

# Was sind Ihre Forschungsdaten?

Interviews mit Wissenschaftlern der Humboldt-Universität zu Berlin



### Zitierungsvorschlag:

Simukovic, Elena; Thiele, Raphael; Struck, Alexander; Kindling, Maxi; Schirnbacher, Peter (2014): Was sind Ihre Forschungsdaten? Interviews mit Wissenschaftlern der Humboldt-Universität zu Berlin. Bericht, Version 1.0. Online verfügbar unter: [urn:nbn:de:kobv:11-100224755](https://nbn-resolving.org/urn:nbn:de:kobv:11-100224755)

Computer- und Medienservice  
Institut für Bibliotheks- und Informationswissenschaft  
Humboldt-Universität zu Berlin

### Schlagwörter:

Forschungsdaten, Forschungsdatenmanagement, Umfrage, Interview

### Autoren<sup>1</sup>:

Konzeption, Durchführung und Auswertung von Interviews	Elena Simukovic
Erster Entwurf des Berichts	Elena Simukovic, Raphael Thiele
Visualisierung mit R / Text mining	Alexander Struck
Aktualisierung des Berichts	Elena Simukovic
Revision des Berichts	Maxi Kindling, Peter Schirnbacher
Finale Version	Elena Simukovic, Maxi Kindling

Die Autoren bedanken sich bei allen Interviewpartnern für ihre Bereitschaft und Teilnahme an den Interviews. Für hilfreiche Hinweise sind wir Frau Ulrike Schenk und für eine rechtliche Hilfestellung Herrn Thomas Hartmann dankbar.

Dieser Bericht setzt die Umfrage zum Umgang mit Forschungsdaten an der Humboldt-Universität zu Berlin<sup>2</sup> fort und ist auf Grundlage von Interviews mit den Wissenschaftlern der Universität entstanden. Zur besseren Lesbarkeit wird im Bericht ausschließlich die männliche Form verwendet. Es sollen sich jedoch beide Geschlechter gleichermaßen angesprochen fühlen.



Dieses Werk bzw. Inhalt steht unter einer [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/).

1 Im Einklang mit „San Francisco Declaration on Research Assessment (DORA)“ werden individuelle Beiträge von Autoren detailliert aufgeführt. Zur Entwicklung einer standardisierten Taxonomie für Rollen der Mitwirkenden (eng. *contributor role taxonomy*) siehe auch Allen et al. (2014).

2 Siehe <https://edoc.hu-berlin.de/docviews/abstract.php?lang=ger&id=40341>

# Inhaltsverzeichnis

Zusammenfassung.....	4
Executive summary.....	5
Vorwort.....	6
1 Einleitung .....	6
1.1 Motivation.....	6
1.2 Hintergrund.....	6
2 Methodik.....	7
2.1 Durchführung der Interviews.....	7
2.1 Visualisierung mit R.....	7
3 Ergebnisse.....	8
3.1 Interviewpartner.....	8
3.2 Arbeitsweisen und Typen von Forschungsdaten.....	8
3.3 Metadaten, Dokumentation, Beschreibung.....	9
3.4 Speicherung, Sicherung, Archivierung.....	9
3.5 Nachnutzung und Veröffentlichung.....	10
3.6 Anforderungen an Serviceleistungen.....	10
4 Diskussion.....	12
5 Literaturverzeichnis.....	15
6 Auflistung der Worthäufigkeiten.....	17
7 Anhänge.....	19
7.1 Interview I: Klinische Psychologie.....	19
7.2 Interview II: Experimentelle Physik.....	21
7.3 Interview III: Bodenkunde und Geomorphologie.....	23
7.4 Interview IV: Geomatik.....	25
7.5 Interview V: Biologische Psychologie.....	27
7.6 Interview VI: Systemimmunologie.....	29
7.7 Interview VII: Züchtungsbiologie.....	31
7.8 Interview VIII: Theoretische Biologie.....	33
7.9 Interview IX: Zeitgenössische Kunstgeschichte.....	35
7.10 Interview X: Deutsche Literatur.....	37
7.11 Interview XI: Kunstgeschichte und Visualisierung.....	39
7.12 Interview XII: Bodenkunde und Standortlehre.....	41
7.13 Interview XIII: Gartenökonomie.....	43
7.14 Interview XIV: Völkerrecht.....	45
7.15 Interview XV: Information Retrieval.....	46
7.16 Interview XVI: Mittelalterliche Geschichte.....	48
7.17 Interview XVII: Steuerlehre.....	50

## Zusammenfassung

Im Zeitraum vom Juli 2013 bis Juli 2014 fanden 17 Interviews mit Wissenschaftlern der Humboldt-Universität zu Berlin zum Thema Forschungsdaten statt. Das Ziel der Interviews war, tiefere Einblicke in den Umgang mit Forschungsdaten in verschiedenen Fachbereichen zu gewinnen und offene Fragen aus der vorangegangenen Umfrage zu beantworten. Zu diesem Zweck wurde ein spezieller Fragebogen erstellt, der als ein Leitfaden für halbstrukturierte Interviews diente. Der vorliegende Bericht stellt die Ergebnisse der Interviews vor.

Die Typen von Forschungsdaten haben sich wiederholt als sehr unterschiedlich gezeigt. Diese reichen von historischen Quellen wie Inkunabeldrucken oder Briefen aus dem 19. Jahrhundert über Bilder aus einer Satellitenmission bis zu Hochdurchsatzdaten zur Erkennung von Antikörpern durch Moleküle. Dementsprechend große Diversität herrscht auch bei den verwendeten Methoden und Datenformaten. Zur Dokumentierung von Forschungsdaten werden oft Notizen nach eigenständig entwickeltem Muster gemacht. Zu finden sind aber auch Community-Standards wie Gene Ontology oder das Europeana Data Model.

Hinsichtlich der langfristigen Archivierung von Forschungsdaten zeichnet sich ebenfalls ein heterogenes Bild ab. Während die Sicherung von Rohdaten und prozessierten Daten im laufenden Betrieb auf mehreren Speichermedien erfolgt, werden oft nur Endergebnisse langfristig archiviert. Zum Einsatz kommen diverse Datenträger wie CDs, DVDs, Festplatten, USB-Sticks, lokale Laufwerke, Server, NAS und verschiedene Kombinationen von diesen. Ein institutionelles Repositorium bzw. Archiv muss aus Sicht mehrerer Interviewpartner zentral angeboten werden, um die Verfügbarkeit der Forschungsdaten auch über zehn Jahre hinaus (wie von den Regeln guter wissenschaftlicher Praxis gefordert) garantieren zu können.

Weiterhin besteht ein großer Bedarf an einem akademischen Online-Speicher, um die Forschungsdaten ortsunabhängig mit Kooperationspartnern aus anderen Institutionen austauschen zu können. Eine Kontaktstelle für Fragen rund um den Umgang mit Forschungsdaten und entsprechende Schulungen zur effizienten Gestaltung des eigenen Forschungsdatenmanagements werden ebenso als sehr vorteilhaft angesehen. Die strategische Bedeutung des Themas wird erkannt, was den notwendigen Ausbau der technischen und personellen Infrastruktur zur Folge haben sollte. Insgesamt lässt sich zusammenfassen, dass die Interviews die Umfrageergebnisse bestätigt haben.

## Executive summary

During the period from July 2013 to July 2014, 17 researchers of the Humboldt-Universität zu Berlin were interviewed about research data topics. This aimed at gathering an in-depth look into different ways of handling research data in various research areas as well as to answer open questions remaining from the preceding online survey.<sup>3</sup> For this purpose, a special questionnaire was developed for a guided semi-structured interview. The present report summarizes the results of the interviews.

The types of research data appeared again as highly diverse. They range from historic sources such as incunabula or correspondence from the 19<sup>th</sup> century to satellite imagery to high-throughput screening to identify antibodies in molecules. The methods and data formats used are accordingly very diverse. Customized templates are used to take notes of data and processes. There are also community standards such as Gene Ontology or Europeana Data Model in use.

Long-term preservation of research data is organized heterogeneously. There is a sharp distinction between storage and backup of raw data and active data during a research project phase and preserving its final outcomes. Different storage media are employed such as CDs, DVDs, HDD, USB sticks, local drives, servers, network attached storage and various combinations of these. From the interviewees' point of view, an institutional repository or archive must be provided centrally. This should ensure accessibility of research data over and above ten years as required by the rules of good scientific practice.

Furthermore, there is a need for an academic online storage cloud that could be used for exchanging research data with cooperation partners based at other institutions. A contact point around research data management and related training to effective planning of personal research data management are favoured. The strategic importance of this topic was recognized, therefore further development of technical and staff infrastructure is required. In summary, it can be stated that outcomes of the interviews are in line with the results of the online survey.

---

<sup>3</sup> See Simukovic, Elena et al. (2013). Humboldt-Universität zu Berlin Research Data Management Survey Results. ZENODO. DOI: [10.5281/zenodo.7448](https://doi.org/10.5281/zenodo.7448)

# Vorwort

Im Frühjahr 2013 wurde an der Humboldt-Universität zu Berlin (HU) eine universitätsweite Umfrage zum Forschungsdatenmanagement durchgeführt. Ein ausführlicher Bericht gibt einen Überblick über den aktuellen Stand zum Umgang mit Forschungsdaten und die Anforderungen an zentrale Serviceleistungen.<sup>4</sup> Teilnehmer der Umfrage hatten die Möglichkeit, sich anschließend für ein persönliches Interview anzumelden. Der vorliegende Bericht fasst die Ergebnisse aus 17 Interviews zusammen und verbindet diese mit bisherigen Erkenntnissen.

## 1 Einleitung

### 1.1 Motivation

Die Umfrage zum Umgang mit digitalen Forschungsdaten an der HU hatte zum Ziel, einen ersten Überblick über die an der HU produzierten bzw. verwendeten Forschungsdaten und den aktuellen Stand im Umgang mit diesen zu geben. Die Ergebnisse dienten als Quelle für informierte Entscheidungen in der Planung von weiteren Schritten hin zum institutionellen Serviceangebot für Forschungsdatenmanagement. Dabei setzte allerdings die Art und Weise einer universitätsweiten Online-Umfrage einige Grenzen hinsichtlich der Formulierung der Fragen und Antwortoptionen. So wurden aus Gründen der Praktikabilität beispielsweise nur die Typen von Forschungsdaten (z.B. Bilder, Audio-Aufzeichnungen oder GIS-Daten) abgefragt und auf die Auflistung von konkreten Dateiformaten und verwendeter Software verzichtet.

Trotz zahlreicher Freitext-Kommentare in der Umfrage sind einige Detailfragen offen geblieben. Die Durchführung der Interviews bot daher eine Möglichkeit, tiefere Einblicke in Arbeitsweisen verschiedener Wissenschaftsdisziplinen und Fachbereiche zu bekommen und bestehende Wissenslücken möglichst zu schließen.

### 1.2 Hintergrund

Seit der Durchführung der Umfrage und der Veröffentlichung des dazugehörigen Berichts im Oktober 2013 haben einige in diesem Kontext bedeutsame Ereignisse stattgefunden. Im Folgenden soll auf die drei wichtigsten eingegangen werden.

#### ***Open Research Data Pilot in Horizon 2020***

Im Dezember 2013 hat die Europäische Kommission Leitfäden für Open Access zu wissenschaftlichen Publikationen und Forschungsdaten im neuen Rahmenprogramm für Forschung und Innovation „Horizon 2020“ herausgebracht.<sup>5</sup> Von besonderer Bedeutung für das Forschungsdatenmanagement ist dabei das „Open Research Data Pilot“, demnach bewilligte Projekte in zunächst sieben Förderbereichen zusätzliche Anforderungen an Zugänglichmachung von Forschungsdaten erfüllen müssen. Neu ist zudem der Datenmanagementplan (DMP), der bei der Beantragung der Mittel mit eingereicht werden muss und auch in die Bewertung einfließt.

#### ***Empfehlung der Hochschulrektorenkonferenz***

Im Mai 2014 hat die Hochschulrektorenkonferenz (HRK) eine Empfehlung der 16. Mitgliederversammlung der HRK „Management von Forschungsdaten – eine zentrale strategische

---

4 Der Umfragebericht ist online zugänglich unter [urn:nbn:de:kobv:11-100213001](http://nbn-resolving.org/urn:nbn:de:kobv:11-100213001)

5 Für nähere Informationen siehe European Commission (2013a, 2013b).

Herausforderung für Hochschulleitungen“ veröffentlicht. Darin werden die Hochschulleitungen aufgefordert, Leitlinien zum Umgang mit digitalen Forschungsdaten abzustimmen, über die Grenzen der Hochschule hinweg zu kooperieren, die Informationskompetenz der Hochschulmitglieder zu stärken und die institutionellen Infrastrukturen zum Forschungsdatenmanagement auszubauen.<sup>6</sup>

### ***Forschungsdaten-Policy der HU***

In Übereinstimmung mit der Empfehlung der HRK und dem Beispiel der Hochschulen im angelsächsischen Raum folgend hat der Akademische Senat der HU im Juli 2014 die „Grundsätze zum Umgang mit Forschungsdaten an der Humboldt-Universität zu Berlin“ beschlossen. In vier formulierten Grundsätzen werden alle forschenden HU-Angehörigen aufgefordert, die in ihrer wissenschaftlichen Tätigkeit entstehenden Forschungsdaten angemessen aufzubereiten, zu dokumentieren und langfristig aufzubewahren, sowie nach Möglichkeit öffentlich zugänglich zu machen. Die Grundsätze werden durch praktische Handlungsempfehlungen ergänzt. Die HU hat sich verpflichtet, die Voraussetzungen für die Erfüllung der Grundsätze zu schaffen.<sup>7</sup>

## **2 Methodik**

Im folgenden Kapitel wird die Durchführung und Auswertung der Interviews näher beschrieben.

### ***2.1 Durchführung der Interviews***

Zur Durchführung der Interviews wurde ein spezieller Fragebogen mit acht offenen Fragen entwickelt. Der Fragebogen diente als Leitfaden für ein halbstrukturiertes Interview, wobei die Reihenfolge der Fragestellung im Gesprächsverlauf variierte. Diese Form ermöglichte es, auch weitere Aspekte über die acht genannten Fragen hinaus zu besprechen, die Besonderheiten in jedem Fachbereich aufzuspüren und gleichzeitig vergleichbare Antworten zu erhalten.<sup>8</sup>

Zur Teilnahme an einem Interview wurden Wissenschaftler der HU eingeladen, die ein entsprechendes Interesse in der Umfrage bekundet haben (s. Vorwort). Der Fragebogen wurde in Vorbereitung auf das jeweilige Interview dem Interviewpartner im Vorfeld zur Verfügung gestellt.

Die meisten Interviews fanden am Arbeitsplatz des Interviewpartners statt und dauerten im Durchschnitt 60-80 Minuten. Zum Zweck der Transkribierung wurden sie aufgezeichnet. Anschließend wurden die Antworten aus dem jeweiligen Interview im Fragebogen zusammengefasst. Der ausgefüllte Fragebogen wurde vom Interviewpartner kontrolliert und bei Bedarf ergänzt und/oder aktualisiert. Die Ergebnisse befinden sich am Anhang.

### ***2.1 Visualisierung mit R***

Ein Transkript umfasste durchschnittlich 5-8 Seiten Text, woraus sich eine Gesamtzahl von 95 Seiten bzw. eine Hochrechnung von ca. 57.000 Wörtern ergeben hat.<sup>9</sup> Der Umfang des

---

6 Für nähere Informationen siehe Hochschulrektorenkonferenz (2014).

7 Die Grundsätze und die Handlungsempfehlungen zum Umgang mit Forschungsdaten an der Humboldt-Universität zu Berlin sind verfügbar unter <http://www.cms.hu-berlin.de/dataman/policy>

8 Zur Methodik eines halbstrukturierten Interviews siehe beispielsweise Bock (1992): „Beim halbstrukturierten-leitfadenorientierten Tiefeninterview wird der Kompromiß zwischen z.T. vorgegebenen Fragen und dem Erzählenlassen, d.h. dem flexiblen Eingehen auf nicht-antizipierte Äußerungen der Befragten gesucht, um sowohl Reichweite als auch Tiefe des Themas abzudecken und um vielfältiges und vergleichbares Material zu erhalten.“ (ebd., S. 94).

9 Aus technischen Gründen wurden 15 von 17 Interviews transkribiert.

vorliegenden Textmaterials legte die Vermutung nahe, dass sich mithilfe von Text Mining-Methoden weitere Einsichten gewinnen lassen können. Ein erster Versuch wird auf der Titelseite des vorliegenden Berichts als eine Schlagwortwolke („word cloud“)<sup>10</sup> veranschaulicht.

Zur Erstellung der Word Cloud wurden zunächst einzelne Wörter mithilfe des R Programms in RStudio-Umgebung<sup>11</sup> nach deren Häufigkeit aufgezählt. Dies ergab 4261 unterschiedliche Einzelwörter, die vorab normalisiert werden mussten: Substantive in Plural wurden weitgehend in Singular gesetzt (z.B. „Projekte“ → „Projekt“), Verben und Adjektive in deren Grundform umgewandelt (z.B. „publiziert“ → „publizieren“, „klinischen“ → „klinisch“). Anschließend wurde eine Stoppwortliste erstellt, um die häufig verwendeten aber an sich nicht informativen Wörter auszufiltern (z.B. „bestimmt“, „dadurch“, „sollen“). Aus den verbleibenden 200 am häufigsten benutzten Wörtern (siehe Kapitel 6) wurde eine Word Cloud generiert. Der dafür verwendete Code in R ist öffentlich zugänglich.<sup>12</sup>

Die Auswertung des Textkorpus nach Wortähnlichkeiten zwischen den Fachbereichen stellt eine weitere interessante Forschungsfrage dar. So könnten beispielsweise mithilfe einer Abstandsmatrix oder eines Clusteralgorithmus weitere, bisher ungesehene Zusammenhänge identifiziert werden. Eine solche Visualisierung ist allerdings mit einem großen Zeitaufwand verbunden und wird derzeit eingehend untersucht.

