Workshop zum Forschungsdatenmanagement in der Romanistik

Inhalt

1. Einleitung 2
2. Panel I: Bestehende Infrastrukturen für ein nachhaltiges Forschungsdatenmanagement (in den Geisteswissenschaften) 2
2.1. Elisabeth Burr: Zur Geschichte des Datenmanagements (in der Romanistik) und zu den verfügbaren Infrastrukturen 2
2.2. Elke Teich: *CLARIN-D*: Unterstützung von Projekten aus den Philologien 3
2.3. Mirjam Blümm/Lisa Klaffki: Services von *DARIAH-DE/TextGrid* 3
2.4. Mirjam Blümm/Lisa Klaffki: DOI-Service für die Geisteswissenschaften und *DARIAH-DE*-Repository 5
2.5. Lisa Klaffki: Aktivitäten der DHd *AG Datenzentren* 5
2.6. Diskussion 6
3. Panel II: Verschiedene Lösungen für ein nachhaltiges Forschungsdatenmanagement in der Romanistik 7
3.1. Hanno Ehrlicher/Nanette Rißler-Pipka: Die virtuelle Forschungsumgebung *Revistas Culturales* 7
3.2. Christoph Bürgel/Sascha Diwersy: *Sketch Engine* als Arbeits- und Suchsystem für linguistische Korpora am Beispiel des *Corpus de référence du français contemporain* (CRFC) 8
3.3. Diskussion 10
4. Panel III: Spezielle Instrumente zur Verbesserung der Nachhaltigkeit, Nachnutzbarkeit, Auffindbarkeit von Forschungsdaten 10
4.1. Georg Rehm: *META-SHARE* als System zur Sicherung und Suche romanistischer Sprachressourcen 11
4.2. Heike Renner-Westermann: Aktivitäten und Perspektiven des Linguistik-Portals mit Blick auf Forschungsdaten 12
4.3. Jan Rohden: Perspektiven zur Verbesserung der Suchbarkeit romanistischer Forschungsdaten im Rahmen des *FID Romanistik* 13
4.4. Diskussion 14
5. Panel IV: Beratungsdienste und Informationsangebote 15
5.1. Patrick Helling: Das Dienstleistungs- und Beratungsangebot des Kölner *Data Centers for the Humanities* 15
5.2. Diskussion 16
6. Schlusswort 17

1. Einleitung

Am 01. Dezember 2017 fand an der Universitäts- und Landesbibliothek (ULB) Bonn ein von der *AG Digitale Romanistik* und dem *Fachinformationsdienst* (*FID) Romanistik* veranstalteter Workshop zum Forschungsdatenmanagement in der Romanistik statt. Ziel war die Entwicklung konkreter Unterstützungsmaßnahmen, ausgehend von den auf einem vorherigen Workshop ermittelten fachspezifischen Bedarfen.[1]

In seiner Begrüßung betonte der Direktor der ULB Bonn Ulrich Meyer-Doerpinghaus die Bedeutung von Forschungsdaten für die heutige Wissenschaftslandschaft und das daraus erwachsende Engagement der ULB Bonn in diesem Bereich.

Zum Einstieg in das Programm situierte die FID-Referentin Doris Grüter das Thema im Kontext des *FID Romanistik* und stellte kurz die Vorhaben zum Forschungsdatenmanagement in der aktuellen Projektphase vor. Im Rahmen dessen verwies sie auf das vom *FID Romanistik* gemeinsam mit der *AG Digitale Romanistik* entwickelte und auf *romanistik.de* angesiedelte Meldesystem,[2] das Romanistinnen und Romanisten die Möglichkeit bietet, der Wissenschaftscommunity Forschungsdaten als publikationsäquivalente Leistung aufzuzeigen.

Anhand von Vorträgen zu vier thematischen Panels wurden anschließend bedarfsgerechte Instrumente und Möglichkeiten zur Unterstützung der Romanistik beim Forschungsdatenmanagement diskutiert.

2. Panel I: Bestehende Infrastrukturen für ein nachhaltiges Forschungsdatenmanagement (in den Geisteswissenschaften)

Moderation: Christof Schöch

In seiner Einführung zum ersten Panel skizzierte Christof Schöch disziplinäre Besonderheiten der Romanistik. Er wies auf den Umstand hin, dass es für sie in Deutschland anders als etwa für die Germanistik keinen nationalen Dokumentationsauftrag gebe und sie aufgrund ihres Forschungsgegenstandes generell ein internationales Publikum bediene. Daraus ergäben sich auch spezifische Anforderungen für den Umgang mit Forschungsdaten: So müsse nicht zuletzt eine internationale Sichtbarkeit durch den Rückgriff auf Standards sowie eine mehrsprachige Dokumentation der Metadaten und Oberflächen gewährleistet werden. Da die Romanistik zudem einige Berührungspunkte mit anderen Fächern und Disziplinen aufweise, sei es nicht unbedingt sinnvoll, eigene Infrastrukturen aufzubauen. Vielmehr gehe es v.a. darum, die spezifischen romanistischen Anforderungen und Erfahrungen in allgemeine Angebote für die Geisteswissenschaften einzubringen.

2.1. Elisabeth Burr: Zur Geschichte des Datenmanagements (in der Romanistik) und zu den verfügbaren Infrastrukturen

Elisabeth Burr bot in ihrem Vortrag einen Überblick über die Anfänge der computergestützten quantitativen Linguistik. Anhand praktischer Beispiele und eigener wissenschaftlicher Erfahrungen erörterte sie deren lange Tradition seit den prädigitalen Anfängen und schilderte

[1] https://www.ulb.uni-bonn.de/de/fid-blog/downloads/workshop-bericht (12.12.17).
[2] https://www.romanistik.de/res (12.12.17).

dabei auch die früheren Schwierigkeiten bei der Recherche, Verwendung und Nachnutzung relevanter Forschungsdaten. Die damaligen Rahmenbedingungen hatten Auswirkungen auf das Datenmanagement: So waren für die Forschung relevante Korpora zumeist nicht frei verfügbar. Aufgrund fehlender Standards zur Auszeichnung bzw. Modellierung der Daten war die Vergleichbarkeit von Sprachressourcen untereinander oftmals nicht gegeben. Software zur Auszeichnung bzw. Analyse linguistischer Daten war selten und kostspielig. Werkzeuge zur kollaborativen Forschung existierten nicht; die einzige Möglichkeit zum fachspezifischen Informationsaustausch boten wissenschaftliche Kongresse.

Im Fazit ihres Vortrags leitete Elisabeth Burr aus den Problemen des frühen Datenmanagements Handlungsempfehlungen für die Zukunft ab: So sei es mit Blick auf wissenschaftliche Nachhaltigkeit wichtig, die eigenen Forschungsdaten, soweit möglich, zur Nachnutzung bereitzustellen und angemessen zu dokumentieren. Darüber hinaus sollte weitgehend auf freie und von Software unabhängige Formate und Standards zurückgegriffen werden. Der enge Austausch mit anderen Wissenschaftlerinnen und Wissenschaftlern sowie die wechselseitige Beratung zu Fragen des Forschungsdatenmanagements seien ebenfalls von großer Bedeutung.

2.2. Elke Teich: *CLARIN-D*: Unterstützung von Projekten aus den Philologien

Mit Blick auf die Philologien stellte Elke Teich die Virtuelle Forschungsinfrastruktur *Common Language Resources and Technology Infrastructure* (CLARIN) vor. Dort werden die Phasen eines Digital Humanities-Projekts in drei Bereichen von Serviceangeboten zusammengefasst, die sich auf das Auffinden, das Auswerten und schließlich das Aufbereiten sowie Aufbewahren von Daten beziehen.

