indem z.B. spezifische Terminologien dokumentiert und externe Daten/Dienste bewusst abgetrennt oder mit archiviert werden

Erhaltung der Daten (logische Nachnutzbarkeit) (> Content, Structure)

d) Dokumentation der Significant Properties und der wissenschaftlichen Anforderungen (nach der hier vorliegenden Struktur) als Basis für Migrationsschritte bei der Registrierung der

[68] Investigating the Significant Properties of Electronic Content Over Time (INSPECT). Jisc. 14. Juni 2014. http://www.webarchive.org.uk/wayback/archive/20140614074606/http://www.jisc.ac.uk/whatwedo/programmes/reppres/inspect.aspx

[69] Vgl. The significant properties of digital objects. Jisc. 15. Juni 2014. http://www.jisc.ac.uk/whatwedo/programmes/preservation/2008sigprops.aspx sowie NARA: Significant Properties. (2009); siehe speziell Annex C. http://www.archives.gov/era/acera/pdf/significant-properties.pdf

[70] Siehe Kapitel 2.1. Die weitere Anforderung "Individuelle Beratung" wird bei den Abläufen des Datenzentrums besprochen.

[71] Diese Informationen sind üblicherweise in Projektanträgen und Projektabschlussberichten enthalten, insofern könnte eine Verknüpfung dieser Dokumente mit den Forschungsdaten ausreichen (z.B. zur Auswahl und Vorverarbeitung der Quellen; Beschreibung des aktuellen Wissensstandes; Annahmen, Methoden und Forschungsfragen).

Daten im HDC (*migration on ingest*) und bei Gefahr von technischer Obsoleszenz nach gängigen Ansätzen der LZA[72]

Zitierbarkeit (Visibilität, wissenschaftliche Anforderungen)

e) Zitierbarkeit auf einer für wissenschaftliche Zwecke adäquaten Granularitätsebene bis hin zu Textabschnitten und Ton-/Videosequenzen

Vernetzbarkeit (Nachnutzbarkeit, wissenschaftliche Anforderungen)

f) Schnittstellen, um Daten und Funktionen direkt in aktuelle Virtuelle Forschungsumgebungen zu überführen bzw. einzubetten

g) Datenmodelle, die eine direkte Analyse und Verknüpfung der Daten unterstützen[73]

Interpretation (intellektuelle Nachnutzbarkeit) (> Appearance, Behaviour)

h) Darstellung und Interaktion, sodass die wissenschaftlichen Erkenntnisse des Projektes unmittelbar einsehbar und nutzbar sind[74]

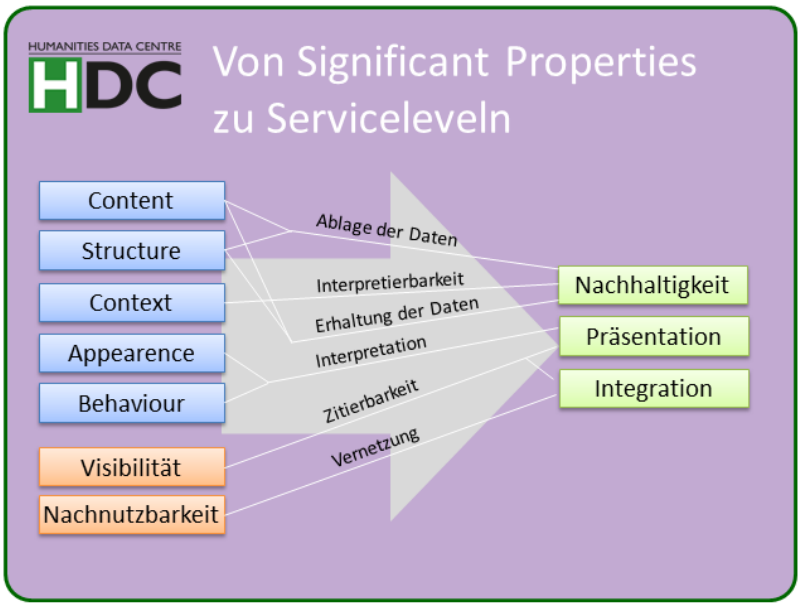


Abbildung 6. Abbildung der Significant Properties und der wesentlichen Nutzeranforderungen auf Angebotsebene.

4.1.2 Beispiele und normalisierte Darstellung

In diesem Absatz können alle bereits existierenden Anwendungen und Anleitungen aufgelistet werden, die prototypisch für einen Forschungsdatentyp stehen oder ihn definieren. Beispielhaft sind im Bereich Digitalisierung und Transkription – der mit dem Digitalisat den wohl bereits am intensivsten bearbeiteten und (zumindest für Textdigitalisierung) am klarsten strukturierten Datentyp in der geisteswissenschaftlichen Forschung hervorgebracht hat – folgende Referenzen

[72] Dazu zählen auch z.B. die Extraktion der Daten aus Microsoft Access 2.0 in CSV oder XML oder die Dokumentation von interaktiven Anwendungen in Story-Boards und Videos.

[73] So kann z.B. eine Edition zur Ablage im ursprünglichen Microsoft Word-Format gehalten bleiben, zur Erhaltung nach XML transferiert und zur Vernetzbarkeit in ein spezifisches XML/TEI umgesetzt werden.

[74] Dieser Punkt ist stark abhängig von den individuellen Significant Properties: für manche Datensätze und deren Autoren mag eine Archivierung der Daten ausreichend sein, während in anderen Situationen die Interpretation der Daten erst durch eine Darstellung oder Funktionalität möglich wird.

einschlägig: Referenzen auf Portale mit Digitalisaten[75], die DFG Praxisregeln[76] sowie der DFG-Viewer[77] als ein bereits existierendes und verbreitetes Werkzeug zur normalisierten Darstellung.

4.1.3 Verbreitung und Nachfrage

Dieser Absatz soll einen groben Eindruck über die Landschaft der Forschungsprojekte im Bereich des beschriebenen Forschungsdatentyps geben und u.a. folgende Fragen beantworten: Ist Forschung dazu begrenzt auf spezielle Institutionen bzw. spezielle Disziplinen? Kann man die Anzahl der Projekte bzw. das Projektvolumen abschätzen? Gibt es in (einem, einigen oder allen) Projekt(en) Besonderheiten im Hinblick auf Größe, Projektdauer oder die Ausstattung mit speziellen Ressourcen?

Forschungsdatentyp		Beispiele
Digitale Editionen	Marx-Engels-Gesamtausgabe	http://telota.bbaw.de/mega/
	Theodor-Fontane-Notizbücher	https://www.textgrid.de/community/fontane/
	Kant Opus Postumum	http://telota.bbaw.de/kant_op/
Datenbanken	Berliner Klassik	http://berlinerklassik.bbaw.de/BK/index_html
Visualisierungen	Global Migration Flows	http://flow.mmg.mpg.de/
Bilddaten*	IMEJI	http://imeji.org/
Interviews*	Global Divercities	http://www.mmg.mpg.de/subsites/globaldivercities/about/

Tabelle 1: In der HDC-Designphase untersuchte Forschungsdatentypen, die prototypisch umgesetzt werden. Das Konzept ist um weitere Forschungsdatentypen erweiterbar bzw. ist auf eine Struktur von mehreren HDCs ausgerichtet, die jeweils unterschiedliche Spezialisierungen aufweisen können.

* Diese beiden Forschungsdatentypen werden im Rahmen dieses Dokumentes nicht detailliert ausgeführt, werden hier aber beispielhaft ausgeführt, um zu verdeutlichen wie Erweiterungen des Konzepts aussehen können.

4.2 Digitale Editionen

In Editionen werden literarische und historische Quellen gesichtet und erschlossen.[78] Die Edition bzw. die Textkritik ist eine verbreitete wissenschaftliche Methode in der Arbeit mit schriftlichen Quellen und wird in den Literaturwissenschaften, der Geschichte, der Musikwissenschaft, den Religionswissenscha3ften, der Rechtsphilosophie und anderen geisteswissenschaftlichen Disziplinen genutzt. Editionen werden typischerweise von Einzelpersonen oder kleinen Gruppen in jahre-, oft jahrzehntelanger Arbeit erstellt.

Editionen unterscheiden sich z.B. in ihrem Umfang (bspw. ein Brief, ein Buch, eine Gesamtedition des Nachlasses eines Autors), der Form der Quellen (bspw. eine Handschrift, Textzeugen

[75] Z.B. eines der großen Digitalisierungszentren, ein großes und langlaufendes Digitalisierungsprojekt, etc.
[76] DFG Praxisregeln „Digitalisierung“. DFG-Vordruck 12.151 – 02/13: http://www.dfg.de/formulare/12_151/12_151_de.pdf
[77] DFG-Viewer: www.dfg-viewer.de
[78] Bodo Plachta: Editionswissenschaft: Eine Einführung in Methode und Praxis der Edition neuerer Texte. Reclam, Philipp, jun. GmbH, Verlag (1997)

unterschiedlicher Fassungen, Übersetzungen) und den der Edition zu Grunde liegenden Forschungsfragen (Arbeitsweise des Autors, Textgenese, Rezeption eines Textes im geschichtlichen Kontext). Auch Editionen derselben Quellen mit ähnlichen Forschungsfragen unterscheiden sich in den wissenschaftlichen Ergebnissen oft so weit, dass die Daten und Darstellungen der Editionen nicht untereinander vergleichbar sind. Eine Normalisierung der Daten oder der Darstellung würde daher aus Sicht der Wissenschaftler einen Informationsverlust und somit die Minderung des wissenschaftlichen Wertes einer Edition bedeuten.

Überwiegend werden auch heute noch Editionen für den Druck erstellt. Zunehmend werden aber auch vernetzte und nicht-lineare Darstellungsformen in digitalen Medien als ideal für Inhalt und Zweck von Editionen angesehen.[79] Viele Aspekte digitaler Editionen sind allerdings nur mit Informationsverlust in die Druckform überführbar.

Trotz der großen Unterschiede zwischen Editionen in Bezug auf den wissenschaftlichen Inhalt und den Kontext, gibt es bei der Erstellung von digitalen Editionen eine gewisse Konvergenz in Bezug auf die verwendeten (technischen) Werkzeuge – vom Datenmodell bis hin zu den Darstellungsmechanismen.

4.2.1 Significant Properties

Diese Eigenschaften sind an Diskussionen in der Community angelehnt[80] und bilden eine Ergänzung zu den bereits herausgearbeiteten generellen, in Kapitel 4.1.1 beschriebenen, Significant Properties. Dabei wird keine umfassende Darstellung der diversen unterschiedlichen Formen (z.B. Briefedition, historisch-kritische Edition) angestrebt, sondern es soll ein Eindruck der Minimalanforderungen vermittelt werden.

Ablage der Daten (Bit Preservation)

- Die Primärdaten (iS.v. Rohdaten) umfassen Texte inkl. Annotationen (in XML/TEI), damit verknüpfte Digitalisate (Bildformate, ggf. Multimedia-Formate) sowie Apparate und Register (in XML). Die Archivierung der Daten ist als Mindestanforderung ausreichend. Es liegen üblicherweise keine zusätzlichen Informationen in den Algorithmen der Anwendung oder in externen Diensten vor, die nicht auf Basis der Daten und Dokumentation rekonstruiert werden können. Davon ausgenommen sind mögliche semantische Auszeichnungen (z.B. Referenzen auf Normdatenbanken oder projektspezifische Thesauri).

Interpretierbarkeit (intellektuelle Nachnutzbarkeit)

- Bei der weit verbreiteten Verwendung des TEI-Standards ist keine zusätzliche Dokumentation der Datenstruktur notwendig. Nichtsdestotrotz ist eine zusätzliche Dokumentation der editorischen Entscheidungen im Datenmodell nützlich.
- Mögliche semantische Vernetzungen müssten explizit herausgearbeitet werden, z.B. Referenzen auf Personennormdatenbanken, in verknüpften Nachlässen etc.

