5 Psycholinguistik

Peter Wittenburg, Sebastian Drude, Daan Broeder

5.1 Einführung in den Forschungsbereich

Die Psycholinguistik ist der Bereich der Linguistik, der sich mit dem Zusammenhang zwischen menschlicher Sprache und dem Denken und anderen mentalen Prozessen beschäftigt, d.h. sie stellt sich einer Reihe von essentiellen Fragen wie etwa (1) Wie schafft es unser Gehirn, im Wesentlichen akustische und visuelle kommunikative Informationen zu verstehen und in mentale Repräsentationen umzusetzen? (2) Wie kann unser Gehirn einen komplexen Sachverhalt, den wir anderen übermitteln wollen, in eine von anderen verarbeitbare Sequenz von verbalen und nonverbalen Aktionen umsetzen? (3) Wie gelingt es uns, in den verschiedenen Phasen des Lebens Sprachen zu erlernen? (4) Sind die kognitiven Prozesse der Sprachverarbeitung universell, obwohl die Sprachsysteme derart unterschiedlich sind, dass sich in den Strukturen kaum Universalien finden lassen?[1]

Um diese Fragen beantworten zu können, bedient sich die Psycholinguistik verschiedener Methoden wie z.B. beobachtender Feldstudien und verschiedenartigster Experimente, die letztlich alle Modalitäten umfassen, die in der menschlichen Kommunikation verwendet werden bzw. etwas über den Zustand der kognitiven Verarbeitung aussagen können. In verstärktem Maße sind dies auch *brain imaging*-Methoden (EEG, MEG, fMRI), um Aussagen über die funktionelle Architektur des Gehirns machen zu können, und Simulationen kognitiven Verhaltens. Die Psycholinguistik als eine noch relativ junge Teildisziplin ist demzufolge von Beginn an eine datenorientierte Wissenschaft, die in starkem Maße naturwissenschaftliche Methoden übernommen hat.

1 Vgl. Evans; Levinson (2009).

Konsequenterweise setzt die Psycholinguistik ein ganzes Arsenal von Technologien ein, um Daten zu messen und auszuwerten. Essentiell war und ist dabei der frühe Einsatz aller Möglichkeiten der digitalen Innovation. Deshalb wurden, soweit dies von den Kapazitäten her möglich war, frühzeitig alle Video- und Audioaufnahmen digitalisiert und analysiert, um skalierbare und quantitative Auswertungsmethoden einzusetzen. Ebenso war es selbstverständlich, immer wieder auch neue dynamische Paradigmen wie z.B. *virtual reality* einzusetzen.

Mithin werden in der Psycholinguistik sehr viele verschiedene Datentypen erzeugt und ausgewertet. Diese Daten lassen sich grob in Primär- und Sekundärdaten unterteilen. Sehr grob verallgemeinert gesagt bilden Primärdaten direkt realweltliche Ereignisse ab, wohingegen Sekundärdaten gewisse Eigenschaften der Ereignisse unter Bezug auf die Primärdaten explizit machen. Typische Primärdaten sind etwa Audio- und Videoaufnahmen, die durch Sekundärdaten annotiert (transkribiert, übersetzt, glossiert, kommentiert etc.) werden. Sekundärdaten umfassen wegen der immer mehr diversifizierten Verarbeitungsmethoden ein immer größeres Spektrum.

Die Primärdaten lassen sich wiederum in zwei Kategorien unterteilen, für die unterschiedliche Zeitfenster der Erhaltung gelten: (1) Experimentelle Daten werden in einer Laborumgebung gewonnen und haben zumeist nur eine kurze Relevanz für die Forschung und für die Gesellschaft. (2) Beobachtungsdaten werden zumeist in natürlichen Umgebungen (im „Feld“) aufgenommen. Sie können eine große Relevanz für die Forschung und auch die Gesellschaft haben, da sie immer auch eine Art „Momentaufnahme“ des Zustandes von Sprachen und Kulturen darstellen. So gilt z.B. für alle Daten, die im DOBES Programm der Volkswagenstiftung zur Dokumentation bedrohter Sprachen[2] aufgenommen und erzeugt werden, dass sie zukünftigen Generationen zur Verfügung stehen sollen, um ihnen eine Einsicht über die Vielfalt der Sprachen und deren linguistische Systeme zu vermitteln. Das bedeutet, hier besteht die Erwartungshaltung, dass diese Daten zeitlich unlimitiert vorgehalten und genutzt werden können. Dies gilt auch für die dazugehörigen Sekundärdaten, die das linguistische Wissen über die aufgenommenen Sprachen enthalten.

2 Vgl. DOBES (2011).

Die Situation bezüglich der Datenmengen und der Datenorganisation in der Psycholinguistik ist sehr unterschiedlich, sie ist abhängig von der Ausstattung der Institutionen bzw. Abteilungen bezüglich der technischen und personellen Ressourcen. Die Situation am Max-Planck-Institut (MPI) für Psycholinguistik[3], die hier vornehmlich als Ausgangspunkt genommen werden soll, gibt daher nicht die allgemeine Realität wieder, kann aber als Ankerpunkt dafür genommen werden, wie derartige Institute/Abteilungen sich in Zukunft verhalten könnten und idealerweise sollten, wenn sie die entsprechenden Mittel zur Verfügung haben.

Da das MPI eine sehr breite Vielfalt an Methoden und Technologien einsetzt, können die Datenmengen und auch die Komplexität der Relationen zwischen den Objekten durchaus als exemplarisch angesehen werden. Wegen der zunehmenden Datenmengen musste das MPI frühzeitig (in den 1990er-Jahren) mit der Entwicklung von Methoden der Datenorganisation beginnen. Das Daten-Repository umfasst heute ein organisiertes Online-Archiv von 74 Terabyte, weitere ca. 200 Terabyte sind zu verwalten. Die jährliche Zunahme lässt sich derzeit mit ca. 17 TB angeben. Diese Zahlen sind heute sicherlich in vielerlei Hinsicht untypisch für die meisten psycholinguistischen Institute. Im Folgenden werden wir die Konsequenzen darstellen, die sich aus dieser Datenvielfalt und -menge ergeben.

5.2 Kooperative Strukturen

Institutsübergreifende Zusammenarbeit

Zur Beantwortung der Frage nach der Existenz kooperativer Strukturen bezüglich Daten muss man die verschiedenen Datentypen betrachten. Experimentelle Daten werden im Allgemeinen unter ganz bestimmten Rahmenbedingungen mit ganz bestimmten Personengruppen zumeist in einer Reihe von Versuchen gewonnen, um eine bestimmte wissenschaftliche Hypothese abzusichern. Ihre Austauschbarkeit ist daher relativ gering und die Datenmengen sind zumeist auch beschränkt. Es

3 Vgl. MPI Homepage: http://www.mpi.nl.

besteht eine Vorhaltepflicht von zehn Jahren, um Argumentationen, die in wissenschaftlichen Artikeln auf der Basis von Daten geführt werden, nachvollziehbar zu machen. Demzufolge wird die Lebensdauer der Daten im Wesentlichen von dieser Vorhaltepflicht bestimmt. Dies hat sich mit der Einführung der *brain imaging*-Techniken insofern etwas geändert, als dass die Erzeugung der Daten recht kostenintensiv ist und es dadurch einen zunehmend erhöhten Druck gibt, den Datenaustausch zwischen den Instituten zu ermöglichen. Das Gleiche mag auch gelten für neue Fachrichtungen wie z.B. der linguistischen Genetik. Man kann im Allgemeinen davon ausgehen, dass nach zumeist weniger als zehn Jahren neue Sensortechnologien zum Einsatz kommen, die dann zumeist genauere Messungen ermöglichen; d.h. abgesehen von den Fällen, bei denen z.B. historisch wertvolle Anomalien festgestellt werden, macht es keinen Sinn, die Masse der experimentellen Daten länger aufzubewahren.

Experimente bedingen oftmals geeignete Stimuli, deren Erzeugung ebenfalls recht kostenintensiv ist – vor allem bei den *virtual reality*-Experimenten. Hier gibt es einen regen Austausch zwischen kollaborierenden Gruppen. Da das Geschick der Stimuluserzeugung die Qualität der wissenschaftlichen Resultate beeinflusst, wird ein offener Austausch jedoch nur selten angestrebt, es sei denn, die Experimente dienen der Studentenausbildung.

