# eHumanities - können wir die Komplexität beherrschen?

P. Wittenburg, D. Broeder
MPI für Psycholinguistik
peter.wittenburg@mpi.nl

T. Zastrow, E. Hinrichs
Universität Tübingen

## 1 Einführung

Sprachressourcen und Werkzeuge entstehen in einem distribuierten und nicht reglementierten Prozess. Diese verteilte Entstehung ist zugleich ein Segen als auch ein Fluch. Denn einerseits wird garantiert, dass nur die Ressourcen und Werkzeuge entstehen, die für die direkten wissenschaftlichen Fragestellungen benötigt werden. Andererseits kann man vermuten, dass verschiedene Ressourcen und Werkzeuge für gleiche Aufgabenstellungen dauernd neu erfunden werden, und dass hinsichtlich Datenformaten und Programmierschnittstellen wenig zusammenpasst. Will man z.B. eine virtuelle Kollektion aus verschiedenen Ressourcen von verschiedenen Linguisten oder eine Verarbeitungskette von Sprachtechnologie-Komponenten verschiedener Teams bilden, so werden diese Vorhaben in den allermeisten Fällen an Inkompatibilitäten oder unterschiedlichen Datenstrukturen scheitern.

e-Humanities, abgeleitet aus dem Begriff e-Science, beschreibt nun genau das daten-basierte Paradigma, in dem Daten und Operationen verschiedenen Ursprungs in einfacher Weise durch jeden Wissenschaftler zusammengefügt werden können, in dem die existierenden Beschränkungen und Hürden - wenn nicht vollkommen abgebaut - so doch erheblich reduziert werden, und in dem die Wissenschaftler kollaborative Umgebungen finden, die eine Annotation, eine Verlinkung und Anreicherung der Daten ermöglichen. Ziel des ganzen Unternehmens ist die Hoffnung, dass sich viele der sogenannten kleinen wissenschaftlichen Fragestellungen schneller beantworten lassen und dass sich darüber hinaus die großen Herausforderungen dieser Zeit, die einerseits mit der Instabilität der Gesellschaften und des Lebens und andererseits mit der zunehmenden Verzahnung unserer Steuerungsmechnismen und der Komplexität unserer Modelle zu tun haben, besser angehen lassen.

Basis eines erfolgreichen e-Humanities Szenarios ist demzufolge eine einheitliche, harmonisierende Infrastruktur, die die beschriebene Fragmentierung überwindet, ohne dabei die verteilten, kreativen und durch eine lange Tradition gefestigten Erzeugungsprozesse zu blockieren. Die sich stellende Frage ist, wie man diese zwei Pole miteinander verknüpfen kann. In den vergangenen Jahren haben wir bereits erleben können, wie umfangreiche Projekte in der Wissenschaft und auch große Firmen diese Frage sehr zum Vorteil der Benutzer bzw. der Gesellschafter haben lösen können.

Im Perseus Digital Library Projekt[1] wird z.B. umfangreiches Material von verschiedenen Quellen über die Geschichte, Literatur und Kultur der griechisch-römischen Zeit in einem uniformen technologischen Rahmen zusammengebracht, so dass es ohne Zweifel zu einer großartigen Fundgrube für die Wissenschaft geworden ist. Von großen Firmen wie FEDEX[2] ist bekannt, dass auch sie mit vielen verschiedenen Partnern zusammenarbeiten und nunmehr einen einheitlichen semantischen und syntaktischen Rahmen definiert haben, um die komplexer werdenden Management- und Austausch-Prozesse auch in Zukunft bewältigen zu können. Bei beiden Beispielen handelt es sich um projektartige Vorgehensweisen, die natürlich jedes in sich selbst wiederum Inseln erzeugen. Allerdings haben Lösungen dieser Art den Vorteil, dass der Rahmen für Vereinbarungen beschränkt ist und im Falle der Firma auch von oben durchgesetzt werden kann.

Eine komplementäre Vorgehensweise ist die, die der ESFRI Roadmap Prozess[3] ausgewählt hat und somit der Definition von J. Taylor folgt: "eScience is about global collaboration in key areas of science and the next generation of infrastructures that will enable it". Dieser Interpretation zufolge bedarf es großer Anstrengungen, um zunächst disziplinspezifisch integrierende und interoperable Infrastrukturen aufzubauen. Sie sind nicht mehr projektgetrieben, sondern sollen disziplinweit die erforderlichen Harmonisierungen erreichen. Auch hier haben wir es mit neuen "Inseln" zu tun, die allerdings breiter gefasst sind. Dabei ist das, was als Disziplin umschrieben wird, durchaus sehr heterogen – man kann vermuten, dass der ESFRI Prozess zunächst erst einmal gute Teams arbeiten lassen will, um dann zu schauen, was letztlich erfolgreich ist und was nicht. In einem Gebiet, in dem Neuland beschritten wird, ist dies sicherlich keine schlechte Strategie. Für den Bereich der Humanities werden im Rahmen der ESFRI Roadmap zwei Infrastrukturprojekte gefördert.