Erhaltung der Daten (logische Nachnutzbarkeit)

[79] Patrick Sahle: Digitale Edition (Historischer Quellen) - Einige Thesen. 1997. http://www.uni-koeln.de/~ahz26/dateien/thesen.htm

[80] Patrick Sahle, et al.: Kriterien für die Besprechung digitaler Editionen, Version 1.1 (Juni 2014). Institut für Dokumentologie und Editorik. http://www.i-d-e.de/aktivitaeten/reviews/kriterien-version-1-1

- Keine zusätzlichen Punkte

Zitierbarkeit (Visibilität, wissenschaftliche Anforderungen)

- XML: Referenzierbarkeit von Teilen der Edition (z.B. auf Absatz oder Wortebene) für Zitation im wissenschaftlichen Diskurs, bzw. Einbindung in externe Anwendungen
- Digitalisate: Referenzierbarkeit von Bildausschnitten

Vernetzbarkeit (Nachnutzbarkeit, wissenschaftliche Anforderungen)

- Suche unter Ausnutzung der Möglichkeiten in den Daten (z.B. XQuery)
- Auszeichnungen der Daten und (externe) Verlinkungen (z.B. Personen oder Orte in Normdatenbanken und Thesauri)
- Direkte Einlesbarkeit der Daten in XML (z.B. in oXygen)
- Bei Briefeditionen: Publikation in TEI Correspondence

Interpretation (Intellektuelle Nachnutzbarkeit) (> Appearance, Behaviour)

- Synchronisierte Darstellung von unterschiedlichen Fassungen und Formen zum direkten Vergleich (z.B. Text/Bild, Übersetzung, Handschrift vs. redigierte Druckfassung)
- XSLT als Formatierungswerkzeug

4.2.2 Beispiele und normalisierte Darstellung

Wie oben (Kap. 4.1) ausgeführt, gibt es Editionen in unterschiedlichen Formen und aus unterschiedlichen wissenschaftlichen Kontexten: von der Brief- zur Musikedition[81], als historisch-kritische oder als Regestenedition[82]. Für die Zwecke dieses Dokumentes und als Startpunkt für die Entwicklung eines HDC-Angebotes greifen wir folgende Beispiele heraus, die in ihrer Form und technischen Umsetzung ebenso weit verbreitet wie flexibel anpassbar sind:

Marx-Engels Gesamtausgabe (MEGA). http://telota.bbaw.de/mega/
Historisch-kritische Edition der Veröffentlichungen, der Manuskripte und des Briefwechsels von Karl Marx und Friedrich Engels. Begonnen in den 70er Jahren in Berlin und Moskau, ist das Projekt aktuell ein Akademievorhaben an der BBAW. Ein Abschluss der Arbeiten wird für 2025 erwartet; von den 114 geplanten Bänden sind bisher 60 erschienen, wobei derzeit nur Auszüge digital zugänglich sind.

Theodor-Fontane Notizbücher. https://www.textgrid.de/community/fontane/
Genetisch-kritische Hybrid-Edition der Notizbücher aus u.a. Tagebucheinträgen, Skizzen, Briefkonzepten und Vortragsmitschriften. Das DFG-Projekt an der Georg-August-Universität Göttingen läuft voraussichtlich von 2011-2017.

[81] TextGrid Interview mit Prof. Dr. Joachim Veit. https://www.textgrid.de/ueber-textgrid/materialien/interview-joachim-veit
[82] Rohrschneider, Michael: Editionstypen, aus: Tutorium Quelleneditionen analog und digital, in: historicum-estudies.net, http://www.historicum-estudies.net/etutorials/tutorium-quelleneditionen/definition/editionstypen/

KANT: Online-Edition des Opus Postumum. http://telota.bbaw.de/kant_op/

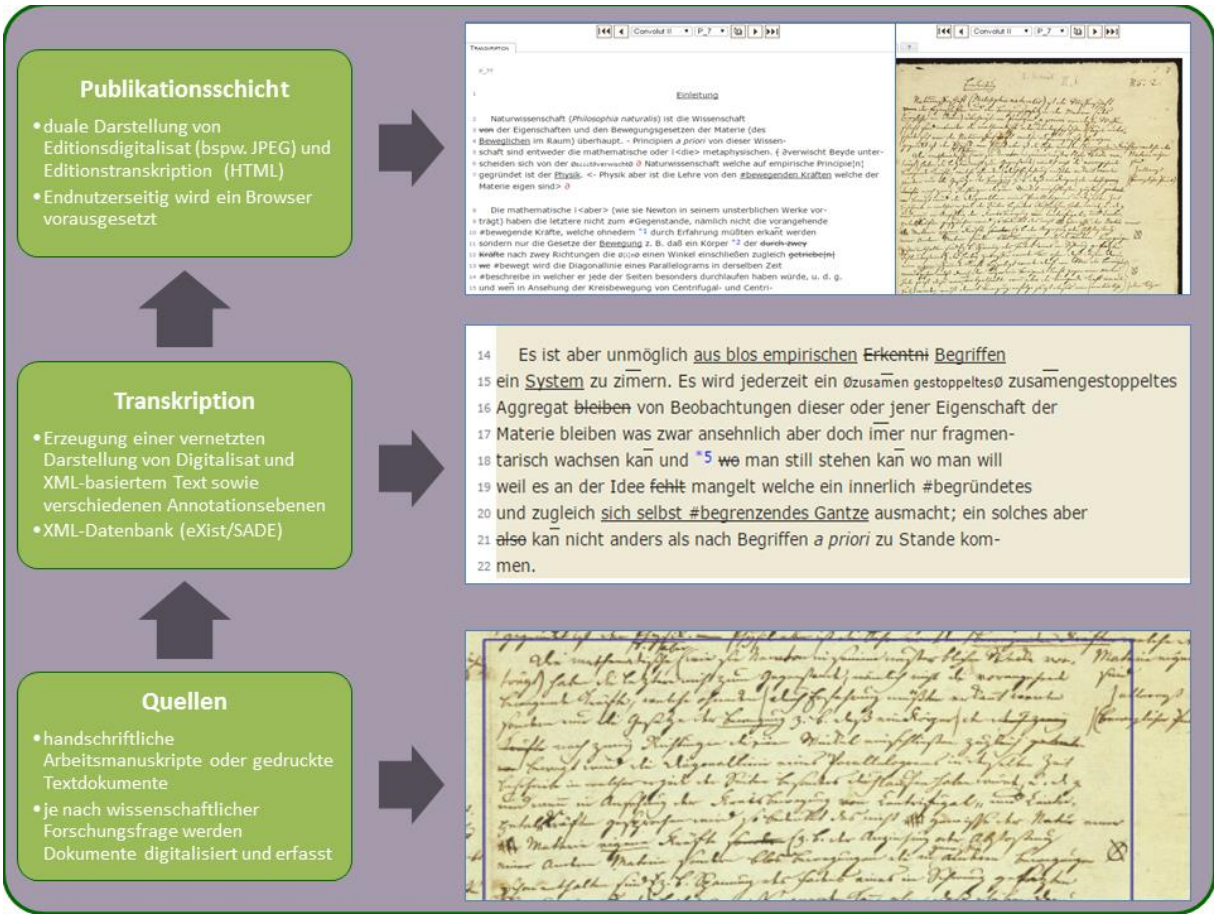


Abbildung 7: Archivierung von Editionen am Beispiel der Kant-Edition der BBAW

Die Online-Edition präsentiert die Rohtranskription in wechselseitiger abschnittsweiser Verbindung mit den digitalisierten Faksimiles in der diplomatischen Abfolge des Manuskriptes und enthält gegenwärtig die Konvolute II und III. Vorgesehen ist, alle weiteren Konvolute sukzessive zur Verfügung zu stellen.

Technisch gesehen sind alle drei Projekte separat entwickelt worden, bauen aber auf ähnlichen technischen Strukturen auf: Bildverwaltung, eine XML-Datenbank sowie XSLT-Skripte zur web-basierten Darstellung. Diese grundlegende Architektur könnte aktuell eine technische Referenz für die große Mehrzahl von existierenden Editionen sein und die Basis für einen generischen Viewer zur normalisierten Darstellung von Editionen bilden.

4.2.3 Verbreitung und Nachfrage

Etwa 70% der Projekte in Akademien sind Editionen, insgesamt etwa 170 Einzelvorhaben in unterschiedlichen geisteswissenschaftlichen Disziplinen.[83] Bei einem jährlichen Gesamtbudget der Akademien[84] von 60 Mio. Euro bedeutet das 40 Mio. Euro jährlich.

[83] Wissenschaftsrat: Stellungnahme zum Akademieprogramm. 2009. www.wissenschaftsrat.de/download/archiv/9035-09.pdf
[84] http://www.akademienunion.de/fileadmin/redaktion/user_upload/Leporello.pdf

4.3 Datenbanken

Unter einer „Datenbank“ verstehen wir im Kontext des HDC (geordnete) Listen von (semi-) strukturierten Einträgen. Eine Datenbank kann das primäre Ergebnis eines Forschungsprojektes sein, z.B. in der Prosopographie eine Personendatenbank mit (strukturierten) Einträgen zu einem bestimmten Personenkreis. Eine Datenbank kann aber auch ein Werkzeug sein, das z.B. lediglich als Hilfsmittel zur Texterschließung in einem Editionsvorhaben aufgebaut wird. Hier wie da ist eine Datenbank ein wesentliches Ergebnis einer Forschungsarbeit und liegt üblicherweise zentral eingebettet in weitere Anwendungen (vgl. Referenzen auf Normdaten) vor.

4.3.1 Significant Properties

An dieser Stelle fokussieren wir auf die einfachste mögliche Form einer Datenbank, ohne weitere technische Ausprägungen vorwegzunehmen. In der technischen Umsetzung ist es ebenso berechtigt zu argumentieren, dass ein gängiges Wiki den Nutzen des Katalogs erfüllt, wie sich andererseits von z.B. TEI oder auch SKOS[85] aus der W3C RDF-Community inspirieren zu lassen und seinen Katalog technologisch aufzurüsten.

Ablage der Daten (Bit Preservation)

- Einträge mit eingebetteten oder externen Quellen (Zitation, Digitalisat, Verknüpfung). Einträge können strukturiert sein, sind in der Umsetzung aber letztlich zumeist semi-strukturiert (wo z.B. Geburtsdaten von Personen fehlen oder nur ungefähr bekannt sind).

Interpretierbarkeit (intellektuelle Nachnutzbarkeit)

- Eine genaue Beschreibung der Quellen und des wissenschaftlichen Vorgehens ist unerlässlich, um die Einträge der Datenbank verstehen zu können. So ist z.B. eine Personendatenbank, die aus den fiktionalen Charakteren eines Buches aufgebaut ist, nicht vergleichbar mit einer Datenbank, die Informationen zu Personen enthält, die aus historischen Zeitungsartikeln extrahiert wurden. Auch wo sich unterschiedliche Quellen widersprechen, muss eine Datenbank Möglichkeiten zur Abbildung der Widersprüche bieten.[86]

Erhaltung der Daten (logische Nachnutzbarkeit)

- Browsen nach Eintrag

Zitierbarkeit (Visibilität, wissenschaftliche Anforderungen)

- Referenzierbarkeit einzelner Einträge

Vernetzbarkeit (Nachnutzbarkeit, wissenschaftliche Anforderungen)

- Spezialsuche (ggf. Synonyme/Übersetzungen)
- Einbettung von RDFa bzw. FOAF zur Vernetzung mit anderen Personendatenbanken.
- Explizite Vernetzung bei Überlappung mit gängigen Normdatenbanken (z.B. PND, VIAF)

Interpretation (Intellektuelle Nachnutzbarkeit)

- Spezialsuche (ggf. Synonyme/Übersetzungen)

[85] SKOS, Simple Knowledge Organization System. http://www.w3.org/TR/2009/NOTE-skos-primer-20090818/
[86] vgl. z.B. die Projektziele des Personendatenrepositoriums (PDR). http://www.personendatenrepositorium.de/

4.3.2 Beispiele und normalisierte Darstellung

Obwohl der Begriff „Datenbank“ an sich auf keine wissenschaftliche Methode verweist, kann man in den unterschiedlichsten Projekten Bedarf nach einem vergleichbaren Werkzeug feststellen.

Berliner Klassik. www.berliner-klassik.de
Das Akademievorhaben (2000-2013) hat eine Reihe von Katalogen über die Berliner Kultur im Zeitraum von 1786-1815 aufgebaut (6700 Personeneinträge, 9500 Bibliografien, 8500 Veranstaltungen). Seit Projektende wird eine technische Nachhaltigkeitsstrategie gesucht, die die Erhaltung der Navigationsmöglichkeiten und der Referenzierbarkeit der Einträge sicherstellt.