### **3 Ergebnisse**

Im Folgenden werden Ergebnisse der Interviews zu den einzelnen Fragen kurz vorgestellt.

#### **3.1 Interviewpartner**

Im Zeitraum vom Juli 2013 bis Juli 2014 fanden 17 Interviews mit Wissenschaftlern aus den geistes- und sozialwissenschaftlichen aber auch aus natur- und lebenswissenschaftlichen Fakultäten und Instituten statt. So wurden Fachwissenschaftler aus den Bereichen experimentelle Physik, Bodenkunde, klinische, biologische und theoretische Psychologie, Geomatik, Systemimmunologie, Züchtungsbiologie, Kunstgeschichte, deutsche Literatur, Gartenökonomie, Völkerrecht, Information Retrieval, mittelalterliche Geschichte und Steuerlehre befragt. Darunter waren sowohl wissenschaftliche Mitarbeiter aus dem akademischen Mittelbau als auch Professoren vertreten. Die Breite des Spektrums und der aktuellen Forschungsfragen lässt sich in den einzelnen Interviews im Anhang (Frage 1) nachlesen.

#### **3.2 Arbeitsweisen und Typen von Forschungsdaten**

Verschiedenen Forschungsfragen zufolge zeichneten sich die Arbeitsweisen in Erhebung bzw. Bearbeitung von Forschungsdaten und deren Typen durch eine große Diversität aus. Mehrfache Nennungen waren bei gängigen Office-Anwendungen, statistischer Auswertung (SPSS, R, SAS, Matlab), Interviews (Audio-Aufzeichnungen, Transkripte), Fragebögen, Patientendaten, GIS-Daten (ArcGIS), Bildern (TIFF, JPEG), Annotationen, fMRT-Messungen (Brain Vision Analyser) zu treffen. Bemerkenswert ist auch, dass ein großer Teil der Befragten explizit die Nutzung von etablierten Open Source-Produkten und eigenständig bzw. in der Community entwickelten Software genannt hat.

---

10 Vgl. <http://de.wikipedia.org/wiki/Schlagwortwolke>

11 Für nähere Informationen siehe <http://www.rstudio.com/>

12 Siehe Struck, Alexander; Simukovic, Elena (2014). Generating a word cloud for interview transcripts with R. ZENODO. DOI: [10.5281/zenodo.12842](https://doi.org/10.5281/zenodo.12842)

Erkennbar wurden dabei aber auch Fachbereiche, in denen keine eigenen Primärdaten erhoben werden und hauptsächlich oder ausschließlich Sekundärdaten und historische Quellen nachgenutzt werden (z.B. Völkerrecht, Mittelalterliche Geschichte, Steuerlehre). Die Forschungsergebnisse mitsamt der Referenzen zu den zugrundeliegenden Quellen werden üblicherweise in wissenschaftlichen Publikationen dokumentiert. Dieser Unterschied zeigte sich in Folge als ein maßgeblicher Umstand, der urheberrechtliche Implikationen mit sich bringt und die Nachnutzungsmöglichkeiten der Forschungsergebnisse nachhaltig beeinflusst. Die vollständigen Beschreibungen von Forschungsdaten im jeweiligen Fachbereich, deren Typen, Formaten und verwendeter Software lassen sich im Anhang (Frage 2 und 3) nachlesen.

### **3.3 Metadaten, Dokumentation, Beschreibung**

In Bezug auf Metadaten und die Dokumentation von Forschungsdaten ließ sich eine ausgeprägte Tendenz erkennen, Metadatenstrukturen an Bedürfnisse im konkreten Forschungsprojekt oder für eine bestimmte Forschungsfrage individuell anzupassen. So wurde eine strukturierte Erfassung von Notizen zum Inhalt oder durchgeführten Vorgängen nach einem eigenständig entwickelten Muster in einer Textdatei mehrfach genannt. Oft wurden dafür einzelne Subsets von großen, umfangreichen Standards übernommen bzw. für spezielle Ansprüche weiterentwickelt. Zu finden sind aber auch etablierte Standards in der Fachcommunity wie z.B. die Bodenkundliche Kartieranleitung zur normierten Beschreibung und Klassifizierung von Böden oder Gene Ontology für eine standardisierte Repräsentation und Annotation von Genen.

In manchen Fachbereichen (z.B. Kunstgeschichte) gibt es zudem eine Tradition, in welcher Reihenfolge die Angaben in der Bildunterschrift gemacht werden. Im Fall der Sekundärnutzung von durch spezielle Datenzentren bereitgestellten Daten (z.B. Statistikdaten) werden diese oft bereits mit der zugehörigen Dokumentation geliefert. Erwähnung fand auch eine direkte Implementierung von Metadatenfeldern in der zur Bearbeitung von Forschungsdaten verwendeten Software. Bibliographische Angaben werden zudem mithilfe von Literaturverwaltungsprogrammen erfasst. Alle Nennungen zur Beschreibung von Forschungsdaten sind im Anhang (Frage 4) zu finden.

### **3.4 Speicherung, Sicherung, Archivierung**

Die weitere Frage bezog sich auf die langfristige Aufbewahrung von Forschungsdaten gemäß der Regeln guter wissenschaftlicher Praxis. Auffällig wurde dabei eine klare Trennung von Daten in Rohdaten, prozessierte Daten oder Zwischenergebnisse und Ergebnisdaten. Während die Rohdaten und deren Analysen für den laufenden Betrieb auf diversen Speichermedien gespeichert und gesichert werden, gelangen oft nur Endergebnisse zu einer langfristigen Archivierung. Zum Einsatz kommen dabei CDs, DVDs, USB-Sticks, lokale Laufwerke, externe Festplatten, Server, netzgebundene Speicher (NAS) und verschiedene Kombinationen von diesen. Sensible und datenschutzrechtlichen Anforderungen unterliegende Daten werden zudem teilweise in einem Stahlschrank oder einem Safe abgeschlossen (z.B. im Bereich klinische Psychologie oder Steuerlehre).

Weiterhin kam es mehrmals zum Ausdruck, dass die Archivierung aller Daten nicht praktikabel wäre, da die Rohdaten und Zwischenergebnisse durch mehrere Prozessierungsschritte oft sehr speicherintensiv und für die Nachvollziehbarkeit der publizierten Forschungsergebnisse nicht zwingend erforderlich sind. Gleichzeitig wurde angemerkt, dass die Aufbewahrung relevanter Forschungsdaten auch über längere Zeiträume hinaus (25 oder mehr Jahre) gewährleistet werden müsste. Zudem ist bei Sekundärnutzung keine zusätzliche Archivierung mehr nötig, weil diese Datenbestände von den sie liefernden Datenzentren bereits langfristig archiviert werden. Die für

die Forschungsfrage verwendeten Datensätze werden lediglich in der Publikation dokumentiert. Zum Teil werden zugrundeliegende Forschungsdaten als Supplemental Online Materials bei einer Fachzeitschrift zusammen mit dem Manuskript eingereicht.

Insgesamt lässt sich zusammenfassen, dass die Art und Weise der Aufbewahrung und Archivierung von Forschungsdaten sich von Fall zu Fall unterscheidet und an individuelle Praktiken in Arbeitsgruppen oder von einzelnen Personen gebunden ist. Gleichzeitig wird darauf verwiesen, dass klare und verbindliche Regelungen dazu oft nicht vorhanden sind. Die Antworten der Interviewpartner finden sich im Anhang (Frage 5).

### **3.5 Nachnutzung und Veröffentlichung**

Grundsätzlich lässt sich die Nachnutzung von Forschungsdaten in zwei Perspektiven unterscheiden: Nachnutzung von Forschungsdaten anderer Forscher oder Forschungsprojekte für eigene Forschungszwecke und Nachnutzung eigener Forschungsdaten durch andere Forscher. Das Erstere wird durch öffentlich verfügbare Datenquellen wie Satellitenbilder veranschaulicht (z.B. in Geomatik). Öffentliche Zugänglichmachung von Forschungsdaten in einer eigenständig entwickelten oder bereits existierenden Datenbank stellt dabei die letztere Nachnutzungsart dar (z.B. Systemimmunologie oder Züchtungsbiologie). Allerdings wird in solchen Fällen nicht die Zitierung des Datensatzes selbst, sondern des dazugehörigen Artikels empfohlen (z.B. Systemimmunologie, Information Retrieval).

Zudem zeigte sich eine informelle Kultur des Datenaustausches basierend auf persönlichen Kontakten innerhalb der Community oder eines Verbundprojekts (siehe z.B. Gartenökonomie, Information Retrieval, theoretische Biologie). Hingewiesen wurde aber auch auf Einschränkungen einer Veröffentlichung von Forschungsdaten wie die Richtlinien des Datengebers oder der Kollaboration, die die Weitergabe von Daten untersagen können (z.B. Experimentelle Physik oder Steuerlehre).

Manche Fachzeitschriften verlangen oder empfehlen es, zugrundeliegende Forschungsdaten zugänglich zu machen (z.B. Klinische Psychologie). Einige Versuche, konkrete Forschungsergebnisse anhand der publizierten Information zu reproduzieren, haben sich jedoch aufgrund der unzureichenden Dokumentation als problematisch erwiesen (siehe z.B. Theoretische Biologie oder Bodenkunde). Limitierte Darstellungsmöglichkeiten von Daten in einem wissenschaftlichen Artikel wurden aber auch als einer der Gründe zur Bevorzugung von fachspezifischen Repository-Lösungen genannt. Dadurch kann unter anderem eine standardisierte Grundannotation aller Datensätze sichergestellt werden (z.B. Züchtungsbiologie).

Selbst programmierte Software oder Code wird in der Entwickler-Community gewöhnlich auf GitHub bereitgestellt (siehe z.B. Theoretische Biologie, Information Retrieval). Eine weitere Nachnutzungsmöglichkeit schließt die Lehrzwecke ein (z.B. Klinische Psychologie oder deutsche Literatur). Alle Beispiele sind im Anhang (Frage 6) dokumentiert.

### **3.6 Anforderungen an Serviceleistungen**

Mit Blick auf notwendige Serviceleistungen werden Anforderungen sowohl an eine technische als auch an eine personelle Infrastruktur gestellt. So wird eine feste Kontaktstelle oder ein Servicezentrum rund um Fragen zu Forschungsdaten und Publikationen gefordert. Gleichzeitig sind solche Fragen mit einem hohen Grad an Fachspezifik verbunden, weshalb immer wieder hinterfragt wurde, welche Serviceleistungen zentral und welche doch lokal an Instituten oder Arbeitsgruppen angeboten werden sollen. Eng damit verknüpft ist eine rechtliche Beratung, die

bereits am Anfang bzw. in der Planungsphase eines Forschungsprojekts notwendig ist und durchgehend verfügbar sein muss. Insbesondere Fragen zu Copyright, Bildrechten und konkreten Verantwortlichkeiten bedürfen einer Klärung.

Das Beratungsangebot soll idealerweise Hand in Hand mit der technischen Umsetzung gehen. Eine Lösung zur Langzeitarchivierung von Forschungsdaten umfasst daher ein professionell aufgebautes Archiv inkl. Metadatenerschließung für eine optimale Sichtbarkeit und Auffindbarkeit der Inhalte. Ein solches Archiv soll es ermöglichen, sowohl die Daten mit geschütztem Zugriff sicher abzulegen (als sog. „Dark Archive“), als auch diese bei Bedarf öffentlich zugänglich zu machen, um beispielsweise individuelle Anfragen nicht einzeln beantworten zu müssen oder wenn dies von Journals verlangt wird. Ein institutionelles Online-Repository muss aus der Sicht mehrerer Interviewpartner unbedingt zentral angeboten werden. Als Hauptgrund dafür werden kurzfristige Förderung von Forschungsprojekten und häufiger Wechsel zwischen den Institutionen innerhalb der laut Regeln guter wissenschaftlicher Praxis vorgeschriebenen 10-Jahres-Frist genannt.

Um den Datenaustausch mit Kooperationspartnern zu unterstützen, wird eine „akademische Alternative“ zu kommerziellen Anbietern wie Dropbox oder Google benötigt. Ein entscheidendes Kriterium für einen akademischen Online-Speicher ist, dass auch sensible Daten dort sicher ablegt und der Zugang den Kooperationspartnern aus externen Einrichtungen gewährt werden können. Die Möglichkeit, die Daten zwischen verschiedenen Geräten und Betriebssystemen zu synchronisieren und „von überall“ unterwegs damit zu arbeiten, bringt einen weiteren Vorteil.

Erwähnt wurde auch der Bedarf an mehr Speicherplatz und zentralem Backup von Forschungsdaten. Die Projektverantwortlichen waren zudem oft bereit, die Kosten für die dafür erforderlichen technischen und personellen Ressourcen in Projektanträge miteinzubeziehen. Eine Bezifferung der Kosten oder ein Leitfaden für notwendige Grundausstattung seitens des CMS wird dabei vorausgesetzt. Als eine optimale Gestaltung des technischen Supports wurde mehrmals der Wunsch geäußert, die Basis-Administration wie beispielsweise die Wartung eines Servers dem CMS zu überlassen, bei gleichzeitig flexiblen Konfigurationsmöglichkeiten für Arbeitsgruppen an den einzelnen Instituten.

Die strategische Bedeutung des Umgangs mit Forschungsdaten muss nach Auffassung einiger Interviewpartner auf der Leitungsebene der Universität verdeutlicht werden. Die Leiter der Arbeitsgruppen sollten ebenso sensibilisiert werden. Die Bewusstseinssteigerung hätte zur Folge, dass die benötigte technische und personelle Infrastruktur eher ausgebaut wird. Weiterhin wurde die Bedeutung von Schulungen und Weiterbildung hervorgehoben. Als Themenbereiche wurden exemplarisch die Organisation und Qualitätssicherung von Forschungsdaten, Versionierung oder Nutzung spezieller wissenschaftlicher Software genannt. Nicht zuletzt wurde zudem die Notwendigkeit einer adäquaten Methodenausbildung für Studenten thematisiert.

Manche Interviewpartner haben explizit angegeben, an der möglichst freien Verfügbarmachung ihrer Publikationen und Forschungsdaten interessiert zu sein. Eine solche Vorgehensweise soll den Fortschritt der wissenschaftlichen Erkenntnis insgesamt befördern und eine weitere Exploration von alten und neuen Forschungsfragen ermöglichen. Als vorteilhaft wurde ferner ein „Datenkatalog der HU“ vorgeschlagen, um interne Kooperationspartner zu finden und Synergieeffekte durch Know-How-Austausch zu schaffen. Grundsätzlich sollen solche Serviceleistungen stets nutzerfreundlich gestaltet werden und eine intuitive Bedienung bieten.

Vereinzelt waren auch Vorschläge für virtuelle Maschinen für Studenten, Latex-Support zur Migration von Dokumenten wie Doktorarbeiten in PDF/A-Format, elektronische Laborbücher oder eine zentrale Archivierung der E-Mailkommunikation zu treffen. Die Vielfalt aller Antworten lässt sich im Anhang (Frage 7 und 8) nachlesen.

## 4 Diskussion

In der Online-Umfrage unter den Wissenschaftlern der HU in Frühjahr 2013 sind Textdokumente als meistgewählte Antwortoption sowohl als Quellen für die Forschung wie auch als Datentyp, der während des Forschungsprozesses entsteht, gewählt worden.<sup>13</sup> Die Vermutung lag nahe, dass Textdokumente nicht nur als Forschungsobjekte selbst, sondern auch in Form von Publikationen als Grundlage der wissenschaftlichen Arbeit enthalten waren.

In den Interviews wurde eine Vielzahl verschiedener Forschungsdaten von XML-Dokumenten über Bilder hin zu Rohdaten aus SNP-Arrays als Grundlage wissenschaftlicher Arbeit genannt. Textdateien gehören zwar auch zu den Forschungsdaten, sie liegen aber ungefähr gleichauf mit anderen mehrfach erwähnten Typen. Eine Erklärung für diese Abweichung wird durch die Interviews selbst nahegelegt, denn Texte als Forschungsdaten werden hauptsächlich von Wissenschaftlern aus dem Bereich der Geistes- und Kulturwissenschaften erwähnt. Die Mehrheit der Interviewpartner forscht allerdings im Bereich der Natur- und Sozialwissenschaften, in dem Texte eine weniger prominente Position einnehmen. Ein weiterer Grund könnte sein, dass in einem Interview mit Rückfragen jederzeit interveniert werden konnte und Unklarheiten wie zum Stellenwert von Textdokumenten umgehend geklärt wurden.

Von besonderer Bedeutung bei der Auswertung der Interviews waren die Anforderungen an Serviceleistungen (Fragen 7 und 8). Die Fragen wurden bewusst sehr offen formuliert und gaben den Interviewpartnern die Möglichkeit, Anregungen und Desiderate basierend auf den eigenen Erfahrungen als Wissenschaftler, aber auch als Vertreter der Scientific Community ihres Fachbereichs zu formulieren. Zu der thematischen Diversität der Antworten tritt auch eine perspektivische: Manche Antworten heben auf das große Ganze ab und äußern eher allgemeine Wünsche und Anregungen in Bezug auf etwaige Serviceleistungen, andere dagegen formulieren sehr spezifische Wünsche und Vorstellungen.

