Oliver Schonefeld / Andreas Witt

FORSCHUNGSINFRASTRUKTUREN AM IDS: GEGENWART UND ZUKUNFT

Zusammenfassung
Im Programmbereich Forschungsinfrastrukturen, der 2009 etabliert wurde, waren von Beginn an die Forschungsprojekte angesiedelt, die sich mit dem Aufbau und der Nutzung von Sprachressourcen beschäftigt haben. Untersucht wurden anfangs primär Fragen zur Informationsanreicherung durch Metadaten und Annotationen, zur Langzeitarchivierung und zur Interoperabilität, d.h. z.B. die Definition und Bereitstellung von Schnittstellen zur Bereitstellung der Sprachressourcen in virtuellen Forschungsumgebungen. Damit einher gingen auch wissenschaftliche Analysen der rechtlichen Situation beim Umgang mit Sprachdaten sowie Aktivitäten im Zusammenhang mit der Standardisierung von Repräsentations- und Austauschformaten. Seit dem Jahr 2013 sind in dem Programmbereich Forschungsinfrastrukturen auch zwei Servicebereiche des Instituts verortet, nämlich die Bibliothek und die Arbeitsstelle Zentrale Datenverarbeitung. Der Programmbereich Forschungsinfrastruktur bündelt somit Informationsangebote, Informationstechnik und Informationswissenschaft. Das IDS hat diese Form der organisatorischen Zusammenfassung von Forschungstätigkeiten und Dienstleistungen auch verschiedenen anderen außeruniversitären Forschungsinstituten vorgestellt. Die im IDS gewählte Umsetzung ist dort nicht nur auf reges Interesse gestoßen, sondern es wurde zum Teil sogar als Vorbild für eigene Restrukturierungen angesehen. Es steht daher zu erwarten, dass die zukünftigen Entwicklungen entlang der jetzt schon bereiteten Wege erfolgen werden. Die konkreten Planungen werden nachfolgend dargestellt.

Der Programmbereich Forschungsinfrastrukturen in der Organisationseinheit Zentrale Forschung ist aus dem BMBF gefördertem Projekt „Aufbau eines Zentrums für digitale Forschungsressourcen für die germanistische Sprachwissenschaft“ (2009-2010) und dem Nachfolgeprojekt „Forschungsinfrastrukturen in wissenschaftlichen Einrichtungen: Implementierung eines Prototyps am Institut für Deutsche Sprache“ (2011-2013) hervorgegangen.
Mit der Absicht, auf neue oder besondere Herausforderungen schnell und flexibel reagieren zu können, hat das IDS Anfang 2009 die direkt dem Direktor zugeordnete Organisationseinheit Zentrale Forschung eingerichtet. Dieser wurden die beiden Bereiche Korpuslinguistik (siehe Kupietz in diesem Band) und Forschungsinfrastrukturen als Programmbereiche zugeordnet. Mit dem neuen Programmbereich Forschungsinfrastrukturen wurde so eine organisatorische Einheit geschaffen, in der Projekte angesiedelt werden konnten,

die im Institut nicht einer der drei Forschungsabteilungen Lexik, Grammatik oder Pragmatik zuordenbar waren, sondern vielmehr einen Schnittstellencharakter aufweisen.[1]

Zu diesen Projekten zählen insbesondere D-SPIN (2008-2011, vgl. Bankhardt 2009) und sein Folgeprojekt CLARIN-D[2] (seit 2012), TextGrid (seit 2006, vgl. Zielinski et al. 2009), WissGrid (2009-2012, s. Ludwig/Enke 2013), Verwertung Geist[3] (seit 2011) sowie das Projekt „Beobachtung des Schreibgebrauchs mit computerlinguistischen Methoden" (seit 2013).