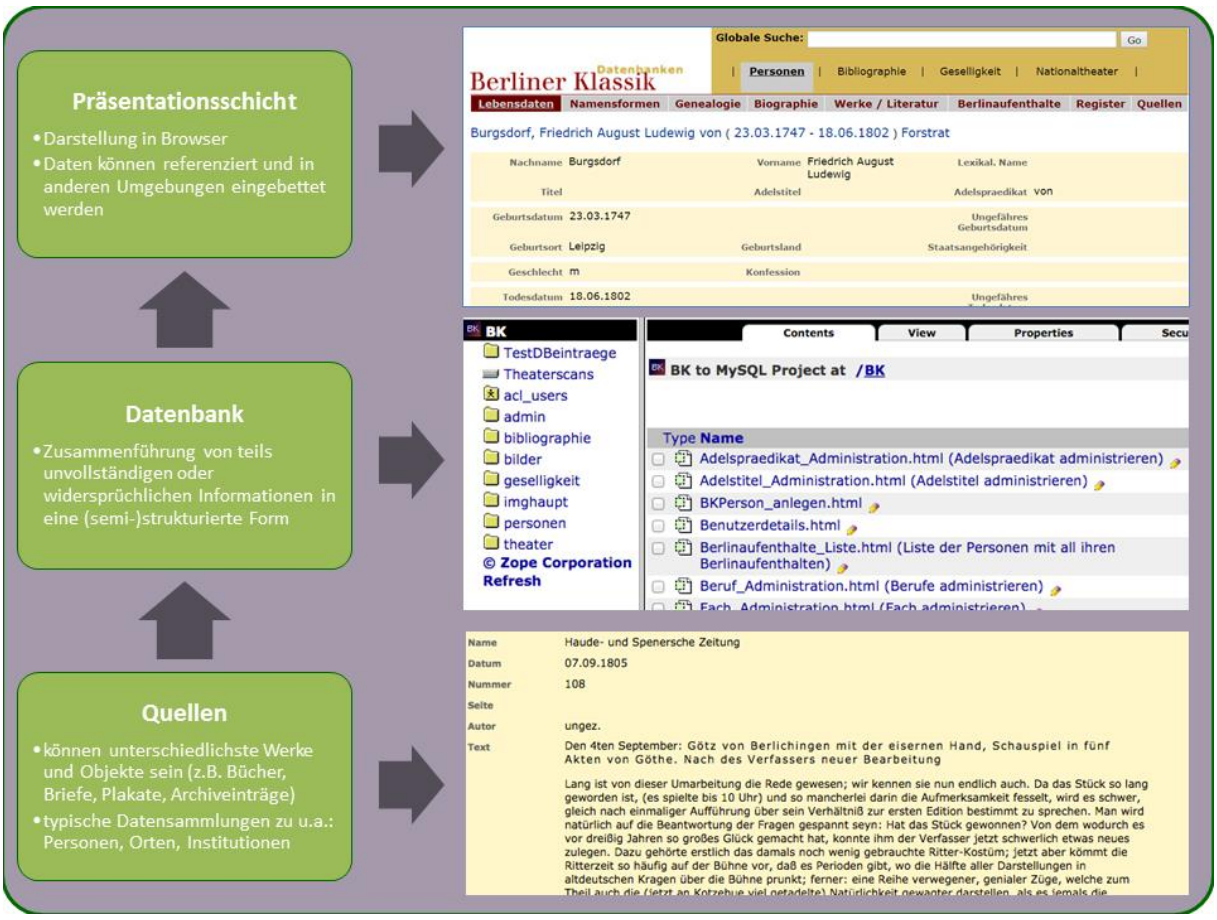


Abbildung 8: Archivierung von Datenbanken am Beispiel der Datenbank Berliner Klassik der BBAW

Eine Vielzahl an Projekten baut ad hoc Personendatenbanken als Werkzeug für ihre Forschungen, darunter die Leibniz-Edition (www.leibniz-edition.de) und die Alexander von Humboldt Forschungsstelle (avh.bbaw.de). Allein diese beiden Projekte nutzen unterschiedliche Paradigmen zur Datenhaltung, nämlich offene SQL- und XML-Datenbanken einerseits und lokale Microsoft-Lösungen andererseits. NoSQL-Datenbanken oder Angebote wie CoNE[87] sind derzeit noch nicht verbreitet in geisteswissenschaftlichen Kontexten aufgegriffen worden. Allen genannten Projekten ist die Wichtigkeit für nachhaltige Lösungen bewusst, die derzeit vorhandenen Lösungen sind aber oft gewachsen und aus einem dringenden Bedarf heraus entstanden. Einzig das Leibniz-Projekt hat beim

[87] CoNE. Control of Named Entities. http://pubman.mpdl.mpg.de/cone/

Aufbau seiner Datenbank bereits früh Nachhaltigkeitsaspekte mit bedacht und die Lösung gleich direkt in Form eines Datenbankclusters an einem Rechenzentrum umgesetzt (bei der GWDG, mit Beratung und Entwicklung durch Telota/BBAW).

4.3.3 Verbreitung und Nachfrage

Eine genaue Abschätzung der Nachfrage nach Datenbanken ist nur schwer möglich, zumal der Anwendungsfall nicht genau abgegrenzt werden kann. Vermutlich erstellt jedes geisteswissenschaftliche Forschungsprojekt irgendeine Form von Verzeichnis (zu Personen, Orten, Geschehnissen, Werken etc.). Der Aufbau einer eigenen Datenbank rentiert sich aber erst bei größeren Projekten, in denen ggf. mehrere Personen gemeinsam an der Datenbank arbeiten und sie in unterschiedlichen Kontexten referenzieren. Doch selbst, wenn sich nur bei jedem zehnten oder zwanzigsten Projekt der Aufbau einer eigenen Datenbank rentieren würde, ist das Aufkommen nach wie vor sehr hoch.

Als vernetzte Werkzeuge sollten Datenbanken eigentlich Teil einer geisteswissenschaftlichen Infrastruktur sein. Obwohl DARIAH Dienste für Normdaten und Personendaten in ihrer Architektur aufführt, gibt es aber derzeit noch keinen generischen Katalogdienst.[88]

4.4 Visualisierungen

Mit der vermehrten Verfügbarkeit digitaler Forschungsdaten auf der einen Seite und der zunehmenden Zugänglichkeit (im Sinne von Bedienbarkeit, Standardisierung) von Werkzeugen zur Aufbereitung und Darstellung dieser Daten auf der anderen, hat das Thema (Daten-)Visualisierung an Bedeutung gewonnen. Prinzipiell können alle Formen von Forschungsdaten in irgendeiner Form aufbereitet dargestellt werden. Für unsere Zwecke – als Forschungsdatentyp für die Definition des HDC-Angebots – grenzen wir Visualisierungen ein auf Datenbanken beruhende Präsentationen von Forschungsdaten für den menschlichen Nutzer. Der Präsentations- und Explorationsaspekt steht im Vordergrund und hebt diesen Forschungsdatentyp von anderen ab. Visualisierungen stehen somit Publikationen näher als Forschungsdaten im Sinne von Primärdaten.

Ein besonderer Aspekt der in unserem Kontext interessanten Visualisierungen ist, dass sie die Interaktion des Nutzers mit der Anwendung ermöglichen (schematisch in der unten stehenden Abbildung beschrieben). Der Nutzer kann entsprechend wechselnder, individueller Fragestellungen die Daten immer wieder neu zueinander in Beziehung setzen (vergleichbar zu Datenbankabfragen). Damit ist eine Visualisierung, anders als bspw. textbasierte Forschungsdaten, aber besonders an ihre Präsentationsumgebung gebunden. Daraus folgt, dass die Langzeitarchivierung der o.g. Mehrschichtigkeit von Visualisierungen Rechnung tragen muss, wenn sie den hier beschriebenen besonderen Nutzwert von Visualisierungen erhalten möchte.

Vereinfacht gesprochen, bestehen interaktive oder explorative Visualisierungen in der Regel aus mindestens drei Schichten:

[88] Siehe DARIAH-DE Confluence:
Die Säulen der DARIAH-DE Infrastruktur. Fachwissenschaftliche und Generische Dienste.
https://dev2.dariah.eu/wiki/pages/viewpage.action?pageId=30376358

- Den Daten, die gesammelt und aufbereitet in einer Datenbank vorliegen. Die **Datenbank** kann unterschiedlich komplex sein – das Spektrum reicht von einfachen Excel-Tabellen, bis zu komplexen Datenbankausführungen. In der Regel wird es sich nicht mehr um reine Primärdaten, sondern um bereits aufbereitete und normalisierte Datenbestände handeln.
- Der auf der Datenbank aufsetzenden **Prozessierungsschicht** (Middleware). Hier werden die Daten aufbereitet an die Client-Anwendung des Nutzers weitergeben. Während die Datenbank in den meisten Fällen noch von den Wissenschaftlern selbst erstellt wird, beginnt bei der Prozessierungsschicht ein Bereich, der häufig externe IT-Expertise erfordert. Dieser Aspekt trifft auch auf die u.g. Präsentationsschicht zu.
- Auf der Prozessierungsschicht setzt die **Präsentationsschicht** auf, die - beispielsweise in einem Browser – die Daten dem Nutzer präsentiert und Interaktionsmöglichkeiten bietet. Visualisierungen sind somit in aller Regel nicht datei-, sondern datenbank- bzw. präsentationsbasiert und stellen Relationen von Informationsobjekten zueinander dar.

Daraus folgt auch eine andere Form der Bereitstellung und Verbreitung von Visualisierungen als Forschungsdatentyp. Anders als Texte, Bilder oder Videos, jedoch ähnlich wie die oben genannten Editionen, waren sie aufgrund ihrer spezifischen Eigenschaften bisher kein Gegenstand bibliothekarischer oder informationswissenschaftlicher Aufbereitung. Praktisch bedeutet das für die Übernahme in ein Forschungsdatenzentrum, dass Visualisierungen nicht einfach aus einem Repository übernommen werden können und auch in der Präsentation/ Archivierung in einem (üblicherweise dateibasierten) Repository nicht unbedingt gut aufgehoben sind.

4.4.1 Significant Properties

Streng genommen sind unter dem Gesichtspunkt der Archivierung nur die einer Visualisierung zugrunde liegenden Forschungsdaten für eine spätere Nachnutzung notwendig. Eine derart minimalistische Sicht auf Forschungsdaten ist nicht ungewöhnlich und wird im vorliegenden Dokument bspw. im Bereich Datenbanken durchaus ähnlich gehandhabt[89]. Im Fall der hier beschriebenen, interaktiven Visualisierungen kann dies jedoch nicht Sinn und Zweck der Archivierung sein, denn für die spätere Nachnutzung sind ja gerade der Erhalt des Mehrwertes und der Interaktionsmöglichkeiten essenziell. Um dies zu gewährleisten ist die Archivierung nur der Forschungsdaten nicht ausreichend – es muss auch der Erhalt entweder der Anwendung selbst oder der Parameter, die ihre spätere Rekonstruktion ermöglichen, gewährleistet sein.

Daten (Bit Preservation und logische Nachnutzbarkeit)

a) Die Primärdaten von Visualisierungen gliedern sich entsprechend der oben genannten Schichten auf, d.h. wie im obigen Beispiel angefangen von den in einer mySQL-Datenbank gespeicherten Forschungsdaten (statistische Daten aus Längsschnitterhebungen), dem Sourcecode für das Processing, die Informationsobjekte im JSON-Format zur Weitergabe an die javascriptbasierte Präsentationsumgebung und der Sourcecode für die Ausgabe mittels ProcessingJS.

[89] Dieser minimalistischen Perspektive wird auch mit den HDC-Serviceleveln Rechnung getragen. D.h. auf einer basalen Ebene wird es immer eine Archivierung der einfachen Forschungsdaten ohne die Prozessierungs- und Darstellungsumgebung geben.

b) Primärdaten können auch in einem recht rudimentären Zustand vorliegen (Excel- oder Word-Tabellenformate) und wurden ggf. aufbereitet bevor sie für eine Prozessierung zur Verfügung standen.

Interpretierbarkeit (intellektuelle Nachnutzbarkeit)

a) Ohne eine Dokumentation sind die oben genannten Einzelbestandteile nicht direkt nutzbar, bestenfalls träfe das noch auf die Primärdaten zu. Die Dokumentation muss somit parallel zu den Rohdaten nachgeführt werden und den gesamten Prozess bis zur Ausgabe der Visualisierung für den Nutzer abdecken.
b) Ggf. sind ausführlichere Dokumentationsbestandteile für die Darstellungsumgebung oder die bis dahin notwendige Prozessierung der Daten notwendig. Im Zweifel hängt die Dokumentationstiefe aber sicher auch von der Nachnutzbarkeit des Codes ab. Gegebenenfalls würde man es bei einer Dokumentation des Sourcecodes belassen.