Ganz anders sieht es bei Beobachtungsdaten aus, die normalerweise für ein bestimmtes analytisches Ziel aufgenommen werden, aber immer häufiger auch für andere Zielsetzungen mit anderen beobachtenden Daten zu neuen virtuellen Korpora verbunden werden. Im DOBES-Archiv[4] bedrohter Sprachen ist die vielfältige Verwendbarkeit sogar eines der Kriterien für die Aufnahmen und davon abgeleitete Daten. Die Erzeugung von Korpora, die neben den Aufnahmen auch Transkriptionen und weitere Annotationen umfassen, ist eine sehr kostspielige Aufgabe, so dass deren Austausch und Wiederverwendbarkeit eine Notwendigkeit ist. Da Observationen eine Momentaufnahme von sich ständig verändernden Kulturen und Sprachen darstellen, haben sie oftmals eine kulturgeschichtliche Bedeutung für zukünftige Generationen.

4 Vgl. DOBES Homepage: http://www.mpi.nl/dobes.

Konsequenterweise ist der Grad an Organisation der Daten des MPI im Bereich der beobachtenden Daten der am weitesten Ausgearbeitete. Es ist seit Jahren im MPI allgemein akzeptiert, dass alle Beobachtungsdaten und davon abgeleitete Daten mittels Metadaten beschrieben und in das DOBES-Archiv gestellt werden. Bezüglich der experimentellen Daten gibt es jetzt eine Vereinbarung, die *brain image*-Daten mittels erweiterter Metadaten ebenfalls organisiert in das Archiv zu stellen, um sie damit zumindest für die geforderten zehn Jahre zu sichern und sie in einfachem Maße auch außerhalb des Instituts wiederverwendbar vorzuhalten.

Dieser Unterschied lässt sich durchaus verallgemeinern. Insbesondere durch Projekte wie DOBES hat sich mittlerweile ein breites Bewusstsein über die Notwendigkeit des sorgfältigen Erhaltens und des Austauschens vieler beobachtender Daten herausgebildet. Im experimentellen Bereich gibt es diese Notwendigkeit zum langfristigen Erhalt im Allgemeinen nicht.

Die Art der kooperativen Strukturen entspricht der Art der Daten. Bezüglich der kulturell wichtigen Beobachtungsdaten gibt es eine wachsende Bereitschaft zur zentralen Speicherung und Pflege in dafür geeigneten Zentren. Wiederum ist das DOBES-Programm ein sehr gutes Beispiel, da es auch für andere nicht direkt vom Projekt finanzierte Wissenschaftler ein Signal gab, mit ihren Daten so umzugehen, dass ihr Erhalt und damit die Nachnutzung wahrscheinlicher wird. Es besteht – auch international – ein reger Austausch über Methoden der Datenaufnahme über ihre Verarbeitung und Analyse bis hin zur Langzeitarchivierung. Im experimentellen Bereich erfolgt der Austausch aus verschiedenen Gründen zumeist über die in den Publikationen beschriebenen Resultate, d.h. die Daten sind die Basis der Resultate; sie selbst werden im Allgemeinen aber nicht ausgetauscht.

Die CLARIN Forschungsinfrasturkturinitiative[5] ist ein gutes Beispiel für eine derartige kooperative Struktur. An die 200 Institute in Europa arbeiten zusammen und haben sich z.B. darauf geeinigt, dass es starke Zentren als Rückgrat der Infrastruktur geben muss, die insbesondere auch für die langfristige Vorhaltung und Pflege der Daten sorgen. Der Grad

5 Vgl. CLARIN (2011).

der Organisierung der wissenschaftlichen Institute mit Beobachtungsdaten und auch Korpora in dieser datenorientierten Infrastrukturinitiative ist weitaus höher als der der experimentell arbeitenden Wissenschaftler.

Langzeitarchivierungsdienste und Zusammenarbeit mit Dienstanbietern

Die Langzeitarchivierung (LZA) hat zwei Aspekte: (1) Zum einen müssen die *bitstreams* und (2) zum anderen muss ihre Interpretierbarkeit gesichert werden. Wie im Report der High Level Expert Gruppe zu Scientific Data (HLEG)[6] festgestellt wurde, ist es die Aufgabe aller Beteiligten, vom Erzeuger bis zum Service-Anbieter, für die Interpretierbarkeit der Daten zu sorgen, d.h. immer wieder geeignete Erzeugungs-, Transformations- und Organisationsmaßnahmen anzuwenden. Dies kann nur von den Experten initiiert werden, die die Inhalte zu einem gewissen Grad kennen und die Zulässigkeit von Operationen und damit die Authentizität der neu erzeugten Repräsentationen beurteilen können. Die Ausführung der u.U. rechenintensiven Aufgaben (man muss hierbei z.B. an die Transformation von Terabytes von Video-Streams von einem Codec[7] auf einen anderen denken) kann natürlich bei einem externen Serviceprovider stattfinden. Für die *bitstream preservation* ist diese Fachkenntnis nicht erforderlich, sie kann demzufolge an andere Institutionen ausgelagert werden.

Die untenstehende von der HLEG entwickelte Grafik (Abb.1), veranschaulicht die drei involvierten Schichten recht gut und deutet auch an, dass die Etablierung des gegenseitigen Vertrauens und die Datenpflege Aufgaben aller sind. Das MPI hat bereits eine solche „Architektur" in Ansätzen realisiert: (1) Es erhält von Wissenschaftlern – auch externen – Daten zur Aufbewahrung und ermöglicht ihnen auch den Zugriff darauf. (2) Es bietet umfangreiche Software an, um auf die Daten zuzugreifen, sie zu manipulieren etc., wobei Domänenwissen in die Gestaltung der Zugriffstechnologien einfließt. (3) Es tauscht mit dem Rechenzentrum (RZ) Garching und der Gesellschaft für wissenschaftliche Datenverarbeitung (GWDG) Göttingen alle Daten aus, um die *bitstream* Sicherung zu gewährleisten. Hinzu

6 Vgl. High Level Expert Group on Scientific Data (2010).

7 Mit dem Begriff „Codec" werden Verfahren bezeichnet, die Daten oder Signale digital kodieren und dekodieren.

kommt, dass Abkommen auf allen Ebenen geschlossen und Code of Conducts[8] etabliert wurden bzw. werden, um Vertrauen zu schaffen. Momentan wird daran gearbeitet, den Software Stack des MPI auch z.B. im RZ Garching zu installieren, um einen redundanten Pfad für Zugriffe anbieten zu können.

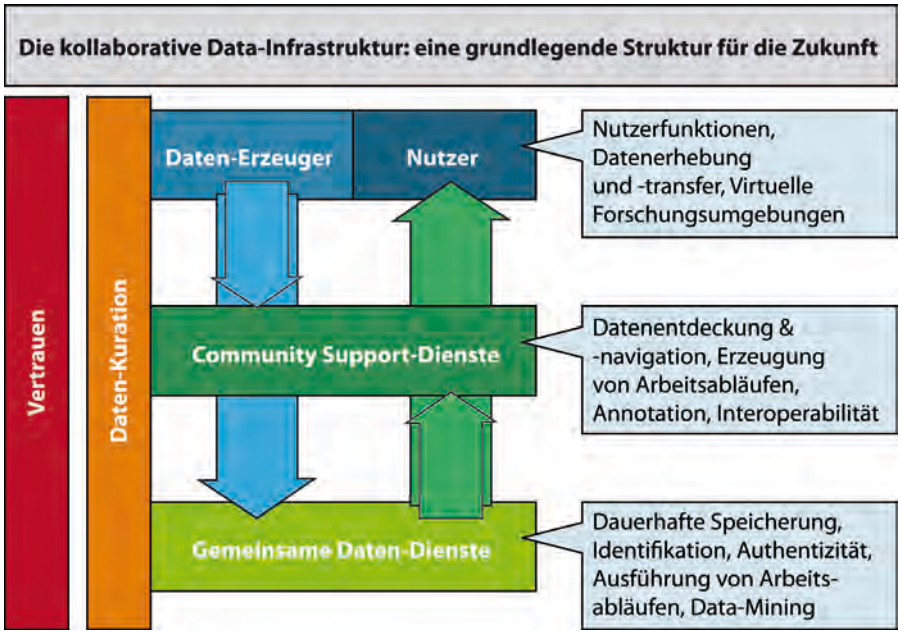

Abb. 1: Die kollaborative Dateninfrastruktur[9]