Auf der einen Seite steht das CLARIN Projekt[4], das zunächst nur die vergleichsweise überschaubare Schar von Wissenschaftlern aus der Linguistik und der Sprachtechnologie zu gemeinsamen Absprachen bringen will, die sich mit der Erzeugung von Sprachressourcen und deren Analyse mittels State-of-the-Art Werkzeugen befassen und die bereits auf eine lange Tradition von Meetings und Konferenzen verweisen können. Auf der anderen Seite steht das DARIAH Projekt[5], das sich zum Ziel setzt, die Wissenschaftler aus den Geisteswissenschaften zusammenzubringen, eine ungleich größere und viel heterogenere Gruppe. Zu Anfang wurde auch die EROHS Initiative diskutiert, die direkt alle Sozialwissenschaften und die Humanities zusammenbringen wollte – allerdings haben viele Entscheidungsträger sehr schnell begriffen, dass die Heterogenität bezüglich Terminologie, des Einsatzes der IT, der Organisations-Traditionen und vielerlei mehr zu gross ist.

[1] www.perseus.tufts.edu
[2] www.fedex.com
[3] http://ec.europa.eu/research/infrastructures/index_en.cfm?pg=esfri
[4] www.clarin.eu
[5] www.dariah.eu

Allerdings ist auch allen klar, dass es viele überlappende Aktivitäten geben würde, wenn es keinen Austausch zwischen den verschiedenen Forschungs-Infrastrukturen und den sogenannten horizontalen e-Infrastrukturen (GEANT-Netzwerke[6], EGI-Grids[7], DEISA/PRACE[8]-High-Performance Computing, Data Preservation and Access) geben würde. Dies ist der Anlass für die europäische Kommission und ESFRI, einerseits sogenannte Cluster Initiativen auf den Weg zu bringen, in denen die Initiativen aus den verschiedenen Gebieten (Social Sciences and Humanities, Life Sciences, Environmental Sciences, Particle Physics) zusammenarbeiten und überlappende Ansätze identifizieren und gemeinsam lösen. Alle diese Aktivitäten, wie auch die Zusammenarbeiten mit den e-Infrastrukturen dienen natürlich dem Ziel, dem Ideal eines Öko-Systems der Infrastrukturen näher zu kommen. Der Preis, der zu zahlen ist, ist ein erheblich höherer Synchronisierungs-Aufwand und bisher kann keine Aussage über Erfolg oder Misserfolg dieser Ansätze gemacht werden.

Die entscheidende Frage, die sich in Zusammenhang mit dem Zitat von J. Taylor ergibt, ist demzufolge die nach dem "Scope" dieser Initiativen oder in anderen Worten – die nach dem Grad an Komplexität, der für uns sinnvoll beherrschbar ist. Es gibt genug Projekte, die als Beispiele genommen werden können, um zu zeigen, dass Harmonisierungen erfolgreich in die Tat umgesetzt werden können. Je kleiner der Umfang derartiger Projekte ist, umso einfacher lassen sich Vereinbarungen erzielen, desto höher wird allerdings die Fragmentierung außerhalb der Projekte sein. Auf dem anderen Ende der Skala steht die Anforderung, allgemein gültige Standards zu vereinbaren, an die sich jeder halten muss.

Am Beispiel von CLARIN werden wir zeigen, welche Anforderungen es zu lösen galt und gilt und welche Komplexität damit verbunden ist.

## 2 Ebenen der Fragmentierung

In der Vergangenheit wurden sprachwissenschaftliche Tools als Kommandozeilen- oder Desktop-Anwendungen zur Verfügung gestellt. Der Anwender musste diese herunterladen und auf seinem lokalen Computer installieren, was durch fehlende Bibliotheken, Inkompatibilitäten etc. zu vielfältigen Problemen führen konnte. Heute bietet das Web 2.0 mit seinen mannigfaltigen Optionen zur Interaktion Mittel und Wege, Applikationen und Daten online zur Verfügung zu stellen, so dass sie vom Benutzer ohne umfangreiche Vorarbeiten direkt im Browser benutzt werden können:

- Webservices stellen verteilt und über das Internet zugänglich einzelne Funktionen zur Verfügung.

[6] http://www.geant.net/pages/home.aspx
[7] http://web.eu-egi.eu/
[8] http://www.deisa.eu/

- Die von den Webservices bereitgestellten Funktionen werden in Webapplikationen gebündelt und mit einer benutzerfreundlichen Oberfläche versehen.