Zugänglichkeit (Intellektuelle Nachnutzbarkeit)

c) Allgemein: transparente und nachvollziehbare Darstellung von Datenmaterial in visualisierter, interaktiver Form. D.h. Quellen und/oder Primärdaten müssen referenziert sein, die Methode der Prozessierung muss nachgewiesen sein.
d) Quellen und/ oder Primärdaten sollten als eigenständiger Datensatz (Datenbank, Tabellen o.ä.) bereitgestellt werden, idealerweise in einem Format, das zur Nachnutzung geeignet ist.
e) Die Art und Weise der Interaktionsmöglichkeiten sollte dokumentiert sein, d.h. der Nutzer sollte in die Lage versetzt werden, individuelle Abfragen bzw. Nutzungsmöglichkeiten selbst anzuwenden.

4.4.2 Beispiele und normalisierte Darstellung

Global Migration Flows: http://flow.mmg.mpg.de/
Die Anwendung ermöglicht es dem Nutzer in einem herkömmlichen Browser verschiedene Datensets individuell zusammenzustellen und die Daten zu visualisieren. Gegenstand der Darstellung sind Migrationsbewegungen zwischen 1970 und 2011 zwischen verschiedenen Staaten. Es können in einfach verständlicher Weise Wanderungsbewegungen in Relation zueinander gesetzt werden, Zuwanderungsgewinne oder Abwanderungsverluste verdeutlicht werden.

DivCon Intergroup Contact: http://divcon.mmg.mpg.de/
DivCon zeigt die Häufigkeit gruppenübergreifender sozialer Kontakte zwischen autochthoner und zugewanderter Bevölkerung. Diese beiden Grobkategorien können nach bestimmten Parametern verfeinert werden, so dass bspw. sozial- oder demographiespezifische Zusammenstellungen möglich sind. Diese Datenvisualisierung demonstriert in besonderer Weise den Aspekt der Interaktivität. Der Nutzer stellt individuell sein eigenes Set an Parametern zusammen und ihm wird erst im Anschluss daran die eigentliche Datenvisualisierung präsentiert.
Global Migration by Origin: http://stock.mmg.mpg.de/origin

Auf dem gleichen Datensatz wie Global Migration Flows basierend, werden hier die Herkunftsstaaten von Migrationsbewegungen detailliert dargestellt. Dieses Beispiel demonstriert einen weiteren Aspekt von Datenvisualisierungen, wie mit Hilfe ein und desselben Datensatzes unterschiedliche Zusammenhänge dargestellt werden können.

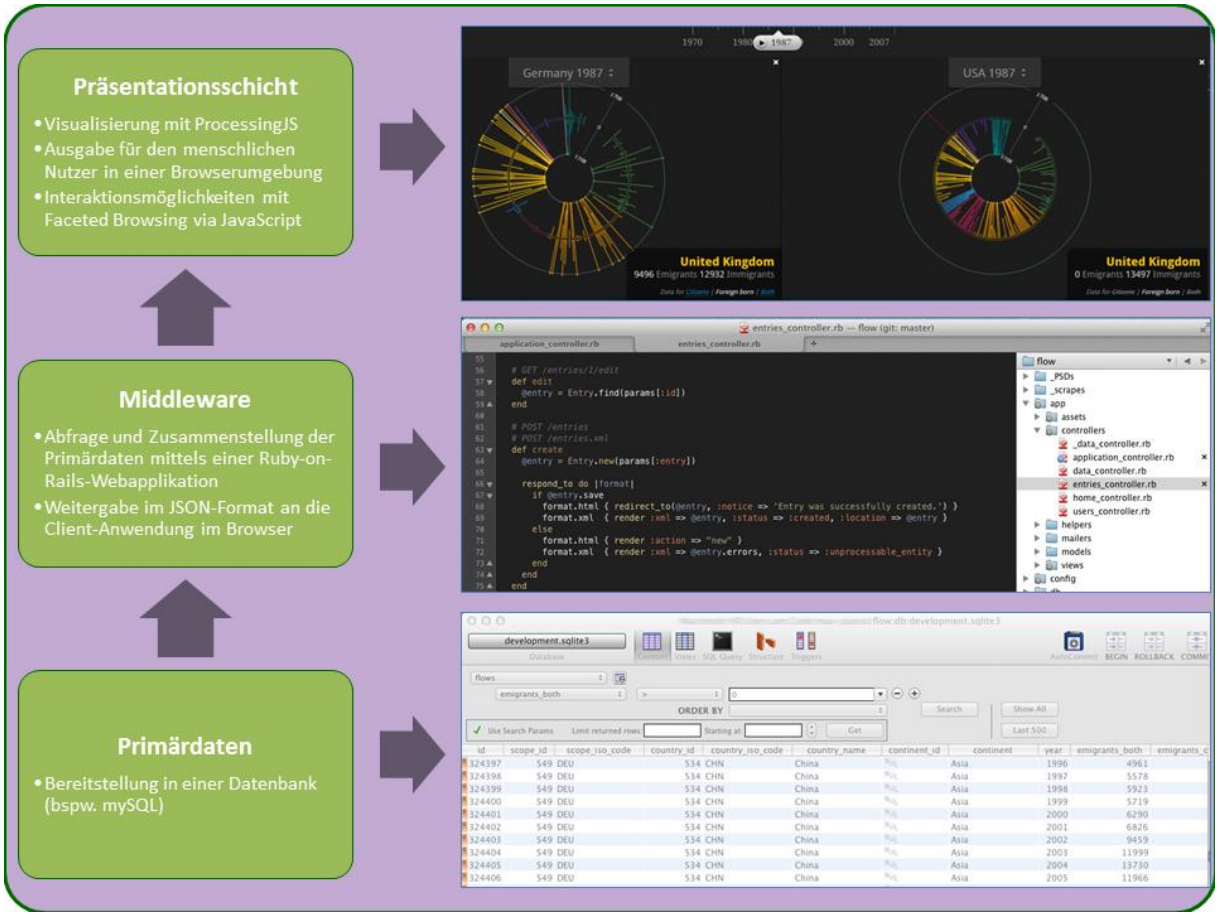


Abbildung 9: Schematische Darstellung einer dreischichtigen Visualisierung von Daten zu Migrationsbewegungen (http://flow.mmg.mpg.de/)

Die oben genannten Beispiele für interaktive Datenvisualisierungen stammen aus einem sozialwissenschaftlichen Forschungskontext des Partners MPIMMG aus dem HDC-Konsortium. Das Thema Migrationsbewegungen ist vor dem Hintergrund seiner aktuellen und gesellschaftspolitischen Bedeutung besonders geeignet, die Vorteile einer ansprechenden, interaktiven und auf eine breitere Öffentlichkeit gerichteten Darstellungsweise von Forschungsdaten und -ergebnissen zu demonstrieren. In verschiedenen Forschungsprojekten werden zahlreiche Fragestellungen zum Thema kulturelle, religiöse und ethnische Diversität untersucht, z.B.: Wie hat sich Diversität entwickelt? Auf welche Weise wurde gesellschaftliche Vielfalt geregelt oder könnte sie geregelt werden? Wie entfalten sich unterschiedliche Prozesse inter-ethnischer oder inter-religiöser Begegnungen? Wie verändern sich Muster und Bilder von Vielfalt und wie entwickeln sich die Bedeutungen der und die Relationen zwischen den Begriffen Ethnizität und Religion? Die Forschungen widmen sich nicht allein migrationsbezogenen Zusammenhängen in den sogenannten Herkunfts-, Transit- und Aufnahmegesellschaften, sondern auch Gesellschaften, die durch verschiedene Arten von Vielfalt charakterisiert sind, wie zum Beispiel Südafrika, die Balkanländern, Indien und Südostasien.

4.4.3 Verbreitung und Nachfrage

Zweifellos wird die Verbreitung und Nutzung von Werkzeugen zur (interaktiven) Visualisierung von Forschungsergebnissen künftig zunehmen. Durch den Einsatz neuer Technologien und der Möglichkeit zur Erschließung und Vernetzung verteilter Datenbestände werden attraktive Darstellungen von Zusammenhängen – gerade auch für eine nichtwissenschaftliche, interessierte Öffentlichkeit – möglich. Diese Visualisierungen können bisher klassisch textgebundene Publikationen nicht nur anreichern, sondern qualitativ aufwerten und entsprechen so einer Nachfrage von Wissenschaftlern und Förderern.

Derzeit ist für die Umsetzung von Visualisierungen wie in den genannten Beispielen noch spezifische Expertise notwendig, aber es ist wahrscheinlich, dass es hier in naher Zukunft zur Durchsetzung von Werkzeugen kommen wird, die auch der informationstechnische Laie gut und zielgerichtet einsetzen kann. Spätestens ab diesem Zeitpunkt müssen auch Informationsinfrastrukturen in der Lage sein, mit diesem Forschungsdatentyp nicht nur umgehen und ihn darstellen zu können, sondern auch Mehrwerte anzubieten. Die Langzeitarchivierung und Bereitstellung in einem Forschungsdatenzentrum wird ein solcher Mehrwert sein.

5 Angebotskategorien

Mit dem Konzept der Forschungsdatentypen und dem Fokus auf Innovation ist ein HDC für neue technische Möglichkeiten und sich ändernde Nutzeranforderungen gewappnet. Bei der Definition von Angebotskategorien spielen aber nicht nur technische Aspekte und Nutzeranforderungen eine Rolle. Die Angebotskategorien stehen auch in engem Zusammenhang mit dem Geschäftsmodell und der Sichtbarkeit (stärker: die „Marke“) eines HDC (siehe Ergebnisse des Teilprojekts 3). Da ein HDC kein etabliertes, sondern ein gerade erst **entstehendes Angebotsfeld** ist, werden sich das Verständnis und auch die Möglichkeiten in den ersten Jahren stark weiterentwickeln. Die hier beschriebenen Angebotskategorien sind daher lediglich ein **Startpunkt**, von dem aus sich ein HDC über die Zeit weiterentwickeln kann.

Gemeinsam mit einem sich ändernden Angebotsfeld wird sich auch das HDC strukturell und in Bezug auf seine Ausrichtung ändern. Die Lean Startup-Ansatz[90] unterscheidet in den ersten Jahren der Geschäftsentwicklung Phasen, in denen jeweils unterschiedliche Ziele im Fokus stehen: **empathy** (ein genaues Verständnis des Problems und der Nutzeranforderungen), **stickiness** (Nutzer kommen zurück und nutzen ein Angebot auch mehrmals, Nutzertreue), **virality** (ein Angebot verbreitet sich und die Nutzerzahl wächst), **revenue** (erst dann kann die Kosten- und Einnahmenstruktur stabilisiert werden), und **scale** (erst am Ende der Geschäftsentwicklung liegt der Fokus auf der operativen Handhabung eines großen Wachstums). Die im Folgenden beschriebenen Angebotskategorien fokussieren auf Nutzertreue: In dieser Phase sind das Problem und die technischen Lösungsmöglichkeiten hinreichend gut verstanden; Ziel ist primär die Entwicklung eines nützlichen und attraktiven Angebots, das Nutzer zurückkommen lässt.

Auch eine Einteilung in **Nutzersegmente**[91] stabilisiert sich erst mit einer hinreichend großen Nutzerbasis. Wie in Kapitel 3.3 ausgeführt, kann die Zielgruppe eines HDC in der gesamten Forschungscommunity liegen.[92] Innerhalb dieser Nutzergruppe kann man segmentieren z.B. in:

(a) passive Nutzer, die z.B. durch wissenschaftliche Zitationen in ein HDC kommen
(b) aktive Nutzer, die z.B. existierende Forschungsdaten analysieren und nachnutzen
(c) einmalige Nutzer, die z.B. lediglich Projektdaten zur Erfüllung der Auflagen von Förderern archivieren wollen
(d) wiederkehrende Nutzer, wie z.B. institutionelle Datenkuratoren, die eine Vielzahl von Projekten in ein HDC übertragen

[90] Startups don’t really know what they are at the beginning. An interview with Alistair Croll and Benjamin Yoskovitz. March 2013. http://radar.oreilly.com/2013/03/startups-dont-really-know-what-they-are-at-the-beginning.html
[91] Customer Segments. Siehe das Business Model Canvas. In: Business Model Generation: Ein Handbuch für Visionäre, Spielveränderer und Herausforderer. August 2011. http://www.businessmodelgeneration.com/
[92] wobei ein institutionelles oder disziplinäres HDC stärker auf eine bestimmte Nutzergruppe fokussiert.