Die Struktur von *CLARIN-D* ist vernetzt und kooperativ, wobei einzelne integrierte Tools (wie z.B. *WebLicht*, *WebMaus*, *CWB/CQP-Web*, *EXMARaIDA*) bzw. Sammlungen (z.B. *Deutsches Textarchiv*, *Deutsches Referenzkorpus*) von unterschiedlichen Akteuren bereitgestellt und gepflegt werden.

Es ist möglich, eigene Sammlungen mit Daten und Metadaten anzureichern und auf diese Weise zu linguistischen Korpora aufzubereiten. Die Nachhaltigkeit wird dabei durch den Rückgriff auf Standards, die Vergabe persistenter Identifikatoren und die Sicherung der Daten in zertifizierten Datenzentren gewährleistet. Als Tool zur Anreicherung der Daten dient die *IMS Corpus Workbench* (CWB), mit deren Hilfe sowohl automatisierte als auch manuelle Annotationen auf der Ebene der Token, der Textstruktur, der Metadaten und in semantischer Hinsicht vorgenommen werden können. Die angereicherten Daten können dann mittels der Programmiersprache *R* statistisch ausgewertet und visualisiert werden.

Zum Abschluss ihres Vortrags resümierte Teich den Nutzen und die Möglichkeiten von *CLARIN-D* mit Blick auf die Philologien am Beispiel von Projekten zur Erforschung von Phänomenen der Sprachvariation und des Sprachwandels.

2.3. Mirjam Blümm/Lisa Klaffki: Services von *DARIAH-DE/TextGrid*

Mirjam Blümm und Lisa Klaffki präsentierten das Dienstleistungsspektrum der Digitalen Forschungsinfrastruktur für die Geistes- und Kulturwissenschaften *DARIAH-DE*.

Im ersten Teil stellte Mirjam Blümm den Aufbau und die Kooperationspartner von *DARIAH-DE* vor und fasste die vier Kernelemente des Dienstleistungsspektrums zusammen (Lehre, Forschungsdaten, Forschung, technische Infrastruktur), die alle auf wissenschaftliche Nachhaltigkeit abzielen. In der Nachfolge ging die Referentin auf einzelne Dienste und Werkzeuge von *DARIAH-DE* ein.

Sie begann mit *Textgrid*, einer insbesondere für digitale Editionen geeigneten virtuellen Infrastruktur zur (kollaborativen) Bearbeitung, Annotation, nachhaltigen Sicherung und Publikation von Text- und Bilddateien. *Textgrid* besteht im Wesentlichen aus drei Modulen: dem *Textgrid Laboratory*, einer Software zur Auszeichnung bzw. Annotation, dem *Textgrid Repository*, einem Repositorium zur Sicherung bzw. Veröffentlichung der Dateien, und dem Community-Bereich mit Hilfetexten und Hinweisen. Da mit *Textgrid* bisher keine spezifisch romanistischen Projekte realisiert wurden, veranschaulichte Blümm einen exemplarischen Workflow zur Erstellung digitaler Texteditionen am Beispiel der *Fontane Notizbücher*[3] .

Anschließend stellte Lisa Klaffki einige über *DARIAH-DE* nutzbare Werkzeuge vor. Sie begann mit dem *DARIAH-DE Geobrowser*[4], einem Webtool, das es erlaubt, Daten in ihrer historisch-geographischen Verteilung zu visualisieren. Die Formatierung und Eingabe der dazugehörigen Daten ist über ein weiteres Tool möglich: den *DARIAH-DE Datasheet Editor*[5], der zur Normierung der Daten den Rückgriff auf den *Getty Thesaurus of Geographical Names (TGN)* unterstützt. Erwähnung fanden auch weitere Programme und Arbeitstechniken: Das *Cosmo-Tool*[6], ein Werkzeug zur Visualisierung biografischer Informationen, die *Voyant Tools*[7], ein Toolset zur Erhebung und Visualisierung quantitativer Merkmale von Textdaten, und ein Workflow zur Ermittlung von semantischen Strukturen auf Basis der distributionellen Semantik namens *Topics*[8]. Außerdem wurden mit *Etherpad*, den *Chili-Projects* und dem *DARIAH-DE-WIKI* einige Hilfsmittel zum kollaborativen Arbeiten vorgestellt.

Nicht zuletzt wurde auf das umfangreiche Informationsangebot auf den Webseiten von *DARIAH-DE* hingewiesen, das unter anderem einschlägige Erläuterungen zu Rechtsfragen und Lizenzen[9], fachspezifische Empfehlungen für Metadaten[10], eine kollaborative Bibliographie zu den Digital Humanities[11] und die *DARIAH-DE Working Papers* umfasst, den zentralen Publikationsort für Beiträge aus dem DARIAH-DE-Kontext[12]. Zum Abschluss ihrer Präsentation resümierte Klaffki die unterschiedlichen Maßnahmen zur Öffentlichkeitsarbeit, u.a. die Kommunikationstools *DHd-Blog*[13] und *DHd-Kanal*.

---

[3] https://fontane-nb.dariah.eu/index.html (05.12.17).
[4] https://geobrowser.de.dariah.eu/ (05.12.17).
[5] https://geobrowser.de.dariah.eu/edit/ (05.12.17).
[6] https://cosmotool.de.dariah.eu/cosmotool/personsearch/ (05.12.17).
[7] https://voyant-tools.org/ (05.12.17).
[8] https://de.dariah.eu/topics (05.12.17).
[9] https://de.dariah.eu/weiterfuhrende-informationen; http://forschungslizenzen.de/ (19.02.18).
[10] https://wiki.de.dariah.eu/pages/viewpage.action?pageId=20058160; https://wiki.de.dariah.eu/pages/viewpage.action?pageId=38080370 (19.02.18).
[11] https://de.dariah.eu/bibliographie (19.02.18).
[12] https://de.dariah.eu/working-papers (19.02.18).
[13] http://dhd-blog.org/ (19.02.18).

2.4. Mirjam Blümm/Lisa Klaffki: DOI-Service für die Geisteswissenschaften und *DARIAH-DE*-Repository

Die zweite Präsentation von Mirjam Blümm und Lisa Klaffki kreiste um die nachhaltige Sicherung und persistente Adressierbarkeit von geisteswissenschaftlichen Forschungsdaten. Die Referentinnen erläuterten zu Beginn den Aufbau, die Vorteile und die Funktionsweise von DOIs (‚Digital Object Identifier'). Danach führte Blümm die verschiedenen Möglichkeiten für Geisteswissenschaftlerinnen und Geisteswissenschaftler in Deutschland auf, DOIs vergeben bzw. beziehen zu können: über eine kostenpflichtige direkte Mitgliedschaft ihrer jeweiligen Einrichtung bei *CrossRef* oder *Datacite*, durch Sicherung ihrer Forschungsdaten im *DARIAH-DE Repository* oder – bei bereits bestehender anderweitiger Datenspeicherung – über einen einschlägigen Service der Niedersächsische Staats- und Universitätsbibliothek (SUB) Göttingen[14].

Anschließend führte Lisa Klaffki das seit kurzem verfügbare *DARIAH-DE Repository* vor und erläuterte die Unterschiede zum *Textgrid Repository*. Letzteres wurde konzipiert mit Blick auf die Speicherung von Daten, die im Rahmen des *Textgrid Laboratory* erstellt worden sind, und ist vor allem auf Textdaten im XML-, TXT- oder HTML-Format ausgelegt. Das *DARIAH-DE Repository* erlaubt demgegenüber die Sicherung aller Arten und Formate geisteswissenschaftlicher Forschungsdaten. Voraussetzungen dafür sind Maschinenlesbarkeit, Interpretierbarkeit, Prozessierbarkeit, Referenzierbarkeit sowie die Bereitschaft zur nachhaltigen Speicherung bzw. Langzeitarchivierung und zur Veröffentlichung im Open Access. Die Referentin schilderte, wie Forschungsdaten mittels des *DARIAH-DE Publikators*[15] in das *DARIAH-DE Repository* hochgeladen und mit Metadaten, einer geeigneten Nutzungslizenz sowie einer DOI versehen werden können.