In den Ergebnissen spiegeln sich einige Tendenzen wieder, die bereits in der Umfrage deutlich geworden waren. Als grundlegend muss hier die Bereitstellung von ausreichendem Speicherplatz bereits während des Forschungsprozesses angesehen werden, die zu den am häufigsten geäußerten Wünschen gehört. Die vorangegangene Umfrage hatte bereits gezeigt, dass die Forschungsdaten in einigen Bereichen und Disziplinen sehr umfangreich sind, jedoch im Durchschnitt zwischen 20 und 50 GB pro Forschungsprojekt liegen.<sup>14</sup> Allerdings wurde in den Interviews mehrmals bestätigt, dass die Datenmengen bereits jetzt und auch zukünftig wachsen werden. Manche Fachbereiche haben zudem den Datendurchsatz als ein mögliches Problemfeld identifiziert.

Ein weiterer Punkt ist die Vermittlung von Informationskompetenz. Als Adressaten dieses Services werden sowohl Lehrende als auch Studierende gesehen. Für aktiv forschende Wissenschaftler schweben den Interviewten hier vor allem Schulungen vor, die auf effiziente Gestaltung der Arbeitsabläufe bei der Speicherung, Archivierung und Veröffentlichung von Forschungsdaten abzielen. Dieser Wunsch steht offensichtlich in Zusammenhang mit dem Wunsch nach einem umfassenden technischen Support. Die Schulungen hätten idealerweise den Effekt, die Teilnehmer zur eigenständigen Verwaltung der Forschungsdaten anzuregen und könnten so dazu beitragen, den technischen Support zu entlasten. Allerdings besteht hier ein gewisser Zweifel, wo ein solcher Service in organisatorischer Hinsicht am besten zu verorten wäre – am jeweiligen Institut oder der Arbeitsgruppe oder einer zentralen Serviceeinrichtung wie dem CMS. Für die Studierenden, aber

---

13 S. Umfragebericht, S. 45

14 S. Umfragebericht, S. 17

auch die Nachwuchswissenschaftler wird vorgeschlagen, die neuen digitalen Arbeitsweisen in die methodische Ausbildung zu integrieren.

Im Hinblick auf die Schaffung einer Forschungsdateninfrastruktur wird auf eine benutzerfreundliche Oberfläche Wert gelegt, die eine intuitive Bedienung möglich macht. Dadurch könne ein Beitrag zur Vereinfachung wissenschaftlichen Arbeitens geleistet werden. Detailliertere Wünsche zielen auf Folgendes ab:

- die Möglichkeit der On-Site-Administration, durch die ein schneller Datentransfer und flexible Anpassungen gewährleistet werden sollen,
- Migration von Daten aus obsoleter Software oder Hardware,
- webbasierte Lösungen, die das Verlinken von Forschungsdaten mit den dazugehörigen Onlinepublikationen des Autors ermöglichen,
- Interaktionsmöglichkeiten, um beispielsweise Datensätze im Repositorium nicht nur abspeichern, sondern auch kommentieren zu können.

Die Interviewpartner wurden zudem explizit danach gefragt, ob die drei aus der Umfrage hervorgegangenen Spitzenpositionen für die gewünschten Serviceleistungen (Speicherplatz für Forschungsdaten, rechtliche Beratung und technische Beratung)<sup>15</sup> für sie zutreffen. Eine überwiegende Mehrheit bestätigte, dass durch diese drei Serviceleistungen die akutesten Bedarfe gedeckt werden könnten. Immer wieder wurde aber auch die Unterscheidung von Datentypen in Abhängigkeit von der jeweiligen Phase eines Forschungsprojekts thematisiert.<sup>16</sup> Für dynamische, noch zu bearbeitende Daten werden geschützte Speicherorte und Austauschsysteme zwischen Kooperationspartnern benötigt. Nach dem Abschluss des Projekts sollen statische Daten langfristig archiviert und/oder publiziert werden. Zu diesem Zweck ist eine umfassende Dokumentation und ggf. Verknüpfung mit den dazugehörigen Publikationen notwendig.

Interessant ist zudem, dass detaillierte und verbindliche Regelungen zum Umgang mit Forschungsdaten in bestimmten Situationen als hilfreich angesehen und sogar vermisst werden. So könnte es beispielsweise helfen, organisatorische Verantwortlichkeiten der Leiter und der Mitarbeiter einer Arbeitsgruppe festzulegen (siehe z.B. Systemimmunologie), institutionelle Regelungen zur sicheren Aufbewahrung sensibler wettbewerbsrelevanter Informationen zu treffen und dadurch Interviewbereitschaft zu steigern (siehe z.B. Gartenökonomie) oder Anschaffung eines Safes zur Ablage von Steuerstatistikdaten zu erleichtern (z.B. Steuerlehre).

Weiterhin äußerten sich mehrere Interviewpartner implizit und explizit zum Zusammenhang zwischen dem Aufbau eines institutionellen Forschungsdatenrepositoriums und dem Open-Access-Gedanken. Begrüßt wird vor allem die Möglichkeit, große Datenmengen in unterschiedlichsten Datenformaten einem möglichst breiten Publikum zugänglich zu machen. Es wird auch auf den positiven Effekt verwiesen, den ein solches Repositorium auf die Bekanntheit weniger populärer Fachbereiche haben könnte.

Gleichzeitig wird auf die positiven Auswirkungen eines Forschungsdatenrepositoriums auf die gute wissenschaftliche Praxis hingewiesen. Die Zugänglichmachung von Forschungsdaten wird als ein Beitrag zur Steigerung der wissenschaftlichen Integrität empfunden, indem die Veröffentlichung der Daten einerseits zu mehr Transparenz in der wissenschaftlichen Arbeit und andererseits auch ein wirkungsvolles Instrument der Selbstkontrolle darstellt. Allerdings werden die Bemühungen zur öffentlichen Zugänglichmachung von Forschungsdaten erst dann erfolgreich sein, wenn diese als

---

15 S. Umfragebericht, S. 31

16 Siehe dazu auch Feijen (2011), S. 4

eine „ganz normale Publikation“ entsprechende wissenschaftliche Anerkennung erfahren. Hinsichtlich des ebenfalls mehrmals erwähnten Publikationsdrucks in konventionellen Medien erscheint aber die Daten-Publikation als eine Form der wissenschaftlichen Kommunikation im Moment noch unattraktiv.

Insgesamt lässt sich die Schlussfolgerung ziehen, dass die Ergebnisse der Umfrage durch Interviews nochmals bekräftigt wurden. Die Diversität der Forschungsdaten und Arbeitsweisen der Wissenschaftler in verschiedenen Fachbereichen und daraus resultierende komplexe Anforderungen an Service-Infrastruktur stellen überdies in weiteren einschlägigen Studien die größten Herausforderungen im Forschungsdatenmanagement dar.<sup>17</sup>

---

17 Vgl. Feijen (2011), Meier zu Verl and Horstmann (2011), Sánchez Solís (2014)

## 5 Literaturverzeichnis

Achard, Pablo; Ayris, Paul; Fdida, Sergio; Gradmann, Stefan; Horstmann, Wolfram; Labastida, Ignasi et al. (2013). LERU Roadmap for Research Data. Advice Paper No.14 – December 2013. League of European Research Universities (LERU). Online zugänglich unter: [http://www.leru.org/files/publications/AP14\\_LERU\\_Roadmap\\_for\\_Research\\_data\\_final.pdf](http://www.leru.org/files/publications/AP14_LERU_Roadmap_for_Research_data_final.pdf).

Allen, Liz; Scott, Jo; Brand, Amy; Hlava, Marjorie; Altman, Micah (2014). Publishing: Credit where credit is due. In Nature, Volume 508, Number 7496, Comment. Online zugänglich unter: <http://www.nature.com/news/publishing-credit-where-credit-is-due-1.15033> , zuletzt geprüft am 11.08.2014.

Bock, Marlene (1992). Das halbstrukturierte-leitfadenorientierte Tiefeninterview: Theorie und Praxis der Methode am Beispiel von Paarinterviews. In: Hoffmeyer-Zlotnik, Jürgen H. P.(Ed.): Analyse verbaler Daten : über den Umgang mit qualitativen Daten. Opladen : Westdt. Verl. (ZUMA-Publikationen). - ISBN 3-531-12360-2, pp. 90-109. URN: <http://nbn-resolving.de/urn:nbn:de:0168-ssoar-25663>

Deutsche Forschungsgemeinschaft (2013). Vorschläge zur Sicherung guter wissenschaftlicher Praxis. Empfehlungen der Kommission "Selbstkontrolle in der Wissenschaft". Denkschrift. Wiley-VCH Verlag, Weinheim. ISBN 3-527-27212-7. Online zugänglich unter: [http://www.dfg.de/download/pdf/dfg\\_im\\_profil/reden\\_stellungnahmen/download/empfehlung\\_wiss\\_praxis\\_0198.pdf](http://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/download/empfehlung_wiss_praxis_0198.pdf) , zuletzt geprüft am 11.08.2014.

European Commission (2013a). Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020. Version 1.0, 11 December 2013. Online zugänglich unter: [http://ec.europa.eu/research/participants/data/ref/h2020/grants\\_manual/hi/oa\\_pilot/h2020-hi-oa-pilot-guide\\_en.pdf](http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf) , zuletzt geprüft am 11.09.2014.

European Commission (2013b). Guidelines on Data Management in Horizon 2020. Version 1.0, 11 December 2013. Online zugänglich unter: [http://ec.europa.eu/research/participants/data/ref/h2020/grants\\_manual/hi/oa\\_pilot/h2020-hi-oa-data-mgt\\_en.pdf](http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf) , zuletzt geprüft am 11.09.2014.

Feijen, Martin (2011). What researchers want. A literature study of researchers' requirements with respect to storage and access to research data. Stichting SURF. Online zugänglich unter: [http://www.surf.nl/binaries/content/assets/surf/en/knowledgebase/2011/What\\_researchers\\_want.pdf](http://www.surf.nl/binaries/content/assets/surf/en/knowledgebase/2011/What_researchers_want.pdf)

Hartmann, Thomas (2013). Zur urheberrechtlichen Schutzfähigkeit von Forschungsdaten. InTeR – Zeitschrift zum Innovations- und Technikrecht. 1. Jahrg., 4/2013. S. 173–236.

Hochschulrektorenkonferenz (2014). Management von Forschungsdaten – eine zentrale strategische Herausforderung für Hochschulleitungen. Empfehlung der 16. Mitgliederversammlung der HRK am 13. Mai 2014 in Frankfurt am Main. Online zugänglich unter: [http://www.hrk.de/uploads/tx\\_szconvention/HRK\\_Empfehlung\\_Forschungsdaten\\_13052014\\_01.pdf](http://www.hrk.de/uploads/tx_szconvention/HRK_Empfehlung_Forschungsdaten_13052014_01.pdf) , zuletzt geprüft am 11.09.2014.

Humboldt-Universität zu Berlin (2014): Satzung der Humboldt-Universität zu Berlin zur Sicherung guter wissenschaftlicher Praxis und zum Umgang mit Vorwürfen wissenschaftlichen Fehlverhaltens. Online zugänglich unter: [https://www.amb.hu-berlin.de/2014/06/06\\_2014\\_20140130%20Beschlussversion%20Satzung%20Wissenschaftliches%20Fehlverhalten\\_DRUCK.pdf](https://www.amb.hu-berlin.de/2014/06/06_2014_20140130%20Beschlussversion%20Satzung%20Wissenschaftliches%20Fehlverhalten_DRUCK.pdf) , zuletzt geprüft am 11.08.2014.

Jones, Sarah; Pryor, Graham; Whyte, Angus (2013): How to Develop Research Data Management Services – a guide for HEIs. DCC How-to Guides. Edinburgh: Digital Curation Centre. Online zugänglich unter: <http://www.dcc.ac.uk/resources/how-guides/how-develop-rdm-services> , zuletzt geprüft am 06.06.2013.

Kindling, Maxi; Schirnbacher, Peter; Simukovic, Elena (2013): Forschungsdatenmanagement an Hochschulen: das Beispiel der Humboldt-Universität zu Berlin. In: LIBREAS. Library Ideas, 23, S. 43-63. Online zugänglich unter: <urn:nbn:de:kobv:11-100212700>

Meier zu Verl, Christian; Horstmann, Wolfram (2011): Subject-Specific Requirements for Open Access Infrastructure - Attempt at a Synthesis. Chapter H. In: Studies on Subject-Specific Requirements for Open Access Infrastructure. Meier zu Verl C, Horstmann W (Eds); Bielefeld: Universitätsbibliothek: 359–381. DOI: [10.2390/PUB-2011-3](https://doi.org/10.2390/PUB-2011-3)

Quirk, Rachel; Olver, Martin; Hammond, Max; Davies, Claire (2008): The Guide to Researching Audiences. Version 1.1. The Strategic Content Alliance. Online zugänglich unter: <http://www.jisc.ac.uk/media/documents/publications/general/2009/scaaudiencesfullguide.pdf> , zuletzt geprüft am 26.09.2014.

San Francisco Declaration on Research Assessment (DORA). Putting science into the assessment of research (2012). Available online at <http://www.ascb.org/dora/> , zuletzt geprüft am 26.09.2014.

Sánchez Solís, Barbara (2014): Factors for Enabling Sharing and Reuse of Research Data. A study performed by NOAD Austria. Version 1.0, March 2014. Online zugänglich unter: <http://phaidra.univie.ac.at/o:343651>

Schreinermacher, Björn; Buchner, Benedikt (2013): Qualitative Interviews online stellen. In Datenschutz und Datensicherheit – DuD, August 2013, Volume 37, Issue 8, pp 537-541. DOI: [10.1007/s11623-013-0215-x](https://doi.org/10.1007/s11623-013-0215-x)

Simukovic, Elena; Kindling, Maxi; Schirnbacher, Peter (2013a): Umfrage zum Umgang mit digitalen Forschungsdaten an der Humboldt-Universität zu Berlin. Umfragebericht, Version 1.0. Online verfügbar auf dem edoc-Server der Humboldt-Universität zu Berlin. URN: <urn:nbn:de:kobv:11-100213001>

Simukovic, Elena; Kindling, Maxi; Schirnbacher, Peter (2013b): Ergebnisse der Umfrage zum Umgang mit digitalen Forschungsdaten an der Humboldt-Universität zu Berlin. ZENODO-Repository. Online zugänglich unter DOI: [10.5281/zenodo.7446](https://doi.org/10.5281/zenodo.7446)

Simukovic, Elena; Kindling, Maxi; Schirnbacher, Peter (2013c): Ergebnisse der Umfrage zum Umgang mit digitalen Forschungsdaten an der Humboldt-Universität zu Berlin, Vergleich der Antworten zwischen Teilnehmergruppen "Professor(in)" und "wissenschaftliche(r) Mitarbeiter(in)". ZENODO-Repository. Online zugänglich unter DOI: [10.5281/zenodo.7447](https://doi.org/10.5281/zenodo.7447)

## 6 Auflistung der Worthäufigkeiten

> word.table[1:200]

all\_new\_words

daten	jahr	projekt	bild	wissenschaftler	leute
384	128	118	101	100	90
problem	nutzen	forschungsdaten	frage	server	arbeit
78	77	66	61	61	57
verschiedene	person	software	bereich	experiment	text
57	55	52	51	48	48
schwierig	metadaten	zugang	dokumentation	interessant	artikel
47	43	43	41	41	40
speichern	format	zeit	bekommen	datenbank	versuchen
40	39	39	38	38	37
institut	nachnutzen	entwickeln	publikation	schreiben	datei
36	36	35	35	35	34
einfacher	information	programm	klein	richtig	cms
34	33	33	32	32	31
sinnvoll	teil	generieren	lang	publizieren	auswertung
31	31	30	30	30	29
buch	datensatz	digital	ergebnis	rohdaten	speziell
29	29	28	28	28	28
standard	webseite	erstellen	tabelle	edition	ende
28	28	27	27	26	26
fachgebiet	kriegen	system	zeitschrift	interview	kollege
26	26	26	26	25	25
student	praktisch	statistisch	analyse	deutschland	methode
25	24	24	23	23	23
mitarbeiter	suchen	direkt	idee	rechner	unterschied
23	23	22	22	22	22
besteht	eigenschaft	support	anfragen	antike	bauen
21	21	21	20	20	20
dropbox	europena	festplatte	konkret	neu	regel
20	20	20	20	20	20
aussehen	open	thema	uni	automatisch	beschreibung
19	19	19	19	18	18
boden	charité	doktorand	geld	geschichte	gigabyte
18	18	18	18	18	18
kunst	anfang	beratung	berlin	jahrhundert	monat
18	17	17	17	17	17
online	schnell	unternehmen	vorstellen	zugreifen	arbeitsgruppe
17	17	17	17	17	16
archivieren	dfg	film	grundlage	handschrift	international
16	16	16	16	16	16
journal	kurz	liegen	praxis	professor	ressource
16	16	16	16	16	16
struktur	tier	zentral	array	aufnehmen	feld
16	16	16	15	15	15
konferenz	kontakt	mensch	moor	nachhaltig	planck
15	15	15	15	15	15
quelle	aufgabe	autor	bibliothek	brief	extern
15	14	14	14	14	14

infrastruktur	link	punkt	tag	technisch	version
14	14	14	14	14	14
warten	archiv	bearbeiten	bildn	computer	falsch
14	13	13	13	13	13
gruppe	kooperationspartner	kunstgeschichte	labor	lesen	sicherung
13	13	13	13	13	13
sprache	stark	stelle	wichtig	xml	angebot
13	13	13	13	13	12
beantworten	bundesamt	chef	eeg	entscheiden	erzeugen
12	12	12	12	12	12
genetisch	gut	komplett	kopie	kosten	nah
12	12	12	12	12	12
prometheus	schwer	sequenzieren	signal	technik	wunsch
12	12	12	12	12	12
zustand	aktiv	anschauen	ansprechpartner	backup	bildrechte
12	11	11	11	11	11
cloud	datenmenge	diskussion	entstehen	erfassen	fallen
11	11	11	11	11	11
fehler	frei				
11	11				

# 7 Anhänge

## 7.1 Interview I: Klinische Psychologie

### Angaben zur Person

Name	Mag. rer. nat. Christian Kaufmann
Institut oder Einrichtung	Institut für Psychologie, Bereich Klinische Psychologie
Position	Assoziierter Wissenschaftler

#### 1. Bitte stellen Sie kurz Ihren Arbeitsbereich vor.

Klinische Psychologie hat zum Schwerpunkt psychophysiologische Forschung. Durch den Betrieb der Hochschulambulanz für Zwangserkrankungen werden „gesunde“ und „kranke“ miteinander verglichen. Es werden daher Studien mit Versuchspersonen durchgeführt. Typischer Ablauf ist z.B. alle 2 Sekunden ein Bild von Gehirnaktivität in einer Zeitreihe von 30 Minuten zu machen, dazu motorische Reaktion und Verhaltensdaten aufzunehmen.