Zur Erreichung von Nutzertreue ist das Ziel, Nutzer sukzessive von Nutzersegment (a) über (b) und (c) nach (d) zu führen. [93]

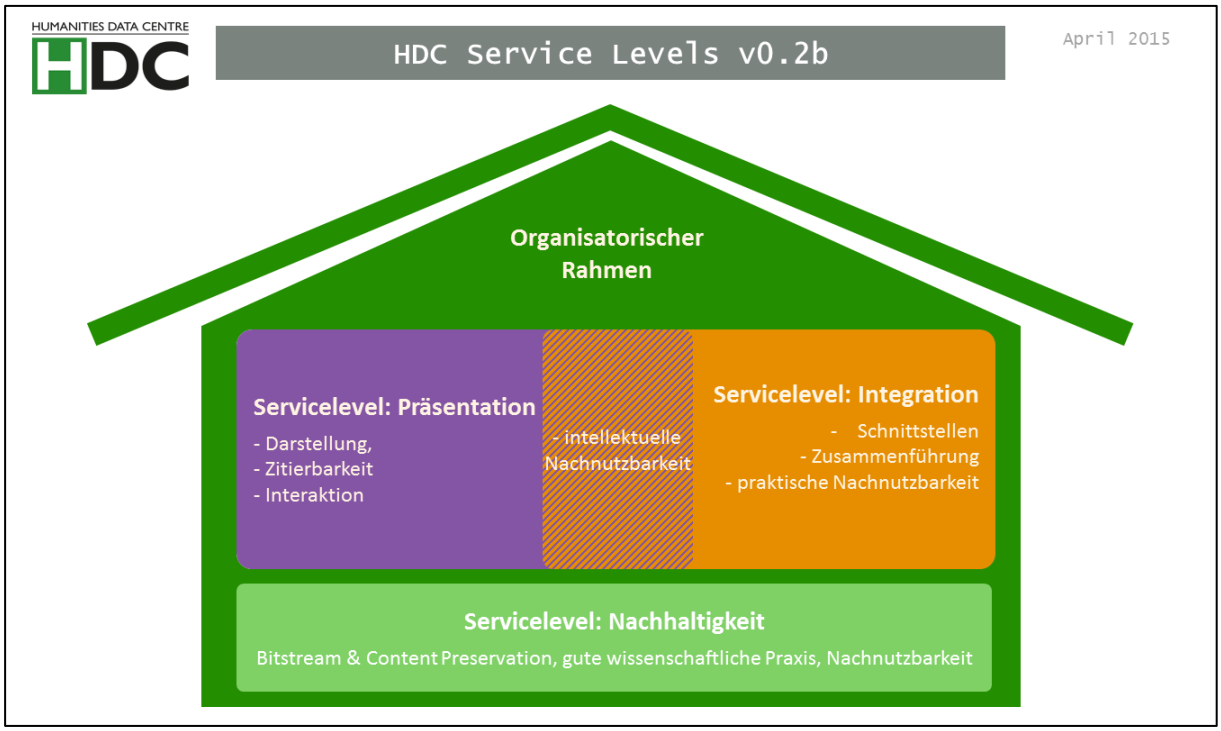


Abbildung 10: Anforderungen an ein HDC (vgl. Kapitel 4.1.1) werden abgedeckt durch die Nutzungsversprechen Nachhaltigkeit, Präsentation und Integration. Die zugehörigen technischen Dienste werden im Bericht von TP2 spezifiziert.

Die folgenden Angebotskategorien fokussieren auf die **Nutzungsversprechen**[94] aus einer technischen, nachfrageorientierten Sicht. Dabei verweisen die in Kapitel 4.1.1 identifizierten Significant Properties und wissenschaftlichen Anforderungen auf die generellen Nutzungsversprechen Nachhaltigkeit, Präsentation und Integration. Diese übergreifenden Versprechen werden durch konkrete (technische) Angebote erfüllt, wobei sich diese im Laufe der Zeit ändern werden.

1. **Repository**: Relevante Daten und Dokumentation werden mit Hilfe der Datenerzeuger identifiziert, ggf. in ein stabiles (d.h. nachhaltiges und zitierbares) Format überführt und über ein Repository zum Download (nachnutzbar) angeboten. Das Repository kümmert sich dabei um die Datenverwaltung, die Nutzungswerkzeuge wie die Suche sowie um Migrationen (Active Preservation). Das Repository ist die Basis für alle weiteren Angebote. Obwohl es technisch nicht unbedingt mit anderen Angeboten verknüpft sein muss, stellt das Repository einen Archiv-Mindeststandard[95] für wissenschaftliche Forschungsdaten im HDC dar. Relativierend muss hier allerdings angefügt werden, dass sich die Erfüllung dieses Nachnutzungsversprechens insbesondere auf die Ebenen Bitstream Preservation und

[93] Zur Messung des Erfolgs der HDC-Strategie können die Nutzersegmente kontinuierlich statistisch erfasst werden: aus Nutzungsdaten (z.B. Weblogs für a und b) und administrativen Metadaten (z.B. Vertragsdaten bzw. Projektbeschreibungen für c und d). Dabei werden die Messungen voraussichtlich aufgrund der Natur des Angebots (Langzeitarchivierung) in einem niedrigen Bereich bleiben und sich nur sehr langsam ändern.
[94] Value Proposition. Siehe Business Model Canvas.
[95] In Bezug auf die WissGrid-Ebenen deckt das HDC-Repository alle drei Ebenen ab: Bit Preservation, Logische und Intellektuelle Nachnutzbarkeit.

logische Nachnutzbarkeit bezieht und die intellektuelle, aber teilweise auch logische Nachnutzbarkeit damit nicht abgedeckt wird.

2. **VM-Einfrieren**: Eine Anwendung wird in ihrer aktuellen Umgebung eingefroren und externe Schnittstellen werden gekappt. Migrationsschritte sind dabei nicht vorgesehen, daher hat dieses Angebot nur eine begrenzte Lebensdauer.[96] Zielgruppe sind primär Projekte, die nach Ende der Projektlaufzeit ihre Projektergebnisse noch einige Jahre in der originalen Darstellung erhalten wollen, aber die technische Infrastruktur nicht mehr erhalten können.
3. **Generische Viewer**: Das HDC übernimmt Daten und Anwendungslogik in seine Verantwortung. Dabei bietet das HDC für diejenigen Forschungsdatentypen, die für dieses Vorgehen geeignet sind[97], einen generischen Viewer zur Präsentation der Daten an. Obwohl bei der Migration von Daten und Anwendungslogik in den generischen Viewer ein Informationsverlust entstehen kann, erlaubt dieses Angebot theoretisch eine technisch unbegrenzte Erhaltung der Anwendung inklusive Präsentation, Interaktivität und Schnittstellen.
4. **Schnittstellen**: Daten aus dem Repository werden aktiv in aktuellen Nachweissystemen (z.B. Bibliothekskatalogen, Digitale Infrastrukturen) registriert, mit virtuellen Forschungsumgebungen zur direkten Nachnutzung verknüpft und in den wissenschaftlichen Diskurs eingebettet (z.B. durch Annotationstools, spezialisierte Netzwerke).[98] Schnittstellen werden aktualisiert und ergänzt. Dazu müssen ggf. auch Datenstrukturen angepasst werden.

Für die meisten dieser Angebote wird erwartet, dass eine **persönliche Beratung** zwischen Wissenschaftlern bzw. Datenerzeugern und dem HDC notwendig ist. Selbst bei genauer Einhaltung von Richtlinien ist aus heutiger Sicht eine Automatisierung einzelner Schritte nur schwer vorstellbar bzw. gar nicht erwünscht, da von dieser Vorgehensweise auch das HDC im Hinblick auf das Verständnis der Daten, der Anforderungen der Nutzer und zukünftige Entwicklungsmöglichkeiten profitiert. Schon die Auswahl des richtigen Angebots wird höchstwahrscheinlich nicht ein Wissenschaftler alleine vornehmen, sondern in Abstimmung mit einem wissenschaftlichen Datenkurator für den Nutzungsfall individuell abstimmen. Festzuhalten bleibt aber, dass die Automatisierung der Abläufe unter dem Gesichtspunkt der Effizienz natürlich ein Ziel bleibt, aber zum derzeitigen Zeitpunkt als schwierig eingeschätzt wird.

6 Organisation

Der Diskussion organisatorischer Ansätze für ein geisteswissenschaftliches Forschungsdatenzentrum ist ein eigenständiges Arbeitspaket in der HDC-Designphase gewidmet. Dennoch soll dieses Thema auch im vorliegenden Dokument beleuchtet werden, da die Modellierung der Angebots- und

[96] Serverseitig kann eine relativ lange Lebensdauer erwartet werden, da neben der Datenbank und Anwendungen auch das Betriebssystem eingefroren wird; die Lebensdauer der Anwendung ist also abhängig von der Lebensdauer der Virtualisierung. Allerdings können clientseitig neue Browser-Versionen die Anwendungen schnell dysfunktional machen, wenn z.B. Flash nicht mehr verfügbar ist oder eine neue JavaScript-Version inkompatibel wird. Zitierbarkeit und Vernetzbarkeit (z.B. in aktuelle Virtuelle Forschungsumgebungen) können durch diesen Ansatz auch nur bedingt gewährleistet werden.

[97] Von den in Kapitel 4 betrachteten Forschungsdatentypen kommen hierfür Editionen und Datenbanken in Betracht.

[98] als konkrete Beispiele: eine Bibliografie kann direkt in Zotero importiert werden, eine Edition direkt im TextGridLab nachgenutzt werden, eine Datenbank als Linked Data abgegriffen werden, eine Brief-Edition über TEI-Correspondence vernetzt werden, alle Daten über ein Annotationstool basierend auf der Open Annotation Collaboration adressiert werden.

Ablaufdefinition erhebliche Abhängigkeiten von der und Auswirkungen auf die Organisation des Forschungsdatenzentrums hat. Gleiches trifft ebenfalls auf das Betriebs- und das Geschäftsmodell zu.

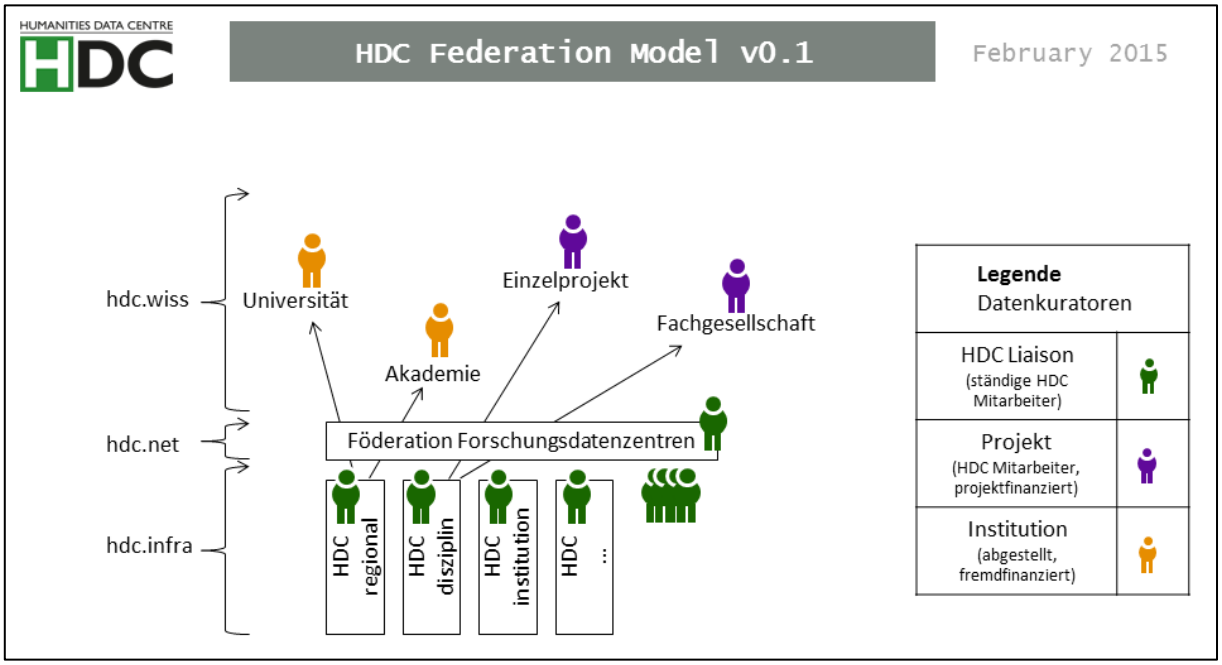


Abbildung 11:Illustration der drei Ebenen der HDC-Organisation: das HDC, die HDC-übergreifende Föderation und die verteilten Datenkuratoren. Die Benennung hdc.infra, hdc.net und hdc.wiss dient lediglich zur Indizierung der Bereiche.