Die so entstandenen Web 2.0-Applikationen und Webservices stellen ganz neue Paradigmen der Entwicklung und Anwendung sprachwissenschaftlicher Daten und Tools dar[9]. Dabei ist der Übergang von rein datengebundenen Angeboten zu online verfügbaren Anwendungen fliessend: Meist treten sie in Kombination miteinander auf, spezielle Tools erlauben den bequemen Zugriff auf die dahinterstehenden Daten. Es ergibt sich die Situation, dass Datenbestände ohne die auf sie zugeschnittenen Tools nicht mehr zugänglich sind und umgekehrt, hochspezialisierte Tools nur im Kontext der jeweiligen Datenbanken anwendbar sind. Hieraus ergeben sich einige neue Fragestellungen, wie die nach einer möglichen Harmonisierung vorhandener Online-Ressourcen und deren Zugänglichkeit bzw. Sichtbarkeit nach aussen.

Bereits heute sieht sich der Benutzer im Netz mit einer Flut verschiedener Anmelde- und Authentifizierungs-Mechanismen konfrontiert. Dies geht einher mit einer immer grösser werdenden Zahl an Benutzernamen und Passwörtern, ohne die der Benutzer keinen Zugang zu seinen personalisierten Daten oder Anwendungen erhält. Hier bietet sich die Anwendung sogenannter "Single-Sign-On"-Systeme (SSO) an: Der Benutzer authentifiziert sich einmalig bei seinem "Identity Provider". Dieser bestätigt nun im Folgenden die Identität des Benutzers gegenüber einer beliebigen Anzahl von "Service Providern", die die eigentliche Funktionalität zur Verfügung stellen. Somit ist gewährleistet, dass der Benutzer nur einen Account mit Username und Passwort zu verwalten hat und gleichzeitig die Service-Provider keine eigenen Nutzer- und Rechte-Datenbanken führen müssen. Soll ein Service Provider auch die Nutzer anderer Identity Provider akzeptieren, so müssen die beteiligten Identity Provider in einer auf gegenseitigem Vertrauen beruhenden Föderation zusammengeschlossen werden.

Ein Nachteil der SSO-Systeme besteht darin, dass ein Nutzer einen Account bei einem registrierten Identity Provider innerhalb der jeweiligen Föderation benötigt, um auf die Angebote der Service Provider zugreifen zu können. Hat er einen solchen nicht, bleiben ihm alle per SSO geschützten Anwendungen verschlossen. Desweiteren stellt ein SSO-System lediglich Funktionen zur Authentifizierung der jeweiligen User zur Verfügung, hiermit ist kein User- oder sonstiges Rechtemanagement verbunden. Ist ein Benutzer einem Service Provider gegenüber authentfiziert, so ist es immer noch diesem überlassen, welche Rechte/Zugriffe dieser dem User einräumt bzw. welche er unterbindet.

[9] Als Beispiel für eine sprachwissenschaftliche, Web 2.0-basierte Service Oriented Architecture sei hier WebLicht genannt: http://www.d-spin.org/weblicht.shtml

Die CLARIN "Service Provider Federation" ist basiert auf dem SAML2[10] Protokoll-Standard, das von Software-Paketen wie z.B. Shibboleth[11] und SimpleSAMLPHP[12] unterstützt wird und einen europaweiten Zusammenschluss nationaler Identitäts-Föderationen ermöglicht[13]. Es wird somit ein grenz- und disziplinübergreifender Zugriff auf europaweit verteilte Tools und Ressourcen ermöglicht. Ist der Benutzer einmalig in der Identitäts-Föderation authentifiziert, stehen ihm eine Vielzahl Web 2.0-basierter Anwendungen und Webservices zur Verfügung. Damit diese nahtlos ineinander greifen können, gilt es weitere Barrieren zu überwinden.

Um Daten zwischen heterogenen Anwendungen austauschen zu können, müssen diese jeweils die gleichen Datenformate verwenden. Diese Problematik umfasst nicht nur den Bereich der "äusseren", die eigentlichen Daten umfassenden (Transport-) Strukturen wie beispielsweise XML-Formate, sondern auch die Beschaffenheit der eigentlichen linguistischen Ressourcen. Dies wird besonders im sprachübergreifenden, europäischen Kontext deutlich, da unterschiedliche Sprachen unterschiedliche Zeichensätze (Encodings) und damit auch Schriftsysteme und -arten zu ihrer Darstellung benötigen. Im sprachwissenschaftlichen Bereich sind in den letzten Jahren viele Datenformate für die unterschiedlichsten Zwecke entwickelt worden. Kann man die Verwendung von UTF-8 als Zeichensatz-Encoding und XML als strukturierendes Containerformat noch als weitgehend akzeptierten Konsens in der Community ansehen, so beginnen die Inkompatibilitäten spätestens bei der Definition ressourcenspezifischer Datenstrukturen. Diese wurden in der Vergangenheit häufig auf einen bestimmten Anwendungszweck bzw. eine bestimmte Ressource hin zugeschnitten und sind dementsprechend nicht kompatibel mit anderen, durchaus ähnlichen Ressourcen.