#### 2. Was sind "Forschungsdaten" (=Grundlage Ihrer wissenschaftlichen Arbeit) in Ihrem Arbeitsbereich?

Während der EEG und fMRT-Messungen werden Gehirnaktivität, Sauerstoffveränderungen über die Zeit (hemodynamisches Signal), Hautwiderstand und Blickbewegungen („Peripherphysiologie“) gemessen, Patientendaten und Verhaltensdaten der Versuchspersonen erhoben (Reaktionszeit und -genauigkeit), standardisierte Fragebögen ausgefüllt, statistische Berechnungen durchgeführt. So entstehen u.a. Bilder, 4-dimensionale Datenreihen, morphometrische Daten.

#### 3. Welche Datentypen sind es? Bitte nennen Sie konkrete Formate und/oder gängige Software.

Die wichtigsten Programme sind Matlab, Brain Vision Analyser (proprietär), Matlab Toolboxes (Open Source). Eine Liste der Software-Tools gibt es auf <http://www.nitrc.org/>, darunter insb. Database applications und <http://www.xnat.org/> zum Austausch der Dateien (aktuell wünschenswert).

#### 4. Werden die Forschungsdaten dokumentiert bzw. mit Metadaten beschrieben? Welche Eigenschaften werden damit erfasst? Nutzen Sie ein standardisiertes Metadatenschema dafür? Wenn ja, welches?

Datenformate DICOM und NIFTY. Beide Formate haben „Platzhalter“ für Metadaten, werden aber nur zum Teil benutzt. Zusätzlich werden in standardisierten Fragebögen die Merkmale der Versuchspersonen abgefragt.

#### 5. Werden die Forschungsdaten nach dem Ende eines Forschungsprojektes archiviert? Wie sieht die übliche Praxis aus?

Die Forschungsdaten werden auf DVDs archiviert und in einem feuersicheren Stahlschrank aufbewahrt.

#### 6. Hatten Sie bereits Erfahrungen mit der Veröffentlichung oder Nachnutzung der Forschungsdaten? Gab es Schwierigkeiten dabei?

Manche Journals im Neuroimaging-Bereich verlangen danach oder empfehlen es, die Forschungsdaten zugänglich zu machen. Konkretes Beispiel dafür ist [Proceedings of the National Academy of Sciences](#), wo mehrere Optionen der Zugänglichmachung erlaubt werden. Nachnutzung fand z.B. durch nähere Auswertung von Forschungsdaten aus dem [Martinus Center for Biomedical Imaging](#) statt, es war aber eher für Lehrzwecke gedacht.

### **7. Welche (zentrale) Serviceleistungen zum Umgang mit digitalen Forschungsdaten sehen Sie als notwendig?**

Zum einen, die Masse an Daten geordnet mit Metadaten zu integrieren, damit diese besser auffindbar und zugänglich sind. Zum anderen, ein effizientes, gesichertes System zum Austausch von Daten mit Kooperationspartnern, damit man es nicht mehr via FTP, Zusenden von CDs per Post oder Nutzung solcher Dienste wie Dropbox machen muss. Auch mehr und mehr Journale werden nach der Zugänglichmachung von Forschungsdaten verlangen, daher wäre eine solche webbasierte Lösung optimal, mit der man per Knopfdruck die dem Artikel dazugehörigen Forschungsdaten verlinken könnte.

### **8. Weitere Hinweise und Anmerkungen**

Die Unterstützung durch den CMS wird sehr geschätzt – die Daten werden redundant gespeichert und sind dadurch sicher, was am wichtigsten ist. Man hofft auch, dass durch solche Initiativen zur Zugänglichmachung von Forschungsdaten der wissenschaftliche Standard steigen wird, weil die Forschung transparenter wird und weniger Daten schlechter Qualität publiziert werden, was aktuell der Fall ist.

## 7.2 Interview II: Experimentelle Physik

### Angaben zur Person

Name	Dr. Oliver Maria Kind
Institut oder Einrichtung	Institut für Physik, Bereich Experimentelle Physik, Experimentelle Elementarteilchenphysik I
Position	Wissenschaftlicher Mitarbeiter

#### 1. Bitte stellen Sie kurz Ihren Arbeitsbereich vor.

Unsere AG arbeitet im ATLAS-Experiment am CERN. Es ist eine weltweite Kollaboration, an der rund 3000 Wissenschaftler beteiligt sind. Es werden Kollisionseignisse untersucht und nach seltenen speziellen Teilchen oder Teilchenreaktionen gesucht, wie bspw. Higgs-Teilchen. Einige unserer Doktoranden sind vor Ort dabei, ansonsten werden die Daten für Auswertungen aus einem verteilten Computing-Netzwerk (LHC-Grid) bezogen.

#### 2. Was sind "Forschungsdaten" (=Grundlage Ihrer wissenschaftlichen Arbeit) in Ihrem Arbeitsbereich?

Zum einen rekonstruieren wir Kollisionseignisse aus Rohdaten. Zum anderen werden Simulationen durchgeführt. So kann man die "echten" Daten mit den simulierten Daten vergleichen. Die Auswertung erfolgt in einzelnen Analyseschritten: aus Rohdaten werden bestimmte Selektionen als Datensätze extrahiert (rausgefiltert), die Datenmenge dadurch reduziert und in neue Formate überführt, danach statistisch ausgewertet und in Histogrammen oder Plots dargestellt.

#### 3. Welche Datentypen sind es? Bitte nennen Sie konkrete Formate und/oder gängige Software.

Vom Lesen und Bearbeiten bis zum Speichern und graphisch Darstellen benutzt man eine von der Hochenergiephysik-Community entwickelte Software [Root](#) (Open Source, zugänglich unter GNU General Public License).

#### 4. Werden die Forschungsdaten dokumentiert bzw. mit Metadaten beschrieben? Welche Eigenschaften werden damit erfasst? Nutzen Sie ein standardisiertes Metadatenschema dafür? Wenn ja, welches?

Es werden sehr viele unterschiedliche Metadaten erfasst: Zeitpunkt und Bedingungen des Experiments, geometrische Daten (Positionierung der Geräte), Kalibration der Detektoren, Temperatur, Gasmischung- und -strömung. Beim Filtern oder bei einer Simulation werden auch die gesetzten Parameter gespeichert. Die Qualität der Rohdaten wird zusätzlich über "Data Quality Monitoring" von der Schichtbesetzung kontrolliert und spezielle Markierungen als Flaggen gesetzt (grün, gelb, rot). Die Metadaten und Rohdaten werden in getrennten Datenbanken gespeichert und später mit Datensätzen zusammengeführt.

#### 5. Werden die Forschungsdaten nach dem Ende eines Forschungsprojektes archiviert? Wie sieht die übliche Praxis aus?

Das wird im Grunde von den Richtlinien des Experiments bestimmt. Aufgrund der sehr großen Datenmengen (mehrere Hundert Petabytes) wäre es schwierig, alle Daten gemäß der guten wissenschaftlichen Praxis 10 Jahre lang zu speichern. Operativ werden Rohdaten in 3 Kopien auf Tapes (Magnetbändern) gespeichert.

#### 6. Hatten Sie bereits Erfahrungen mit der Veröffentlichung oder Nachnutzung der Forschungsdaten? Gab es Schwierigkeiten dabei?

Die Richtlinien des Experiments lassen die Weitergabe von Daten außerhalb der Kollaboration nicht zu. Es werden aber viele Artikel publiziert, die wiederum sehr datenintensiv sind. Zur

Publikation der Ergebnisse sind die von den Fachgesellschaften herausgegebenen Journale wie [Physical Review](#) und [European Journal of Physics](#) von höchstem Renommee. Außerdem wird mit Open-Access-Modellen experimentiert (z.B. im [SCOAP3-Projekt](#) oder Umstieg auf den Goldenen OA-Weg beim [EPL Journal](#)). Ein berühmtes Beispiel zur Reanalyse der Rohdaten aus den 1980er Jahren hat dazu gezeigt, dass die Nachnutzung zwar möglich, aber mit sehr großem Aufwand verbunden ist.

### **7. Welche (zentrale) Serviceleistungen zum Umgang mit digitalen Forschungsdaten sehen Sie als notwendig?**

Da der Fachbereich durch große Kollaborationen mit der Infrastruktur bereits gut versorgt ist, wird keine technische Unterstützung benötigt. Hilfreich wäre aber eine rechtliche Beratung und eine Kontaktperson, an die man sich bei Bedarf wenden könnte.

### **8. Weitere Hinweise und Anmerkungen**

Die Community setzt sehr stark auf offene Standards: hauptsächlich wird mit einer selbst entwickelten und an eigene Bedürfnisse angepassten Software Root gearbeitet, als Betriebssystem wird überall Linux benutzt, Dokumente, Dissertationen und sogar Präsentationen mit LaTeX erstellt. Die primäre Quelle für Literaturrecherche ist arXiv. Außerdem werden in der [INSPIRE-Datenbank](#) (High Energy Physics Information System) sämtliche Journal- und Konferenzveröffentlichungen aus der Hochenergiephysik registriert, was die Suche, aber auch das Erstellen von Literaturverzeichnissen und Publikationslisten wesentlich erleichtert. Schließlich, als Teilchenphysiker programmiert man oft und gerne. Einige speziell entwickelten statistischen Methoden werden sogar von anderen Fachdisziplinen benutzt.

## 7.3 Interview III: Bodenkunde und Geomorphologie

### Angaben zur Person

Name	PD Dr. rer. nat. Mohsen Makki
Institut oder Einrichtung	Geographisches Institut, Abteilung Geomorphologie, Bodengeographie und Quartärforschung
Position	Wissenschaftlicher Mitarbeiter

#### 1. Bitte stellen Sie kurz Ihren Arbeitsbereich vor.

Unser Lehrstuhl beschäftigt sich mit Geomorphologie und Bodenkunde. Dabei arbeiten wir in interdisziplinären Gruppen zusammen mit Archäologen aus dem [DAI](#). Meine zwei Schwerpunkte sind Bodengeographie und Umwelt, insbesondere Stadtboden von Metropolen und sehr schnell wachsenden Städten, und Geoarchäologie. Mein Partner in Berlin ist die [Senatsverwaltung für Stadtentwicklung und Umwelt](#).

#### 2. Was sind "Forschungsdaten" (=Grundlage Ihrer wissenschaftlichen Arbeit) in Ihrem Arbeitsbereich?

Im Themenbereich "Umwelt" führen wir auch Interviews durch (oft informell ohne Audio-Aufzeichnung), die dann abgetippt und auf CDs geliefert werden. Weiterhin benutzen wir Grunddaten wie Satelliten-Bilder oder DGM-Daten ([digitalles Geländemodell](#)). Wenn wir in der Stadt arbeiten, dann sind es natürlich auch topographische Karten und GIS-Daten. GIS ist in unserem Arbeitsbereich sehr wichtig.

#### 3. Welche Datentypen sind es? Bitte nennen Sie konkrete Formate und/oder gängige Software.

GIS-Daten (ArcGIS), Datenbanken (MS Access), Tabellen, Bilder, Textdokumente (MS Word und GoogleDocs). Grafik-Programme wie Photoshop, Corel Draw und Photo-Paint. Satelliten-Bilder werden aus öffentlichen Datenbanken heruntergeladen oder bei kostenpflichtigen Diensten bestellt.

#### 4. Werden die Forschungsdaten dokumentiert bzw. mit Metadaten beschrieben? Welche Eigenschaften werden damit erfasst? Nutzen Sie ein standardisiertes Metadatenschema dafür? Wenn ja, welches?

Als ein Standard für Erstellung von Datenbanken mit genormten Abkürzungen gilt in Deutschland (und teilweise auch in anderen Ländern) die [Bodenkundliche Kartieranleitung](#), die alle wesentlichen Merkmale zur Bodenbeschreibung und -klassifizierung (Bodentypen, Bodenart, Grundwasser, Staunässe, Ausgangsgestein der Bodenbildung, Humusform etc.) sowie umfangreiche Kennwerttabellen als Auswertungsgrundlagen zum Wasser- und Lufthaushalt des Bodens und zur Standortbewertung enthält. Damit ist Bodenkunde ein sehr fortgeschrittener Fachbereich. Für internationale Bodenansprache und Klassifikation benutzen wir WRB ([World Reference Base](#)).

#### 5. Werden die Forschungsdaten nach dem Ende eines Forschungsprojektes archiviert? Wie sieht die übliche Praxis aus?

Die Daten werden auf CDs und auf lokalen Laufwerken gespeichert. Jeder macht sich zudem seine eigene Kopie. Problematisch sind aber z.B. obsolete Datenträger und Software, die mit einer älteren Version erstellte Dateien nicht mehr öffnen oder korrekt darstellen kann.

#### 6. Hatten Sie bereits Erfahrungen mit der Veröffentlichung oder Nachnutzung der Forschungsdaten? Gab es Schwierigkeiten dabei?

Meine Arbeitsergebnisse werden auf der Homepage der Senatsverwaltung zugänglich gemacht. Ein Beispiel für Nachnutzung sind die Satelliten-Bilder, die wir von anderen Datenbanken

herunterladen. Die Fachzeitschriften haben bisher nach Daten nicht gefragt.

### **7. Welche (zentrale) Serviceleistungen zum Umgang mit digitalen Forschungsdaten sehen Sie als notwendig?**

Ich wünsche mir, dass ich große Daten unkompliziert allen zur Verfügung stellen könnte – in unterschiedlichster Form, als Karten, PDFs, Tabellen. Wenn ich meine Ergebnisse in einem Artikel bei einer Zeitschrift veröffentliche, dann stellt es nur den geringsten Teil meiner eigentlichen Arbeit dar, das Wichtigste bleibt dahinter unveröffentlicht. Diese Möglichkeit würde solchen weniger populären Fachbereichen wie Bodenkunde helfen, mehr Leute für ihre Forschungsthemen zu begeistern. Außerdem sollen solche künftige Systeme nutzerfreundlich und plakativ gestaltet werden, um die Arbeit der Wissenschaftler zu unterstützen und nicht zu komplizieren.

### **8. Weitere Hinweise und Anmerkungen**

Ich bin sehr daran interessiert, meine Daten zu veröffentlichen – je mehr Daten zur Verfügung stehen, desto besser. Das führt zu mehr Transparenz und Eigenkontrolle. Auf solcher Weise könnten nicht nur Fehler entdeckt werden, sondern auch die Daten von anderen Forschern neuartig, mit weiteren Ideen nachgenutzt werden. Das kann in anderen Fächern (z.B. wo die Forschung mit Patenten oder kommerziellen Interessen verbunden ist) nicht so einfach sein, in Bodenkunde sind wir nicht so eingeschränkt. Solche Veröffentlichung muss aber auch von der DFG und Universitäten als eine ganz normale Publikation belohnt werden, sonst bleibt es eine Art ehrenamtliche "Öffentlichkeitsarbeit der Wissenschaft" ohne Impactfactor für Drittmittel und Karriere.

## 7.4 Interview IV: Geomatik

### Angaben zur Person

Name	Prof. Dr. Patrick Hostert
Institut oder Einrichtung	Geographisches Institut, Abteilung Geomatik
Position	Professor, stellvertretender Institutsdirektor, Gründungssprecher IRI THESys

#### 1. Bitte stellen Sie kurz Ihren Arbeitsbereich vor.

Die Abteilung für Geomatik beschäftigt sich mit globalen Fragen der Landnutzung. Dafür werden die Methoden der Fernerkundung und GIS-Daten verwendet. Außerdem werden im Labor quantitative Messungen der Spektrometrie durchgeführt. Im Rahmen der Exzellenzinitiative werden im integrativen Forschungsinstitut „Die großen Transformationen von Mensch-Umwelt-Systemen“ (IRI THESys) auch die mit dem Umweltwandel verbundenen gesellschaftlichen Herausforderungen interdisziplinär erforscht.

#### 2. Was sind "Forschungsdaten" (=Grundlage Ihrer wissenschaftlichen Arbeit) in Ihrem Arbeitsbereich?

Räumliche Daten: Satellitendaten (digitale Rasterdaten, „Pixel“), digitale Karten („GIS-Daten“), raumbezogene Statistiken. Nicht-räumliche Daten, wie z.B. Statistikdaten, Interviews, etc.