Es wird eine Organisationsstruktur mit drei unterschiedliche Ebenen angestrebt:

1. ***das* HDC (hdc.infra**[99]) – Der technische Betrieb sowie Management und Verwaltung für ein HDC sind möglichst lokal gebündelt. Dabei wird ein HDC typischerweise an eine existierende Institution oder eng kooperierende Partner angeschlossen, ist aber eine eigenständige Organisation mit eigenen Kompetenzen.
2. **die Datenkuratoren (hdc.wiss)** – Die Nutzerkommunikation ist verteilt und nicht nur am Datenzentrum, sondern auch unabhängig von einem HDC direkt an z.B. Wissenschaftsorganisationen angesiedelt.[100] Dabei sind verteilte Datenkuratoren die Vermittler zwischen den Wissenschaftlern und dem HDC, wofür sie eine intensive fachliche Nähe zum Wissenschaftler brauchen.
3. **die Föderation** (**hdc.net**) – Aus einer Vogelperspektive koexistieren mehrere HDCs mit eigenen (oder überlappenden) Strukturen. Ihre Aufgaben teilen sie sich nach regionalen und inhaltlichen Gesichtspunkten (hypothetisch z.B. das DCH in Köln als regionales Datenzentrum, GAMS in Graz als institutionelles Datenzentrum sowie IANUS als disziplinspezifisches Datenzentrum).

Besonders die Zusammenarbeit zwischen dem HDC und Datenkuratoren ist eng; beide Parteien können ohne die jeweils andere nicht funktionieren. Abgebildet auf ein „virtuelles“ Organigramm, wird in der Kombination dieser Bereiche die Gesamtheit der nötigen Kompetenzen skizziert (vgl.

[99] Oder hdc-local.
[100] Kommunikationskanäle zwischen HDC und “Kunde”, siehe Channels bzw. Customer Relationship im Business Model Canvas. In: Business Model Generation: Ein Handbuch für Visionäre, Spielveränderer und Herausforderer. August 2011. http://www.businessmodelgeneration.com/

Abbildung 12). Dabei meint „virtuell“: nicht notwendigerweise in dieser Form an einem Ort gebündelt, sondern eingebettet in eine bestehende Organisation bzw. bestehende Organisationen.

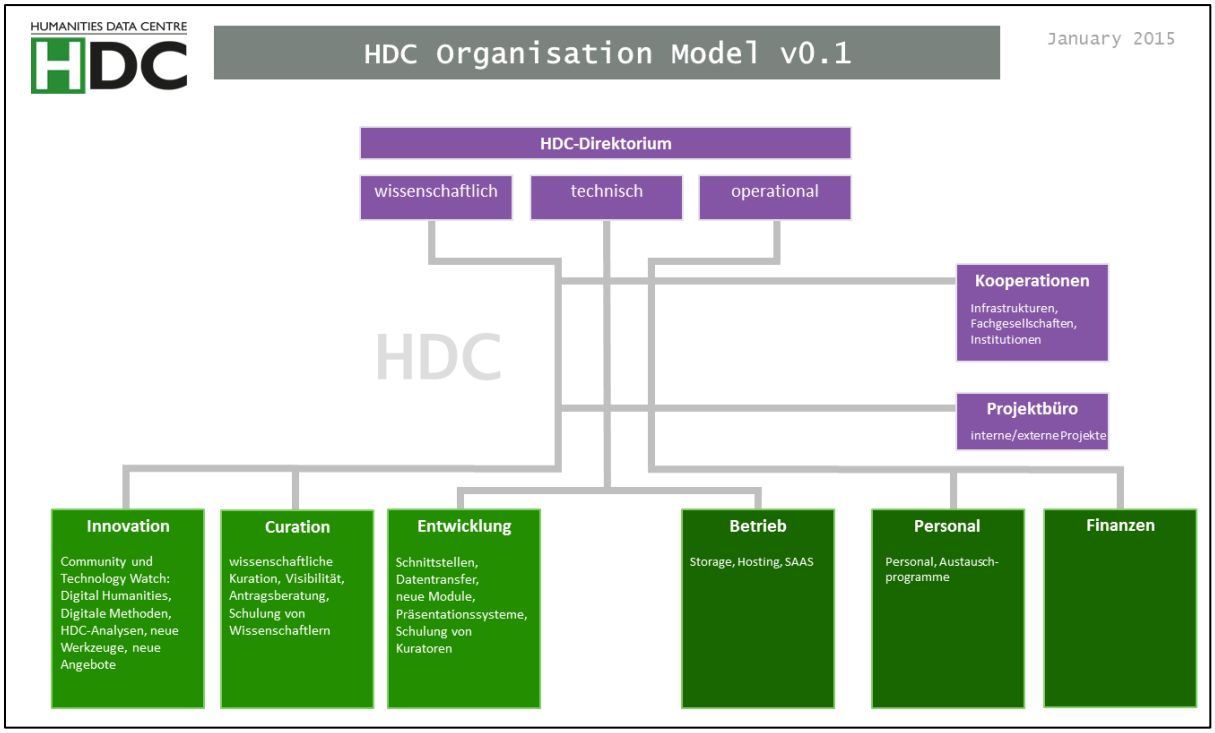


Abbildung 12: Organigramm/Organisationsaufbau des HDC. Dabei sind der Bereich Curation und teilweise die Bereiche Innovation und Entwicklung stark verteilt und typischerweise nicht am Datenzentrum, sondern vielmehr an Wissenschaftseinrichtungen angesiedelt. Dieses Organigramm ist daher lediglich eine Illustration der nötigen Kompetenzen in einer HDC-Organisation.

Das Organigramm[101] soll neben den nötigen Kompetenzen auch einige weitere Aspekte der Organisation des HDC hervorheben:

- In der **Leitung** des HDC muss ein ausgewogenes Verhältnis zwischen wissenschaftlichen, technischen und operationalen Perspektiven herrschen.
- **Kooperationen** sind wesentlich für das HDC, sowohl beim Aufbau des Netzwerks aus Datenkuratoren und von Kooperationen mit Wissenschaftsorganisationen, als auch in der Arbeitsteilung mit anderen HDCs und Infrastrukturen (z.B. gemeinsame Entwicklung von Diensten, Schnittstellen).
- Ein HDC ist projektorientiert organisiert, da Kernabläufe kompetenzübergreifend sind (z.B. Ingest, Migrationen oder eine Modifikation des Angebots)[102]. In der Matrixorganisation des HDC kommt dem **Projektbüro** eine besondere Rolle zu, da es die Verfügbarkeit von Kompetenzen und Ressourcen abteilungsübergreifend – und im Fall der Datenkuratoren auch institutionsübergreifend – sicherstellen muss.
- Die Aufgaben der **Datenkuratoren** beziehen sowohl Kommunikation, Organisation und Datenmodellierung als auch technische Entwicklung mit ein. Standards des HDC (z.B. in Bezug auf

[101] Siehe dazu auch eine Darstellung des Organisationsmodells im Anhang.
[102] z.B. Ingest neuer Ressourcen (Curation → Entwicklung → Betrieb), Angebotsmodifikationen (Innovation → Entwicklung → Betrieb), Migrationen (Betrieb → Entwicklung ↔ Curation)

Datenmodelle und Schnittstellen) müssen vermittelt, aber umgekehrt auch die Angebotspalette des HDC sukzessive erweitert und an die Nachfrage angepasst werden.

- Zur **Innovation** im HDC müssen technische und fachwissenschaftliche Expertise eng zusammenspielen, um (1) gemeinsam mit Wissenschaftlern neue DH-Methoden zu testen und in das Angebot des HDC aufzunehmen sowie (2) die Haltbarkeit von Technologien und die Skalierbarkeit der HDC-Infrastruktur zu erforschen.

7 Abläufe

Für die Darstellung von Aufgaben und Abläufen für Forschungsdatenzentren empfiehlt sich unmittelbar das OAIS, eine verbreitete und in unterschiedlichen Organisationen getestete Referenz. Wir haben uns aus folgenden Gründen dafür entschieden, nicht direkt das OAIS zu übernehmen, sondern eine eigene Struktur zu entwickeln:

- Das OAIS ist auf den Umgang mit dateibasierten Objekten ausgerichtet. Man kann argumentieren, dass das OAIS z.B. lauffähige Anwendungen in seinem funktionalen Modell nicht adäquat abdeckt.
- Die Angebote des HDC gehen mit Beratungs- und Entwicklungsaufgaben über die Aufgaben eines fokussierten Archivs hinaus (z.B. Antragsberatung, Projektbegleitung).
- Obwohl durch die letzten beiden Punkte ein HDC eigentlich mehr Abläufe übernehmen muss, versuchen wir lediglich einen schlanken Überblick über wesentliche Abläufe zu geben (typische Abläufe, die durch OAIS oder TRAC[103] abgedeckt sind, werden hier nicht wiederholt).

In der folgenden tabellarischen Darstellung ist angedeutet, dass das Organigramm lediglich Kompetenzen bündelt, während viele Abläufe kompetenzübergreifend stattfinden.

1	vor Übergabe: Antragsphase, Projektphase	Organigramm*
1.1	(Weiter-)entwicklung von Standards, Datenmodellen und Datenmanagement in Projekten	Curation, Innovation, Entwicklung, Betrieb
1.2	Beratung in Bezug auf Datenstrukturen, technische Architekturen (z.B. Stabilität von Software-Komponenten) und Projektabläufen; Förderung von HDC-Compliance	Curation
1.3	Sichtbarkeit des HDC, Schulung von Datenkuratoren und von Wissenschaftlern	Curation, Entwicklung
1.4	Mitarbeit in Projekten als Teil des Datenmanagements (Embedded Data Curator)	Curation

2	Übergabe und Migration	
2.1	Definition des Angebots und Vertragsabschluss	Curation
2.2	Definition der Anforderungen (inkl. Significant Properties) gemeinsam mit Nutzer: Diskussion und formale Definition	Curation
2.3	Ingest (Überführung in HDC-approved Datenformat, Registrierung der Metadaten, Übertrag je nach Angebotslevel)	Curation, Entwicklung, Betrieb
2.4	Übernahme von Nicht-HDC-abgestimmten Daten (z.B. Feuerwehreinsatz für alte MS Access-Anwendung)	Curation, Entwicklung, Betrieb

[103] TRAC - Trustworthy Repositories Audit + Certification: Criteria and Checklist. http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf

3	Nachnutzung	
3.1	Betrieb und organisatorische Anforderungen (bspw. Datenlizensierung, Moving Wall)	Betrieb
3.2	Langzeitarchivierungsmaßnahmen (Migrationen der Daten, neue Schnittstellen, Weiterentwicklung Viewer)	Entwicklung, Betrieb
3.3	Nachfolgeregelung (z.B. Wissenschaftlicher Vertreter), Vertragsverhandlung und Folgevertrag	(Projektbüro)
3.4	kontinuierliches Anforderungsmanagement: Nutzerkommunikation, Technology Watch, neue Angebote	Innovation, Curation, Entwicklung, Betrieb

**Bezugnahme auf das Organigramm auf S. 39*

7.1 Checklisten

In den Angebotskategorien sowie den Organisationsstrukturen und Abläufen steckt das wesentliche organisatorische Wissen für den Betrieb eines HDC. Als konkrete Handreichungen entwickelt das HDC-Projekt folgende Checklisten/Richtlinien:

- **(TP1) Checkliste für Anforderungen** (inkl. Significant Properties) – siehe Kapitel 4.1.1, und **Angebotskategorien** .
  Sie werden in der Beratung durch die Datenkuratoren herangezogen und dokumentiert. Dies dient als Basis für die Spezifikation des Angebots und den Vertrag, die Umsetzung des Ingest sowie in der Nachnutzungsphase für die Durchführung von Migrationen.
- **(TP2) Liste von HDC-konformen Datenformaten**. Für diese Datenformate wird es dezidierte technische Dienste im HDC geben: Migrationspfade zur Langzeitarchivierung im Repository und wissenschaftliches Zitieren durch Anzeige von Ausschnitten (z.B. direkte Verweise auf Bildausschnitte, Textabschnitte, Videosequenzen).
- **(TP2) technische Schnittstellen und Ingest-Entry Points**.
- **(TP3) Richtlinien (Policy) für ein HDC**. Eine transparente Beschreibung der Werte und des wissenschaftlichen HDC-Kerns, der bei technischen oder strategischen Entscheidungen im HDC der Wissenschaft immer Mitbestimmung lässt (z.B. kontinuierlicher Ausbau der Angebote, möglicher Informationsverlust bei Migrationen).