Diese untereinander zu harmonisieren und verbindliche und tragfähige Datenstandards zu entwickeln, ist Aufgabe der Arbeitsgruppe TC 37/SC4 "Language Resources Management"[14] innerhalb der "International Organization for Standardization" (ISO)[15]. Deren Arbeit geschieht in enger Abstimmung mit weiteren international etablierten Gremien, beispielsweise der "Text Encoding Initiative" (TEI)[16].

Hat der Benutzer mit Hilfe der oben genannten Anwendungen wissenschaftlich relevante Ergebnisse erzielt, so müssen diese anschliessend publiziert bzw. der Community zugänglich gemacht werden. Dies gilt für "klassische" Publikations-Arten wie Papers, Zeitschriftenbeiträge, Monographien und ähnlichem. Neu hinzu kommt die Notwendigkeit, auch Primärdaten und sekundäre Zwischenergebnisse der Untersuchungen nachhaltig zugänglich zu machen: Ohne diese ist eine Wiederverwendung und z.B. die Nachvollziehbarkeit der publizierten Ergebnisse nicht mehr gewährleistet. Aus dieser Anforderung ergibt sich die Notwendigkeit, auch für grössere Datenmengen nachhaltig zugänglichen und entsprechend abgesicherten Speicherplatz zur Verfügung stellen zu müssen. Dieser muss über die reine Storage-

[10] http://saml.xml.org/
[11] http://shibboleth.internet2.edu/
[12] http://rnd.feide.no/simplesamlphp
[13] Zum aktuellen Stand der CLARIN Service Provider Federation, siehe http://www.clarin.eu/node/2965
[14] http://www.tc37sc4.org/index.php
[15] http://www.iso.org/iso/home.html
[16] http://www.tei-c.org/index.xml

Funktionalität hinausgehend eine Versionierung sowie die Möglichkeit, Dokumente im Team bearbeiten zu können, bereit stellen (Collaboration).

Um Ressourcen, Tools und Publikationen in den einschlägigen Such- und Registratursystemen auffindbar zu machen, müssen diese mit persistenten Identifiern (PID) und beschreibenden Metadaten versehen werden. Auch hierfür bedarf es weitgehender Absprachen, denn bisher passte wenig zusammen: (a) Zumeist gibt es überhaupt keine online verfügbaren Metadaten; (b) Oftmals sind für die beschriebenen Ressourcen und Werkzeuge nur wenige Metadaten-Elemente vorhanden; (c) Andere Metadaten sind mittels nicht-standardisierter Elemente und Vokabularien beschrieben und somit nicht kompatibel mit zentralen Metadaten-Registraturen. Sind die entsprechenden Metadaten-Strukturen allerdings standardisiert und registriert, können sie auf Grundlage etablierter Austausch-Protokolle wie OAI-PMH[17] automatisiert von weiteren Registratursystemen übernommen werden (Harvesting).

Das Projekt CLARIN versucht, die in diesem Kapitel dargestellten Aufgabenstellungen exemplarisch im Bereich der Sprachwissenschaft und Linguistik anzugehen und zu lösen. Dabei wurde immer wieder deutlich, dass auch andere Fachdisziplinen mit den gleichen oder ähnlichen Problemen zu kämpfen haben. Gleichzeitig werden sprachwissenschaftliche Ressourcen und Daten durch die erhöhte Zugänglichkeit und Sichtbarkeit auch für Forscher aus anderen Bereichen der Geistes- und Sozialwissenschaften interessant. Diese tragen neue Fragestellungen und Anforderungen an die sprachwissenschaftlichen Daten und Tools heran, so dass die Komplexität der zugrunde liegenden Infrastruktur weiter zunimmt.

## 3 Lösungsansätze von CLARIN

CLARIN kann bereits auf eine lange Diskussion über Standards und Vereinbarungen verweisen, d.h. in weiten Kreisen gab es bereits eine terminologische Abstimmung, die über die Jahre gewachsen ist und zu einer verschärften Bewusstseinsbildung über Begriffe geführt hat. Als ein Beispiel kann man hier auf die im Jahre 1993 begonnenen EAGLES[18] Diskussionen verweisen, die später im ISLE Projekt[19] weitergeführt wurden. Die 2-jährig stattfindende und im Jahre 1998 begonnene LREC[20] Konferenz hat ebenfalls zu einer Kohärenz beigetragen. Bereits auf der ersten LREC Konferenz gab es einen Workshop zu dem Thema “sharing of resources”.