Im Rahmen der beiden Projekte „Aufbau eines Zentrums für digitale Forschungsressourcen für die germanistische Sprachwissenschaft" und „Forschungsinfrastrukturen in wissenschaftlichen Einrichtungen: Implementierung eines Prototyps am Institut für Deutsche Sprache" wurden zunächst verschiedene Konzepte erarbeitet, die den neuen Programmbereich am IDS etablieren und stärken sollten. Grundsätzlich wurde zunächst überlegt, ob am IDS ein Zentrum für Infrastrukturen eingerichtet werden kann, das dann die wissenschaftliche Auseinandersetzung mit abteilungsübergreifenden Fragen, insbesondere zum Einsatz der Informationstechnik in geisteswissenschaftlichen Forschungsinstituten, untersucht. Diese Herangehensweise barg allerdings die Gefahr in sich, nach der Aufbauphase ein monolithisches Zentrum installiert zu haben, das dann relativ isoliert von anderen Abteilungen arbeitet. Stattdessen wurde ein Vorgehen gewählt, bei dem einzelne Infrastrukturkomponenten eingerichtet wurden, von deren Zusammenspiel untereinander und mit den Abteilungen Mehrwerte für das gesamte Institut erwartet werden. Dies sollte erreicht werden, indem ein „infrastruktureller Werkzeugkasten", d.h. eine Sammlung möglicher Aktivitäten, zusammengestellt wurde. Der Nutzen der einzelnen Komponenten soll regelmäßig hinterfragt und überprüft werden. Einzelne Bestandteile können ggf. modifiziert oder wieder abgeschafft werden, neue können hinzukommen. In diesem Zusammenhang wurden beispielsweise Vorschläge für die Einführung und Etablierung kleinerer in der Institution zu verankernden Einrichtungen (z.B. Datenzentren, digitale Archive und Bibliotheken) entwickelt. Es wurde auch vorgeschlagen, aktuelle Entwicklungen in der IT-Landschaft kontinuierlich zu evaluieren, zu diskutieren und ggf. einzuführen. Auch sollen mögliche neue Forschungsschwerpunkte mit Bezügen zu Forschungsinfrastrukturen (z.B. Visualisierungstechniken für Massendaten) oder weitere Maßnahmen, wie grundsätzliche Entscheidungen zur Verwendung von Standards oder auch die Mitwirkung an Standardisierungsaktivitäten vom Bereich

[1] In diesem Sinne ist auch der Terminus „zentral" in *Zentrale Forschung* zu verstehen: Die im Bereich Zentrale Forschung durchgeführten Tätigkeiten lassen sich nicht einer der Säulen des IDS zuordnen.

[2] http://de.clarin.eu.

[3] http://vg.ids-mannheim.de.

Forschungsinfrastrukturen aus angestoßen und gelenkt werden. Durch die Schaffung des Programmbereichs sollte auch den im Januar 2011 vom Wissenschaftsrat veröffentlichten Empfehlungen „Forschungsinfrastrukturen für die Geistes- und Sozialwissenschaften stärken" (Wissenschaftsrat 2011), in denen das IDS bereits mehrfach exemplarisch erwähnt wurde, in besonderer Weise Rechnung getragen werden. Der Wissenschaftsrat empfiehlt in dieser Veröffentlichung unter anderem den geisteswissenschaftlichen Forschungsinstitutionen, dem Bereich der Forschungsinfrastrukturen eine stärkere Relevanz beizumessen. Selbstverständlich kann der Wissenschaftsrat keine konkreten Empfehlungen an einzelne Institute geben, da dies zum einen nicht seinem Auftrag entspricht und zum anderen aber auch, da es in einer heterogenen Forschungslandschaft nicht eine universelle Lösung für diese Aufgaben geben kann. Vor diesem Hintergrund sollten die Forschungsinfrastruktur-Projekte auch als Beitrag angesehen werden, nicht nur das IDS, sondern auch andere Forschungsinstitute in die Lage zu versetzen, den Bereich Forschungsinfrastrukturen institutionell zu verankern.