#### 3. Welche Datentypen sind es? Bitte nennen Sie konkrete Formate und/oder gängige Software.

Erdas/Imagine – img, IDL/Envi – bsq (Pakete für digitale Bildbearbeitung von Satellitendaten), ArcGIS / Quantum GIS (Open-Source-Alternative von ArcGIS) – shp oder komplexere Datenbankformate, alle Statistik-relevanten Formate (R, SPSS, Excel, ...), Matlab. Für viele Ansprüche genügen Open-Source-Produkte, es wird aber auch kommerzielle Software benutzt. Mittlerweile gibt es auch spezielle Ausprägungen von Programmiersprachen für GIS-Daten wie z.B. ArcPython.

#### 4. Werden die Forschungsdaten dokumentiert bzw. mit Metadaten beschrieben? Welche Eigenschaften werden damit erfasst? Nutzen Sie ein standardisiertes Metadatenschema dafür? Wenn ja, welches?

Teils; wir haben eine eigene [Geodateninfrastruktur](#) aufgebaut, die alle raumbezogenen Daten umfasst, bzw. umfassen kann; in einigen Projekten (z.B. BMBF-Projekt [CarBioCial](#)) entstehen auch Metadatenstrukturen, teils auch als ein Teil größerer Datenverbünde (z.B. [GLUES](#)). Es gibt auch nationale und internationale Standards, oft als ISO-Normen direkt in Softwarepaketen implementiert. Diese sind aber sehr umfangreich, daher werden für Ansprüche in konkreten Projekten eher spezielle darauf basierende Subsets entwickelt.

#### 5. Werden die Forschungsdaten nach dem Ende eines Forschungsprojektes archiviert? Wie sieht die übliche Praxis aus?

Ja, in der Regel aber nur Ergebnisdaten, da Rohdaten und Zwischenergebnisse (dynamische Daten während der Projektlaufzeit) zu speicher-intensiv sind. Wir arbeiten auf den SAN-Laufwerken des CMS und den entsprechenden Backup-Strukturen. Die Endergebnisse kommen auch häufig „ins Regal“, auf HDD. Die Aufbewahrungsfrist muss daher zwischen „Endergebnissen“ und „Zwischenergebnissen“ differenziert werden: wenngleich das Erstere oft auch länger als 10 Jahre aufbewahrt werden soll, muss beim Zweiteren entschieden werden, ob diese zur Nachvollziehbarkeit der Endergebnisse nötig sind und überhaupt aufbewahrt werden sollten.

#### 6. Hatten Sie bereits Erfahrungen mit der Veröffentlichung oder Nachnutzung der

### **Forschungsdaten? Gab es Schwierigkeiten dabei?**

Unsere wichtigste Datenquelle ist [Landsat](#), die seit 1970er Jahren fliegende Satellitenmission, und die Datenbanken des [U.S. Geological Survey](#) (USGS) – dies kann als Nachnutzung der Daten angesehen werden. Wir arbeiten aber auch an einer eigenen Open-Access-Lösung für Ergebnisse unserer Satellitendatenanalysen (Veröffentlichung).

### **7. Welche (zentrale) Serviceleistungen zum Umgang mit digitalen Forschungsdaten sehen Sie als notwendig?**

Flexiblere Handhabung von in-house-Server-Lösungen (Wartung durch den CMS, aber eigener Zugriff auf Softwarelösungen, auch im Kontext der Installation von patches, updates, etc.). SAN-artige Lösungen in-house; gerne „gemanagt“ seitens CMS, aber mit direkter Anbindung an unsere Server (z.B. NAS-basiert); der Datendurchsatz ist unser „Flaschenhals“.

### **8. Weitere Hinweise und Anmerkungen**

Der CMS scheint nach meiner Außenansicht drastisch unterfinanziert und lebt „auf Kosten“ der sehr guten Ausstattung von vor 10 Jahren. Es wird aber immer deutlicher, dass notwendige Infrastrukturmaßnahmen und die für einen modernen Universitätsbetrieb nötige Personaldecke knapp werden oder auch schon fehlen (zumindest mit Blick auf große Mengen von Forschungsdaten). Vor dem Hintergrund der Verstärkungsbemühungen zur Exzellenzinitiative sollten eine State-of-the-Art IT-Infrastruktur im Allgemeinen und der Umgang mit Forschungsdaten im Besonderen zu „Flaggschiff-Themen“ auf der Leitungsebene der Universität gemacht werden.

## 7.5 Interview V: Biologische Psychologie

### Angaben zur Person

Name	Prof. Dr. rer. soc. habil. Werner Sommer
Institut oder Einrichtung	Institut für Psychologie, Bereich Biologische Psychologie
Position	Professor

#### 1. Bitte stellen Sie kurz Ihren Arbeitsbereich vor.

Wir interessieren uns für kognitive und emotionale Prozesse, z.B. Emotionsverarbeitung, Identitätsverarbeitung, Gesichtererkennung, Teilprozesse beim Lesen, Doppelaufgaben (Multitasking), Sprachverarbeitung.

#### 2. Was sind "Forschungsdaten" (=Grundlage Ihrer wissenschaftlichen Arbeit) in Ihrem Arbeitsbereich?

Um die o.g. Prozesse zu untersuchen, arbeiten wir mit Verhaltensdaten, elektrophysiologischen Daten (z.B. EEG), Fragebögen, Intelligenz-Tests, Audio-, Video-Aufnahmen von den Personen und teilweise auch genetischen Daten. Diese Daten sind oft sehr sensibel und müssen anonymisiert werden. Aus diesem Grund gibt es auch eine interne Ethik-Richtlinie des Instituts.

#### 3. Welche Datentypen sind es? Bitte nennen Sie konkrete Formate und/oder gängige Software.

Wir benutzen sehr viele Softwarepakete, u.a. Brain Vision Recorder, Brain Vision Analyser, Matlab.

#### 4. Werden die Forschungsdaten dokumentiert bzw. mit Metadaten beschrieben? Welche Eigenschaften werden damit erfasst? Nutzen Sie ein standardisiertes Metadatenschema dafür? Wenn ja, welches?

Das macht jeder eigenverantwortlich. Teilweise ist die Dokumentation schon mit der Software implementiert, eine standardisierte Praxis gibt es aber nicht.

#### 5. Werden die Forschungsdaten nach dem Ende eines Forschungsprojektes archiviert? Wie sieht die übliche Praxis aus?

Vieles wird lokal auf CDs oder DVDs gesichert. Eine professionelle Lösung wie z.B. ein vom CMS gewartetes Datenarchiv wäre vorteilhaft.

#### 6. Hatten Sie bereits Erfahrungen mit der Veröffentlichung oder Nachnutzung der Forschungsdaten? Gab es Schwierigkeiten dabei?

Im Rahmen der DFG-Forschergruppe 868 "[Computational Modeling of Behavioral, Cognitive, and Neural Dynamics \(Mind and Brain Dynamics\)](#)" existiert das [Potsdam Mind Research Repository](#) (PMR<sup>2</sup>), in dem Publikationen der Forschergruppe zusammen mit zugrundeliegenden Daten und Scripts für Analysen zugänglich gemacht werden.

#### 7. Welche (zentrale) Serviceleistungen zum Umgang mit digitalen Forschungsdaten sehen Sie als notwendig?

Am Beispiel der DFG-Forschergruppe wäre es gut, ein solches Repository an der HU zu haben. Die Daten (unabhängig vom Projekt) und Scripts für deren Auswertung könnten für andere interessierte Personen oder die gesamte Öffentlichkeit bequem freigeschaltet werden. So müsste man sich nicht mit jeder Anfrage einzeln beschäftigen. Weiterhin gibt es Daten, die aus historischen Gründen interessant sein können. Deren langfristige Archivierung – auch über 10 Jahre hinaus – sollte zentral und professionell organisiert werden.

#### 8. Weitere Hinweise und Anmerkungen

Ein notorisches Problem sind Publikationen. In manchen Fällen ist es sehr schwierig, auf bestimmte Zeitschrift Zugriff zu bekommen. So bleiben meine Forschungsergebnisse zu einem gewissen Grad unsichtbar. Es wäre gut, wenn ich meine Publikationen an einer Stelle ablegen und von dort verlinken könnte. Allerdings ist es aktuell nicht klar, ob eine Zweitveröffentlichung von Verlagen erlaubt wird.

## 7.6 Interview VI: Systemimmunologie

### Angaben zur Person

Name	Dr. rer. nat. Michal Or-Guil
Institut oder Einrichtung	Institut für Biologie, Nachwuchsgruppe "Systemimmunologie"
Position	Leiterin der Arbeitsgruppe

#### 1. Bitte stellen Sie kurz Ihren Arbeitsbereich vor.

Wir arbeiten an der Entwicklung einer Pipeline für die Identifizierung von Biomarkern. Das Ziel ist, [Biomarker](#) auszusuchen, die später für ein Diagnostik-Produkt eingesetzt werden können. Speziell geht es um Antikörper-Diagnostik für Krankheiten, für die keine [Epitope](#) bekannt sind. Zur Erkennung von Antikörpern durch Moleküle werden [Hochdurchsatz](#)-Daten generiert, gesunde und kranke verglichen und nach systematischen Unterschieden gesucht. Die Ergebnisse können für Diagnostik benutzt werden, auch wenn das [Pathogen](#) unbekannt ist. Es wird sehr interdisziplinär gearbeitet: zum Einsatz kommen neben experimentellen Methoden auch solche aus der theoretischen Physik und der Bioinformatik. Durch klinische Fragestellungen und Entwicklung von Meßplattformen besteht außerdem Zusammenarbeit mit medizinischen Einrichtungen und der Industrie.

#### 2. Was sind "Forschungsdaten" (=Grundlage Ihrer wissenschaftlichen Arbeit) in Ihrem Arbeitsbereich?

Bilder von [Arrays](#), die mit einem Laser-Scanner aufgenommen werden, Tabellen, die aus diesen Bildern gewonnen werden; histologische Bilder und deren Auswertung; Daten aus der Durchflusszytometrie. Um die Hochdurchsatz-Daten generieren und die Ergebnisse evaluieren zu können, werden auch Patientendaten verarbeitet.

#### 3. Welche Datentypen sind es? Bitte nennen Sie konkrete Formate und/oder gängige Software.

Bilder im TIFF- und JPG-Format, Tabellen im CSV- und selbst definierte Textformate. Array-Tabellen werden mit dem Programm „GenePix“ ausgelesen und im [GPR-Format](#) (GenePix Results Format) gespeichert. Für statistische Auswertung wird das Programm R benutzt, früher auch Matlab. [Durchflusszytometrie](#)-Daten werden mit dem Programm FlowJo oder mit R ausgewertet ([FCS-Format](#)).

#### 4. Werden die Forschungsdaten dokumentiert bzw. mit Metadaten beschrieben? Welche Eigenschaften werden damit erfasst? Nutzen Sie ein standardisiertes Metadatenschema dafür? Wenn ja, welches?

Zu jedem Projekt bzw. jeder untergeordneten Fragestellung wird ein sog. Report erstellt (Word-Dokument). Dieser basiert auf einem eigenständig entwickelten Template, nach dem die Angaben zum Experiment (wer, wann, warum), zu Referenzen, Ergebnissen und Auswertungen der Ergebnisse gemacht werden. Experimentelle Arbeiten werden in Laborbüchern dokumentiert.

#### 5. Werden die Forschungsdaten nach dem Ende eines Forschungsprojektes archiviert? Wie sieht die übliche Praxis aus?

Die Daten werden thematisch geordnet (in einer Ordnerstruktur) und zusammen mit den Reports auf einem Server gespeichert. Bei Veröffentlichungen wird ein extra Ordner angelegt, in dem alle zugehörigen Daten archiviert werden. Der Backup wird von der Arbeitsgruppe selbst gemacht.

#### 6. Hatten Sie bereits Erfahrungen mit der Veröffentlichung oder Nachnutzung der Forschungsdaten? Gab es Schwierigkeiten dabei?

Zusammen mit einem Industriepartner wurde eine Datenbank mit über 3.000 histologischen

Bildern erstellt. Diese Datenbank ist öffentlich zugänglich und darf nachgenutzt werden (für Zitierung wird ein zugehöriger Artikel empfohlen). Im Rahmen von BMBF-Verbundprojekte wird üblicherweise ein Datenmanagementplan gefordert (insb. bei biologischen Hochdurchsatz-Daten).

### **7. Welche (zentrale) Serviceleistungen zum Umgang mit digitalen Forschungsdaten sehen Sie als notwendig?**

Ein zentraler Backup-Service und Speicherplatz für Forschungsdaten wird benötigt. Zur Einhaltung der Regeln guter wissenschaftlicher Praxis müssen zudem zahlreiche rechtliche Fragen geklärt werden: wem gehören die Daten, wer darf sie für welche Zwecke nutzen, Regelungen in den Arbeitsverträgen und im Falle von Stipendiaten und Studentischen Hilfskräften sollten bekannt sein bzw. geklärt werden. Regelungen zur Übergabe von Daten beim Verlassen der Universität sollten erstellt werden, auch mit Blick auf die in Zusammenarbeit mit der Industrie üblichen Non-Disclosure-Agreements. Eine rechtliche Aufklärung der grundlegenden Fragen ist daher dringend notwendig.

### **8. Weitere Hinweise und Anmerkungen**

Ein effizientes System für eine vereinfachte Unterstützung der Projektverwaltung scheint angebracht zu sein. Es soll helfen, die im Laufe eines Projekts anfallenden Verwaltungsaufgaben möglichst automatisiert zu erledigen und den damit verbundenen Aufwand zu verringern. Da die E-mailkommunikation eine wichtige Quelle dafür ist, wie die Projekte oder deren Forschungsdaten zustande gekommen sind, wäre außerdem hilfreich, auf verschiedenen Konten vorgehaltene Emails zentral aufbewahren bzw. migrieren zu können.

## 7.7 Interview VII: Züchtungsbiologie

### Angaben zur Person

Name	Dr. rer. nat. Ralf Bortfeldt
Institut oder Einrichtung	Landwirtschaftlich-Gärtnerische Fakultät <sup>18</sup> , Fachgebiet Züchtungsbiologie und molekulare Tierzucht
Position	Wissenschaftlicher Mitarbeiter

#### 1. Bitte stellen Sie kurz Ihren Arbeitsbereich vor.

Das Fachgebiet Züchtungsbiologie und molekulare Genetik erforscht den Zusammenhang zwischen genetischen Eigenschaften und quantitativen Merkmalen in Nutztierspezies (Rind, Schwein, Huhn, Pferd) und Modellorganismen (Maus). In der molekularbiologischen Charakterisierung selektierter Individuen sucht man nach Kandidatengenomen mit dem höchsten Einfluss auf wirtschaftlich interessante Merkmale (z.B. Milchmenge, Milchfett- und Proteingehalt beim Milchrind). Dazu werden komplette Genome von vorselektierten Individuen mit vielversprechenden [QTLs](#) sequenziert, um die unter bestimmten Bedingungen induzierte Genexpression zu bestimmen und neue Marker zu identifizieren.

#### 2. Was sind "Forschungsdaten" (=Grundlage Ihrer wissenschaftlichen Arbeit) in Ihrem Arbeitsbereich?

Genetische Analysen werden heutzutage mehr und mehr mit Hochdurchsatzmethoden wie [SNP-Arrays](#) und [Next Generation Sequencing](#) (NGS) durchgeführt. Dabei werden bis zu mehrere hundert Individuen eines bestimmten Selektionsexperiments auf hunderttausende SNPs gescreent. Rohdaten aus SNP-Arrays werden statistisch analysiert und nach Zusammenhängen zwischen genetischen Unterschieden und quantitativen Merkmalen gesucht.

#### 3. Welche Datentypen sind es? Bitte nennen Sie konkrete Formate und/oder gängige Software.

Es gibt eine ganze Reihe von teilweise frei verfügbaren Werkzeugen, z.B. Bowtie, [BWA](#), Tophat, [Picard](#)-Tools, BCF- und VCF-Tools, [SAMtools](#), eXpress, oder das [Genome Analysis Toolkit](#) (GATK) vom MIT. Als Standards in NGS-Analysen haben sich mittlerweile das [VCF](#) und [SAM/BAM](#)-Format etabliert. Für die Ablage von SNP-Array Daten für weiterführende genetische Analysen wird in unserem Bereich hauptsächlich das PED, MAP und Linkage-Format verwendet. Für die statistische Auswertung benutzen wir SAS und R.

#### 4. Werden die Forschungsdaten dokumentiert bzw. mit Metadaten beschrieben? Welche Eigenschaften werden damit erfasst? Nutzen Sie ein standardisiertes Metadatenschema dafür? Wenn ja, welches?

Die Dokumentation der Vorgänge erfolgt sehr individuell. Die Notizen zur Erhebung der Daten und zum Ordnerinhalt werden oft mit MS OneNote oder als einfache "readme"-Dateien abgespeichert. Für eine standardisierte Repräsentation und Annotation von Genen wird [Gene Ontology](#) (GO) verwendet – ein kontrolliertes Vokabular zur Beschreibung von biologischen Prozessen, molekularer Funktion, zellulärer Lokalisation etc. In der Wirtschaft (z.B. Pharma-Industrie) wird dafür oft sog. Standard Operating Procedure (SOP) definiert, die die Ausführung und Beschreibung routinemäßiger Arbeitsprozesse regelt.

#### 5. Werden die Forschungsdaten nach dem Ende eines Forschungsprojektes archiviert? Wie sieht die übliche Praxis aus?

Die Rohdaten und Analysen werden zusammen mit dem Verzeichnis in Projektordnern auf dem

<sup>18</sup> Das Interview wurde noch vor der Fakultätsreform geführt. Seit April 2014 gehört das neu gegründete Albrecht Daniel Thaer-Institut für Agrar- und Gartenbauwissenschaften zur Fakultät für Lebenswissenschaften.