Das MPI ist wohl gegenwärtig eine der ganz wenigen Institutionen, die allen Wissenschaftlern die Möglichkeit eines „offenen Deposits" für Sprachdaten anbietet, wobei bestimmte Anforderungen an die Qualität gestellt werden und die prinzipielle Verfügbarkeit der Daten für die Forschung gegeben sein muss. Das Institut hat das oben geschilderte dreistufige System zur Sicherung der Daten und zu deren Zugänglichkeit im Rahmen eines Online-Archivs implementiert. Es verfügt selbst über Speicherplatz, um

8 Code of Conducts sind Handlungsanleitungen zu einem ethisch und rechtlich verantwortlichen Umgang mit Daten, die die Beteiligten im Allgemeinen unterschreiben müssen und deren Einhaltung im Allgemeinen der sozialen Kontrolle unterliegt.

9 Vgl. High Level Expert Group on Scientific Data (2010).

befugten Personen einen direkten oder web-basierten Zugang zu den Daten zu geben. Weiterhin verfügt es über ein hierarchisches *storage management*-System, so dass lokal zumindest zwei Kopien aller Objekte zur Verfügung stehen. Darüber hinaus werden Kopien aller Objekte dynamisch an den zwei Rechenzentren der Max-Planck-Gesellschaft (MPG) mittels verschiedener Protokolle[10] angefertigt, die selbst wiederum an ihren Standorten Abkommen mit leistungsstarken Partnern haben, um alle Archivdaten zu spiegeln. Von allen Objekten existieren also mindestens sechs Kopien.

Für die Kopien an den Rechenzentren gilt eine institutionelle Garantie von 50 Jahren, die vom Präsidenten der MPG gegeben worden ist. Basis dieses Langzeitarchierungsservices der beiden Rechenzentren ist eine Empfehlung des beratenden Rechenausschusses für Rechenanlagen der MPG aus dem Jahre 2004.

Das Institut hat insofern ebenfalls eine Ausnahmestellung, als dass es aktive Unterstützung geleistet hat, bisher weltweit 13 regionale Datenarchive aufzubauen, die mit dem digitalen Archiv des Instituts selbst verbunden sind. Der bisher statische Datenaustausch soll im Jahr 2011 dynamisch gestaltet werden. Allerdings sind die bisher zum Teil schlechten Netzwerkanbindungen ein Hinderungsgrund.

5.3 Daten und Metadaten

Datentypen

Die im Archiv gespeicherten Daten umfassen eine Vielfalt von Datentypen. Neben einer großen Menge digitalisierter Fotos, Audio- und Videoaufnahmen umfasst das Archiv verschiedene „Zeitreihen" (*eye tracking*, Gesture-Aufnahmen mit selektiven Techniken, EEG, fMRI), eine Reihe von strukturierten und semi-strukturierten Daten, die textuelle Darstellungen beinhalten (Annotationen der verschiedenen Aufnahmen, Lexika, linguistische Notizen etc.) und kleinere Datensätze

10 Vgl. u.a. Rsync (2011); AFS (2011).

mit statistischen Daten etc. Alle Daten, die im Online-Archiv enthalten sind, sind mittels IMDI-Metadaten[11] beschrieben.

Im Archiv sind alle Daten einschließlich der Metadaten in ihrem Ursprungsformat enthalten, d.h. sie werden nicht in Containern gekapselt. Dies sichert ab, dass nur eine minimale Abhängigkeit von zusätzlicher Software gegeben ist. Im Prinzip reicht ein simpler Metadaten-Browser aus, die auf einem XML-Schema basierten Metadaten „von der Wurzel her zu parsen", um das komplette Online-Archiv mit all seinen Ressourcen zugänglich zu machen. Jedes Objekt ist mit einem persistenten Identifikator (PID registrierter Handle[12]) versehen, der die Identifikation und die Integritätsprüfung erlaubt.

Langzeitarchivierung und Publikation von Daten

Wie bereits erwähnt, werden die Daten auf zweierlei Pfaden kopiert: (1) Zu den zwei Rechenzentren mittels unterschiedlicher Protokolle und (2) Teile des Archivs zu den weltweit verteilten, regionalen Repositorien. Da keines der verwendeten Protokolle gegenwärtig eine zufriedenstellende Integritätsprüfung beinhaltet, ist in einer Kollaboration zwischen der CLARIN Forschungsinfrastruktur und der DEISA Infrastruktur für High Performance Computing[13] an einer sicheren Replikation[14] gearbeitet worden, die in 2011 in den praktischen Dienst übernommen werden soll. Dabei wird bei jedem Vorgang, der mit einer der Instanzen eines Objektes zusammenhängt, überprüft, ob die mit dem persistenten Identifikator verknüpfte Prüfsumme noch zutrifft. Die registrierten PIDs können über ein API (*application programming interface*) entsprechend angesprochen werden. Dieses System ist auf der Basis expliziter *policy rules* realisiert, die festlegen, was mit den Daten geschehen darf und routinemäßig geschehen soll (Zugriffsrechte, Kopien, Konversionen etc.). Dies ist unserem Wissen nach der einzig machbare Weg, um Langzeitarchivierung in verteilten Szenarien in Zukunft vertrauensvoll und überprüfbar zu gestalten.

11 Vgl. IMDI (2007).
12 Vgl. Handle (2010).
13 Vgl. DEISA Homepage: http://www.deisa.eu.
14 Vgl. REPLIX (2011).

Lediglich die im Online-Archiv befindlichen Daten (74 TB) werden gespiegelt, da nur sie die erforderlichen Mindestvoraussetzungen (Metadatenbeschreibung) erfüllen.

Eine Publikation der Forschungsdaten erfolgt mittels der IMDI-Metadaten, die von jedem OAI-PMH[15] basierten Serviceprovider eingeholt werden können (dies geschieht auch tatsächlich) und die persistente Identifikatoren beinhalten. Metadaten sind offen; ein explizites XML-Schema ist über das Web zugreifbar und ein Mapping auf Dublin Core[16] wurde ausgearbeitet, das natürlich viele IMDI-Informationen nicht erfasst. Somit kann über das OAI-PMH-Protokoll auf alle Metadaten im IMDI-Format oder als Dublin Core-Beschreibung zugegriffen werden. Dies wird z.B. vom Virtual Language Observatory,[17] das von CLARIN betrieben wird, durchgeführt, wobei ein gemeinsamer Index für die Metadaten vieler Datenanbieter erzeugt wird und dann mittels z.B. *facetted browsing* durchsuchbar ist. Dies gilt z.B. auch für die Daten des Bayerischen Spracharchivs[18] und zunehmend auch für die Daten anderer deutscher Zentren.[19] Die Metadaten beinhalten zumeist den PID oder eine URL, so dass ein Zugriff auf die Daten selbst möglich ist.

Datenressourcen können beliebig zu neuen „virtuellen" Kollektionen zusammengefasst werden, wobei lediglich ein neuer Metadatensatz erzeugt wird, der die Referenzen auf die Ressourcen umfasst. Auch dieser Metadatensatz erhält eine PID und die typischen Metadatenbeschreibungen, kann also zitiert werden.

Mindestanforderungen an Daten und Formate

Bezüglich der Qualitätsanforderungen müssen zwei Ebenen unterschieden werden: (1) Die Qualität der Daten und Metadaten und (2) die Qualität des Archives.

---

15 Vgl. Open Archives (2011).
16 Vgl. Dublin Core (2011).
17 Vgl. VLO (2010).
18 Vgl. BAS (2005).
19 Vgl. u.a. IDS (2011); BBAW (2010); diverse Universitäten. Diese Zentren sind zumeist Mitglied der CLARIN Forschungs-Infrastruktur-Initiative (vgl. CLARIN (2011)), die Kriterien für Zentren aufgestellt hat, die jetzt von mehreren deutschen Teilnehmern schrittweise erfüllt werden.