[17] http://www.openarchives.org/
[18] http://www.ilc.cnr.it/EAGLES/intro.html
[19] http://www.mpi.nl/ISLE/
[20] http://www.lrec-conf.org/lrec2010/

### 3.1 Standardisierung

Es war daher folgerichtig, dass die Community als Ergebnis der vielen Diskussionen über eine verbesserte Interoperabilität im Jahre 2002 das ISO SubCommittee ISO TC37/SC4 zum Thema "Language Ressource Management" gründete – also 6 Jahre vor dem Beginn der CLARIN Initiative. Die Schwerpunkte lagen auf (a) der Etablierung einer Registratur linguistischer Konzept als eine Referenz-Sammlung, auf die jeder Linguist bei seiner Arbeit verweisen kann und (b) der Absprache von Formaten/Strukturen verschiedener wesentlicher Datentypen wie z.B. Annotationstrukturen und Lexika. Damit wurden die Kernaspekte einer jeden Interoperabilität angesprochen: die Verwendung expliziter und allgemein-verwendbarer Syntax und der Verweis auf weithin akzeptierte semantische Definitionen linguistischer Kategorien.

Damit ist jedoch nur ein Grundstein gelegt und nach 8 Jahren Diskussion können wir auch auf erste verwendbare Ergebnisse verweisen. Die auf dem allgemeinen Modell ISO 12620 basierende Konzept-Registratur[21] ist im Einsatz und bereits mit vielen Kategorien gefüllt. Die Tatsache, dass das Modell relativ eingeschränkt ist, ist einerseits seine große Schwäche und andererseits sein großer Vorteil. Die Definition von Kategorien ist zeitlich überschaubar und damit machbar, allerdings kritisieren einige Experten seine Beschränktheit und hätten anstelle dessen z.B. lieber ein flexibles erweiterbares System von RDF Assertions gesehen. Auch hier galt es wiederum, die Frage nach dem zu stellen, was innerhalb eines begrenzten Zeitfensters machbar ist, um dennoch einen riesigen Schritt vorwärts in Richtung auf Interoperabilität machen zu können.

Ebenso wurde mit der Arbeit an flexiblen und daher generischen Formaten begonnen. Das Lexical Markup Framework[22] kann am ehesten mit einem Baukastensystem verglichen werden, da es dem Wissenschaftler erlaubt, jede gewünschte lexikalische Struktur konstruieren zu können. Einige Experten sagen zurecht, dass uns diese Flexibilität bezüglich der Interoperabilität nicht weiterbringen kann, solange semantische Kategorien in jeden möglichen syntaktischen Kontext eingebunden werden können und sich damit einer einheitlichen Interpretation entziehen. Sie argumentieren für mehr strukturelle Beschränkungen, um den Interpretationsaufwand gering zu halten und folgen damit einer Kritik, die bereits gegenüber dem äußerst flexiblen TEI Modell geäußert wurde. Der neue Schritt in ISO ist jedoch die Kontext-unabhängige Definition von Kategorien, die möglichst weitgehend z.B. in Lexika verwendet werden sollen bzw. auf die aus dem lexikalischen Schema heraus referiert werden soll. Nicht immer werden sich dadurch Interpretations-Unterschiede aufgrund der kontextuellen Einbettung vermeiden lassen. Daher können Empfehlungen für bestimmte Komponenten für Sub-Communities durchaus hilfreich sein.

Insgesamt ist der ISO Prozess ein sehr langwieriger und aufwendiger, und zudem ist vollkommen unklar, ob die entwickelten Standards sich in der Community durchsetzen werden. Das bedeutet auch, dass das Entwickeln von Tools, die auf den sich entwickelnden Standards beruhen, ebenfalls langwierig und nicht ohne Risiken ist.

[21] http://www.isocat.org/
[22] http://www.lexicalmarkupframework.org/

### 3.2 Temporäre Lösungen

Aufgrund des innovativen Charakters vieler wissenschaftlicher Projekte und auch des hohen Zeitdruckes, dem sie sich stellen müssen, werden sich immer wieder neue Lösungen z.B. auch für Formate ergeben. Dies konnte z.B. in den vergangenen Monaten im WebLicht[23] Teilprojekt [Hi10] von CLARIN gesehen werden, wo man ein einfaches Format benötigte, das bei Prozessketten mitgeführt wird und alle für die einzelnen Schritte erforderlichen Informationen enthält. Früher haben diese zumeist um Tools oder spezielle Ressourcen herum entstandenen Formate nahezu selbstverständlich zu einer eigenen User Community und zu einer eigenen Dynamik geführt, wenn das Tool breit angenommen wurde. Gute Beispiele sind hier z.B. die CHILDES Initiative[24], die bei vielen Sprach-Akquisitionsforschern und darüberhinaus populär war und ist, und das Toolbox Tool[25], welches für Feldforscher sehr attraktiv ist, da es Annotationen und Lexikon sehr eng zu koppeln gestattet.