Server abgelegt und im Netzwerkspeicher (NAS) gespiegelt. Für Manuskripte von Publikationen wird ein Unterordner angelegt, die dazugehörigen Daten liegen i.d.R. als "Supplement" auf den Webseiten von Journals.

#### **6. Hatten Sie bereits Erfahrungen mit der Veröffentlichung oder Nachnutzung der Forschungsdaten? Gab es Schwierigkeiten dabei?**

Die Gensequenzen werden gewöhnlich in [GenBank](#) eingestellt – das ist die einfachste Form die Daten öffentlich zugänglich zu machen. Außerdem wird durch Submission-Formulare eine gewisse Grundannotation sichergestellt. Da die Darstellung der Daten in einem wissenschaftlichen Artikel nur in beschränktem Umfang möglich ist, müssen weitere, die Arbeit unterstützende und belegende Datensätze in einem Supplement untergebracht werden. So werden sie aber nicht indiziert, daher hätte ich auch nichts dagegen, die Daten in einem Online-Repository abzulegen. Letztes Jahr haben wir zudem Software veröffentlicht und dafür den WWW2-Server der HU benutzt, wobei die langfristige Verfügbarkeit dieser Veröffentlichung nur schwerlich gesichert werden kann.

#### **7. Welche (zentrale) Serviceleistungen zum Umgang mit digitalen Forschungsdaten sehen Sie als notwendig?**

Man müsste erstmal schauen, welche Arten der Informationsvermittlung von Zentraleinrichtungen angeboten werden können und sollen und wer der richtige Ansprechpartner dafür ist. Spezielle Schulungen oder Kurse wie z.B. "Datenorganisation- bzw. Strukturierung für wissenschaftliche Projekte", wo gute Beispiele vermittelt werden, wie man eigene Arbeitsabläufe effizienter gestalten kann, wären Gold wert. Allerdings sollten solche Angebote vielleicht doch besser direkt an den Fakultäten oder Departments angeboten werden, um einen engen Kontakt zu den Wissenschaftlern herzustellen. In jedem Fall müssten AG-Leiter für das Thema sensibilisiert werden, um die Umsetzung aktiv zu unterstützen.

#### **8. Weitere Hinweise und Anmerkungen**

Die Fakultätsreform der HU, von der auch die Landwirtschaftlich-Gärtnerische Fakultät betroffen ist (zukünftig Fakultät für Lebenswissenschaften, zusammen mit dem Institut für Biologie und Institut für Psychologie), bietet sich als ein günstiger Zeitpunkt an, notwendige Strukturen für Koordination des Forschungsdatenmanagements und relevanter Belange aufzusetzen. Weiterhin wäre sehr hilfreich, einen durchsuchbaren "Datenkatalog" der HU zu haben. So könnten die Arbeitsgruppen mit ähnlichen Fragestellungen gefunden und deren Ansprechpartner gezielt angesprochen werden, interne Kooperationen entstehen und Synergieeffekte durch Know-How-Austausch geschaffen werden.

## 7.8 Interview VIII: Theoretische Biologie

### Personal details

Name	Dr. rer. nat. Tiziano Zito, Owen Mackwood, Thomas McColgan
Institute or department	Institute for Theoretical Biology Berlin (ITB)
Position	System Administrator, PhD Student, PhD Student

#### 1. Could you please briefly describe your field of work and main research questions.

Research groups at ITB are working on theoretical approaches in the biological sciences, from modeling of circadian clocks or evolution of living systems to neurophysiology and neural networks. Research topics such as auditory systems of birds (e.g. barn owl) or how inhibitory plasticity affects dynamics and information processing in recurrent networks are being explored using simulations, mathematical models, methods of computational neuroscience and electrophysiology.

#### 2. What is "research data" (=basis for your research activities) in your field of work?

Functional magnetic resonance imaging (fMRI) to measure brain activity, audio recordings, electrophysiology data (measurements of voltage change or electrical current flow used to study the electrical properties of biological cells and tissues). Software code written in order to process/analyse data might be considered as a research object, too, but does not constitute primary "research data". In the purely theoretical groups "research" data are the theoretical and computational implementations of mathematical models.

#### 3. What types are these data? Please give some examples of specific formats and/or software you use.

Raw data deriving from computational simulations/mathematical models; in vivo and in vitro data from animal testing experiments; numerical data. Custom-written software "xdphys" (from California Institute of Technology) is used for stimulus synthesis and calibration. Matlab. Mathematica. LaTeX for typesetting, self-programming in Python; [HDF5 for Python](#) and various relational Databook for storage of the data. There is also a large open-access database of program codes of published computational neuroscience models called [ModelDB](#) (maintained by Yale University).

#### 4. Do you describe or annotate your research data? Is there any in-house or standard template for that (e.g. metadata schema, controlled vocabulary, ontology)? What properties are captured by that?

[HDF5](#) is used as a data model, library, and file format for storing and managing data. Self-written free format text files are the most typical data annotation format. There is an interesting project in UK called [CARMEN](#) where standards for annotation of data and data sharing are being developed. By using its [MINI](#) reporting guideline for electrophysiology (Minimum Information about a Neuroscience investigation) characteristics on general features, study subject, task, stimulus, behavioral event, recording and time series data can be captured.

#### 5. Do you preserve or archive research data after completing a research project? What does the common practice look like?

NAS at ITB (on tape; professionally curated by system administrator) with regular backup on the daily, weekly and annual basis; private copies.

#### 6. Have you ever published or re-used research data? Did you experience any difficulties?

There is some common practice in exchanging data between theoretical and experimental biology

research groups working on similar questions. However, when trying to reconstruct a model in a published paper claiming wonderful results one researcher had to end up calling the authors of the paper because of insufficient details to reproduce the results. Therefore science should ideally be an open enterprise where researchers provide full and accurate information and get credit for sharing their results. It's already common to publish the code on [GitHub](#) and both researchers are willing to share data after finishing their PhD thesis.

**7. What (central) support or services in managing research data do you consider as necessary?**

Research groups need storage space for managing active research data. This service should allow on-site administration in order to provide flexible solution and rapid transfer of data. Support in software / scientific computing and version control would be also helpful.

**8. Further comments or suggestions**

As most research groups here use LaTeX for typesetting and preparing documents, support for migrating them (e.g. PhD thesis) to PDF/A format as required for publishing and archiving on edoc-Server is vitally needed. We are also interested in using electronic lab notebooks instead of paper-based ones, but there is no spare time to develop own tools for that. The pressure to publish is high so that's all still up in the air.

## 7.9 Interview IX: Zeitgenössische Kunstgeschichte

### Angaben zur Person

Name	Stefanie Gerke, M.A.
Institut oder Einrichtung	Institut für Kunst- und Bildgeschichte, Lehrstuhl für Kunst und neue Medien
Position	Wissenschaftliche Mitarbeiterin

#### 1. Bitte stellen Sie kurz Ihren Arbeitsbereich vor.

In meiner Doktorarbeit beschäftige ich mich mit Architektur in Werken zeitgenössischer Kunst. Zu meinen Forschungsschwerpunkten gehören auch Geschichte und Theorie der Fotografie, Raumtheorie und aktuelle Ausstellungspraxis.

#### 2. Was sind "Forschungsdaten" (=Grundlage Ihrer wissenschaftlichen Arbeit) in Ihrem Arbeitsbereich?

Bilder und bewegte Bilder (Filme und Videos), Textdokumente, persönliche Gespräche mit Künstlern.

#### 3. Welche Datentypen sind es? Bitte nennen Sie konkrete Formate und/oder gängige Software.

Bilder, Fotografien (JPEG), diverse Videoformate (VLC Player zum Abspielen), Texte aus Literaturdatenbanken (PDFs), Audioaufnahmen (MP3) und Interview-Transkripte, ansonsten gängige Office-Pakete. Damit ich von vielen verschiedenen Orten mobil arbeiten kann, nutze ich Dropbox als meinen persönlichen Speicher. So habe ich Zugang zu meinen Forschungsdaten von überall. Für Annotationen von Bildern und persönliche Notizen benutze ich auch [Evernote](#).

#### 4. Werden die Forschungsdaten dokumentiert oder beschrieben? Erfolgt es nach einem eigenentwickelten oder standardisierten Muster (z.B. Metadatenschema, kontrolliertes Vokabular, Ontologie)? Welche Eigenschaften werden damit erfasst?

In der Kunstgeschichte gibt es eine lange Tradition, in welcher Reihenfolge die Angaben in der Bildunterschrift gemacht werden. Für Bilder gilt es: Künstler, Werktitel, Jahr, Material, Maße, Standort, Rechteinhaber für Bildrechte. Bei Bewegtbildern kommen noch technische Metadaten hinzu, wie z.B. Länge des Videos oder Films, 16mm oder 32mm Film, [NTSC](#) oder [PAL](#)-Kodierung, HD oder nicht. Es hängt natürlich davon ab, welches Material man vor sich liegen hat.

#### 5. Werden die Forschungsdaten nach dem Ende eines Forschungsprojektes archiviert? Wie sieht die übliche Praxis aus?

Ich beschäftige mich mit Kunstwerken, die bereits an die Öffentlichkeit getragen worden sind. Insofern muss ich keine originären Daten archivieren. Meine Analysen dokumentiere ich vor allem in meiner Doktorarbeit.

#### 6. Hatten Sie bereits Erfahrungen mit der Veröffentlichung oder Nachnutzung der Forschungsdaten? Gab es Schwierigkeiten dabei?

Material für meine Arbeit finde ich u.a. im [prometheus-Bildarchiv](#), in Katalogen oder auf Online-Videodatenbanken. Meist bekomme ich es aber vor allem von den Künstlern selbst oder ihren Galerien zur Verfügung gestellt. Beim Veröffentlichen oder Nachnutzen von Bildern muss man aber sorgfältig mit Bildrechten umgehen, also diese zuvor einholen bzw. kaufen. Dafür gibt es aber bereits etablierte Abläufe, so bin ich auch in meinem eigenen Aufsatz für eine polnisch-englischsprachige Online-Zeitschrift [Widok](#) vorgegangen.

#### 7. Welche (zentrale) Serviceleistungen zum Umgang mit digitalen Forschungsdaten sehen Sie als notwendig?

Die drei am meisten nachgefragten Serviceleistungen aus der Umfrage (Speicherplatz, rechtliche Beratung, technische Beratung) finde ich sehr sinnvoll und notwendig. Es wäre tatsächlich begrüßenswert, wenn die Universität ein Servicezentrum für alle Fragen rund um Publikationen hätte, auch wenn diese stark mit fachspezifischen Besonderheiten verbunden sind. Bei konkreten Serviceangeboten (z.B. Online-Speicher) wäre es von Vorteil, diese intuitiv zu gestalten und Kommentierbarkeit der Daten zu ermöglichen. Sehr wichtig ist aber auch die Kompatibilität mit verschiedenen Geräten – gerade in den Geisteswissenschaften arbeiten viele sowohl mit PCs als auch mit Macs.

#### **8. Weitere Hinweise und Anmerkungen**

Es ist spannend, wie in unterschiedlichen Fachbereichen mit ähnlichen Fragen umgegangen wird. Die Initiative der Universitätsleitung, durch Kontakt mit den Wissenschaftlern Desiderate in puncto Forschungsdaten zu erfahren, finde ich sehr gut.

## 7.10 Interview X: Deutsche Literatur

### Angaben zur Person

Name	Dr. phil. Anne Baillot
Institut oder Einrichtung	Institut für deutsche Literatur, Nachwuchsgruppe "Berliner Intellektuelle 1800-1830" (Emmy-Noether-Programm)
Position	Leiterin der Nachwuchsgruppe

#### 1. Bitte stellen Sie kurz Ihren Arbeitsbereich vor.

Die Nachwuchsgruppe ist ein Team bestehend aus sieben Personen. Zum Teil arbeiten wir traditionell literaturwissenschaftlich, es wird aber auch eine digitale Edition ["Briefe und Texte aus dem intellektuellen Berlin um 1800"](#) aufgebaut. Die Idee dabei ist, die Korrespondenzen von Autoren wie August Boeckh oder Adelbert von Chamisso digital zu erschließen und miteinander zu verknüpfen, um Entstehungs- und Rezeptionsgeschichte von bestimmten Werken zu beleuchten.

#### 2. Was sind "Forschungsdaten" (=Grundlage Ihrer wissenschaftlichen Arbeit) in Ihrem Arbeitsbereich?

Wir produzieren Bücher, Aufsätze, Tagungsbände. Wir erschließen Werkmanuskripte, Vorlesungsmitschriften und andere Archivalien. Briefe und Handschriften werden dank einer Kooperationsvereinbarung speziell von Archiven und Bibliotheken zur Verfügung gestellt. Diese werden dann manuell transkribiert und als XML-Dokument ediert.

#### 3. Welche Datentypen sind es? Bitte nennen Sie konkrete Formate und/oder gängige Software.

Briefe bekommen wir als TIFF-Bilder, nach den Richtlinien der "Text Encoding Initiative" ([TEI P5](#)) wird ein XML-Dokument pro Brief bzw. pro Handschrifteinheit erstellt und zusätzlich eine HTML- und eine PDF-Version generiert. Die technische Umsetzung erfolgt im XML-Editor [Oxygen](#) und mithilfe von Skripten in der Programmiersprache [Perl](#). Bei der Gestaltung der Buchcover, Tagungsplakate und -flyer sowie Poster benutzen wir regelmäßig [Adobe InDesign](#).

#### 4. Werden die Forschungsdaten dokumentiert oder beschrieben? Erfolgt es nach einem eigenentwickelten oder standardisierten Muster (z.B. Metadatenschema, kontrolliertes Vokabular, Ontologie)? Welche Eigenschaften werden damit erfasst?

Aufbauend auf TEI haben wir ein eigenes Metadatenschema entwickelt (s. [Kodierungsrichtlinien](#) des Projekts), das an unsere Forschungsfrage speziell angepasst ist. Wir haben zwei Auszeichnungsschichten: Einmal die textgenetische – was ist geschrieben worden, von wem usw., und eine Netzwerkannotation – wie z.B. Personen, Orte, Werke. Dabei greifen wir auch auf Schnittstellen wie die Gemeinsame Normdatei ([GND](#)) der Deutschen Nationalbibliothek oder das [Personendaten-Repository](#) der Berlin-Brandenburgischen Akademie der Wissenschaften zurück.

#### 5. Werden die Forschungsdaten nach dem Ende eines Forschungsprojektes archiviert? Wie sieht die übliche Praxis aus?

Im Moment werden die gesamte digitale Edition auf einem Server und eine Kopie auf einer Festplatte gespeichert. Eine zuverlässige Lösung für die Langzeitarchivierung und -verfügbarkeit der Daten auch nach 10 oder 20 Jahren muss noch gefunden werden.

#### 6. Hatten Sie bereits Erfahrungen mit der Veröffentlichung oder Nachnutzung der Forschungsdaten? Gab es Schwierigkeiten dabei?

Unsere Edition von [E.T.A. Hoffmanns „Sandmann“](#) wird beispielsweise in manchen Gymnasien bei thematischen Stadtführungen benutzt. Alle von uns erstellten Inhalte der Edition wie XML- und PDF-Dateien werden unter einer Creative Commons-Lizenz ([CC BY 3.0](#)) veröffentlicht. Die Rechteinhaber bei Handschriften und Digitalisaten (die wir selbst nachnutzen) sind aber die

jeweiligen Bibliotheken und Archive.

### **7. Welche (zentrale) Serviceleistungen zum Umgang mit digitalen Forschungsdaten sehen Sie als notwendig?**

Für unsere digitale Edition ist es wichtig, deren Langzeitverfügbarkeit zu sichern. Es geht dabei nicht nur um Speicherplatz, sondern auch um Manpower: Dass jemand dafür eindeutig zuständig und ansprechbar ist, auch wenn z.B. Updates durchgeführt oder die benötigten technischen Ressourcen in einem Projektantrag beziffert werden müssen.

### **8. Weitere Hinweise und Anmerkungen**

In Literaturwissenschaften arbeitet man vorwiegend noch sehr traditionell. Ich finde es aber wichtig, digitale Arbeitsweisen in Methodenausbildung der Studenten zu integrieren, damit künftige Wissenschaftler bereits früh lernen können, wie sie wissenschaftliche Quellen im Netz finden und damit umgehen. Ein solches Seminar wird es an unserem Institut in diesem Sommersemester<sup>19</sup> zum ersten Mal geben.

---

<sup>19</sup> Gemeint ist das Sommersemester 2014.

## 7.11 Interview XI: Kunstgeschichte und Visualisierung

### Angaben zur Person

Name	PD Dr. Dr. Erna Fiorentini
Institut oder Einrichtung	Institut für Kunst- und Bildgeschichte
Position	Heisenberg Fellow (DFG-Projekte „Visualisation. A Critical Survey of the Concept“ und „Induction of Visibility“)

#### 1. Bitte stellen Sie kurz Ihren Arbeitsbereich vor.

Im Projekt zur Induktion von Sichtbarkeit geht es um Visualisierungsprozesse in der Kunst- und Wissenschaftsgeschichte aber auch in verschiedenen anderen Bereichen wie Ökonomie und Literaturwissenschaft. Es ist eine theoretische und methodische Frage, inwieweit der Begriff der Induktion von Sichtbarkeit alle Facetten umfassen kann, die ich an Beispielen wie der Arbeiten vom Histologen [Santiago Ramón y Cajal](#) oder dem bildenden Künstler [William Kentridge](#) erforsche.