Ein Forschungsinstitut lebt von der Innovation und der Kreativität seiner Wissenschaftler. Eine zu strenge Standardisierung bzw. Qualitätskontrolle bezüglich der Daten wäre demzufolge kontraproduktiv. Dies ist anders bei den Metadaten, wo die Einhaltung des IMDI-Schemata gefordert wird. Das Repository eines Instituts hat immer verschiedene Bereiche, für die unterschiedliche Kriterien gelten: (1) ein Bereich, der durch Standardformate abgesichert ist und (2) ein anderer, indem der Innovation Raum gegeben wird. Grundsätzlich ist dabei zu beachten, dass eine Kontrolle der Dateninhalte kaum stattfindet. Wir vermuten sogar, dass eine inhaltliche Kontrolle mittels eines traditionellen *peer to peer review*-Systems für die meisten unserer Daten nicht einfach möglich ist.

Bezüglich der formalen Kontrolle gilt für das Online-Archiv, dass zunächst alle Objekte mittels einer IMDI Metadaten-Beschreibung assoziiert sein und dass bestimmte akzeptierte Formate (zumeist offene, standardisierte Formate wie z.B. MPEGx[20], mJPEG2000[21], XML[22], linear PCM[23] etc.) eingehalten werden müssen. Beim Hochladen von Daten mittels bereitgestellter Werkzeuge des LAMUS Systems[24] werden entsprechende Überprüfungen der Daten und Metadaten vorgenommen und somit auch eine relativ hohe Formatkonsistenz des Archivs erzeugt. Wir müssen jedoch auch immer wieder Kompromisse schließen und andere Formate wie z.B. Resultatdateien, die mittels statistischer oder mathematischer Programme (SPSS[25], Matlab[26] etc.) erzeugt werden, akzeptieren, obwohl sie z.T. proprietär sind. Nicht immer sind formale Prüfmethoden verfügbar bzw. vom Archiv schnell erzeugbar. Daher können für derartige Daten oft keine Garantien bezüglich langfristiger Pflege und Unterstützung gegeben werden.

Eine hohe Formatkonsistenz vor allem bezüglich der Primärdaten wird also angestrebt, obwohl immer wieder auch in kleinem Maßstab, insbesondere bei den Sekundärdaten, Ausnahmen gemacht werden müssen.

---

20 Mit MPEGx wird eine Klasse von Codecs beschrieben, die von der MPEG Initiative (http://www.mpeg.org/) entwickelt wurden wie z.B. MPEG1, MPEG2, MPEG4/H.264.

21 Vgl. JPEG 2000 (2011).

22 Vgl. XML Homepage: http://www.w3.org/XML/.

23 Vgl. PCM (2011).

24 Vgl. LAT (2008).

25 Vgl. SPSS (2011).

26 Vgl. Mathworks (2011).

Die Qualität des Archivs kann hauptsächlich an der Umsetzung seiner Kernaussagen und Zusagen gemessen werden. Hier hat das Institut entschieden, sich dem Qualitätskontrollprozess „Data Seal of Approval“[27] zu unterziehen. Diese Überprüfung wird regelmäßig vorgenommen. Die erste Überprüfung wurde gerade erfolgreich abgeschlossen. Wir sehen die Berechtigung, das *seal* als eine der erforderlichen vertrauensbildenden Maßnahmen zu verwenden.

Datenvolumen

Das Online Archiv umfasst gegenwärtig 74 Terabyte an Daten. Darüber hinaus werden noch weitere 200 Terabyte gespeichert. Diese Daten sind jedoch nicht mittels Metadaten beschrieben und ihre Formate sind nicht überprüft. Sie gehören zumeist den Katergorien „experimentelle Daten mit geringer Lebensdauer“ oder „noch nicht-erschlossene Beobachtungsdaten“ an. Der zunehmende Umfang dieser schlecht gepflegten Daten gibt zur Sorge Anlass. Momentan verzeichnet das Archiv einen Zuwachs von 17 TB pro Jahr, der durch neue Verfahren (uncompressed mJPEG2000, fMRI Scanner mit höherer Resolution, Anforderungen aus der Linguistic Genetics) in den kommenden Jahren sicherlich um ein Vielfaches zunehmen wird.

Nutzungsbeschränkungen

Der Zugriff auf Daten ist in jedem Fall kostenfrei. Allerdings handelt es sich bei vielen Daten um Aufnahmen, bei denen der rechtliche Aspekt des Personenschutzes zu garantieren ist. Wir unterscheiden vier Kategorien:[28]

- Daten, die ohne Bedenken über das Web zugänglich sind,
- Daten, die zunächst eine Registrierung und das elektronische Unterschreiben eines *code of conduct* erfordern,

27 Vgl. Data Seal of Approval Homepage: http://www.datasealofapproval.org/.

28 Im Jahre 2002 wurde ein Workshop mit mehreren namhaften deutschen und niederländischen Rechtsexperten im Rahmen des DOBES Projektes durchgeführt. Es konnte jedoch bezüglich der Rechtslage eines international operierenden Online-Archivs keinerlei klare Rechtsauskunft gegeben werden.

- Daten, die eine elektronische Anfrage bei dem verantwortlichen Wissenschaftler oder dem Archivleiter erfordern,
- Daten, die grundsätzlich nicht zugänglich sind – außer für den Archivbetreiber, der immer (technischen) Zugang hat.

Ein großer Teil der Daten fällt derzeit in die dritte Kategorie. Dies hat zum Teil auch mit der fehlenden Praxis des korrekten Zitierens von Web-Ressourcen zu tun, das technisch noch nicht zufriedenstellend, d.h. hochgradig automatisch, gelöst ist. Wissenschaftler befürchten oft, dass ihre in der Datenerzeugung und -pflege enthaltene Arbeit nicht entsprechend gewürdigt wird.

Die gespeicherten Ressourcen umfassen auch z.T. sehr sensible Inhalte wie z.B. Daten über *split brain*-Patienten oder Aufnahmen geheimer Zeremonien. Daher werden einige der Daten der vierten Kategorie angehören und nicht verfügbar sein. Allerdings haben wir uns in allen Fällen das Recht auf Archivierung gesichert, was die Möglichkeit des Kopierens zu vertrauenswürdigen Partnern für Archivierungszwecke einschließt.

Zugriff auf ältere Daten

Für die Beobachtungsdaten, die oft nicht nur z.B. Sprechakte, sondern eine Kultur zu einem gegebenen Zeitpunkt dokumentieren, ist der Unterschied zwischen „alten" und „neuen" Daten fraglich, da auch alte Daten ihre Relevanz nicht verlieren (oft ist das Gegenteil der Fall). Es gibt jedoch einen Unterschied zu machen zwischen der originären Forschungsfrage (für die eine bestimmte Ressource aufgenommen wurde) und neuen Forschungsfragen, die den Aufbau immer wieder neuer virtueller Kollektionen durch die Wissenschaftler erfordert. Dabei werden „ältere Daten" im Allgemeinen wieder neu annotiert, d.h. die neuen Annotationen gehören letztlich wieder zu einer Kollektion. Somit ist eine strenge Unterscheidung in alte und neue Beobachtungsdaten immer weniger sinnvoll.

Ganz anders ist es im Bereich der experimentellen Daten, wobei allerdings von einer hohen Verfallsrate ausgegangen werden muss: Alte Daten verlieren schnell an Relevanz. Zumeist führt der Weggang eines Wissenschaftlers dazu, dass seine Daten kaum noch weiter analysiert werden.

Ein anderer Grund liegt darin, dass immer neue Sensor-Technologien zum Einsatz kommen, die Daten einer höheren Qualität bieten.

Metadaten und persistente Identifikatoren

Aufgrund der extremen Datenzunahme in den letzten 15 Jahren hat sich das MPI frühzeitig um die Etablierung eines Metadaten-Standards bemüht. Zusammen mit anderen europäischen und auch einigen außereuropäischen Wissenschaftlern konnte der IMDI Best-Practice-Standard etabliert werden. Er basiert auf einem festen, strukturierten XML-Schema, in dem die relevanten Elemente in definierte Kontexte eingebunden sind und z.T. offene Vokabulare haben. Einige Flexibilitäten sind vorgesehen, spezielle Profile können gebildet werden und IMDI erlaubt es, ganze Hierarchien von (virtuellen) Kollektionen zu bilden.