Natürlich besteht heute die Sorge, dass sich ein neues Format wiederum verselbständigt und das, obwohl die Community sich gerade im Prozess der Einigung auf Standards befindet. Wir erkennen immer wieder diesen Zwiespalt zwischen einerseits dem Druck, der Innovation nachzugeben, und andererseits der theoretischen Einsicht, dass wir die Vielfalt eingrenzen müssen, um die Fragmentation zu überwinden, die Gesamtkomplexität handhaben zu können und die Nutzer nicht mit einer unüberschaubaren Flut von Konvertoren zu überfordern. Natürlich müssen wir als Community den allgemeinen vorherrschenden Einstellungen entgegenwirken, die sich in verschiedenen Formen ausdrücken können: (1) Wir wissen es sowieso besser; (2) Der Aufwand, um alle Dokumente über einen Standard zu lesen, ist zu hoch; (3) Der Standard entspricht nicht ganz meinen Erwartungen; etc.

Kurzfristig ist das Arbeiten mit neuen ad-hoc Formaten sicherlich vertretbar, insbesondere dann, wenn es sich um Tool-interne Formate handelt und wenn das interne Austauschformat mit relevanten ISO Standards voll kompatibel ist, wie im Fall von WebLicht mit dem ISO/DIS 24612 Standard LAF[26]. In dem Augenblick, in dem Formate eine Infrastruktur bereichern sollen, bedarf es einer sorgfältigen Überprüfung ihrer potentiellen Bedeutung im Vergleich zu den bereits existierenden. Erschwert wird eine derartige Überprüfung dadurch, dass wir die wissenschaftliche Innovation nicht behindern sollten.

### 3.3 Bedeutung generischer Lösungen

Für viele Bereiche gibt es sogenannte generische Lösungen, d.h. Lösungen die auf abstrakten IT-Prinzipien und Überzeugungen beruhen. Um die Problematik zu erläutern, werden wir Beispiele aus dem Bereich Metadaten verwenden.

[23] http://weblicht.sfs.uni-tuebingen.de/weblicht.shtml
[24] http://childes.psy.cmu.edu/
[25] http://www.sil.org/computing/toolbox/
[26] www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=37326

Die im Bibliotheks-Umfeld entstandene DublinCore Initiative[27] hat sich sehr stark um eine Vereinbarung über orthogonale Elemente bemüht, die die Beschreibung aller wesentlichen Eigenschaften von web-basierten Ressourcen erlaubt. Herausgekommen war zunächst ein Satz von 15 Elementen, deren semantische Definitionen notwendigerweise sehr weitgefasst und unpräzise waren. Schon bald wurde der Nachteil derart unspezifischer Definitionen deutlich und es wurden eine Vielzahl von sogenannten Qualified DCMI Elementen hinzugefügt. Obwohl dies sicherlich ein notwendiger Schritt war, konnte dies aus der Sicht wissenschaftlicher Anwendungen nicht zufriedenstellen. Wissenschaftler sehen Metadaten nicht nur als Informationen, mit denen man Ressourcen finden kann, sondern sie sehen sie vor allem als ein Mittel um daten-basierte Forschung betreiben zu können. Sie wollen ihre über viele Jahre gewachsenen Klassifikations-Systeme und Terminologien behalten und sie möchten auch vermehrt mittels automatischer Methoden Selektionen durchführen können, wie z.B. zu einem gegebenen komplexen Ressource-Bündel das passende Werkzeug finden. Folgerichtig entstanden nicht nur in der Linguistik spezifische, auf die Diszplin zugeschnittene Element Sets, die dann natürlich wieder semantisch auf DCMI abgebildet werden mussten, um eine allgemeine Interoperabilität zu erhalten.

Mit CERIF[28] wurde ein recht weitgehendes Modell für Metadaten entwickelt, das auch kontextuelle Informationen z.B. über Personen, Projekte und Publikationen aufnehmen kann. Dieser interessante Ansatz ist momentan für viele Communities nicht primär von Bedeutung, da sie ihre Lösungen zur Beschreibung ihrer Ressourcen und Werkzeuge weitgehend gefunden haben. Man muss hier abwarten, welche Methoden verwendet werden, um z.B. Beschreibungen von Organisationen, Personen und dergleichen zu modellieren und mit denen der Ressourcen zu verbinden. Noch ist nicht klar, ob man aus Komplexitäts-Gründen nicht getrennte Systeme pflegen wird.

Ein weiteres Beispiel ist der UDDI[29] Vorschlag, der in Zusammenhang mit Web Services vom W3C ausgearbeitet wurde. Auch hier gilt, dass die Communities mit ihren spezifischen Ansätzen weit vorangeschritten sind und den generischen UDDI Rahmen als zu komplex und ungeeignet ansehen.

Wie auch andere disziplin-basierte Initiativen hat auch CLARIN diese “generischen” Lösungen nicht übernommen und damit nicht im Sinne einer allgemeinen Interoperabilität gehandelt, sondern die Eigeninteressen, die sich aus der konkreten Forschungspraxis ergeben, höher gewichtet.