Schließlich wies Lisa Klaffki auf die generische Suche[16] von *DARIAH-DE* hin, über die nicht nur die im *DARIAH-DE Repository* abgelegten Datensammlungen recherchierbar sind, sondern auch auf anderen Servern liegende Kollektionen, sofern sie im Collection Registry[17] von *DARIAH* verzeichnet sind.

2.5. Lisa Klaffki: Aktivitäten der DHd *AG Datenzentren*

Mit Blick auf die inzwischen zahlreichen Initiativen zum Forschungsdatenmanagement in den Geisteswissenschaften präsentierte Lisa Klaffki die Aktivitäten der *AG Datenzentren* des Verbands *Digital Humanities im deutschsprachigen Raum* (DHd). Bei der AG handelt es sich um einen offenen Zusammenschluss von Zentren, die Infrastrukturen für Forschungsdaten in den Geisteswissenschaften anbieten oder perspektivisch anbieten werden. Sie versteht sich als Kommunikationsplattform zum Austausch über Begriffe, rechtliche Fragen, spezifische Problemfelder und Aufgaben des geisteswissenschaftlichen Forschungsdatenmanagements und verfolgt das Ziel, die diversen Erfahrungen zu bündeln, die Standardisierung

---

[14] Kontakt: info@eresearch.uni-goettingen.de, Kontaktperson ist Jan Brase.
[15] Informationen unter: https://de.dariah.eu/publikator; Tool: https://repository.de.dariah.eu/publikator/ (19.02.18).
[16] Informationen unter: https://de.dariah.eu/generische-suche; Tool: https://search.de.dariah.eu/search/ (19.02.18).
[17] Informationen unter: https://de.dariah.eu/collection-registry; Collection Registry: https://colreg.de.dariah.eu/colreg-ui/ (19.02.18).

voranzutreiben und die Sichtbarkeit zu erhöhen.[18] In diesem Sinne wurde u.a. mit der Erfassung der von den einzelnen Zentren angebotenen Dienste begonnen. Angestrebt sind auch Services zur Evaluation bzw. Zertifizierung von Infrastrukturen für geisteswissenschaftliche Forschungsdaten.

Für die kommende Jahrestagung des DHd-Verbandes 2018 in Köln wird die AG mit einem eigenen Panel präsent sein.[19]

Die AG unterstützt insbesondere die vom *Rat für Informationsinfrastrukturen* (RfII) angeregte Schaffung einer *Nationalen Forschungsdateninfrastruktur* (NFDI) in Form eines föderalen Netzwerkes.[20]

2.6. Diskussion

Die Diskussion begann mit der Frage, wie die bestehenden virtuellen Infrastrukturen passgenauer auf die Anforderungen der Romanistik ausgerichtet werden können, damit Romanistinnen und Romanisten den spezifischen Nutzen besser einschätzen und so leichter das geeignete Instrument für ihre Forschungsarbeit identifizieren können. Dazu wurde angemerkt, dass es aufgrund jeweils unterschiedlicher wissenschaftlicher Praktiken, Kenntnisstände und Nutzungsgewohnheiten generell nicht immer einfach sei, den richtigen Grad an Spezialisierung des Angebots zu finden und unerfahrene Nutzerinnen und Nutzer sowie Expertinnen und Experten gleichermaßen anzusprechen. Für fächerübergreifende Angebote wie *DARIAH-DE* und *CLARIN-D* sei es kaum möglich, die Dienste jeweils für alle fachspezifischen Bedarfe und Anforderungen transparent darzustellen. Daher brauche es zum einen eine klare Unterscheidung zwischen generischen und speziellen Komponenten und zum anderen eine gezielte Kommunikation mit Blick auf die jeweiligen Zielgruppen über die fachlich geeigneten Kanäle. Hilfreich könnten auch gut vernetzte und erreichbare Ansprechpartnerinnen und Ansprechpartner sein.

Um das Auffinden von geeigneten Infrastrukturen und Instrumenten zu unterstützen, wurde darüber hinaus vorgeschlagen, mithilfe von wissenschaftspraktischen Anwendungsszenarien (z.B. Erstellung digitaler Editionen, Durchführung computerlinguistischer Analysen, Annotation von Bilddaten etc.) konkrete Einstiegspfade und fachspezifische Leitwege für Nutzerinnen und Nutzer zu schaffen. Man müsse transparent machen, welche Funktionen die jeweiligen Tools mitbringen und wozu man sie nutzen könne. Dabei seien sowohl disziplinäre als auch methodische Einstiege denkbar. Wichtig sei aber in beiden Fällen die Verwendung von passenden, standardisierten Formaten, die über die Romanistik hinaus Anwendung finden. Ergänzend wurde angeregt, die jeweiligen Instrumente in den Nutzungsszenarien durch Screenshots zu veranschaulichen. Auch eine umfangreiche Dokumentation existierender Vorhaben, wie sie z.B. mit den Rezensionen von digitalen Editionen und Ressourcen in der elektronischen und kostenfrei zugänglichen Zeitschrift

---

[18] Vgl. DHd-AG Datenzentren*: Geisteswissenschaftliche Datenzentren im deutschsprachigen Raum. Grundsatzpapier zur Sicherung der langfristigen Verfügbarkeit von Forschungsdaten*, Hamburg 2017, DOI: 10.5281/zenodo.1134760 (19.02.18).

[19] Titel: "Die Summe geisteswissenschaftlicher Methoden? Fachspezifisches Datenmanagement als Voraussetzung zukunftsorientierten Forschens".

[20] Vgl. https://dig-hum.de/stellungnahme-dhd-nfdi (05.12.17)

RIDE erfolge[21], könne Forschenden helfen, die Eignung spezifischer Instrumente mit Blick auf ähnliche Vorhaben einzuschätzen.

Des Weiteren wurden Ansätze zur Langzeitarchivierung von romanistischen Forschungsdaten diskutiert. Im Plenum wurde empfohlen, dafür kein eigenes Repositorium anzulegen, sondern bereits bestehende Infrastrukturen nachzunutzen. In diesem Zusammenhang sei auch zu klären, inwieweit das *DARIAH-DE-Repository* als technische Grundlage für ein Forschungsdatenrepositorium der Romanistik dienen könne.

Schließlich wurde mit Blick auf den oft dynamischen Charakter von Forschungsdaten über Mittel und Wege zur Gewährleistung einer langfristigen Zitierbarkeit mittels persistenter Identifikatoren diskutiert. Es wurde erläutert, dass eine DOI sich immer auf eine spezifische Version eines Datensatzes bezieht und dynamische Inhalte dadurch dauerhaft zitierbar gemacht werden, dass für verschiedene Versionen jeweils eigene DOIs vergeben werden. Eine solche Versionierung werde z.B. im Rahmen der digitalen Speicherinfrastruktur *Zenodo* praktiziert. Mit Blick auf dynamische Inhalte, bei denen frühere Versionen obsolet werden (z.B. bei kontinuierlich erweiterten Datensammlungen und Korpora), wurde auf einen anderen persistenten Identifikator namens *Handle* hingewiesen, der im Unterschied zur DOI bei Bedarf auch gelöscht werden kann.