So konnte sich IMDI neben dem Dublin Core-basierten „Open Language Archives Community Standard“[29] in der Fachwelt weltweit etablieren. Das MPI Archiv basiert ausschließlich auf IMDI, um einen einheitlichen Standard zur Beschreibung und zur Organisierung der Daten zu bekommen. Das Archivmanagement erfolgt im Prinzip auf der Basis von IMDI-Kollektionen, wobei eine der möglichen Hierarchien mit den Archivbetreibern als kanonisch abgesprochen wird, d.h. zum Beispiel auch zur Spezifizierung von Rechten und Kopiervorgängen verwendet wird.

Im Rahmen von CLARIN[30] wird gegenwärtig ein flexiblerer Metadaten-Standard entwickelt, dessen Konzepte in der offenen auf ISO 12620[31] basierten ISOcat-Registratur[32] enthalten sein müssen, um semantische Interoperabilität in einem durch Komponenten definierten Raum von Metadaten-Beschreibungen zu ermöglichen, die von verschiedenen Experten erzeugt worden sind. Dieser neue CMDI Standard (*component based metadata infrastructure*)[33] wird den alten IMDI Standard ablösen, wenn seine Infrastrukturmodule (Editor, Browser, Suche etc.) fertig gestellt und getestet sind und die volle Kompatibilität zum IMDI- und zu anderen Standards (z.B. OLAC) sichergestellt ist.

29 Vgl. OLAC (2011).
30 Vgl. CLARIN (2011).
31 Vgl. ISO (2004).
32 Vgl. ISOCAT (2011).
33 Vgl. CMDI (2011).

Alle Ressourcen im Online-Archiv müssen mit einer Metadatenbeschreibung versehen sein, wobei beim Hochladen die Korrektheit geprüft wird und regelmäßig Langzeitarchivierungsschritte vorgenommen werden. Allerdings ist der Füllungsgrad der Metadaten sehr unterschiedlich und auch die Verwendung von weniger gut spezifizierten Elementen wie z.B. „Genre" ist sehr verschieden, was genaue Suchen und damit präzise Treffermengen behindern kann. Einen „akzeptierten Metadatenelementsatz" für experimentelle Daten in der Psycholinguistik gibt es noch nicht, denn die Bedürfnisse sind a) bisher nirgendwo durch eine Gruppe formuliert worden und b) derzeit nicht durch einen starken Wunsch nach Austausch motiviert. IMDI erlaubt es jedoch, auch experimentelle Daten zu beschreiben, da jeder Wissenschaftler seine eigenen *key value*-Paare hinzufügen kann. Im MPI wurde wegen des höheren Drucks zur verbesserten Sichtbarkeit der *brain imaging*-Daten in Abstimmung mit den lokalen Wissenschaftlern ein IMDI-basiertes Profil erzeugt, das in Zukunft verwendet wird. Dieser Satz ist bisher noch nicht mit anderen Instituten abgestimmt. Einige Informationen werden dabei automatisch z.B. aus Daten im DICOM[34]-Datenstandard extrahiert, die durch fMRI-Scanner erzeugt werden. Dies bedeutet, dass IMDI und in Zukunft auch CMDI verstärkt zur Beschreibung experimenteller Daten verwendet werden.

5.4 Interne Organisation

Etablierte Langzeitarchivierungsstrategien

Die oben beschriebenen Mechanismen, inklusive der Maßnahmen zur Langzeitarchivierung, operieren nunmehr seit etwa zehn Jahren. Die beteiligten Softwarekomponenten wurden in dieser Zeit systematisch verbessert, so dass wir von einer robusten mehrschichtigen Archivierungsinfrastruktur ausgehen können, die es uns sogar ermöglicht, Daten externer Wissenschaftler zu integrieren. Dort, wo offensichtliche Lücken identifiziert wurden, werden diese in Zusammenarbeit mit anderen Partnern behoben.

34 Vgl. DICOM (2010).

Momentan werden die Qualitätsprozeduren, die bei dem Hochladen auszuführenden Operationen (Konversionen) und die Kommandos zur Replikation in Konfigurationsdateien spezifiziert, wobei alle Operationen auf physikalischer Ebene (Directory-Struktur) festgelegt werden. Im Replixprojekt[35] wurde mit einem *policy rules*-basierten Ansatz experimentiert, indem alle Schritte in Form von deklarativen Regeln ausformuliert und daher sowohl schneller verifizierbar als auch veränderbar werden. Dadurch sollen LZA-Maßnahmen auf logischer, d.h. Kollektions-/Metadatenebene, spezifiziert werden. Alle Schritte sollen über Pre- und Post-Verarbeitungsoperatoren evaluiert werden, so dass z.B. eine Integritätsprüfung mittels der mit dem PID assoziierten *checksum* ermöglicht wird.

Finanzierung des Datenarchivs

Die Finanzierung des Datenarchivs wurde im Wesentlichen durch die MPG und das MPI geleistet, wobei zur Software- und Standardentwicklung immer wieder auch externe Mittel eingeworben wurden. Da derartige Finanzierungen zumeist nur die Entwicklung neuer Methoden erlauben, mussten immer wieder alternative Wege gesucht werden, die Pflege des existierenden Programmcodes abzusichern.

Vier wesentliche Merkmale können genannt werden, bei denen sich neue Finanzierungsmöglichkeiten eröffneten: (1) Mit dem offiziellen Auftrag des MPG Präsidenten an die zwei Rechenzentren, den Max-Planck-Instituten einen Archivierungsdienst anzubieten, konnten ohne Schwierigkeiten seit ca. 2005 externe Kopien der Daten erzeugt werden. (2) Mit dem Start des durch die VolkswagenStiftung finanzierten DOBES-Programmes im Jahre 2000 wurde die Archivierung zum ersten Mal in den Mittelpunkt gerückt. (3) Mit der Finanzierung des CLARIN Infrastrukturprojektes seit 2008 wurde u.a. das Datenmanagement zu einem offiziellen Förderziel der EU und nationalen Regierungen, was z.B. der Entwicklung von Standards und dem Publizieren von Metadaten sehr geholfen hat. (4) Seit September 2010 wurde offiziell eine neue Einheit am MPI mit dem Namen „The Language Archive“[36] gebildet, so dass die

35 Vgl. REPLIX (2011).
36 Vgl. TLA (2011).

Archivierung und das Datenmanagement einen festen Platz im Institut bekommen haben und damit eine Langzeitperspektive gegeben ist.

Kosten des Datenarchivs

Bezüglich der Kosten müssen zwei Vorbemerkungen gemacht werden: (1) Die LZA-Kosten bei der Übernahme und Eingliederung neuer Daten sind z.T. erheblich und nicht vorher abschätzbar, d.h. sie können in einer Kostenübersicht nicht erscheinen und müssen über spezielle Projektfinanzierungen abgewickelt werden. (2) Ein digitales Archiv kann nur dann kosteneffizient arbeiten, wenn es über ein robustes Repository-System und ein Arsenal an Standard-Softwaretechnologien verfügt[37]. Um derartige Software lebensfähig zu halten, bedarf es einer regelmäßigen Pflege des Codes und einer Erweiterung der Funktionalität. Letzteres kann ebenfalls sehr teuer sein und nicht abgeschätzt werden, d.h. die Integration neuer Funktionalität kann ebenfalls lediglich durch gesonderte Projektmittel erfolgen. Im Folgenden werden diese zwei Bereiche (LZA neuer Datenbestände, Funktionelle Software-Erweiterungen) nicht in die Kostenübersicht einbezogen.