[27] http://dublincore.org/
[28] http://cordis.europa.eu/cerif/src/annexes/annex3.htm
[29] http://www.uddi.org/pubs/uddi_v3.htm

## 4 Ungelöste Probleme

Initiativen wie z.B. Europeana[30] oder auch das Virtual Language Observatory (VLO)[31], als Teil von CLARIN, zeigen uns, wo die großen praktischen Probleme der Integration und Interoperabilität liegen. Beide Initiativen befassen sich mit dem Bilden breit angelegter Portale, auf denen interessierte Wissenschaftler, Studierende, Bürger etc. nach digitalen Ressourcen suchen können. Europeana will ein Portal sein, das auf die verschiedensten Kollektionen von Bibliotheken, Archiven und Museen verweisen kann. Es hat damit einen weitaus umfassenderen Anspruch als das VLO Portal, das auf alle möglichen Sprach-Ressourcen und –Werkzeuge verweisen will.

Beide Initiativen haben das technische Problem des Harvestens von Metadaten und deren Zusammenfügen in einem großen Index weitgehendst gelöst – auch im Sinne einer Skalierbarkeit. VLO umfasst nunmehr bereits Informationen über mehr als 270.000 Ressourcen und Tools, die aus vielen Teilen der Welt kommen. Europeana spricht sogar über mehr als 6 Millionen Digitaler Items, die registriert und mittels Verlinkungen erreichbar sind.

Zumindest bei der VLO können wir nicht von einem ausgereiften Portal sprechen, das wissenschaftlich umfassend nutzbar ist. Die Mängel, mit denen wir konfrontiert sind, sind vielfältiger Art:

- Die Granularität der Beschreibungen ist sehr unterschiedlich. Zu einem großen Teil referieren die enthaltenen Metadaten auf Kollektionen (Korpora) und nicht auf einzelne Ressourcen, obwohl gerade in der heutigen Zeit einzelne Objekte aus Kollektionen in verschiedener Weise und in verschiedenen Kontexten wiederverwendet werden. Noch gibt es keinerlei weithin akzeptierter Empfehlungen und natürlich ist es für Ressourcen-Anbieter einfacher, nur eine Beschreibung auf Korpus Niveau zu erzeugen. Dadurch aber ist der VLO Inhalt unbalanciert und der Benutzer kann zu einem Teil nicht zu dem individuellen Objekt gelangen und es mittels Playern betrachten.

- Die Qualität der Beschreibungen ist sehr unterschiedlich. So werden z.B. Organisations-Namen in verschiedener Weise geschrieben, was eine Suche erheblich erschwert. Eine Kuration derartiger Felder in einem dynamischen Umfeld kommt jedoch einer Sysiphus-Arbeit gleich, solange die Quellen nicht verbessert werden. Eine andere Möglichkeit für VLO besteht darin, Synonymlisten zu führen, was einen nicht unerheblichen Aufwand darstellt, denn aufgrund von Statistiken und manuellen Überprüfungen müssen immer wieder Änderungen vorgenommen werden. Noch schwieriger wird der Umgang mit Elementen, bezüglich derer es keine weithin akzeptierten Vokabulare gibt wie z.B. für “Genre”. Hier bedürfte es einer Ontologie im Hintergrund, die Begriffe aufeinander abbildet etc. Auch dies ist ein ziemlicher Aufwand und es ist von Beginn an absehbar, dass diverse Abbildungen nicht von jedem akzeptiert werden.

[30] http://www.europeana.eu/portal/
[31] http://www.europeana.eu/portal/

- Der Ausfüllungsgrad der Beschreibungen ist sehr unterschiedlich. Viele Ressourcen und Tools werden nur minimal beschrieben (Sprache, Repository, Kontaktnamen) und entziehen sich damit jeder genaueren Analyse, die z.B. eine Longitudinalstudie umfasst, bei der es interessant ist zu wissen, welches Alter die aufgenommenen Personen haben.

- Der Fehlergrad der Beschreibungen ist sehr hoch. Natürlich wird der Service Provider mit sehr vielen Tipp-Fehlern konfrontiert.

Moderne Verfahren wie z.B. der Einsatz von Facetted Browsing oder ein automatischer Profil-Vergleich lassen sich nur sinnvoll anwenden, wenn die oben genannten Mängel behoben werden. Leider sind wir von einem Zustand weit entfernt, indem wir avancierte Verfahren aus dem Semantic Web einsetzen können, um z.B. die Qualität und den Füllungsgrad von Metadaten Beschreibungen automatisch auf der Basis von Kontext-Informationen verbessern können. Momentan sind nur traditionelle nicht skalierende Verfahren einsetzbar.