Durch die Entwicklung eines Methodeninventars sollten demnach nicht nur Lösungsansätze aufgegriffen oder entwickelt werden, die im Anschluss nur vom IDS selbst verwendet werden, sondern vielmehr war von Beginn an beabsichtigt, die Erfahrungen auch mit anderen geistes- und sozialwissenschaftlichen Forschungsinstituten zu besprechen. Falls gewünscht, werden die entwickelten Methoden dann auch anderen Einrichtungen zur Verfügung gestellt. Selbstverständlich müssten vor einer möglichen Implementierung an anderen Standorten diese Methoden an die lokalen Gegebenheiten angepasst werden. Dies kann so erfolgen, dass aus dem Inventar des „Werkzeugkastens" interessierte Forschungseinrichtungen passende Komponenten auswählen können, die dann in adaptierter Form dort implementiert werden können. Im Verlauf des Projektes „Forschungsinfrastrukturen in wissenschaftlichen Einrichtungen: Implementierung eines Prototyps am Institut für Deutsche Sprache" wurden deshalb auch eine Vielzahl von Interviews mit sehr verschiedenen geistes- und sozialwissenschaftlichen Forschungseinrichtungen geführt. Hierdurch konnten sich die Projektmitarbeiter nicht nur über den Stand der Implementierung digitaler Forschungsinfrastrukturen an anderen Instituten informieren, sondern es konnten den befragten Expertinnen und Experten auch die Ansätze des IDS vorgestellt werden.

Heute sind im Programmbereich Forschungsinfrastrukturen nicht mehr nur Forschungsprojekte angesiedelt, sondern auch zwei zentrale Serviceeinrichtungen. Eine dieser organisatorischen Umgestaltung zu Grunde liegende Überlegung bestand darin, dem bisherigen Forschungsbereich mit Fokus auf der Forschung einen Servicebereich mit einem entsprechenden Dienstleistungsauftrag gegenüberzustellen. Durch diesen Verbund sollen jedoch

nicht nur bereits vorhandene transdisziplinäre Aufgaben, die sowohl eine forschungsorientierte Expertise verlangen als auch für eine praktische Umsetzung am Institut vorgesehen sind, enger miteinander verzahnt werden, sondern auch neue Synergien geschaffen werden, indem methodologisches Wissen aus dem bisherigen, rein forschungsorientierten Programmbereich sowie praktiziertes Vorgehen aus den Servicebereichen mit infrastrukturorientierter Komponente besser aneinander herangeführt und dadurch Gemeinschaftsarbeiten ermöglicht werden.

Seit 2013 sind dazu die wissenschaftliche Hausbibliothek (vgl. Pohlschmidt in diesem Band) und die zentrale Arbeitsstelle für Datenverarbeitungsdienste mit dem Programmbereich Forschungsinfrastrukturen zu einer gemeinsamen Einheit zusammengelegt worden. Das Profil des neuen Programmbereichs umfasst dann Forschungsbereiche zu speziellen Aspekten beim Umgang mit linguistischen Forschungsdaten sowie die Servicebereiche Zentrale Datenverarbeitung (ZDV) und Bibliothek.

Im Bereich der Forschung plant der Programmbereich Forschungsinfrastrukturen, verschiedene Aktivitäten weiterzuführen bzw. neue Aktivitäten zu beginnen. Hierzu zählen derzeit die Bereiche Juristische Aspekte beim Umgang mit Sprachdaten (Ketzan/Kamocki 2012), Standards für Sprachressourcen (Stührenberg/Werthmann/Witt 2012), Informationsmodellierung (Witt/Metzing (Hg.) 2010; Mehler et al. (Hg.) 2011) und Infrastrukturforschung. Der Servicebereich wird ebenfalls um weitere Aktivitäten ergänzt. Weiterhin ist im Programmbereich Forschungsinfrastrukturen auch die Geschäftsstelle des Vereins TextGrid e.V.[4] angesiedelt. In naher Zukunft sollen außerdem ein Drittmittelbüro und eine Anlaufstelle zur Beratung im Zusammenhang mit dem Wissens- und Technologietransfer geschaffen werden. Einzelne dieser Aufgaben werden nachfolgend kurz ausgeführt.