Auf dieser Basis können wir die Kosten für das MPI Archiv benennen (siehe folgende Tabelle). (1) Das MPI verwaltet selbst zwei Kopien aller wesentlichen Daten in einem *hierarchical storage management* (HSM) System, was auch den ständigen Zugriff und die Erweiterbarkeit ermöglicht[38]. Das Online-Archiv umfasst dabei gegenwärtig 74 TB. Daneben gibt es weitere zu verwaltende ca. 200 TB, d.h. das Online-Archiv umfasst etwa 25% des Gesamtbestandes. Die wesentlichen Hardware-Komponenten werden alle vier Jahre ausgetauscht, was einen Kostenaufwand von ca. 80.000 Euro pro Jahr ausmacht. (2) Die Kosten für die externen Kopien (lediglich *bitstream* Verwaltung) machen etwa 10.000–20.000

37 Im Bereich der Archivierung digitaler Bestände wird es auch sehr stark auf *economy of scale*-Faktoren ankommen. Dies wird vor allem durch die Erfahrung und das Wissen der involvierten Experten und durch die Verfügbarkeit von robusten Software-Techniken ankommen.

38 Es sollte hier hinzugefügt werden, dass einer der Grundsätze bei der Erweiterung natürlich der ist, dass existierende Ressourcen nicht manipuliert werden, d.h. alle Erweiterungen werden in den Metadaten und weiteren Indizes eingetragen, um die externen Verweise zu erhalten. Alte Versionen werden selbstverständlich bewahrt und sind weiterhin zugreifbar.

Euro aus, schlagen also insgesamt kaum zu Buche.[39] Dies hängt damit zusammen, dass die 74 TB im Rahmen der sowieso in den Rechenzentren zu speichernden PetaBytes vernachlässigbar klein sind und die Datenverwaltung sehr effizient erfolgt.[40]

Kostenfaktor	**Kosten pro Jahr [€] (2011)**	**Kommentar**
IT & Speicher Infrastruktur	80.000	Erneuerung alle 4–8 Jahre
Vier Kopien in Datenzentren	10.000–20.000	Kopien erfolgen vollautomatisch
Lokale Systemverwaltung	40.000	für verschiedene Aktivitäten gemeinsam
Archiv-Verwaltung	80.000	Archivmanager und Hilfskräfte
Repository-Software Pflege	60.000	Grundlegende Pflege des Codes
Pflege der Zugangs-/ Nutzungssoftware	> 120.000	Diese Kosten können sehr leicht stark steigen

Tab. 1: Kostenfaktoren des Archivs des TLA am MPI

Ca. 40.000 Euro sind für einen Systemmanager angesetzt, der zusätzlich benötigt wird, um z.B. alle Archiv- und Web-Applikationen lauffähig zu halten und deren Funktionsweise zu überwachen. Für das Management des Archivs bedarf es eines guten Managers, der einen Überblick über die Daten hat, die datenerzeugenden Nutzer kennt, mit ihnen Verträge schließt und sich mit den Formaten und den Metadaten sehr gut auskennt. Er wird im MPI von einigen Assistenten unterstützt, die fortwährend die Konsistenz prüfen und Verbesserungen durchführen. Um den existierenden Code, der für die Archivierung erforderlich ist (Metadaten, Content

39 Diese Angabe relativiert auch die gegenwärtige Diskussion über Cloud-Lösungen, die effizient sind, aber nur wirklich in der dritten Schicht sinnvoll sind und daher kostenmäßig nicht so stark ins Gewicht fallen.

40 Es sei an dieser Stelle nochmals betont, dass die MPG von Live Archives ausgeht, d.h. es wird kein getrennter, besonders zu behandelnder Datenbereich verwaltet, sondern eher auf mehrere Kopien gesetzt. Dieses Konzept wird nicht von allen Archivaren unterstützt, wird sich aber wohl aus Kostengründen und Gründen der Dynamik wissenschaftlicher Daten als ein akzeptables Modell in der Wissenschaft etablieren.

Management), lediglich zu pflegen, würde im Prinzip ein guter Entwickler ausreichen.[41]

Die Pflege der Zugangssoftware, die über den grundlegenden Zugang über den Metadatenkatalog hinausgeht, kann ein „Fass ohne Boden" sein. Man könnte ganz auf derartige web-basierte Komponenten verzichten, was die Attraktivität des Archivs jedoch stark beeinträchtigen würde, oder aber versuchen, alle Nutzerbedürfnisse, die sehr heterogen sind, befriedigen zu wollen. Wir sehen hier als Minimum zwei Entwickler für die Pflege der existierenden Zugriffskomponenten auf die verschiedenen Datentypen vor.

Personal

Die in den letzten Jahren aufgebaute TLA-Gruppe besteht momentan aus 22 Experten, wobei die meisten über Drittmittelprojekte mit einem weitgefächerten Aufgabengebiet eingebunden sind. Diese Aufgaben reichen von der funktionellen Erweiterung existierender Software-Komponenten bis hin zu einem Projekt, das Audio- und Videoinhalte (semi-) automatisch annotieren und indexieren will. Die Kerngruppe besteht aus sieben Personen (u.a. der Archivmanager), die bezüglich des allgemeinen System- und Netzwerkmanagements von der technischen Gruppe des MPI unterstützt wird. Diese Mischung wird als ausreichend für die langfristige Absicherung der Gruppe und des Archivs angesehen, wobei bezüglich des Archivmanagements *„economy-of-scale-Faktoren"* eingeplant werden können: die Verwaltung eines großen Archivs braucht üblicherweise nicht proportional mehr Verwaltungsaufwand als die eines kleinen.

Die TLA-Kerngruppe ist nunmehr für fünf Jahre abgesichert, wobei allerdings eine Planung für die kommenden 25 Jahre besteht. Von den sieben Kernmitarbeitern haben drei momentan unbefristete Verträge. Die über Drittmittel angeworbenen Mitarbeiter haben zumeist zwei- oder drei-Jahres-Verträge.

41 Hierbei muss berücksichtigt werden, dass bei der Übernahme externer Software auch erhebliche Kosten anfallen würden, da es sich a) nicht um Standard-Software handelt and b) da eine solche Software doch wiederum durch zu pflegende Zusätze auf die speziellen Bedürfnisse abgestimmt werden müsste.

Der Archivierungsservice der GWDG und des RZ Garching ist zeitlich nicht befristet. Alle Daten werden bereits dort repliziert. Wenn der MPI Software-Stack in diesen RZ installiert worden ist, ist auch der Zugriff relativ unabhängig von der TLA-Gruppe. Allerdings bedingt die Pflege der Zugangssoftware die Existenz von Entwicklern.

Externe Dienste

Wie bereits ausgeführt nehmen wir im Moment von den zwei MPG-Rechenzentren Archivierungsdienste an. In Zukunft wird das bisher in eigener Regie gepflegte Handle System zur Verwaltung der über eine Million registrierten Handles auf das „European PID Consortium" (EPIC)[42] übertragen. Dieser EPIC-Service wird von der GWDG, CSC und SARA[43] betrieben. Im Rahmen des geplanten EUDAT Projektes[44] soll untersucht werden, inwieweit z.B. das RZ Jülich, das nationale RZ Finnlands (CSC) und das nationale RZ der Niederlande (SARA) in die Langzeitarchivierung auf niedriger Kostenbasis eingebunden werden können.

5.5 Perspektiven und Visionen

Spezifische Herausforderungen: Abbau von Barrieren

Wie auch viele andere Wissenschaften steht die Forschung im Bereich der Psycholinguistik vor großen Umbrüchen. Diese werden insbesondere durch die technologische Innovation beeinflusst, die die Sensor-, die Kommunikations- und die Informationstechnologie umfassen. Es werden immer mehr Daten generiert und die Fähigkeit, diese Daten zu beherrschen und sie zur Beantwortung von Forschungsfragen zu verwenden, verbessert sich kontinuierlich. Auch allgemein wird in den Geistes- und Sozialwissenschaften („*humanities*") die Forschung zunehmend mehr und mehr datenorientiert („*data driven science*"). In dem aktuellen zunehmend kompetitiven Forschungsszenario wird u.a. der

42 Vgl. EPIC (2011).
43 CSC ist das nationale finnische Rechenzentrum, SARA das nationale niederländische.
44 Vgl. EUDAT (2011).

Wissenschaftler neue wesentliche Publikationen erzeugen können, der es versteht, die existierenden Daten geschickt zu kombinieren und auszuwerten.