Diese und weitere Checklisten sind ein erster Ausgangspunkt und werden während der Aufbau- und Betriebsphase kontinuierlich angepasst und erweitert. Außerdem werden sie ergänzt mit weiteren Konzepten z.B. zur Rechtsform eines HDC und Vertragsvorlagen.

7.2 Kontinuierliches Anforderungsmanagement

In einem kontinuierlichen Anpassungs- und Erweiterungsprozess wird ein HDC seine Angebote, Strukturen, Abläufe und Technologien ständig erweitern müssen. Dies ist (1) im Sinne gängiger Ansätze der Langzeitarchivierung, nach denen existierende Daten und Dienste in (mehr oder weniger) regelmäßigen Abständen überarbeitet und migriert werden müssen, (2) im Sinne des noch jungen Bereichs der Digital Humanities, in dem sich die Methoden und somit auch die Anforderungen an ein Forschungsdatenzentrum ständig weiterentwickeln, und (3) ist dies auch in anderen Bereichen üblich, da jeder Wissenschafts- und Geschäftsbereich Entwicklungen durchläuft und sich Organisationsstrukturen oft schneller ändern als Informationsinfrastrukturen.

Der Bereich „Innovation“ (siehe Organigramm, Kapitel 6) ist dafür verantwortlich, dass sich ein HDC ständig in diesem Erneuerungsprozess bewegt:

- gemeinsam mit Datenkuratoren und Wissenschaftlern, um das bestehende Angebot auf Basis von aktuellen, in den Digital Humanities genutzten Methoden und Technologien anzupassen und zu erweitern.
- gemeinsam mit den Entwicklern und dem Betrieb des HDC, um die technische Infrastruktur des HDC zu erweitern und um einen Beitrag zur internationalen Forschung im Bereich Datenmanagement zu leisten.
- gemeinsam mit der Leitung des HDC, um durch Erweiterung des Angebots neue Zielgruppen für das HDC zu gewinnen.

Um diesem Auftrag gerecht zu werden, müssen gleichermaßen technische und wissenschaftliche Kompetenzen an diesem Erneuerungsprozess beteiligt sein. Zudem muss der Informationsfluss hin zu und zwischen den oben genannten Parteien (Datenkuratoren, Wissenschaftler, Entwickler, Betrieb und HDC-Leitung) gesichert sein.

Dieser Informationsfluss kann durch statistische Auswertungen von Nutzungs- und Betriebsdaten auch mit Zahlen unterfüttert werden. Die Auswertung dieser Zahlen muss Teil der HDC-Infrastruktur sein und auch wesentlichen Einfluss auf Management-Ebene haben.

In Ausarbeitung der Optionen für das „kontinuierliche Angebotsmanagement“ haben wir – ähnlich den *Humanities Indicators*[104] – versucht, einen globalen Blick auf die Digital Humanities zu bekommen.[105] Solch ein Überblick ermöglicht es einem HDC zu jedem Zeitpunkt, die Angebote des HDC mit den aktuellen Anforderungen aus der DH-Forschungslandschaft abzugleichen, um ggf. Lücken zu identifizieren. Unterstützt werden könnte dieses Vorgehen durch die Verwendung von (anonymisierten) Daten aus Datenmanagementplänen, bspw. aus DFG-Anträgen, was allerdings aus Datenschutzgründen sehr schwierig umzusetzen wäre. Die Nutzung einer anderen Quelle für eine Kartierung der DH aus HDC-Sicht wäre unserer Ansicht nach zu aufwändig. Stattdessen schlagen wir vor, dass sich unterschiedliche HDCs auf ein gemeinsames Statistikformat ihrer Nutzungsdaten einigen und diese untereinander austauschen. Der Vergleich des Erfolgs von Angeboten in unterschiedlichen HDCs kann wesentliche Anhaltspunkte für die Weiterentwicklung der Anforderungen geben.

[104] Humanities Indicators - a project of the American Academy of Arts & Sciences. http://archive201406.humanitiesindicators.org/humanitiesData.aspx#aboutHI

[105] Eine Dokumentation der Analyse kann vom HDC direkt angefragt werden, und ist in Form von Folien gesammelt (HDC-nachfrage_komplett.pptx)

8 Zusammenfassung: Ergebnisse und Prototypen

Dieser Bericht deckt den kompletten Aufgabenbereich des Teilprojekts Service (TP1) der Designphase des Humanities Data Centre ab und behandelt daher auch gesammelt alle in diesem Teilprojekt vorgesehenen Deliverables/Meilensteine.

Die oben genannten Ergebnisse sind in enger Abstimmung mit den anderen Teilprojekten TP2-Technik und TP3-Organisation entstanden. Speziell sind gemeinsam mit TP2 entwickelt worden: das Konzept der Standardisierung für Forschungsdatentypen, die Identifikation der dafür nötigen technischen Dienste (vgl. TP2-Deliverable) sowie die Abläufe. Dabei war es ein Balanceakt, die in der Einleitung beschriebene Spannung zwischen individualisierten Angeboten und Skalierbarkeit aufzulösen; die Wahrung dieser Balance wird auch in der Aufbau- und der Betriebsphase des HDC eine kontinuierliche Aufgabe bleiben. In Zusammenarbeit mit TP3 sind die Angebotsebenen als Basis für die Nutzungsversprechen, die Organisation und die Abläufe – als Grundlage für das HDC-Organisationsmodell – sowie die Struktur der Dienste und Abläufe – als Basis für Kostenberechnungen – entstanden. Struktur und Terminologie der TP1-Ergebnisse sind dabei immer im Hinblick auf die Arbeiten in TP3 entstanden. Dazu zählen die in Anlehnung an das Business Model Canvas[106] entstandenen Aufschlüsselungen von Akteuren (Kapitel 3.3, Akteure), Nutzersegmenten

Deliverable/ Milestone	Projektmonat Kalendermonat	Kapitel in diesem Bericht
TP 1.1 – Definition des Angebots (Leitung: BBAW)		
Führen von Experteninterviews	6 Oktober 2014	Allgemeine Grundlage aller Arbeiten
Auswertung der Interviews und Workshops; Übersetzen in techn. Anforderungen	12 April 2015	Kapitel 2.1: Erhebung von Nutzeranforderungen Kapitel 4.1.1: Anforderungen / Significant Properties
Definition der Angebote im Rahmen der technischen und ökonomischen Rahmenbedingungen, Erstellen der Policy	24 April 2016	Kapitel 4: Vom Datenmodell zu Forschungsdatentypen Kapitel 5: Angebotskategorien
TP 1.2 – Definition Abläufe (Leitung: MPI-MMG)		
Übersicht der notwendigen Arbeitsabläufe	12 April 2015	Kapitel 6: Organisation Kapitel 7: Abläufe
Übersicht über notwendige Checklisten und Vorlagen in den einzelnen Arbeitsabläufen	18 Okt 2015	Kapitel 7.1: Checklisten
Skizzen einiger ausgewählter Arbeitsabläufe	24 April 2016	Kapitel 7: Abläufe Kapitel 8: Prototypen
Definition eines Verfahrens zum kontinuierlichen Anforderungsmanagement	24 April 2016	Kapitel 7.2: Kontinuierliches Anforderungsmanagement

[106] Business Model Canvas. In: Business Model Generation: Ein Handbuch für Visionäre, Spielveränderer und Herausforderer. August 2011. http://www.businessmodelgeneration.com/

und Nutzungsversprechen (Kapitel 0,) sowie die an das Lean Canvas angelehnte Identifikation der Erfolgsfaktoren und Advantages (Kapitel 3.4, Erfolgsfaktoren).

Als wichtigste TP-übergreifende Aktivität haben sich die Partner auf die Entwicklung und Umsetzung von **Prototypen** geeinigt, für die TP1 die Anforderungen spezifiziert und abstimmt. Jeder dieser Prototypen steht idealtypisch für eine größere Menge an existierenden Projekten, sodass die Erkenntnisse von einem Prototyp auf andere existierende Projekte übertragbar sind. Jeder dieser Prototypen hat eine relevante Sichtbarkeit in seiner Community, was hilfreich für Außendarstellung und Sichtbarkeit des HDC-Projektes ist.

Die Teilprojekte sind in folgender Weise in die Prototypen involviert:

- TP1 lotet die Übersetzung von Anforderungen aus der Wissenschaft in Angebote auf ihre technische und organisatorische/finanzielle Machbarkeit hin aus.
- TP2 validiert existierende technische Dienste spezifisch auf geisteswissenschaftliche Anforderungen hin und identifiziert Lücken für die Entwicklung neuer Dienste.
- TP1 und TP2 können eine gemeinsame Ebene der Standardisierung finden, die wissenschaftliche Anforderungen und technische Gegebenheiten gleichermaßen mit einbezieht.
- TP3 erhält von den Prototypen konkrete Erfahrungen über Abläufe und technische Ressourcen, aus denen es Kostenrechnungen ableiten kann.
- TP3 bekommt (umfangreicher als durch rein theoretische Überlegungen) einen Eindruck von möglichen (langfristigen) Risiken aus technischer oder operationaler Sicht.

<table>
<tr><th>Projekt</th><th>Forschungsdatentyp</th><th>Angebotsform</th><th>HDC-Partner</th><th>techn. Lernziel</th></tr>
<tr><td>Kant</td><td>Edition</td><td rowspan="2">Generischer Viewer</td><td>BBAW/ZIB</td><td rowspan="2">Architekturabgleich RZ
(eXist)</td></tr>
<tr><td>Fontane</td><td>Edition</td><td>SUB/GWDG*</td></tr>
<tr><td>GloDiv</td><td>Interviews</td><td>Repository</td><td>MMG/GWDG*</td><td>Zugangsmodell, Rechte</td></tr>
<tr><td>Berliner Klassik</td><td>Datenbank</td><td>Generischer Viewer</td><td>BBAW/ZIB</td><td rowspan="2">Übernahme System (Daten + Präsentation)</td></tr>
<tr><td>Global Migration Flows</td><td>Visualisierung</td><td>Hosting</td><td>MMG/GWDG</td></tr>
</table>

Abbildung 13: Überblick über die Prototypen, die kooperativ zwischen TP1, TP2 und TP3 erarbeitet werden. Prototypen decken einen möglichst breiten Bereich aus unterschiedlichen Forschungsdatentypen und Angebotsformen ab. (*) Anwendung läuft bereits bei der GWDG. Wo dies nicht der Fall ist, wird der Prototyp nur konzeptionell durchgesprochen, aber nicht technisch umgesetzt.

9 Anhang III: Von Forschungsdatentypen zum Angebot

HUMANITIES DATA CENTRE

HDC

Über Forschungsdaten-typen zum Angebot

Andreas Aschenbrenner, BBAW Berlin
Sven Bingert, GWDG Göttingen
Stefan Buddenbohm, MPI MMG Göttingen
Claudia Engelhardt, SUB Göttingen
Elias Oltmanns, ZIB Berlin
Ulrike Wuttke, AdW Göttingen

1. Warum ein Forschungsdatenzentrum für die Geisteswissenschaften?

Die Nachnutzung von Forschungsdaten gewinnt zunehmend an Bedeutung und Wissenschaftler dringen auf anwenderfreundliche Dienste. Anders als bei Publikationsrepositorien gibt es für geisteswissenschaftliche Forschungsdatenzentren (FDZ) noch keine etablierten Standards, die der Heterogenität der Daten und Methoden gerecht werden, gerade auch in Verbindung mit der langfristigen Archivierung und Bereitstellung.
In der Designphase des Humanities Data Centre (HDC) werden daher die Grundlagen für den Aufbau eines solchen FDZ geschaffen. Ein wesentlicher Bestandteil ist die Angebotsdefinition für die Nutzer, gespiegelt in einem Entwicklungskonzept für die konkrete technische Infrastruktur.

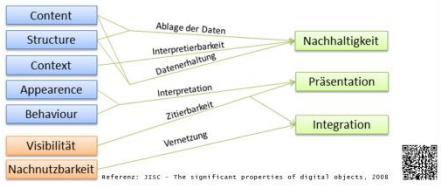
2. Prämissen für Datenmodelle in einem FDZ

Anders als bei Publikationsrepositorien genügen einfache, objektorientierte Datenmodelle bei einem Großteil der geisteswissenschaftlichen Forschungsdatentypen (FT) nicht. FDZ haben es mit Sammlungen verschiedener Objekte unterschiedlicher Formate zu tun. Von großer Bedeutung sind häufig auch die Beziehungen der Objekte zueinander oder umgebungsabhängige Darstellungen (bspw. Datenbanken oder Visualisierungen). Die signifikanten Eigenschaften von FT sind individuell abhängig von Forschungsfrage und -methode. In der Regel ist nur ein administrativer Kern einfach standardisierbar.