Bei der Arbeit mit sprachlichen Daten sind die Wissenschaftler/innen immer wieder mit potenziellen rechtlichen Problemen konfrontiert. So ist es beispielsweise dem IDS nicht möglich, die von ihm verwalteten Daten öffentlich oder gar zum Download frei zugänglich anzubieten, da das IDS zwar im Besitz von Kopien der Daten, nicht jedoch deren Eigentümer ist. Vielmehr wurden mit allen Rechteinhabern separate Lizenzabkommen getroffen, die einen limitierten Umgang mit den gelagerten Daten erlauben. Hierbei ist allen Lizenzen gemein, dass gesammelte Textdaten ausschließlich zu den beschriebenen wissenschaftlichen Zwecken eingesehen werden dürfen. Der Zugriff durch authentifizierte Nutzer darf hingegen nur über eine ausschließlich zu diesem festgeschriebenen Zweck konzipierte Software erfolgen, mithilfe derer jedoch keine Vollversion der zugrunde liegenden Texte rekonstruiert,

[4] Vgl. http://www.textgrid-verein.de.

sondern lediglich ein auf die Suchparameter beschränktes Abfrageergebnis dargestellt werden darf. Zunächst scheinen diese recht restriktive Vorgaben für einen potenziellen Nutzer abschreckend (zu) wirken, jedoch stellen sie doch auch einen gangbaren Mittelweg dar, der zwischen dem Interesse der Besitzer der Ressourcen, d.h. der Lizenzgeber, an der Wahrung ihrer Urheber- und Leistungsschutzrechte und dem Bestreben der Wissenschaft, nach einem möglichst umfassenden Zugriff auf ein breites Spektrum an Primärdaten ohne allzu hohe Lizenzkosten, abwägt. Während in diesem Fall die Rechtslage eindeutig ist, da das IDS Lizenzverträge mit den Besitzern der Ressourcen abgeschlossen hat, ist die juristische Situation bei der Nutzung von z.B. im Internet „frei zugänglichen" Sprachressourcen weit weniger klar. Neben dem Urheberrecht kommen bei diesen Ressourcen z.B. auch Restriktionen der europäischen Datenbankdirektive zum Tragen, die das Kopieren und Weiterverwenden von Ressourcen beschränken, die in elektronischen Speichersystemen liegen und deren Erstellung mit einem erheblichen Aufwand erfolgte. Aber auch der Status von Daten, die bei Experimenten erhoben worden sind, ist nicht immer klar, da Datenschutz und Persönlichkeitsrechte zu beachten sind. Letztendlich lassen sich die genannten Probleme auf den Grundrechtskonflikt zwischen der wissenschaftlichen Freiheit einerseits und dem Recht auf Eigentum andererseits zurückführen. Diese Schwierigkeiten sind keineswegs speziell mit den Arbeiten im IDS verbunden, sondern betreffen alle Forscher/innen, die mit sprachlichen Daten arbeiten.

Am IDS wurde und wird innerhalb von Drittmittelprojekten, insbesondere in den Projekten D-SPIN und CLARIN-D, deshalb intensiv an der Thematik „Rechtliche Aspekte beim Umgang mit Sprachdaten" gearbeitet (vgl. Ketzan/Kamocki 2012). Hier geht es vor allem um Forschung zu diesem Aspekt und der Erstellung von Leitfäden, Musterlizenzverträgen etc., die Forscher/innen helfen können, rechtliche Klippen zu umschiffen. Perspektivisch sollen diese Aktivitäten im IDS im Programmbereich Forschungsinfrastrukturen fest verankert werden, um bei diesem wichtigen Themenkomplex nicht von der Finanzierung durch Drittmittelprojekte abhängig zu sein.

Alle maßgeblichen Empfehlungen zum guten Umgang mit Sprachdaten empfehlen die Verwendung von Standards. Das IDS ist schon seit vielen Jahren in den entsprechenden Standardisierungsgremien vertreten. Auf der nationalen Ebene ist das IDS im Deutschen Institut für Normung (DIN) sowie international in der International Organization for Standardization (ISO) in den einschlägigen Arbeitsgruppen aktiv. Zusätzlich engagiert sich das IDS auch in der, besonders für die eHumanities bedeutenden, Text Encoding Initiative (TEI). Diese Aktivitäten zielen auf die Fortentwicklung und damit Beeinflussung des Standardisierungsprozesses durch die mannigfaltige Erfahrung des IDS im Umgang mit Sprachressourcen sowie der allgemeinen Verbreitung

der Standards, beispielsweise durch Publikationen, Vorträge, Tutorials oder Workshops auf Konferenzen. Der Programmbereich wird in Zukunft in diesem Bereich weiter aktiv sein und die Standardisierungsarbeiten bzw. die Forschung zu Standards durch Schaffung eines eigenen Forschungsbereichs weiter stärken.