Mithin wird der Wunsch größer, bestehende Barrieren bezüglich des Datenzugriffes abzubauen. Es sind dabei insbesondere nicht-technische Barrieren, die einer besseren Nutzung im Wege stehen: (1) Viele Forscher sind noch immer nicht bereit, ihre Daten einem Zentrum zu übergeben, so dass sie für andere in einer geeigneten Form zugänglich sind. Die Gründe sind vielfältig – z.B. fehlendes Vertrauensverhältnis, keine Zeit zur Strukturierung der Daten, keine Zeit zur Erzeugung von Metadaten etc. (2) Viele Forscher haben keine geeignete Möglichkeit, ihre Daten einem von ihnen akzeptierten Zentrum anzuvertrauen, weil es kein systematisches bzw. kostengünstiges Angebot gibt. (3) Nicht in allen Disziplinen ist der Gewinn durch eine „zentrale" Speicherung bereits deutlich. (4) Die Zugriffskonditionen sind momentan noch zu divergent und limitierend für Wissenschaftler.

Spezifische Herausforderungen: Werkzeuge

Natürlich gibt es auch viele technische Gegebenheiten, die zu Barrieren führen. Im Allgemeinen unterstützen die zum Einsatz kommenden Tools keine Workflows, die die Archivierung als einfach zu benutzenden Teil umfassen. Hier sehen wir eines der größten Hemmnisse: es fehlt vielfach an Werkzeugen, die bereits bei der Erstellung von Daten deren Life-Cycle-Management als integralem Bestandteil berücksichtigen. Wir wissen, dass jede Operation im Nachhinein a) nicht gerne ausgeführt wird und b) mehr Zeit kostet.

Wegen der verschiedenen Hemmnisse nehmen wir eine heterogene Situation wahr, die viele Forscher von der datenorientierten Wissenschaft ausschließt. Es müssen dringend breitflächige Maßnahmen getroffen werden, um diese Unterschiede zu beseitigen.

Spezifische Herausforderungen: DOBES und CLARIN

Das DOBES-Programm hat in der Linguistik eine breite und nachhaltige Wirkung gehabt, da es viele der erwähnten Zielstellungen bereits seit 2001 verfolgt und die Strategien optimiert hat. Es hatte zum Ziel, die eminent

bedrohten Sprachen zu dokumentieren und somit einen Teil unseres kulturellen Erbes der heutigen Wissenschaft und zukünftigen Generationen verfügbar zu machen. Gut fünfzig Teams arbeiteten und arbeiten weltweit an individuellen Dokumentationen. Einem Archiv – jetzt das TLA am MPI – wurde die Aufgabe gegeben, die nachhaltige Verfügbarkeit der Daten sicherzustellen.

Die folgenden wichtigen Ergebnisse können genannt werden: Die Materialien sollen vielfältig und interdisziplinär verwendet werden; es wurde eine Vertrauensrelation zwischen den Linguisten und dem Archiv aufgebaut; Linguisten erkennen nun, dass man auch „unfertige" Ressourcen in ein Archiv stellen muss; es besteht Einigkeit, dass die Daten in akzeptierten Formaten und mit Metadaten ausgezeichnet übergeben werden müssen; das Archiv hat gelernt, was es heißt, Daten systematisch und sorgfältig zu behandeln, ohne die Wünsche nach einer kontinuierlichen Anreicherung und einer Online-Zugänglichkeit aufzugeben; etc. Dieses sehr erfolgreiche Projekt hat viele Wissenschaftler im MPI und in anderen Instituten überzeugt, in einer ähnlichen Weise zu arbeiten. Es hat eine weltweite Ausstrahlung und hat sowohl die linguistische als auch die technologische Arbeit wesentlich verändert.

Auch CLARIN soll hier als ein Beispiel genannt werden, das aufgrund neuer innovativer Gedanken einen verbesserten Zugriff auf Sprach-Ressourcen und -Werkzeuge bieten soll. Während DOBES und ähnliche Projekte eine projektspezifische Infrastruktur realisiert haben, ist es die Zielsetzung von CLARIN, eine integrierte, interoperable und persistente Domäne zu bilden, die es den einzelnen Wissenschaftlern erlaubt, möglichst barrierefrei virtuelle Kollektionen und Workflows zu erzeugen, die Daten bzw. Tools von verschiedenen Zentren umfassen. Durch CLARIN ist deutlich geworden, dass eine solche „e-Humanities" erlaubende Infrastruktur nur funktionieren kann, wenn es starke und verteilte Zentren gibt, die sich um das Management der Daten und die Verfügbarkeit von Services kümmern. Dabei spricht auch CLARIN – ganz im Sinne des obigen Drei-Schichten-Modells der HLEG – von einem Ökosystem von Infrastrukturen, an deren Basis gemeinsame horizontale Services (Standards, Protokolle, Indexierung etc., die nicht nur für eine Disziplin gültig sind) z.B. für Netzwerke und Daten stehen.

CLARIN ist noch zu jung, um von Auswirkungen sprechen zu können. Allerdings ist CLARIN genau eine der Initiativen, die eine systematische Lösung auch für das Datenmanagement im Bereich der (Psycho-) Linguistik aufbauen will und bereits wesentliche Pfeiler (PID, Metadaten, AAI[45], Konzept/Schemaregistraturen etc.) implementiert hat.

Spezifische Herausforderungen: Computational Humanities

Wenn über Visionen gesprochen wird, dann müssen die Herausforderungen der *digital humanities* umschrieben werden, die durch mächtige virtuelle Forschungsumgebungen bestimmt werden. Solche Umgebungen werden den Wissenschaftlern Zugriff auf eine große Vielfalt an Daten und Services geben. Der Report der HLEG zu wissenschaftlichen Daten mit dem Titel „Riding the Wave“[46] enthält verschiedene, noch futuristisch anmutende Szenarien. Alle Infrastrukturen, insbesondere solche, die sich mit Daten beschäftigen, müssen sich auf diese Szenarien einstellen. In diesem Zusammenhang wird des Öfteren von der *datenintensiven Wissenschaft* gesprochen, die mit intelligenten Algorithmen in großen virtuell zusammengefügten Datenmengen nach Mustern sucht. Erst vor kurzem konnte im Watson-Experiment der IBM ein „Computer“ menschliche Quiz-Experten schlagen, indem große Mengen Text und Allerweltsdaten entsprechend zusammengefügt wurden.[47] Dies mag als Beispiel für datenintensive Operationen im Bereich der *Humanities* gelten.

Lanzeitarchivierungsvisionen

Die Wissenschaft ist langfristig nur an der Bewahrung und Verwaltung von Daten interessiert, die den wissenschaftlichen Zielsetzungen entsprechen, d.h. für die Forschung verwendet werden. Wenn wir in der Wissenschaft also von Archivierung sprechen, müssen wir nachweisen, dass es sich um ein Online-Zugriffsarchiv handelt, dessen Inhalt auch stetig z.B. durch Annotationen erweitert werden kann. Dieser Aspekt wurde in einer Sitzung

45 Mit der Abkürzung „AAI“ werden die Technologien bezeichnet, die es einem im Internet arbeitenden Nutzer ermöglichen, mittels einem von einer vertrauenswürdigen Institution verliehenen Identifikator auf distribuierte Ressourcen zugreifen zu können.

46 Vgl. High Level Expert Group on Scientific Data (2010).

47 Vgl. IBM (2011).

des beratenden Ausschusses für Rechenanlagen der MPG durch die Aussage auf den Punkt gebracht: „Lassen Sie uns nicht über Archivierung sprechen, sondern über die Langzeitverfügbarkeit von Daten für die Zwecke der Wissenschaft".[48] Die Archivierung muss also Teil des normalen Wissenschaftsbetriebes sein. Nur so wird auch die Bereitschaft gegeben sein, Mittel an zentralen Stellen vorzusehen, um diese Daten auf den zwei Ebenen (*bitstream* Sicherung durch Rechenzentren, Interpretierbarkeit durch Domänen-Spezialisten) für die wissenschaftlichen Ziele verfügbar zu halten und deren kreative Verwendung auch für ursprünglich gar nicht anvisierte Zwecke zu ermöglichen. Bezüglich der *bitstream* Sicherung gilt dabei, dass die jetzt erzeugten Daten nach einem Generationswechsel in der Speichertechnologie nach ca. zehn Jahren nur noch etwa 10% des „normalen" Speichervolumens ausmachen – sie fallen also kostenmäßig nicht mehr ins Gewicht, solange man keine speziellen separaten Strukturen für die Archivierung schaffen muss, die gepflegt werden müssen.