Einige dieser Mängeln lassen sich mittels Scripts unterschiedlicher Art gepaart mit einem großen manuellen Aufwand beheben, was auf Dauer nicht finanzierbar ist. Grundsätzlich lassen sich diese Probleme nur durch eine Änderung des Bewusstseins über die Erfordernisse eines guten Datenmanagement in der Wissenschaft beheben. Bereits bei der Entstehung müssen Tools das Erzeugen von Metadaten unterstützen, wie es z.B. selbstverständlich für einige Elemente bei Kameras ist. Es sind die Repositorien, die bereits bei einem Deposit-Vorgang auf die Qualität und Kuration der Metadaten achten müssen. Das Etablieren einheitlicher Guidelines für eine geeignete Granularität wird viel Zeit und Überzeugung kosten, denn für viele in der Community ist die Wiederverwendbarkeit einzelner Ressourcen für verschiedene Zielsetzungen und mithin verschiedenen Kontexten kein aktuelles Thema, d.h. sie legen keinen Wert auf einen höheren Granularitätslevel zur Identifizierung und Beschreibung von Objekten.

## 5 Zusammenfassung

Der Begriff eHumanities ist vom Begriff eScience abgeleitet und beschreibt eine neue Phase daten-getriebener und computationell unterstützter Forschung, die von allen Möglichkeiten des Web extensiv Gebrauch macht. Wie in dem Zitat von J. Taylor angedeutet, bedingt das eHumanities Szenario einen neuen Typus Infrastruktur, der alle möglichen Barrieren bezüglich der einfachen Verwendung und Kombination von web-basierten Ressourcen und Werkzeugen überwindet.

Das Überwinden von Barrieren lässt sich in Projekt-basierten Ansätzen zielstrebig erreichen, wenn ein effizientes Projekt-Management gegeben ist, da diese im allgemeinen einen beschränkten Deckungsbereich haben. In Infrastrukturen, die notwendigerweise breiter angelegt sind, müssen andere Wege beschritten werden. In CLARIN wird auf breite Absprachen gesetzt, die in ISO Standardisierungsbemühungen münden. Dabei wird zumeist nicht auf Lösungsvorschläge mit einem disziplin-unabhängigen Gültigkeitsanspruch gesetzt, sondern auf Lösungen, die sich aus der Arbeit in der Community ergeben. Damit wird vor allem auch den Traditionen und der Innovationsdynamik in der Disziplin Rechnung getragen.

Der Aufwand, um derartige Lösungen bereits innerhalb einer Community zu definieren und zu breiter Akzeptanz zu führen ist enorm, und der Weg ist mit Risiken behaftet. Auch diese Lösungen müssen in gewisser Weise "generisch" sein, da sie für eine ganze Community gelten sollen und von den verschiedenen Anwendungsszenarien von Teil-Communities abstrahieren müssen. Konkrete Beispiele wie das Virtual Language Observatory zeigen, dass mit jeder Integration ein enormer Aufwand verbunden ist, der nur zu einem kleinen Teil technisch gelöst werden kann, aber langfristige Community-Prozesse umfasst.

Die entscheidende Frage, mit der wir bei dem Überwinden von Barrieren konfrontiert sind, ist die nach dem Abdeckungsbereich, den wir erzielen wollen, und damit nach der Komplexität, die wir innerhalb eines beschränkten Zeitfensters zu bewältigen in der Lage sind. Der Bereich eHumanities umfasst derart viele Disziplinen, die alle mit ihren eigenen Terminologien und Traditionen behaftet sind, dass alles umfassende Lösungen zunächst nur sehr allgemeiner Art sein können, wie es z.B. durch XML und UNICODE beschrieben wird. Es wird größere Zeiträume bedürfen, um gemeinsame Lösungsansätze für den gesamten Bereich der Humanities umzusetzen, die über basale Aspekte hinausgehen und direkt mit den Inhalten zu tun haben – der Form der Repräsentationen und der Codierungen der relevanten Phänomene. Diese Aspekte können von Initiativen, die Projekt-bezogen definiert sind und eine gewisse Breite haben, oder durch Infrastrukturen mit einem breiteren Deckungsgrad vorangetrieben werden.

Im Sinne eines Öko-Systems von Infrastrukturen, das aus Kostengründen letztlich unumgänglich sein wird, bedarf es dann vorsichtiger Versuche, die Erkenntnisse aus den verschiedenen Disziplinen immer wieder zusammenzufassen und nach abstrakteren Lösungen zu schauen. Da diese Abstraktion oft nur in enger Zusammenarbeit mit den wirklichen Experten geleistet werden kann, sind uns auf diesem Weg zeitliche Beschränkungen auferlegt, denn diese Experten sind auch genau diejenigen, die die Integrations- und Interoperabilitäts-Anstrengungen in den Disziplinen vorantreiben.

## Referenzen

[Hi10] Hinrichs,M., Zastrow, T, Hinrichs, T.: WebLicht: Web-based LRT Services in a Distributed eScience Infrastructure, in: Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), Valetta, Malta. Published by European Language Resources Association (ELRA), 2010