Wie oben erwähnt, bildet der Programmbereich Forschungsinfrastrukturen seit 2013 auch den Ort, wo neben den Forschungsaktivitäten auch verschiedene Servicebereiche angesiedelt sind.

Die Zentrale Datenverarbeitung (ZDV) widmet sich ihren Kernaufgaben, also dem generellen Betrieb von Rechnern und Netzwerkinfrastruktur. Dazu zählen unter anderem die Wartung und Betreuung von Betriebssystemen und zentralen Softwarepaketen, eine zentrale Datensicherung von zentralen Servern und Speicherarrays, das Einrichten und Konfigurieren neuer Hard- und Software. Die Mitarbeiter der ZDV unterstützen die Benutzer/innen bei Problemen mit Hard- und Software. Ferner administriert die ZDV die zentralen Netzwerkdienste, die das IDS anbietet, wie z.B. Webserver, bzw. zur täglichen Arbeit benötigt, wie Mail- oder VPN-Dienste. Weiterhin liegt der ordnungsgemäße Betrieb der zentralen Netzwerkkomponenten, wie LAN, Internet-Uplinks sowie WLAN, im Verantwortungsbereich der ZDV. Themen, die mittel- bis langfristig in diesem Bereich relevant werden, sind eine Neuausrichtung der IT-Gesamtstrategie auf aktuelle und zukünftige Bedürfnisse, wie beispielsweise zentrales Identitätsmanagement (IdM) oder die Virtualisierung von Servern und Desktop-Clients. Neben dem Tagesgeschäft ist die ZDV, als eine von zwei Arbeitsstellen im IDS, in der Ausbildung von Nachwuchs aktiv und wird auch weiterhin Ausbildungsstellen zum/zur Fachinformatiker/in anbieten.

Die Bibliothek widmet sich ihren Kernaufgaben, d.h. insbesondere der Anschaffung, Katalogisierung und Bereitstellung von Büchern und Zeitschriften sowohl in herkömmlicher Papierversion als auch in elektronischer Fassung. Traditionell kümmert sich die Bibliothek um die Betreuung der Gäste. Durch die Integration der Bibliothek in den Programmbereich wird es ihr zudem ermöglicht, einen Dokumentenserver zu betreiben, auf dem sie ihre digitalen Angebote, im Speziellen auch die Hauspublikationen, in einer integrierten technischen Umgebung bereitstellen kann. In Zukunft wird auch das Thema Open Access für das IDS immer relevanter werden. Die Bibliothek wird dazu eine Open-Access-Strategie für das Haus ausarbeiten und koordinieren.

Als ein neuer Servicebereich ist das Zentrum für Sprachressourcen geplant. Dieser Bereich soll ein zertifiziertes Langzeitarchiv für das Haus anbieten, das eine zentrale Plattform zur Nutzbarmachung einer empirischen Basis von Primärdaten und Arbeitsmaterialien darstellt (Fischer/Witt 2012). Dieses Repositorium soll auch ein integraler Beitrag des IDS zu verschiedenen Initiativen sein, die eine nationale und internationale Sprachressourcen-Infrastruktur