#### 2. Was sind "Forschungsdaten" (=Grundlage Ihrer wissenschaftlichen Arbeit) in Ihrem Arbeitsbereich?

Es sind die Praktiken dieser Personen, die man aus Zeugnissen wie eigenen Aussagen zu ihrer Arbeit oder aus den Objekten (Bildern) selbst ableiten kann; außerdem Berichte von Zeitgenossen und Kritikern und manchmal auch persönliche Gespräche mit den Künstlern.

#### 3. Welche Datentypen sind es? Bitte nennen Sie konkrete Formate und/oder gängige Software.

Für Vorlesungen, Präsentationen und Publikationen benutze ich digitale Bilder, die meist im TIFF oder JPEG-Format vorliegen. Beim Untersuchen von z.B. Beschaffenheit des Bildes ist es aber manchmal notwendig, sich das Original vor Ort anzuschauen, weil kein Programm es so genau darstellen kann. Ansonsten arbeite ich mit üblicher Office-Software.

#### 4. Werden die Forschungsdaten dokumentiert oder beschrieben? Erfolgt es nach einem eigenentwickelten oder standardisierten Muster (z.B. Metadatenschema, kontrolliertes Vokabular, Ontologie)? Welche Eigenschaften werden damit erfasst?

Die meisten Bilder kommen oft aus Datenbanken wie [prometheus](#), [easyDB](#) oder dem [Marburger Fotoarchiv](#), von daher werden sie bereits dort dokumentiert. Da die meisten Bilder in diesem Forschungsbereich aus Büchern entnommen werden, müssen bibliographische Angaben zum Buch (Autor/Herausgeber, Titel, Erscheinungsort, Jahr), Angaben zum Künstler (Name, ggf. Pseudonym, Lebensdaten) und zum Bild selbst (Autor, Titel, Jahr, Material, Technik, Stelle im Buch, Größe und Standort des Originalbildes) gemacht werden. Am IKB kümmert sich die [Mediathek](#) darum.

#### 5. Werden die Forschungsdaten nach dem Ende eines Forschungsprojektes archiviert? Wie sieht die übliche Praxis aus?

Die benutzten Materialien werden in Publikationen eingebettet, insofern werden sie dadurch archiviert. Zusätzlich mache ich Kopien auf Speichermedien und mir anderweitig zugänglichen Ressourcen (Universitäts-Backup-Server).

#### 6. Hatten Sie bereits Erfahrungen mit der Veröffentlichung oder Nachnutzung der Forschungsdaten? Gab es Schwierigkeiten dabei?

Während meiner Zeit am MPWG Berlin habe ich eine Open Digital Library („[Drawing with optical instruments](#)“) herausgegeben, die im ECHO-Portal („European Cultural Heritage Online“) zur Verfügung steht. Es ist eine öffentlich zugängliche Bilddatenbank, die zum Recherchieren nach Quellen, Abbildungen, Kategorien oder Instrumenten benutzt werden kann. Damals gab es noch

keine Templates für solche Datenbanken, wir haben alles proaktiv selbst aufgebaut. Heutzutage sind solche zentralisierten Konglomerate wie Prometheus natürlich mehr praktisch.

### **7. Welche (zentrale) Serviceleistungen zum Umgang mit digitalen Forschungsdaten sehen Sie als notwendig?**

Ein hilfreicher Service wäre, eine Art zentrales Archiv zu haben, wo Materialien aus Projekten sicher gespeichert werden können. Das würde dem Wissenschaftler die Mühe ersparen, jedes Mal beim Verlassen des Instituts oder Auslaufen eines Projekts alle die Daten migrieren zu müssen.

### **8. Weitere Hinweise und Anmerkungen**

In der Kunstgeschichte müssen wir uns intensiv mit Copyright-Fragen beschäftigen. Es ist ein ganz wichtiges Gebot, auch weil die Rechtsgrundlagen sich verändern. Eine Kontaktstelle, an die man sich mit einem konkreten Problem wenden könnte, wäre natürlich toll. Allerdings sind es immer sehr spezielle Aspekte, daher wären lokale Ansprechpartner direkt am Institut für alle damit verbundenen Fragen vielleicht doch besser geeignet.

## 7.12 Interview XII: Bodenkunde und Standortlehre

### Angaben zur Person

Name	Dipl.-Ing. Paul Schulze
Institut oder Einrichtung	Albrecht Daniel Thaer-Institut für Agrar- und Gartenbauwissenschaften, Fachgebiet Bodenkunde und Standortlehre
Position	Projektmitarbeiter

#### 1. Bitte stellen Sie kurz Ihren Arbeitsbereich vor.

Aktuell arbeite ich im Projekt [ELaN](#) (Entwicklung eines integrierten Landmanagements durch nachhaltige Wasser- und Stoffnutzung in Nordostdeutschland) am Aufbau eines Entscheidungsunterstützungssystems (eng. *decision support system*) für Landwirte, die eine Moorfläche torfschonend bzw. torferhaltend bewirtschaften wollen. Am Beispiel eines früheren Projekts [DSS-WAMOS](#) wird ein webbasiertes Tool entwickelt. Durch einfache Ja-Nein-Fragen wird der Nutzer zu den Themenbereichen Wassermanagement, Restriktionen und Wasserrückhalt befragt. Als Ergebnis erhält er eine druckbare Zusammenfassung über den aktuellen Zustand seines Moores sowie passende Handlungs- und Nutzungsempfehlungen.

#### 2. Was sind "Forschungsdaten" (=Grundlage Ihrer wissenschaftlichen Arbeit) in Ihrem Arbeitsbereich?

Um die Größe einer Moorfläche zu messen, benutzen wir GPS-Koordinaten und Polygone, Luftaufnahmen, oder — wenn keine Daten vorhanden sind — wir laufen die Ränder des Moors vor Ort ab. Wir erheben stratigraphische Daten (Moormächtigkeit, Bodenhorizonte), berechnen den Grundstoffhaushalt, physikalische (Wassergehalt, Lagerungsdichte) und chemische (pH-Wert, C-Gehalt, Carbonat-Gehalt) Boden- bzw. Torfparameter.

#### 3. Welche Datentypen sind es? Bitte nennen Sie konkrete Formate und/oder gängige Software.

Geodaten als GPS-Punkte und Shape-Dateien (\*.gpx, \*.shp) mit [ESRI ArcGIS](#), stratigraphische Daten und andere Parameter als Tabellen mit Microsoft Excel. 3D-Modelle von GIS-Shapes können mithilfe von [Golden Software Surfer](#) visualisiert werden. Beim Ablaufen eines Moors nutzen wir zudem einen Bohrer oder eine Moorklappsonde, um die Eigenschaften des Substrates vor Ort zu analysieren.

#### 4. Werden die Forschungsdaten dokumentiert oder beschrieben? Erfolgt es nach einem eigenentwickelten oder standardisierten Muster (z.B. Metadatenschema, kontrolliertes Vokabular, Ontologie)? Welche Eigenschaften werden damit erfasst?

Die [Bodenkundliche Kartieranleitung](#) dient als Standardwerk zur Bestimmung der Boden- und Standorteigenschaften im Feld. Viele Bodenparameter wie Gefüge, Textur, Kalkgehalt, Farbe des Bodens und weitere werden in Tabellen der Kartieranleitung als Spannbreite für Messwerte vorgegeben. Viele chemische Parameter werden im Anschluss im Labor bestimmt um ggf. die Messwerte der Feldansprache zu verbessern. Die Bodenkundliche Kartieranleitung definiert ganz klar, was wie zu nennen ist, damit andere Kollegen es genauso verstehen. Mit einem Entscheidungsunterstützungssystem innerhalb der Kartieranleitung kann der „deutsche Bodentyp“ und mit dem DSS der [FAO](#) (Food and Agriculture Organization of the United Nations) der internationale Bodentyp bestimmt werden.

#### 5. Werden die Forschungsdaten nach dem Ende eines Forschungsprojektes archiviert? Wie sieht die übliche Praxis aus?

Archivierung erfolgt auf unseren NAS, Projektberichte werden an den Projektträger als DVD übergeben und liegen auch in gedruckter Form im Fachgebiet zur Ausleihe vor. Eine elektronische Version kann auch als PDF veröffentlicht werden wie im Fall eines [Ratgebers für die](#)

[Grünlandbewirtschaftung in Brandenburg](#) auf dem Medien-Repository der HU. Artikel in Fachzeitschriften wie [Mires and Peat](#), [Geoderma](#), [Telma](#) werden auch publiziert.

#### **6. Hatten Sie bereits Erfahrungen mit der Veröffentlichung oder Nachnutzung der Forschungsdaten? Gab es Schwierigkeiten dabei?**

Der Versuch, ein bestehendes Softwareprodukt zu übernehmen und nachzunutzen, kostete mir erhebliche Mühe und Zeit. Die Dokumentation war unvollständig und die Programmlogik ließ sich nicht klar darstellen, weil mir im Wesentlichen nur der Projektbericht und die Quelltexte übergeben wurden. Zum Glück waren die Autoren noch erreichbar und haben auch die Anfragen beantwortet.

#### **7. Welche (zentrale) Serviceleistungen zum Umgang mit digitalen Forschungsdaten sehen Sie als notwendig?**

Eine zentrale Lösung für Langzeitarchivierung, wo man am Ende eines Projekts die Daten sicher ablegen könnte, um beispielsweise die Forschungsfrage auch zum späteren Zeitpunkt nochmals aufzugreifen. Eine OwnCloud als eine bequeme in-house Alternative für Dropbox oder Google Docs mit einem Gast-Zugang für externe Partner. [Etherpad-Schreibmaschine](#) der HU benutze ich für einfache Zwecke wie z.B. zum Erstellen meiner aktuellen To-do-Liste, es eignet sich aber nicht zum Schreiben der Artikel, weil wir sehr viel mit Bildern arbeiten.

#### **8. Weitere Hinweise und Anmerkungen**

Große Unsicherheiten herrschen beim Thema Bildrechte und richtiges Zitieren. Schulungen über [Berufliche Weiterbildung](#) der HU können dabei helfen, aber auch Studenten müssen besser geschult werden — spezielle Module zum Umgang mit Abbildungen aus dem Internet oder Büchern gibt es in Methodenausbildung derzeit nicht.

## 7.13 Interview XIII: Gartenökonomie

### Angaben zur Person

Name	Dr. Bettina König
Institut oder Einrichtung	Albrecht Daniel Thaer-Institut für Agrar- und Gartenbauwissenschaften, Fachgebiet Ökonomik der Gärtnerischen Produktion
Position	Wissenschaftliche Mitarbeiterin

#### 1. Bitte stellen Sie kurz Ihren Arbeitsbereich vor.

Unser Fachgebiet beschäftigt sich mit der Ökonomie von gartenbaulichen Wertschöpfungsketten, Marketing und landwirtschaftlicher Kommunikation und Beratung. Im [Kompetenznetzwerk WeGa](#) werden aktuell effiziente Wissenssysteme für Gartenbau erforscht und die Frage gestellt, wie das in der Wissenschaft erarbeitete Wissen übertragen und für die praktische Anwendung nutzbar gemacht werden kann. Im Projekt [Trans-SEC](#) forschen wir in einem großen Konsortium zum Innovationssystem zur Ernährungssicherung in Tansania. In einer EU-weiten [COST-Aktion](#) zu [Aquaponik](#) vernetzen wir uns u.a. mit internationalen Unternehmen und Startups. Eine neue Nachwuchsgruppe arbeitet zudem transdisziplinär an Innovationen für das nachhaltige Landmanagement.

#### 2. Was sind "Forschungsdaten" (=Grundlage Ihrer wissenschaftlichen Arbeit) in Ihrem Arbeitsbereich?

Es werden Interviews zu Innovationsprozessen mit Unternehmen geführt oder auch Experten-Befragungen in einer zweistufigen [Delphi-Studie](#) gemacht. Daraus entstehen Interview-Aufzeichnungen, Gedächtnisprotokolle, Transkriptionen. In quantitativen Befragungen wie dem Haushalts-Survey in Afrika werden statistische Daten erhoben. Als Workshop-Ergebnisse haben wir Fotos und Workshop-Protokolle. Andere Kollegen im Fachgebiet/ Department arbeiten beispielsweise auch mit Fokusgruppen-Diskussionen oder ökonometrischen oder Klimamodellen.

#### 3. Welche Datentypen sind es? Bitte nennen Sie konkrete Formate und/oder gängige Software.

Interviews werden aufgezeichnet (Audio) und transkribiert (Word-Dokument). Zur qualitativen Auswertung von Interviews nutzen wir die Software [MAX-QDA](#) und Statistiksoftware SPSS, ansonsten Standard-Officepaket. Für Dateiaustausch benutzen wir Dropbox oder [MyDrive](#).

#### 4. Werden die Forschungsdaten dokumentiert oder beschrieben? Erfolgt es nach einem eigenentwickelten oder standardisierten Muster (z.B. Metadatenschema, kontrolliertes Vokabular, Ontologie)? Welche Eigenschaften werden damit erfasst?

Die Eckdaten zum Interview stehen immer in der Kopfzeile im Fragebogen und im Protokoll. Damit wir eine einheitliche Vorgehensweise haben und die Ergebnisse vergleichen können, benutzen wir einen strukturierten Code-Baum.

#### 5. Werden die Forschungsdaten nach dem Ende eines Forschungsprojektes archiviert? Wie sieht die übliche Praxis aus?

Die Daten werden auf externen Festplatten gesichert und im Fachgebiet aufbewahrt. Da es aber keine verbindliche Regelungen oder klare Zuständigkeitszuschreibung dafür gibt, ist die Praxis immer an Personen (Arbeitsvertragslaufzeiten) gebunden.

#### 6. Hatten Sie bereits Erfahrungen mit der Veröffentlichung oder Nachnutzung der Forschungsdaten? Gab es Schwierigkeiten dabei?

Wir arbeiten in inter- und transdisziplinären Verbänden, wo Arbeitspakete zusammen mit Forschungsdaten unserer Partnerorganisationen geteilt werden. Im Haushalts-Survey wurde es beispielsweise so vereinbart, dass derjenige, der eine Frage eingespist hat, auch das Recht hat,

die Ergebnisse zuerst zu nutzen. In Publikationen wird dann der Hauptbearbeiter und der Autor der Frage genannt.

### **7. Welche (zentrale) Serviceleistungen zum Umgang mit digitalen Forschungsdaten sehen Sie als notwendig?**

In Verbundprojekten brauchen wir eine sichere Daten-Cloud, die man an internen Bereich von einer Webseite koppeln könnte. Sie muss aber auch für Praxispartner an anderen Organisationen erreichbar sein. Die rechtliche Beratung ist sehr wichtig und müsste gleich am Anfang der Konzeption von der Datenerhebung erfolgen. Für langfristige Archivierung ist eine Infrastruktur samt technischer und personeller Kapazität notwendig, um die Daten fachkompetent sichern zu können. Da wir teilweise mit sehr sensiblen, wettbewerbsrelevanten Informationen arbeiten, ist es zugleich eine forschungsethische Frage – um die Vertraulichkeit dieser Daten dem Interviewpartner zusichern zu können, muss man einen geschützten und zuverlässigen Speicherplatz haben.

### **8. Weitere Hinweise und Anmerkungen**

Da die Methodenvielfalt in unserer Disziplin größer wird, müssen wir zukünftig auf ein größeres Repertoire an Daten-Expertise zugreifen können. Eine fachliche Beratung zur Qualitätssicherung und dem Aufbau von Forschungsdaten wäre daher hilfreich, um beispielsweise neue Methoden ausprobieren zu können oder die Forschungsdaten nach einem standardisierten Metadatenschema zu beschreiben.

## 7.14 Interview XIV: Völkerrecht

### Angaben zur Person

Name	Prof. Dr. Georg Nolte
Institut oder Einrichtung	Juristische Fakultät, Lehrstuhl für Öffentliches Recht, Völker- und Europarecht
Position	Professor

#### 1. Bitte stellen Sie kurz Ihren Arbeitsbereich vor.

Am Lehrstuhl für Öffentliches Recht, Völker- und Europarecht arbeiten wir zur Zeit hauptsächlich im Bereich völkerrechtliche Verträge und ihrer späteren Praxis sowie allgemein zur Entwicklung des Völkerrechts.

#### 2. Was sind "Forschungsdaten" (=Grundlage Ihrer wissenschaftlichen Arbeit) in Ihrem Arbeitsbereich?

Wir arbeiten mit Standardliteratur und –datenbanken. Die wichtigsten „Rohstoffe“ für unsere wissenschaftliche Arbeit sind offizielle Dokumente von internationalen Organisationen wie United Nations, Gesetzestexte, Gerichtsurteile, Kommentare, Sitzungsprotokolle, Berichte in den Massenmedien, wissenschaftliche Artikel und Monographien.

#### 3. Welche Datentypen sind es? Bitte nennen Sie konkrete Formate und/oder gängige Software.

Wir fragen Datenbanken ab und schreiben unsere Texte in Standardsoftware (Office-Pakete). Die von uns abgespeicherten und geschriebenen Daten sind vergleichsweise nicht sehr umfangreich. Es handelt sich praktisch nur um Textdokumente, ggf. Karten im JPG-Format und Interview-Aufzeichnungen als Audio-Dateien.

#### 4. Werden die Forschungsdaten dokumentiert oder beschrieben? Erfolgt es nach einem eigenentwickelten oder standardisierten Muster (z.B. Metadatenschema, kontrolliertes Vokabular, Ontologie)? Welche Eigenschaften werden damit erfasst?

Wir gehen da eher ad hoc vor; zu Textpublikationen werden Literaturlisten mit üblichen bibliographischen Angaben angelegt.