Diesbezüglich ist es zu einer breiten Diskussion zwischen Fachleuten gekommen. Während das Konzept eines Online-Zugriffsarchivs für einen traditionellen Archivar eine Horrorvorstellung ist, erscheint es für einen IT-Experten der einzig machbare Weg für die Zukunft. Ausgangspunkt ist die fundamentale Tatsache, dass sich das „Nie berühren"-Prinzip[49], das für alle physischen Objekte und analogen Datenaufzeichnungen gilt, in ein „Berühre regelmäßig"-Prinzip bei digitalen Daten ändert. Bei digitalen Daten verzeichnen wir keinen Informations- und Qualitätsverlust mehr, wenn wir auf die Daten zugreifen, im Gegenteil: Wir sollten regelmäßig prüfen, ob die Daten noch gültig und verwendbar sind. Das wird natürlich am besten dadurch erreicht, dass tatsächlich Nutzer mit den Daten arbeiten. Hinzu kommt der Aspekt der Kosteneffizienz. Die Filmindustrie in Hollywood hat einen Kostenvergleich gemacht: Analoge Masterkopien wurden unter Berücksichtigung bestimmter Konditionen in

48 Karl Lackner (2010), Kommission zur Langzeit-Archivierung in der MPG (mündliche Mitteilung).

49 Mit diesem Prinzip umschreiben wir die Tatsache, dass jede Berührung eines Objektes oder auch jedes Abspielen einer analogen Bandaufzeichnung immer mit einer Beeinträchtigung des Objektes durch chemische oder physikalische Prozesse einhergeht. Beim Abspielen analoger Bänder zum Beispiel wird die magnetische Schicht durch Reibungseffekte beeinträchtigt. Aus Gründen der Langzeiterhaltung empfiehlt es sich demzufolge, die Bänder möglichst nicht abzuspielen.

ein ehemaliges Salzbergwerk eingelagert und dort so selten wie möglich angerührt. Dies macht in der komplett digitalen Produktionsumgebung keinen Sinn mehr, d.h. die digitalen „Masterkopien“ werden nunmehr auf getrennten Rechnersystemen aufbewahrt – mit der Konsequenz, dass diese Aufbewahrung um den Faktor zwölf teurer ist.[50] Dies ist einer der wesentlichen Gründe für die Filmindustrie, das herkömmliche Modell der Trennung zwischen Zugriffs- und Archivgeschäft aufzugeben. Natürlich erkennen wir, dass die Abhängigkeit von robuster Software sehr groß wird. Eine Abhilfe kann nur das mehrfache Kopieren der Daten und deren Einbettung in verschiedene Softwaresysteme sein, obwohl wir wissen, dass kein einziges System wirklich fehlerfrei ist.

Langfristig brauchen wir eine systematische Archivierungslösung für alle Wissenschaftler an allen Einrichtungen Deutschlands. Die seit 2004 funktionierende Lösung der MPG ist eine Insellösung, die zunächst nur ihren Wissenschaftlern zugute kommt. Es bedarf aber einer grundsätzlichen Herangehensweise, die alle einschließt. Diese muss vor allem auch bestimmen, wie die kostenintensiven LZA-Aspekte gelöst werden sollen. Der Ansatz von Forschungsinfrastrukturen wie z.B. CLARIN und DARIAH[51] in den Humanities und CESSDA[52] in den Sozialwissenschaften scheint angemessen zu sein. Der europäische EUDAT-Projektvorschlag kann hinsichtlich einer systematischen Herangehensweise Maßstäbe setzen, da er ein Netz von Datenzentren mit den domänenbezogenen Forschungsinfrastrukturen zusammenbringt. Dieser Weg sollte auch national unterstützt werden. Die großen Forschungseinrichtungen müssen sich dazu zusammenfinden, um ein Netzwerk von starken Zentren zu etablieren, die die Langzeitarchivierung in obigem Sinne übernehmen. Ohne Frage bedarf es eines Netzwerkes von Zentren, um z.B. das Kopieren der Daten zu organisieren.

Das MPI kann im Rahmen des MPG-Services von einer bewährten Architektur ausgehen, die auch erhebliche Störungen wie z.B. den zeitweiligen Ausfall eines der zwei momentan für die Langzeitarchivierung verwendeten Zentren (z.B. wegen eines schwerwiegenden Fehlers

50 “To store a digital master record of a movie costs about $12.514 a year, versus the $1,059 it costs to keep a conventional film master” (Academy (2008)).

51 Vgl. DARIAH (2011).

52 Vgl. CESSDA (2011c).

in einer Kernsoftwarekomponente) überstehen kann. Grundlage ist allerdings das Repository-System am MPI selbst, durch das die Organisations-, Management- und LZA-Aufgaben bezüglich der Daten geregelt werden. Es wurde bewusst ein relativ einfaches System entwickelt, das nunmehr seit über zehn Jahren im Einsatz ist, modular umgesetzt und schrittweise optimiert wurde und auf jede Kapselung von Daten verzichtet. Das Organisieren der Daten-Migration zu den zwei Zentren ist dabei ein relativ einfach zu handhabender Teilaspekt.

Im Rahmen von CLARIN konnten wir feststellen, dass die Situation am MPI auch europaweit zu den wenigen Ausnahmen gehört. Soweit wir informiert sind, arbeiten nur wenige der gegenwärtigen Abteilungen im Bereich der (Psycho-) Linguistik auf einem derartigen Niveau der Archivierung. Im Allgemeinen herrschen Abteilungslösungen auf der Basis einfacher File-Strukturen und Abteilungsserver vor. Nur einige große Institute haben systematische Strukturen entwickelt, die das Kopieren von Daten und auch den Export von Metadaten-Informationen umfassen. Einige sind nun im CLARIN Verbund dabei, derartige Strukturen zu entwickeln. Der Aufwand, die Daten in ein *state of the art*-Repository-System zu überführen und damit die Zugriffsfähigkeit zu verbessern, ist erheblich und kann kaum von den kleineren Instituten geleistet werden. Diesbezüglich wäre es sehr wünschenswert, in Deutschland über ein verteiltes Kompetenzzentrum zu verfügen, das derartige Systeme im Detail kennt, diese zeiteffizient installiert und bei der Datenmigration helfen kann.

Eine der großen Herausforderungen ist es, die Metadateninfrastrukturen stetig zu verbessern. Dabei gilt es, diese flexibler zu machen, um neue Aspekte einfach in den normalen Arbeitsablauf integrieren zu können. Beispielsweise sollten *provenance* Informationen (die verwendete Grundlagen nennt und die ggf. schrittweise Generierung der Daten nachvollziebar macht) automatisch hinzugefügt werden, immer wenn eine Konvertierung o.Ä. angewendet wird. Gegenwärtig herrschen noch zu beschränkte Systeme vor, die von festen Schemata ausgehen und damit nur schwer an neue Anforderungen anzupassen sind. Dies ist der Grund, warum in CLARIN die *component*-Metadaten-Infrastruktur ausgearbeitet worden ist. Die Basis ist einerseits die Verwendung eines in ISOcat registrierten Vokabulars, das jederzeit erweitert

werden kann, und andererseits die flexible Erstellung von maßgeschneiderten Schemata durch den Benutzer selbst.

Beides, maßgeschneiderte Schemata und die Integration von Metadatensoftwarekomponenten in die normalen Arbeitsvorgänge der Wissenschaftler, sind essentiell, um die Qualität der Metadaten schrittweise zu verbessern. Diese sind wiederum Voraussetzung für die Anwendung von modernen Verfahren wie automatisches Profile-Matching zum effizienten Auffinden geeigneter Werkzeuge. Gegenwärtig sind die mangelnde Qualität der Metadaten und deren zu geringe Granularität eines der großen Hemmnisse für den Einsatz fortgeschrittener Methoden.