3. Significant Properties und Serviceklassen

Welche FT sind typisch für die Geisteswissenschaften? Wie kann ein Infrastrukturangebot konzipiert werden, das diesen Gegebenheiten gerecht wird, gleichzeitig aber die Realisierbarkeit (Ressourceneinsatz) im Blick behält?

Ausgehend von Diskussionen mit Wissenshaftlern und umfangreichen Sekundäranalysen von Umfragen zum Thema wurde entlang des Significant Properties-Konzepts ein Mapping der wissenschaftlichen Anforderungen auf Serviceklassen durchgeführt. Das Mapping in seiner Granularität angepasst werden. Für das HDC wird von drei basalen Serviceklassen ausgegangen, die die technische Infrastruktur strukturieren werden.



5. Serviceklassen für Dienste und Objekte

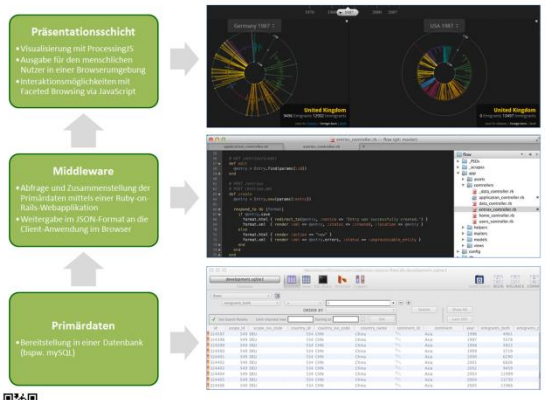
Entlang der Kontinuen Integration und Präsentation können verschiedene Dienste angesiedelt werden. Der Fokus liegt hier auf Nachnutzung und Anwendungen.
Die dafür notwendige Basis wird durch die Nachhaltigkeitsschicht gebildet, die auf Objektebene archiviert.

Ein Archivierungsvorhaben wird i.d.R. alle drei Serviceklassen (in unterschiedlicher Ausprägung und zeitlicher Abfolge) berühren.

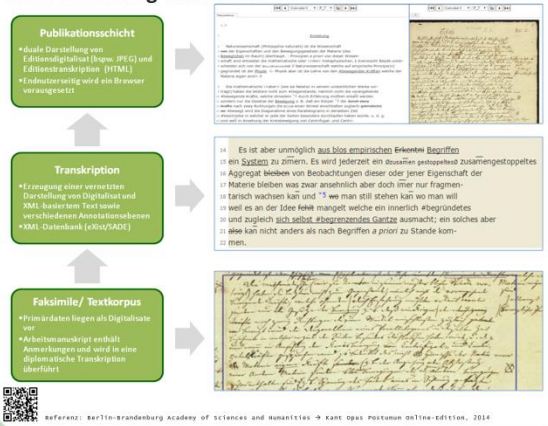
4a. Forschungsdatentypen und Angebotsdefinition

Wie lassen sich im nächsten Schritt die technischen Anforderungen an eine Infrastruktur ableiten? Welche Angebote und Abläufe sind zu definieren?
In der HDC-Designphase spielen für diesen Prozess die exemplarischen FT eine wesentliche Rolle. Dabei handelt es sich um repräsentative Praxisbeispiele aus den Geisteswissenschaften.
Beispielhaft und vereinfacht dargestellt sind hier die FT „Interaktive Visualisierung“ und „Kritische Edition“ anhand zweier Projekte aus dem Konsortialumfeld. So könnte ein FDZ bspw. für interaktive Visualisierungen (4b.) einen Dienst bereitstellen, der die Anwendung zu einem Übergabezeitpunkt „einfriert“ und zur Nachnutzung bereitstellt (Freezing). Für die Archivierung von Editionsprojekten (4c.) wäre ein generischer Viewer ein nützliches Angebot. Im Vordergrund steht immer – unter Beachtung der zur Verfügung stehenden Ressourcen – die einfache Nachnutzbarkeit für die Wissenschaftler.

4b. Archivierung von interaktiven Visualisierungen



4c. Archivierung von Editionen

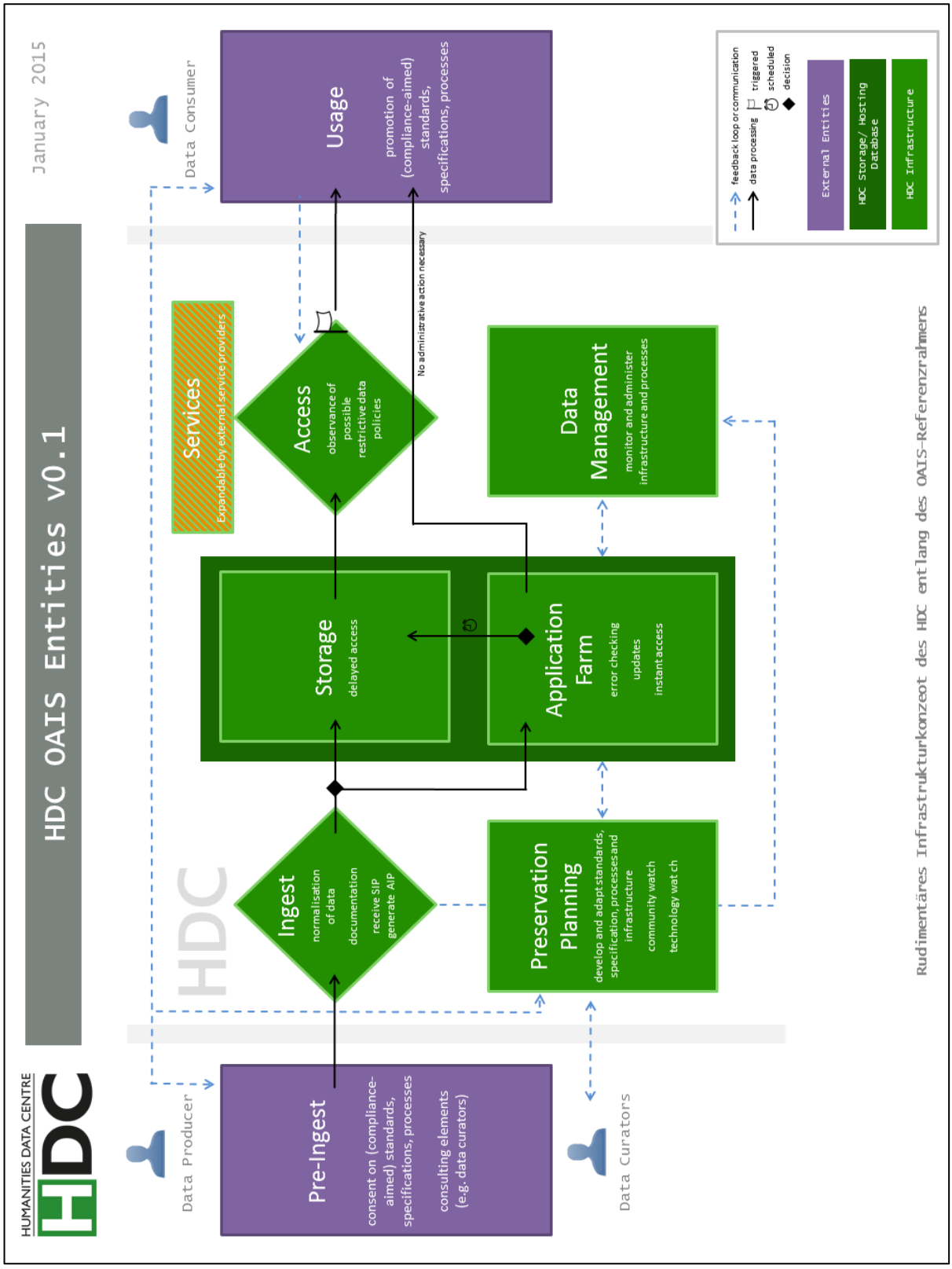


Gefördert von:
Niedersächsisches Ministerium für Wissenschaft und Kultur

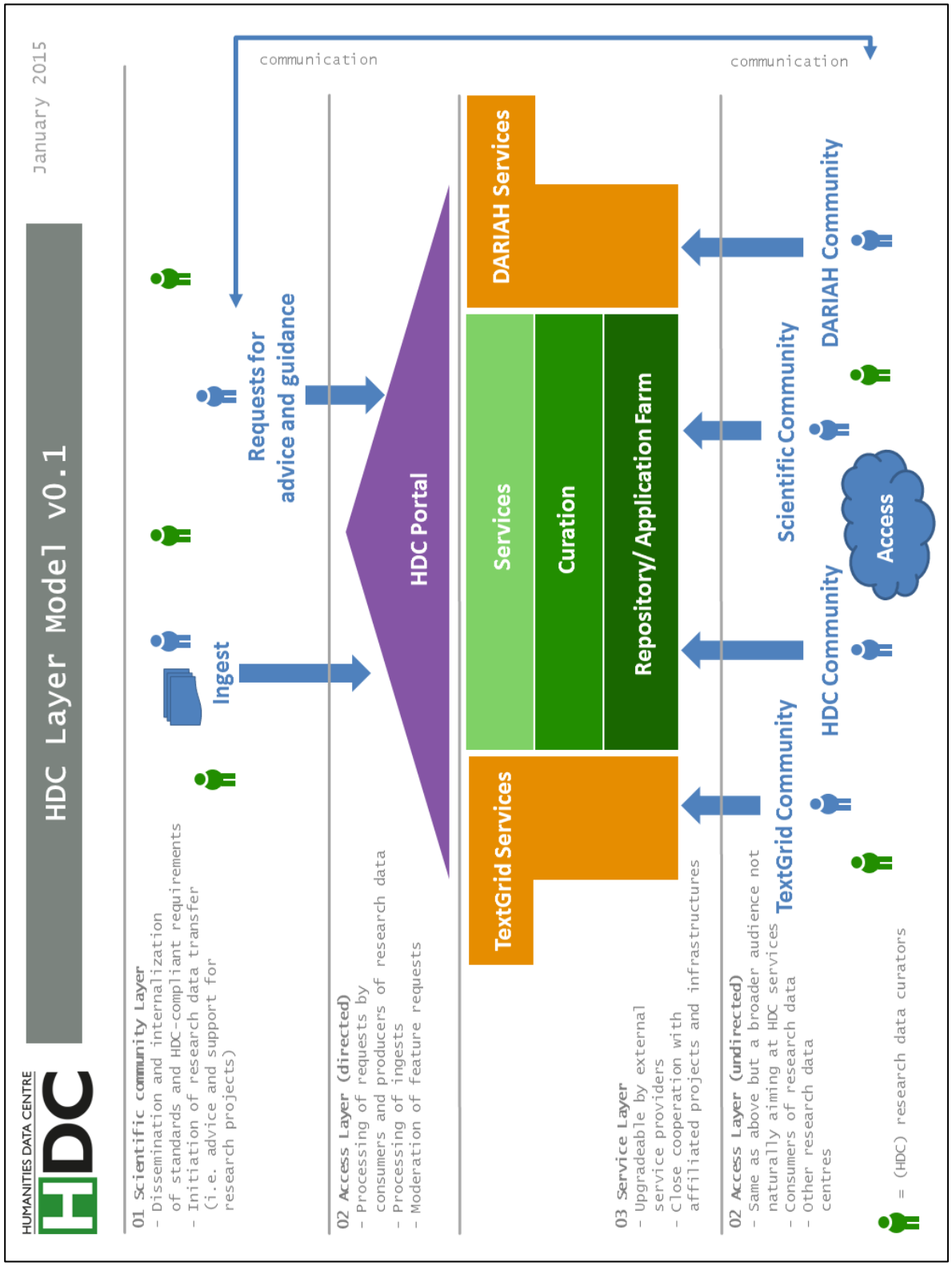
Konsortium
GWDG
mpi mmg
SUB Niedersächsische Staats- und Universitätsbibliothek Göttingen
ZIB

DHd 2015 „Von Daten zu Erkenntnissen – Digitale Geisteswissenschaften als Mittler zwischen Information und Interpretation“, Graz, 23.–27. Februar 2015

10 Anhang IV: OAIS-konforme HDC-Struktur



11 Anhang V: Zielgruppen des HDC



12 Anhang VI: Mögliches Organisationsmodell