3. Panel II: Verschiedene Lösungen für ein nachhaltiges Forschungsdatenmanagement in der Romanistik

Moderation: Christof Schöch

Im nächsten Panel wurden zwei romanistische Projekte vorgestellt, die spezifische Lösungswege für ein nachhaltiges Forschungsdatenmanagement jenseits der zuvor vorgestellten Infrastrukturen repräsentieren. Dabei ging es unter anderem darum, Vor- und Nachteile aufzuzeigen, die sich aus dem Aufbau eigener Plattformen bzw. aus dem Rückgriff auf kommerzielle Angebote ergeben können.

3.1. Hanno Ehrlicher/Nanette Rißler-Pipka: Die virtuelle Forschungsumgebung *Revistas Culturales*

Hanno Ehrlicher und Nanette Rißler-Pipka präsentierten die Virtuelle Forschungsumgebung *Revistas Culturales*.[22] Den Anfang machte Hanno Ehrlicher, der zunächst den Ausgangspunkt des Projekts erläuterte. Da für die zu untersuchenden lateinamerikanischen Kulturzeitschriften der Moderne weder die relevanten Bestände noch die dazugehörigen Digitalisate an einem Ort zentral einsehbar waren und darüber hinaus keine bedarfsgerechte Infrastruktur existierte, wurde eine eigene Plattform aufgebaut, die für ein Netzwerk von Forschenden an verschiedenen Standorten als virtuelle Anlaufstelle zur Untersuchung, Präsentation und kollaborativen Bildannotation lateinamerikanischer Kulturzeitschriften dienen konnte. Technisch basiert das Portal auf dem Content-Management-System *Drupal*, wobei standardisierte Metadaten zu den präsentierten Objekten aus den Beständen des *Ibero-Amerikanischen Instituts* (IAI) in die virtuelle Forschungsumgebung importiert werden

[21] Vgl. http://ride.i-d-e.de/ (05.12.17).
[22] https://www.revistas-culturales.de/ (06.12.17).

können. Die Webdarstellung des Portals ist zweisprachig (Deutsch und Spanisch), teils dreisprachig (Deutsch, Spanisch, Englisch). Inhaltlich umfasst die Forschungsumgebung aktuell 80 Kulturzeitschriften mit über 14.000 Exemplaren, eine umfassende Forschungsbibliographie, eine Linkliste zu Quellen und Informationsseiten über spanischsprachige Kulturzeitschriften der Moderne, eine Darstellung der Publikationen der Mitglieder des Forschungsnetzwerks und Blogs mit aktuellen Informationen.

Im zweiten Teil der Präsentation stellte Nanette Rißler-Pipka den im Rahmen des *eCodicology*-Projekts[23] entwickelten *Software workflow for the automatic tagging of medieval manuscript images (SWATI)* vor. Dieser bildet die Grundlage für die kollaborative Annotation bzw. Analyse der Digitalisate. Zur weiteren Anreicherung der Metadaten wurde zudem ein Formular erstellt, das auch externen Nutzerinnen und Nutzern Bearbeitungsmöglichkeiten einräumt (‚crowd-tagging').

Revistas Culturales wird stetig weiterentwickelt. Geplant sind ein inhaltlicher Ausbau durch Kooperationen mit weiteren Archiven und Bibliotheken, der Aufbau einer internationalen und interdisziplinären Verbundforschung, die Integration von Volltexten und von Möglichkeiten zur Fehlerkorrektur, die Entwicklung fachspezifischer Tools zur Untersuchung der digitalen Zeitschriften sowie die Implementierung bereits extern existierender Tools. In diesem Sinne wurde der Geo-Browser von *DARIAH-DE* bereits integriert. Zudem ist die Einrichtung von Import- und Exportmöglichkeiten über XML bzw. XMS hinaus vorgesehen (u.a. *METS/MODS*, CSV). Zum Abschluss des Vortrags formulierten Ehrlicher und Rißla-Pipka Desiderate, zu deren Umsetzung der *FID Romanistik* einen Beitrag leisten könne. So sei es wichtig, projektbezogene Forschungsinfrastrukturen in der Romanistik zu bündeln und ihre Sichtbarkeit, beispielsweise durch eine fachgerechte Katalogisierung, auf breiter Basis zu verbessern. Ferner wurde für eine systematische Rückkopplung zwischen dem FID und romanistischen Forschungsprojekten plädiert, etwa durch regelmäßig stattfindende Workshops. Außerdem wurde Unterstützung bei der langfristigen Archivierung von Forschungsdaten mit Blick auf eine internationale Zielgruppe gewünscht.

3.2. Christoph Bürgel/Sascha Diwersy: *Sketch Engine* als Arbeits- und Suchsystem für linguistische Korpora am Beispiel des *Corpus de référence du français contemporain* (CRFC)

Christoph Bürgel und Sascha Diwersy präsentierten das *Corpus de référence du français contemporain* (CRFC)**,** ein etwa 310 Mio. Wörter umfassendes Monitorkorpus der französischen Sprache, das derzeit über ein kostenpflichtiges System zur Verwaltung und Analyse namens *Sketch Engine* erreichbar ist.

Im ersten Teil des Vortrags erläuterte Christoph Bürgel den Aufbau des Korpus, das sich in methodischer Hinsicht an angloamerikanischen Vorbildern wie dem *British National Corpus* (BNC) und dem *Corpus of Contemporary American English* (COCA) orientiert, sich allerdings durch eine höhere Anzahl an berücksichtigten Textsorten sowie eine größere Repräsentativität und Ausgewogenheit auszeichnet. Im Unterschied zu den bisherigen

[23] Hierbei handelt es sich um ein vom Bundesministerium für Bildung und Forschung (BMBF) gefördertes Projekt zur computergestützten Annotation und Auswertung des mittelalterlichen Handschriftenbestands der Benediktinerabtei St. Matthias in Trier.

einschlägigen Korpora des Französischen enthält das CRFC nicht nur schriftliche, sondern auch sprechsprachliche bzw. pseudosprechsprachliche Sprachdaten. Eine große Stärke des Korpus beruht auf seinem über 3.000 Stunden umfassenden sprechsprachlichen Teil, der unter anderem aus Fernsehsendungen, Interviews, Parlamentsdebatten und anderen Korpora gewonnen wurde. Der Referent wies darauf hin, dass sich die Bearbeitung dieser großen Menge an linguistischen Daten zunächst auf das Machbare konzentrierte, weswegen nicht alles wissenschaftlich Wünschenswerte bisher umgesetzt werden konnte. So wurde z.B. vorläufig auf die aufwändige Annotation von Teilen des sprechsprachlichen Korpus im Hinblick auf Datum, Turn, semantische und prosodische Merkmale verzichtet.

Mit dem Aufbau des CRFC ist das Ziel verknüpft, eine Grundlage zu schaffen für die Entwicklung von korpusinduzierten Wörterbüchern, Grammatiken, Lehr- und Lernmaterialen sowie für die Bearbeitung linguistischer, fremdsprachendidaktischer und übersetzungswissenschaftlicher Forschungsfragen. In diesem Kontext verwies Bürgel auf einige bereits auf dem Korpus basierende Studien, unter anderem zum Gebrauch des *Subjonctif* in der gesprochenen Sprache und zur Lexiko-Grammatik der Präpositionen, sowie auf sein eigenes aktuelles Forschungsprojekt zur Erstellung einer wissenschaftlichen Grammatik des Französischen.