#### 5. Werden die Forschungsdaten nach dem Ende eines Forschungsprojektes archiviert? Wie sieht die übliche Praxis aus?

Unsere Forschungsdaten sind bereits veröffentlichte Dokumente und somit vom Herausgeber bereits archiviert. Speicherplatz für Rechner wird vom CMS verwaltet, gelegentlich werden Daten auch auf externen Festplatten gespeichert.

#### 6. Hatten Sie bereits Erfahrungen mit der Veröffentlichung oder Nachnutzung der Forschungsdaten? Gab es Schwierigkeiten dabei?

Bei uns stellen sich nur die üblichen Fragen des Copyright für eigene und fremde Publikationen. Ich habe ein Interesse daran, dass meine Publikationen im Netz möglichst frei abrufbar sind.

#### 7. Welche (zentrale) Serviceleistungen zum Umgang mit digitalen Forschungsdaten sehen Sie als notwendig?

Für meine Zwecke reichen die zentralen Serviceleistungen der HU aus. Zur Ablage von Preprints als Open Access vor der eigentlichen Publikation bei einem Verlag nutzen wir [SSRN – Social Science Research Network](#), Software wie „Dropbox“ kommt insbesondere bei Kollaborationen mit anderen Einrichtungen zum Einsatz. Speicherplatz hatten wir bislang genug und ich sehe auch keine Probleme voraus.

#### 8. Weitere Hinweise und Anmerkungen

Dezentrale IT-Beratung vor Ort ist für unsere Zwecke sehr hilfreich und soll nach Möglichkeit ausgebaut werden.

## 7.15 Interview XV: Information Retrieval

### Angaben zur Person

Name	Prof. Vivien Petras, PhD
Institut oder Einrichtung	Institut für Bibliotheks- und Informationswissenschaft, Lehrstuhl für Information Retrieval
Position	Professorin

#### 1. Bitte stellen Sie kurz Ihren Arbeitsbereich vor.

Das Forschungsgebiet Information Retrieval erforscht den Zugang zu Informationen. Meine Spezialgebiete sind dabei das mehrsprachige Retrieval (Suche und Recherche), kontrollierte Vokabulare und Evaluation. Momentan arbeiten wir an [Europeana](#), einer großen digitalen Bibliothek, und am Aufbau von Linked Data-Infrastruktur für Europeana-Daten im [DM2E-Projekt](#) („Digitized Manuscripts to Europeana“). In früheren Projekten wie [Galateas](#) und [Promise](#) wurde beispielsweise Software für automatische Übersetzung und Klassifikation von Suchanfragen in Informationssystemen entwickelt.

#### 2. Was sind "Forschungsdaten" (=Grundlage Ihrer wissenschaftlichen Arbeit) in Ihrem Arbeitsbereich?

Um Informationssysteme testen und miteinander vergleichen zu können, generieren wir sogenannte Test-Korpora – das sind Dokumente, Anfragen und die Bewertung, ob ein Dokument für eine bestimmte Anfrage relevant ist. Wir haben digitale Objektdaten, Metadaten, Annotationen, Log-Dateien. Wir berechnen viele Maßzahlen und bereiten diese dann visualisiert auf.

#### 3. Welche Datentypen sind es? Bitte nennen Sie konkrete Formate und/oder gängige Software.

Für die Auswertung der Evaluationsergebnisse benutzen wir meistens eigenständig entwickelte Skripte, die auf Java oder Python laufen. Ansonsten Statistik-Programme wie SPSS und R, Excel, teilweise auch Datenbanken wie [MongoDB](#) und spezielle Programme für Log-Dateien wie [Piwik](#) und [Awstats](#).

#### 4. Werden die Forschungsdaten dokumentiert oder beschrieben? Erfolgt es nach einem eigenentwickelten oder standardisierten Muster (z.B. Metadatenschema, kontrolliertes Vokabular, Ontologie)? Welche Eigenschaften werden damit erfasst?

Die Evaluationskorpora sind immer dokumentiert, weil davon ausgegangen wird, dass die von anderen Forschern in der Community wieder verwendet werden. Im DM2E-Projekt werden Provenienz-Metadaten für Metadaten von digitalen Objekten automatisch erfasst – wann und von wem sie generiert, auf welches Modell sie gemappt und wie oft sie geändert wurden. In diesem Bereich gibt es [Open Annotation](#) als Standardformat mit vorgegebener Struktur und Community-Formate wie Europeana Datenmodell ([EDM](#)) und TREC ([Text REtrieval Conference](#)), einer Konferenz für Austausch und Bewertung von Ergebnissen in Information Retrieval.

#### 5. Werden die Forschungsdaten nach dem Ende eines Forschungsprojektes archiviert? Wie sieht die übliche Praxis aus?

Daten in laufenden Projekten werden auf Servern gespeichert, Kopien auf mehreren Festplatten gemacht. Ein statisches Paket mit Daten und Dokumentation irgendwo abzulegen ist an sich kein Problem, viel wichtiger ist aber, dass die Software und Infrastruktur weiter gepflegt werden müssen. Angesichts der befristeten Projektförderung sind lange Zeiträume schwer zu garantieren, was vor allem mit den dafür notwendigen Personalkosten verbunden ist.

#### 6. Hatten Sie bereits Erfahrungen mit der Veröffentlichung oder Nachnutzung der

### **Forschungsdaten? Gab es Schwierigkeiten dabei?**

Ein ausgesprochenes Ziel in unserer Community ist die Wiederholbarkeit der Evaluationskorpora, damit auch andere Forscher sie benutzen und sich alle aneinander messen können. Deshalb existiert eine auf persönlichen Anfragen basierte „Data Sharing-Kultur“. Man zitiert in der Regel den Artikel, der die Entwicklung des Datenkorpus beschreibt. Mithilfe des DIRECT-Portals im Promise-Projekt ist es aber möglich, einen Datensatz mit einem DOI direkt zu zitieren. Unsere Software aus dem DM2E-Projekt ist auch auf [Github](#) öffentlich zugänglich.

### **7. Welche (zentrale) Serviceleistungen zum Umgang mit digitalen Forschungsdaten sehen Sie als notwendig?**

Eine zentrale Server-Administration, die man als Leistung und Speicherplatz kaufen könnte, ohne sich um Hardware oder Backup kümmern zu müssen. Eine „akademische Cloud“ oder ein sogenanntes Dark Archive, wo man eigene Daten sicher ablegen und bei Bedarf anderen zugänglich machen könnte. Für spezielle Fragen braucht man aber einen IT-Ansprechpartner vor Ort, mit dem man direkt kommunizieren und das Problem oder was man exakt braucht erklären könnte.

### **8. Weitere Hinweise und Anmerkungen**

Idealerweise würde ich auch meinen Studenten eigene virtuelle Maschinen geben, damit sie Fragen in Information Retrieval selbst ausprobieren können. Aufgrund der Ressourcen-Knappheit ist es aber momentan unrealistisch. In einem EU-Projekt wollen wir zudem mit dem Fördermittelgeber in Dialog treten, wie man Projektergebnisse auch über längere Zeiträume hinaus nachhaltig sichern kann.

## 7.16 Interview XVI: Mittelalterliche Geschichte

### Angaben zur Person

Name	Dr. Stefan Schlelein
Institut oder Einrichtung	Institut für Geschichtswissenschaften, Fachgebiet Mittelalterliche Geschichte II, Sonderforschungsbereich 644 „Transformationen der Antike“
Position	Wissenschaftlicher Koordinator

#### 1. Bitte stellen Sie kurz Ihren Arbeitsbereich vor.

Unser Sonderforschungsbereich beschäftigt sich mit sämtlichen Erscheinungsformen des Nachlebens, der Rezeption und der Bezugnahmen auf die Antike in den nachantiken Epochen Mittelalter und Neuzeit bis in die Gegenwart. Das können Renaissancehumanisten sein, die das Latein des Cicero wiederbeleben wollten, die frühen 'Sandalenfilme' oder die Herausbildung des Wissenschaftssystems im 19. Jahrhundert – also ein sehr breites Spektrum. Ich selbst bin Mittelalterhistoriker mit den Schwerpunkten Humanismus und Renaissancegeschichte und als wissenschaftlicher Koordinator und Geschäftsführer des SFB für die gesamte Koordination, administrative Abwicklung und inhaltliche Planung der Aktivitäten verantwortlich.

#### 2. Was sind "Forschungsdaten" (=Grundlage Ihrer wissenschaftlichen Arbeit) in Ihrem Arbeitsbereich?

Das sind in erster Linie historische Quellen wie z.B. Inkunabeldrucke, historiographische Texte des 16. Jahrhunderts, Gedichte zum Siebenjährigen Krieg oder wissenschaftliche Texte des 17. und 18. Jahrhunderts. In kunsthistorischen Projekten wie zur Erforschung von Ausstellungskonzepten und Aufstellungskontexten im Museum im 19. Jahrhundert dienen auch Bilder oder Abbildungen von Kunstgegenständen und schriftliche Aufzeichnungen von der Projektskizze bis zum Ausstellungskatalog als Forschungsmaterial, ferner (Spiel-)Filme mit Antikethematik.

#### 3. Welche Datentypen sind es? Bitte nennen Sie konkrete Formate und/oder gängige Software.

Hauptsächlich Texte in Form von Digitalisaten oder Bildern (häufig als PDF ohne OCR), Bilddateien in TIFF oder JPEG, Filme, Datenbanken (MySQL und MS Access), bibliographische Daten in Citavi, ansonsten die üblichen Office-Anwendungen.

#### 4. Werden die Forschungsdaten dokumentiert oder beschrieben? Erfolgt es nach einem eigenentwickelten oder standardisierten Muster (z.B. Metadatenschema, kontrolliertes Vokabular, Ontologie)? Welche Eigenschaften werden damit erfasst?

In einem Übersetzungsprojekt werden beispielsweise Metadaten zu Übersetzungen wie Publikationsort, Publikationsdaten, Rezensionen sehr genau in einer eigenentwickelten Datenbank erfasst. Die Praxis ist aber sehr unterschiedlich über Teilprojekte hinweg und wird nicht systematisch umgesetzt.

#### 5. Werden die Forschungsdaten nach dem Ende eines Forschungsprojektes archiviert? Wie sieht die übliche Praxis aus?

Zur Sicherung der Daten für den laufenden Betrieb werden Kopien auf USB-Sticks oder Festplatten gemacht. Mein Eindruck ist, dass eine langfristige Archivierung auch über das Förderungsende eines Drittmittelprojekts wie unseres SFBs hinaus nicht üblich ist. Wir haben dies zwar vor, wissen aber noch nicht genau, wie wir es am besten machen sollen.

#### 6. Hatten Sie bereits Erfahrungen mit der Veröffentlichung oder Nachnutzung der Forschungsdaten? Gab es Schwierigkeiten dabei?

Die Nutzung von historischen Quellen ist an sich eine Art Nachnutzung. Mit der Nachnutzung von unseren Forschungsergebnissen haben wir noch keine Erfahrung, weil es im Grunde an dem

Punkt einsetzt, an dem es den SFB nicht mehr gibt.

### **7. Welche (zentrale) Serviceleistungen zum Umgang mit digitalen Forschungsdaten sehen Sie als notwendig?**

Zum einen wäre ein zentrales, sichtbares, gut kommuniziertes und ständig verfügbares Beratungsangebot wichtig, sowohl für einzelne Forschende als auch für größere Projekte bzw. die Projektverantwortlichen oder -koordinatoren. Diese Beratung sollte dauerhaft verfügbar sein: bereits bei der Vorbereitung der Antragstellung, im Laufe des Projekts und auch dann, wenn es zum Abschluss gebracht werden soll. Das Zweite ist die entsprechende Bereitstellung der technischen Kapazitäten inklusive Betreuungspersonals.

### **8. Weitere Hinweise und Anmerkungen**

Der Knackpunkt bei der Bereitstellung zentraler Serviceleistungen ist die personelle Ausstattung – es reicht nicht, einen Rechner irgendwo hinzustellen, er muss von einer entsprechenden Anzahl der Personen betreut werden. Die Universität muss dies als eine Daueraufgabe erkennen. Rundum ausgebildete Ansprechpartner zu inhaltlichen Fragen zu haben, eine Art digitale Bibliothekare oder Archivare, wäre ideal. Von der Praktikabilität her müsste es aber eher dezentral angeboten werden.

## 7.17 Interview XVII: Steuerlehre

### Angaben zur Person

Name	Prof. Dr. Henriette Houben
Institut oder Einrichtung	Wirtschaftswissenschaftliche Fakultät, Ernst & Young Stiftungs-Juniorprofessur für Quantitative Betriebswirtschaftliche Steuerlehre
Position	Professorin

#### 1. Bitte stellen Sie kurz Ihren Arbeitsbereich vor.

Ich arbeite im Bereich Steuern und forsche viel zur Erbschaftssteuer. Es handelt sich um Quantifizierung der Steuer (Wie hoch werden Unternehmen mit der aktuellen Erbschaftssteuer belastet?) und mögliche Alternativen, wie z.B. Was wäre, wenn die aktuelle Erbschaftssteuer für verfassungswidrig erklärt würde; wie soll sie stattdessen reformiert werden oder wie hoch dürfen Belastungen sein, ohne Unternehmen zu gefährden.

#### 2. Was sind "Forschungsdaten" (=Grundlage Ihrer wissenschaftlichen Arbeit) in Ihrem Arbeitsbereich?

Wir arbeiten mit Steuerstatistiken wie der faktisch anonymisierten Einkommenssteuerstatistik (FAST), [Einkommens- und Verbrauchsstichprobe \(EVS\)](#) oder dem Sozio-oekonomischen Panel (SOEP). Die Steuerdaten werden über entsprechende Forschungsdatenzentren des [Statistischen Bundesamtes](#) bzw. des [DIW Berlin](#) bezogen. In diesen Sekundärdaten liegt ein großer Schatz an Informationen, die von Forschern noch erschlossen werden müssen. Eigene Primärdaten erheben wir nicht.

#### 3. Welche Datentypen sind es? Bitte nennen Sie konkrete Formate und/oder gängige Software.

Das Statistische Bundesamt gibt vor, dass große Datensätze (mehr als eine Million Beobachtungen) über Fernrechnen mit SAS zu bearbeiten sind. In Einzelfällen nutzen wir auch Stata und R, oder schreiben ein eigenes Programm in C++. Die eigentliche Software zur Auswertung der Statistikdaten ist aber SAS.

#### 4. Werden die Forschungsdaten dokumentiert oder beschrieben? Erfolgt es nach einem eigenentwickelten oder standardisierten Muster (z.B. Metadatenschema, kontrolliertes Vokabular, Ontologie)? Welche Eigenschaften werden damit erfasst?

Die Dokumentation der Sekundärdaten wird von Forschungsdatenzentren bereitgestellt. Wir haben aber auch ein eigenentwickeltes Wiki-basiertes System für internen Gebrauch aufgesetzt, damit wir die Datenplausibilisierung vornehmen und die Zusammenhänge nochmals überprüfen können.

#### 5. Werden die Forschungsdaten nach dem Ende eines Forschungsprojektes archiviert? Wie sieht die übliche Praxis aus?

Die Archivierung der Sekundärdaten wird von Forschungsdatenzentren übernommen. Wenn wir ein Projekt abgeschlossen haben, werden zugehörige Daten und Dokumentation auf externen Festplatten gespeichert und im Safe abgelegt.

#### 6. Hatten Sie bereits Erfahrungen mit der Veröffentlichung oder Nachnutzung der Forschungsdaten? Gab es Schwierigkeiten dabei?

Die Statistikdaten unterliegen sehr restriktiven Anforderungen des Statistischen Bundesamtes und dürfen nicht öffentlich publiziert werden. Andererseits wird in Publikationen immer auf Quellenangaben verwiesen. Dementsprechend hat jeder Forscher die Möglichkeit, den Zugang zu diesen Daten beim Forschungsdatenzentrum zu beantragen.

## **7. Welche (zentrale) Serviceleistungen zum Umgang mit digitalen Forschungsdaten sehen Sie als notwendig?**

Rechtliche Beratung gleich verknüpft mit technischen Lösungen wäre für uns sehr hilfreich. Viele Fragen hinsichtlich der praktischen Umsetzung von datenschutzrechtlichen Anforderungen sind offen: Wo dürfen welche Daten gespeichert werden? Darf der Rechner (z.B. Laptop) das Gebäude der Universität verlassen? Müssen die Daten passwortgeschützt (Anforderungen an Passwort?) oder sogar verschlüsselt sein? Kann man sich auf Firewall-Lösung verlassen? Ist ein zentraler Backup erforderlich bzw. ausreichend? Was genau wird von mir erwartet? Ein ganzheitliches Beratungs-Know-How mit sehr konkreten Handlungsanweisungen würde ordnungsgemäße Handhabung von sensiblen Daten maßgeblich erleichtern.

## **8. Weitere Hinweise und Anmerkungen**

Eine Lösung für Langzeitarchivierung im Sinne der Guten wissenschaftlichen Praxis muss unbedingt zentral angeboten werden, weil der einzelne Wissenschaftler oft die Institution innerhalb von deutlich weniger als zehn Jahren wechselt. Zu diesem Zweck wäre es sehr angeraten, einen institutionellen Rahmen zur Verfügung zu stellen, wie die Daten übergeben werden können. Von Datenlieferanten wird zudem erwartet, dass Universitäten und Forschungseinrichtungen selbst Vorschriften haben, wie mit Sekundärdaten umzugehen ist und wie sie zu halten sind. Ein Leitfaden für dafür notwendige Grundausstattung wie beispielsweise eine Safe-Anschaffung wäre daher ebenfalls hilfreich.