aufbauen und betreiben wollen. Im Rahmen der Drittmittelprojekte LIS und CLARIN-D wurde bereits der Grundstein für diesen Servicebereich gelegt. Im April 2013 ist das Repositorium erfolgreich durch den Zertifizierungsprozess für das Data Seal of Approval (DSA)[5] gelaufen.
Das IDS soll sich langfristig zu einem zentralen Archiv für germanistische Forschungsdaten entwickeln. Empirisch arbeitende Germanisten aus Deutschland erhalten somit die Möglichkeit, den Vorgaben der DFG und anderer Forschungsförderungsorganisationen zu genügen, indem sie ihre Daten am IDS abliefern. Dazu wird angestrebt, das Repositorium auch externen und potenziell auch kommerziellen Nutzern zur Verfügung zu stellen. Um den eigenen Aufwand zur Datenaufbereitung in einem vertretbaren Rahmen halten zu können, wird das IDS hierfür Richtlinien ausarbeiten und zur Verfügung stellen. Selbstverständlich kann ein derartiges Serviceangebot nur dann angeboten werden, wenn die Finanzierung gesichert wird. Modelle hierfür sind bereits im Rahmen von existierenden Projekten diskutiert worden, bedürfen aber weitergehender Ausarbeitung z.B. in Nachfolgeprojekten des Projekts Verwertung Geist.
All diese Aktivitäten schaffen unter anderem auch Möglichkeiten, eine Lücke im germanistischen bzw. sprachwissenschaftlichen Wissenschaftsbetrieb zu schließen, die Görl/Puhl/Thaller (2011, S. 177) folgendermaßen beschrieben haben:

In den Gesprächen mit den teilnehmenden Institutionen hat sich ergeben, dass die Regeln der DFG zur Langzeitarchivierung sowohl der Primärdaten aber auch von Publikationen und anderen Dokumenten durch die Hochschulen bisher fast nicht erfüllt werden können, da ein Mangel an hochschulinternen Verfahrensregeln und Strategien aber auch Hilfe bei deren Umsetzung festzustellen ist.

Zusammenfassend bleibt festzuhalten, dass sich das IDS in den vergangenen Jahren nicht nur intensiv mit Fragen der Umsetzung einer zeitgemäßen, verlässlichen und an neue Entwicklungen anpassbaren Forschungsinfrastruktur beschäftigt hat, sondern diese auch sukzessive implementiert hat, so dass sich das IDS – zu Beginn seiner nächsten 50 Jahre – in diesem Bereich in einem sehr guten Zustand befindet.

[5] https://assessment.datasealofapproval.org/assessment_85/seal/html/ (zuletzt abgerufen 30.7.2013).

Literatur

- **Bankhardt, Christina** (2009): D-SPIN – Eine Infrastruktur für Deutsche Sprachressourcen. In: Sprachreport 1/2009, S. 30-31.
- **Fischer, Peter M./Witt, Andreas** (2012): Developing Solutions for Long-Term Archiving of Spoken Language Data at the Institut für Deutsche Sprache. In: Proceedings of the LREC'12 Workshop on Best Practices for Speech Corpora in Linguistic Research.
- **Görl, Simone/Puhl, Johanna/Thaller, Manfred** (2011): Empfehlungen für die weitere Entwicklung der wissenschaftlichen Informationsversorgung des Landes NRW. Berlin.
- **Ketzan, Erik/Kamocki, Paweł** (2012): The CLARIN-D Legal Help Desk and emerging copyright issues for language scientists. In Proceedings of LREC 2012 Workshop on IPR Issues. Istanbul, Turkey, 2012.
- **Mehler, Alexander et al.** (Hg.) (2011): Modeling, learning, and processing of text-technological data structures. (= Studies in computational intelligence 370), Berlin/Heidelberg.
- **Ludwig, Jens/Enke, Harry** (Hg.) (2013): Leitfaden zum Forschungsdaten-Management: Handreichungen aus dem WissGrid-Projekt. Glückstadt.
- **Wissenschaftsrat** (2011): Wissenschaftsrat: Übergreifende Empfehlungen zu Informationsinfrastrukturen (Drs. 10466-11), Januar 2011. Internet: http://www.wissenschaftsrat.de/download/archiv/10466-11.pdf.
- **Stührenberg, Maik/Werthmann, Antonina/Witt, Andreas** (2012): Guidance through the standards jungle for linguistic resources. In: Proceedings of the LREC-12 Workshop on Collaborative Resource Development and Delivery.
- **Witt, Andreas/Metzing, Dieter** (Hg.) (2010): Linguistic Modeling of Information and Markup Languages. Contributions to Language Technology. (= Text, Speech and Language Technology 41). Springer Netherland.
- **Zielinski, Andrea et al.** (2009): TEI documents in the grid. In: LLC 24(3), S. 267-279.