Im zweiten Teil des Vortrags ging Sascha Diwersy auf *Sketch Engine*[24] ein, eine kommerzielle Plattform mit vielfältigen Funktionen zur Analyse und zum Management von linguistischen Korpora. Er präsentierte deren Nutzeroberfläche und stellte die unterschiedlichen Arten von Korpora vor, die über *Sketch Engine* abgerufen und ausgewertet werden können: Zu den 400 momentan verfügbaren Korpora in mehr als 90 Sprachen zählen Referenzkorpora, Webkorpora, Parallelkorpora, Spracherwerbskorpora, Lernerkorpora und diachrone Korpora.

Nach Ansicht von Diwersy bietet ein System wie *Sketch Engine* eine Reihe von Vorteilen. Dazu zählen neben den auf Sprachkorpora zugeschnittenen Analysefunktionen die angebotenen Hilfen mit User-Guides, die intuitive Bedienung, die Möglichkeiten zum Sortieren, Visualisieren und Herunterladen von Daten und das unkomplizierte Verfahren zum Hochladen eigener Korpora. Dem stünden jedoch auch einige Nachteile entgegen: die Kompliziertheit komplexer Suchabfragen aufgrund der je nach Korpus unterschiedlichen CQL-Syntax, die fehlenden Schnittstellen zur Übertragung von numerischen Daten in Statistikprogramme zwecks quantitativ-statistischer Weiterverarbeitung, die wegen der Bindung an den jeweiligen Personenaccount komplizierte Verfügbarmachung von Korpora, die schwierige bzw. mit Kosten verbundene Sichtbarmachung von Korpora für Dritte und die Kosten im Allgemeinen (ca. 3.000 € jährlich für ein Korpus von 1 Mia. Wörtern). Generell sei die Nutzung kommerzieller Systeme mit dem Problem verbunden, dass Wissenschaftlerinnen und Wissenschaftler häufig die Kontrolle über die eigenen Daten einbüßen, da die Nutzungsrechte oft bei den jeweiligen Anbietern liegen. Daraus ergeben sich bisweilen Einschränkungen, etwa für das Text- und Datamining. Da zudem die Nachnutzbarkeit der Daten im Sinne einer Anpassung und Weiterentwicklung i.d.R. nicht vorgesehen ist, wird oft auch weniger auf die Verwendung anerkannter Standards und Formate geachtet.

[24] https://www.sketchengine.co.uk/ (06.12.17).

Als Konsequenz aus den Nachteilen präsentierte Diwersy abschließend einen Ausblick auf seine im Aufbau befindliche Plattform *PrimeStat*, über die das *CRFC* Ende 2018 kostenlos verfügbar gemacht werden soll. Dieses System verbindet die englischen und französischen Traditionen der Korpusanalyse. Es soll neben Suchfunktionen vielfältige Werkzeuge zur Analyse und Modellierung der Daten beinhalten, unter anderem für die Ermittlung von Konkordanzen, Frequenzen, Kookkurrenzen sowie zur Anzeige und zum Export quantitativer lexikographischer Daten. Außerdem soll die Plattform Möglichkeiten zur graphischen Visualisierung bieten, beispielsweise durch sogenannte ‚scatterplots' (Streudiagramme).

3.3. Diskussion

In der anschließenden Diskussion wurde v.a. die Problematik der Nutzungsrechte beim Rückgriff auf ein kommerzielles System thematisiert. Die Rechte an den Daten würden dabei in der Regel an die jeweiligen Anbieter abgetreten, die dadurch entscheidenden Einfluss auf die Modalitäten der Nachnutzung nehmen könnten. Um trotz der Zugriffsbeschränkungen bei kostenpflichtigen Angeboten die Reproduzierbarkeit der jeweiligen Forschungsdaten und -ergebnisse in gewissem Maße zu ermöglichen, wurde vorgeschlagen, frei verfügbare Teile auszugliedern und öffentlich verfügbar zu machen. Mit Blick auf die derzeitige Diskussion über Rechtsfragen des Text- und Dataminings wurde die Notwendigkeit betont, Strategien dafür zu entwickeln, dass Forschungsdaten auf juristisch gesicherter Basis wissenschaftlich nachnutzbar gemacht werden können. In diesem Kontext wurde auch auf die Webseiten des *FID Romanistik* zum Open Access hingewiesen, auf denen unter anderem Informationen zu Rechtsfragen zu finden sind.[25]

Die Frage, wozu man überhaupt auf individuelle Lösungen oder kommerzielle Systeme zurückgreift und nicht von vornherein bestehende freie Infrastrukturen nutzt, wurde an dieser Stelle nicht eigens thematisiert. Im Laufe des Workshops klangen aber immer wieder mögliche Gründe dafür an: Teilweise sind die Funktionalitäten der bestehenden Infrastrukturen den Forschenden nicht so transparent, dass sie als potentielle Lösung wahrgenommen werden; teilweise waren Funktionen zum Start bestimmter Projekte noch im Aufbau; teilweise werden für manche Vorhaben sehr spezifische Funktionalitäten benötigt, welche generische und frei zugängliche Infrastrukturen kaum bereitstellen können. Für diese Fälle wurde die Option genannt, eine nachträgliche Anbindung an eine generische Infrastruktur vorzunehmen und etwa die generierten Forschungsdaten zumindest teilweise dort abzulegen, um eine Nachnutzung zu ermöglichen.

4. Panel III: Spezielle Instrumente zur Verbesserung der Nachhaltigkeit, Nachnutzbarkeit, Auffindbarkeit von Forschungsdaten

Moderation: Doris Grüter und Jan Rohden

Im dritten Panel des Workshops lag der Schwerpunkt auf der Verbesserung der Sichtbarkeit und der Auffindbarkeit von romanistischen Forschungsdaten. Diese sind vielfältig und heterogen, in verschieden Systemen verzeichnet und gesichert bzw. in vielen Fällen überhaupt nur unzureichend nachgewiesen, was eine Reihe von Problemen für die

[25] Vgl. https://www.fid-romanistik.de/open-access/rechtliche-informationen/ (07.12.17).

Transparenz, die Nachhaltigkeit, die Nachnutzbarkeit und nicht zuletzt die systematische Suche mit sich bringt. Mittlerweile gibt es eine Reihe von Initiativen zur Verbesserung dieser Situation. Einige Beispiele wurden im Rahmen der Vorträge zu *META-SHARE*, dem *FID Linguistik* und dem *FID Romanistik* vorgestellt.

4.1. Georg Rehm: *META-SHARE* als System zur Sicherung und Suche romanistischer Sprachressourcen

Georg Rehm präsentierte ein von der EU gefördertes System zur Sicherung und Recherche von Sprachressourcen namens *META-SHARE*. Es wird von dem Exzellenznetzwerk *META-NET* aufgebaut, das sich u.a. zur Aufgabe gemacht, für die vielsprachige europäische Informationsgesellschaft technologische Lösungen zur Überwindung von Sprachgrenzen zu entwickeln. *META-SHARE* fungiert als dezentrale, vernetzte, offene und interoperable Infrastruktur, die sich auf Bedarfe und Probleme hinsichtlich der Sichtbarkeit, Dokumentation, Identifizierung, Verfügbarkeit, Langzeitspeicherung und Interoperabilität von Sprachdaten und Tools konzentriert. Dabei liegt ein besonderer Fokus auf juristischen Aspekten und der Entwicklung und Bereitstellung von Workflows. Favorisiert werden offene Daten und Open Source, eine Verwendung des Systems für kostenpflichtige Ressourcen ist gleichwohl nicht ausgeschlossen. *META-SHARE* zählt derzeit 35 Mitgliedsorganisationen in 25 Ländern. Wie am Beispiel der nationalen CLARIN-Infrastruktur in Griechenland erläutert wurde, kann die zur Einrichtung von *META-SHARE*-Repositorien eingesetzte und frei zugängliche Software auch für andere virtuelle Infrastrukturen genutzt werden.

META-SHARE enthält derzeit über 3.000 Sprachressourcen auf 28 Repositorien. Dabei sind v.a. die verbreiteten Sprachen mit zahlreichen Ressourcen vertreten (aktuell z.B. Spanisch mit 611, Französisch mit 600, Portugiesisch mit 469, Italienisch mit 457, Katalanisch mit 69, Rumänisch mit 68, Galicisch mit 27, Okzitanisch mit 2 Ressourcen). Da für kleinere Sprachen oft nur eine sehr geringe Anzahl an digitalen Sprachressourcen überhaupt existiert, sieht es *META-SHARE* auch als seine Aufgabe an, einen Beitrag zu deren Sicherung zu leisten.[26]

Das zur Beschreibung der Sprachressourcen und verwandter Objekte eingesetzte Metadatenschema besteht aus obligatorischen und optionalen Elementen. Dabei können Ressourcentypen im Rahmen von zwei Klassifikationsachsen (resourceType“[27] und „mediaType“[28]) differenziert beschrieben werden. Eine Besonderheit stellt die Unterstützung unterschiedlicher Lizenzsysteme dar, darunter *Creative Commons*, *META-SHARE Commons*, *META-SHARE “No Redistribution”*, Standard-Open-Source- und kommerzielle Lizenzen.

META-SHARE unterstützt durch seine Architektur einen barrierefreien Austausch und eine automatische Einspielung von Metadaten. Die Open Source Repository Software ermöglicht

[26] Die aktuelle Situation von 31 europäischen Sprachen ist im Detail in der frei zugänglichen *META-NET White Paper Series* zusammengefasst, die nach wie auf vor große Resonanz stößt. Siehe: http://www.meta-net.eu/whitepapers/overview (07.12.17).
[27] Pro Ressource kann nur eine der momentan zur Auswahl stehenden vier Kategorien („Corpus“, „Lexical Conceptual Resource“, „Tool Service“ und „Language Description“) angegeben werden.
[28] Folgende Medientypen können derzeit gewählt werden: „Text“, „Audio“, „Video“, „Textngram“, „Image“, „Text numerical“ und „Textnumerical“. Mehrfachnennungen sind möglich.

eine unkomplizierte Einrichtung und Einbindung neuer Repositorien. Darüber hinaus ist es möglich, auch Daten aus Quellen zu verzeichnen, die außerhalb des *META-SHARE*-Netzwerks liegen.

Für Nutzerinnen und Nutzer bietet *META-SHARE* u.a. verschiedene Funktionen zur Suche nach Sprachressourcen, die Möglichkeit zur Auswahl und Nutzung von Lizenzen sowie einen Helpdesk. Mit Blick auf die Usability ist die Oberfläche von *META-SHARE* so angelegt, dass sie den Zugriff auf die gewünschte Ressource nach maximal fünf Klicks ermöglicht (Suche, Auswahl der Ressource, Beschreibung, Lizenzbedingungen, Download der Ressource).

4.2. Heike Renner-Westermann: Aktivitäten und Perspektiven des Linguistik-Portals mit Blick auf Forschungsdaten

Heike Renner-Westermann stellte in ihrem Vortrag die Aktivitäten und Perspektiven des Linguistik-Portals (www.linguistik.de) bzw. des *Fachinformationsdienstes Linguistik* vor. Zunächst präsentierte sie das in den Webauftritt eingebettete Linkverzeichnis, in dem neben anderen Internetressourcen auch Forschungsdaten und linguistische Korpora nachgewiesen werden. Danach schilderte sie das neue Projekt, in dessen Rahmen als Linked-Open-Data (LOD) vorliegende Sprachressourcen (Korpora, Wörterbücher) in das Suchmodul des Linguistik-Portals eingebunden werden. Die Basis dafür bildet der Thesaurus der *Bibliography of Linguistic Literature* (BLL), der für LOD aufbereitet wurde und mit dem Referenzmodell der *Ontologies of Linguistic Annotations* (OLiA), dem gewählten Einstiegspunkt in die *Linguistic Linked Open Data* (LLOD) *cloud*, verlinkt wurde. Dadurch können Metadaten von Sprachressourcen aus der LLOD-Cloud über einen eigens dafür entwickelten *Web Crawler* automatisiert in den Suchindex eingespeist und somit recherchierbar gemacht werden.

Dieser Ansatz wird in einem nächsten Schritt auf Sprachbezeichner ausgeweitet. Als Anknüpfungspunkt für die ca. 2300 im BLL-Thesaurus vorhandenen Sprachbezeichner wurden dabei in der LLOD-Cloud die Metadaten-Repositorien *Glottolog* und *Lexvo* gewählt. Dadurch können alle Ressourcen sichtbar und recherchierbar gemacht werden, die deren Sprachbezeichner verwenden. Die vorhandene Vernetzung von *Glottolog* und *Lexvo* (z.B. mit *GeoNames*, *DBpedia*, *YAGO*) ermöglicht zusätzliche Suchfunktionen.

Eine andere Maßnahme des *Fachinformationsdienstes Linguistik* besteht darin, bibliographische Nachweise von Publikationen mit den oft schwer auffindbaren Forschungsprimärdaten zu verknüpfen, auf die sie sich beziehen. Begonnen wurde hier mit den über 3.500 Publikationen, die in der BLL mit Schlagwörtern zu spezifischen Korpora versehen sind. Mit dem Ziel, die entsprechenden Nachweise mit Metadaten zu den behandelten Korpora zu verlinken, wurden bislang 1.018 dieser Korpora auf ihre Referenzierbarkeit mit folgendem Ergebnis geprüft: 19 besitzen eine *International Standard Language Resource Number (ISLRN)* und 595 eine gültige URL (183 einen persistenten Identifikator, 412 einen eigenen Internetauftritt).

Des Weiteren berichtete Renner-Westermann von den Bemühungen des *FID Linguistik*, kostenpflichtige Sprachressourcen zu lizenzieren. Dazu zählen insbesondere Sprachkorpora des *Linguistic Data Consortium* (LDC) und der *European Language Resources Association* (ELRA). Die Verhandlungen, die für die Fachinformationsdienste allgemein vom

Kompetenzzentrum für Lizenzierung durchgeführt werden, erweisen sich bisher allerdings als schwierig.

Abschließend erläuterte die Referentin einen vom *FID Linguistik* entworfenen Workflow, der dazu dienen soll, auch nicht in der *Linguistic LOD Cloud* enthaltene frei verfügbare Sprachressourcen besser auffindbar zu machen. Zu diesem Zweck wird ein Verfahren entwickelt, um die Metadaten entsprechender Sprachressourcen automatisiert mit inhaltlichen Informationen (Schlagwörtern) anzureichern.

4.3. Jan Rohden: Perspektiven zur Verbesserung der Suchbarkeit romanistischer Forschungsdaten im Rahmen des *FID Romanistik*

Jan Rohden präsentierte die Überlegungen des *FID Romanistik* zur Verbesserung der Recherchierbarkeit von romanistischen Forschungsdaten. Einleitend verwies er auf die Ergebnisse des ersten Workshops, bei dem die problematische Nachweissituation thematisiert wurde.

Vor dem Hintergrund dieser Ausgangslage wurde zunächst gemeinsam mit der *AG Digitale Romanistik* und *romanistik.de* ein Meldesystem eingerichtet, das die Möglichkeit eröffnet, relevante Forschungsdaten für die Fachcommunity sichtbar zu machen. Des Weiteren wurden diverse existierende Suchinstrumente für Forschungsdaten untersucht und mit Blick auf ihre Relevanz und Eignung für die Romanistik geprüft. Die derzeitigen einschlägigen Nachweissysteme werden auf den Webseiten des *FID-Romanistik* vorgestellt. Nicht zuletzt wurden die zahlreichen Forschungsdaten, die zur Entwicklung des Meldesystems gesichtet wurden, katalogisiert, um sie unmittelbar im Rahmen des FID recherchierbar zu machen. Als pragmatischste Lösung dafür wurde eine kooperativ nutzbare Datenbank zur Verzeichnung von Internetquellen namens *Academic LinkShare* (ALS) gewählt, die zum einen über ein Metadatenschema verfügt, mit dem die wichtigsten Informationen zu verschiedenen Typen von Forschungsdaten in standardisierter Weise erfasst werden können,[29] zum anderen einfache Präsentationsmöglichkeiten im Web bietet und perspektivisch in das FID-Suchportal eingebunden werden kann. Den aktuellen Stand veranschaulichte Rohden am Beispiel eines Datensatzes aus der Sammlung der *FID-Internetressourcen*[30] sowohl aus Perspektive der Nutzenden als auch aus jener der Katalogisierenden.

Im Anschluss nannte der Referent, ausgehend von den auf dem ersten Workshop[31] und im Positionspapier der *AG Digitale Romanistik*[32] formulierten Bedarfen, einige Anforderungen, die Suchinstrumente für romanistische Forschungsdaten erfüllen sollten: eine enge Anbindung an die romanistische Fachcommunity, Offenheit der Metadaten mit Blick auf Mehrsprachigkeit und Interdisziplinarität, Nachhaltigkeit (durch den Rückgriff auf standardisierte Metadaten, persistente Identifikatoren und die Vernetzung mit Systemen zur nachhaltigen Sicherung von Forschungsdaten), Vermeidung redundanter Doppelstrukturen durch Abstimmung bzw. Kooperation mit bestehenden Infrastrukturen. Ausgehend von den

[29] Beispielsweise durch kontrollierte Vokabulare (etwa die Gemeinsame Normdatei), Klassifikationen (z.B. mittels Dewey decimal classification), etc.
[30] https://www.fid-romanistik.de/forschungsdaten/suche-nach-forschungsdaten/fid-internetressourcen/ (08.12.17).
[31] https://www.ulb.uni-bonn.de/de/fid-blog/downloads/workshop-bericht (08.12.17).
[32] http://www.deutscher-romanistenverband.de/der-drv/agdr/positionspapier/ (08.12.17).

vier Anforderungen skizzierte Rohden ein mögliches Modell zur Verbesserung der Suche nach romanistischen Forschungsdaten. Denkbar wäre es, die über *romanistik.de* eingehenden Meldungen über Schnittstellen oder manuell in ein geeignetes Katalogisierungssystem wie etwa ALS zu übertragen, dort die Metadaten mit Normdaten anzureichern und sie dann in einem für die Romanistik einschlägigen Suchportal zur Verfügung zu stellen. Darüber hinaus sollte die Möglichkeit geschaffen werden, die Metadaten von romanistischen Forschungsdaten aus nicht fachspezifischen Nachweissystemen und Infrastrukturen über Schnittstellen in dasselbe Suchportal einzuspeisen.

Abschließend wies Rohden auf die Vielfalt der Maßnahmen hin, die zur Umsetzung eines solchen Modells notwendig wären. Diese würden technische Arbeiten (u.a. zur Einrichtung von Schnittstellen), Vorkehrungen zur Sicherung der Datenqualität (etwa Arbeitsschritte zur Anreicherung bzw. Normierung der Metadaten), die Schaffung geeigneter Kommunikationswege für die erforderliche Vernetzung umfassen.

4.4. Diskussion

Die anschließende Diskussion kreiste zunächst um die Frage, wie die Metadaten zu Forschungsdaten in standardisierter Form in einschlägige Nachweissysteme einfließen können. Dabei wurde nicht nur an spezifische Suchinstrumente für die gezielte Recherche gedacht, sondern auch an allgemeine Kataloge, von denen man sich, nicht zuletzt durch deren Einbindung in den *Karlsruher Virtuellen Katalog* (KVK), eine weitreichende Sichtbarkeit verspricht. Allerdings wurden die Möglichkeiten, Forschungsdaten auf breiter Basis direkt in den traditionellen Katalogen zu verzeichnen, von den anwesenden bibliothekarischen Vertreterinnen derzeit als gering eingeschätzt. Ein praktikabler Ansatz könnte jedoch darin bestehen, die in einem spezifischen Nachweissystem erfassten Metadaten z.B. in den Suchindex der *Bielefeld Academic Search Engine* (BASE) einzuspielen und sie darüber auch im KVK recherchierbar zu machen.

Anknüpfend an die Vorträge zum *FID Linguistik* und zum *FID Romanistik*, wurde anschließend auch über verschiedene Vorgehensweisen zur Verbesserung der Nachweissituation von Forschungsdaten diskutiert und die Frage aufgeworfen, inwieweit eine händische Verzeichnung einerseits und die Entwicklung von automatisierten Systemen zur virtuellen Einspeisung der Daten (LOD, Web-Crawler) andererseits zielführend sind. Dabei wurde unterstrichen, dass verlinkte Datenstrukturen sehr hilfreich seien, um gerade angesichts der großen Menge der für die allgemeine Sprachwissenschaft relevanten, weltweit existierenden Korpora einen sinnvollen Sucheinstieg anbieten zu können. Allerdings seien automatisierte Verfahren mit Blick auf den Anspruch an die Datenqualität häufig nur schwer umzusetzen. Zum gegenwärtigen Zeitpunkt seien beide Ansätze hilfreich. Für die Romanistik, wo die Voraussetzungen für ein automatisiertes Verfahren über LOD angesichts sehr disparater Erschließungssysteme derzeit noch kaum gegeben sind, wurde von Seiten der romanistischen Fachwissenschaft noch einmal die Idee eines vernetzten Workflows vorgebracht, der u.a. vorsieht, dass die über *romanistik.de* gemeldeten Metadaten von Expertinnen und Experten geprüft, normiert, angereichert und in ein einschlägiges Suchsystem überführt werden.

Mehrfach wurde im Verlauf des Workshops die Bedeutung internationaler Sichtbarkeit für romanistische Forschungsdaten thematisiert. Am Beispiel der *Taxonomy of Digital Research Activities in the Humanities (*TaDiRAH)[33] wurde betont, wie aufwändig und komplex die Vergabe und Pflege von multilingualen Metadaten sind. Erwähnt wurde auch, dass es im Rahmen von *META-NET* und der EU einige vielversprechende Initiativen und Projekte gebe, die sich mit der computergestützten Übersetzung von Webressourcen beschäftigten. Schließlich wurde auch die europäische Ausrichtung von *DARIAH-DE* als Chance gesehen. Derzeit fungiere *DARIAH-EU* für die jeweiligen nationalen Mitglieder und Partner zwar v.a. als Knotenpunkt nebeneinanderstehender nationaler Angebote, es gebe aber aktuelle Bemühungen, die wechselseitige Sichtbarkeit zur verbessern.

5. Panel IV: Beratungsdienste und Informationsangebote

Moderation: Doris Grüter und Jan Rohden

Auf dem ersten Workshop zur Bedarfsermittlung war die Notwendigkeit von Informations- und Beratungsangeboten in vielfacher Hinsicht betont worden, und zwar hinsichtlich grundlegender Standards, wesentlicher Arbeitsmethoden, geeigneter Formate, Tools, Repositorien, Virtueller Forschungsumgebungen und Rechtsfragen. Idealerweise wurde ein mehrstufiges Angebot gewünscht: von grundlegenden, frei im Web zugänglichen Informationen, z.B. zu Infrastrukturen, Formatstandards und Tools, über Informationsveranstaltungen zu einschlägigen Einzelthemen, wie etwa computerlinguistischen Analysemethoden oder TEI-Annotation, bis hin zu individueller Antrags- und Projektberatung.

In diesem Sinne wurde auf der FID-Webseite begonnen, für die Romanistik ein virtuelles Informationsangebot aufzubauen. Workshops zu einschlägigen Einzelthemen werden bei Bedarf von der *AG Digitale Romanistik* durchgeführt. Antrags- und Projektberatung wird auf der Ebene einiger Hochschulen bereits praktiziert, es existiert aber noch kein flächendeckendes Angebot. Allerdings haben sich dabei einzelne Kompetenzzentren entwickelt, die nicht nur lokale Services anbieten, sondern darüber hinaus eine überregionale Bedeutung haben. Ein Beispiel dafür ist das *Data Center for the Humanities* in Köln.

5.1. Patrick Helling: Das Dienstleistungs- und Beratungsangebot des Kölner *Data Centers for the Humanities*

Im letzten Vortrag des Workshops stellte Patrick Helling das Dienstleistungs- und Beratungsangebot des Kölner *Data Centers for the Humanities* (DCH) vor. Nach einer kurzen Vorstellung des Zentrums und der Rahmenbedingungen an der Philosophischen Fakultät der Universität Köln präsentierte er das aus drei Bausteinen bestehende Angebot des DCH: Antragsberatung, Projektentwicklung sowie Betrieb und Betreuung von Softwarelösungen/Repositorien zur Erhaltung digitaler Ressourcen. Das Angebot des DCH richtet sich in erster Linie an die Wissenschaftlerinnen und Wissenschaftler der Philosophischen Fakultät an der Universität zu Köln, ist jedoch nicht auf diese beschränkt, sondern kann deren Projektpartner und Projekte mit starkem Bezug auf in Köln beheimatete

[33] Dabei handelt es sich um ein mehrsprachiges kontrolliertes Vokabular, um für die Digital Humanities relevante Forschungspraktiken und Ressourcen in standardisierter Form zur klassifizieren.

Bestände miteinschließen. Das DCH entwickelt und betreibt auch fachbezogene Lösungen, die überregional offenstehen und in nationale Forschungsinfrastrukturen eingebunden sind. Hier sind insbesondere die Angebote rund um die Archivierung von audiovisuellen (Sprach-) Daten zu nennen (CLARIN K-Centre CKLD, Language Archive Cologne, KA³).

Den ersten Dienstleistungsbereich veranschaulichte Helling am Beispiel des *Consulting Workflow for Humanities Research Data*. Dieser sieht zunächst Gespräche zwischen den anfragenden Wissenschaftlerinnen bzw. Wissenschaftlern und dem DCH vor, auf dessen Grundlage die geäußerten Bedarfe und Vorhaben analysiert würden. Davon ausgehend wird ein Kurzbericht mit konkreten Handlungsempfehlungen angefertigt. In einer Teamsitzung trifft das DCH dann die Entscheidung, inwieweit und in welcher Form die erwünschten Beratungsdienstleistungen durch das DCH selbst erbracht werden können oder ob auf einen geeigneten Kooperationspartner verwiesen wird.

Der zweite Servicebereich betrifft Entwicklungsprojekte, an denen sich das DCH als Kooperationspartner beteiligt, sofern es bislang keine einschlägigen Services bzw. Tools für das geplante Vorhaben gibt. Hilfreich sind dabei zum einen die enge Zusammenarbeit mit dem *Cologne Center for eHumanities* (CCeH) sowie dem *Regionalen Rechenzentrum der Universität zu Köln* (RRZK) und zum anderen die Clusterstruktur des DCH, die Kompetenzen aus drei Bereichen des geisteswissenschaftlichen Arbeitens zusammenführt: zu Objekten und Bildern, zu Texten und Dokumenten und zu audiovisuellen Daten.

Danach ging Helling auf den dritten Dienstleistungsbereich ein, den Betrieb bzw. die Betreuung von Software-Lösungen und Repositorien. In diesem Kontext verwies er auf das vom DCH betriebene zur *CLARIN*-Infrastruktur gehörende *Language Archive Cologne* (LAC), das zur Archivierung von linguistischen Daten dient.

Abschließend fasste Helling aktuelle Herausforderungen des Forschungsdatenmanagements aus der Perspektive des DCH zusammen. Dabei wies er auf die Unterschiede in der Zielsetzung von Wissenschaftlerinnen und Wissenschaftlern auf der einen Seite und Betreibern von Systemen zur Sicherung von Forschungsdaten auf der anderen Seite hin: Während Forscherinnen und Forscher in der Regel recht spezifische und individuelle Daten produzieren, für deren Dokumentation allerdings kaum Ressourcen aufwenden können, erwarten Archive in der Regel gut erschlossene und standardisierte Daten. Vor dem Hintergrund des steigenden Bedarfs an Dienstleistungen zum Forschungsdatenmanagement betrachtet Helling es als wichtige Aufgabe, Wissenschaft und Serviceangebote passgenau zusammenzubringen.

5.2. Diskussion

Im Anschluss an den Vortrag konnte angesichts der fortgeschrittenen Zeit lediglich auf einige Rückfragen eingegangen werden. Dabei wurde erläutert, in welcher Relation das Kölner DCH zu den Initiativen zur Schaffung nationaler bzw. internationaler Infrastrukturen für die Geisteswissenschaften steht. Das DCH ist Mitglied der *AG Datenzentren* ist und teilt deren Ziele zum Aufbau einer *Nationalen Forschungsdateninfrastruktur*, bietet jedoch gleichzeitig auf lokaler und überregionaler Ebene spezifische Dienstleistungen für geisteswissenschaftliche Forschungsprojekte an. Gerade angesichts der Tatsache, dass individuelle Projektberatung von überregionalen Infrastrukturen nicht abgedeckt wird und es

an vielen Hochschulen noch kein lokales Angebot gibt, stellen die diesbezüglichen Services von Kompetenzzentren wie dem DCH eine wichtige Ergänzung des bestehenden Informationsangebotes für die Geisteswissenschaften dar und können damit auch der Romanistik zu Gute kommen.

6. Schlusswort

Im Rahmen der zahlreichen Vorträge und Diskussionsbeiträge wurden viele Initiativen vorgestellt, die auf die drängenden Herausforderungen des Forschungsdatenmanagements in den Geisteswissenschaften und insbesondere in der Romanistik reagieren. Daran anknüpfend wurde eine Reihe von Ansätzen sichtbar, die es künftig weiter zu verfolgen gilt. Sie betreffen u.a. den bedarfsgerechten Ausbau der Infrastrukturen, die Verbesserung der Nachweissituation, die Vernetzung der Infrastrukturen untereinander, die Vernetzung zwischen den Forschenden und den Anbietern der Infrastrukturen, den Ausbau von fachspezifischen Informationsangeboten zur Vermittlung zwischen generischen Infrastrukturen und spezifischen wissenschaftlichen Vorhaben. Ziel wird es sein, die einzelnen Ansätze in einer koordinierten Strategie zusammenzuführen und auf die Bedarfe der Romanistik auszurichten, um die Attraktivität existierender Initiativen für die Forschenden zu erhöhen und Fachwissenschaftlerinnen und Fachwissenschaftlern unterschiedlicher Disziplinen und Kenntnisstände den Zugang zu den für ihre Arbeit notwendigen Diensten und Infrastrukturen zu ermöglichen.