Standards in der geisteswissenschaftlichen Textdatenverarbeitung

Über die Zukunftssicherung von Sprachdaten

Vortrag beim Workshop der Union der deutschen Wissenschaftsakademien

Mannheim, den 6. Oktober 2008

C. M. Sperberg-McQueen

---

1. Einleitung

Die Überschrift dieses Vortrags lautet: “Standards in der geisteswissenschaftlichen Textdatenverarbeitung”. Es wäre verständlich, würde man verblüfft auf diese Überschrift gucken und sich fragen “Standards? Normierung? In den *Geisteswissenschaften*? Kann das als Witz gemeint sein?”[1]

Denn Normieren heißt, regelkonform machen. Und Dilthey hat uns ja gelehrt, daß die Geisteswissenschaften sich mit dem Verstehen, nicht mit den Regelmäßigkeiten, der Dinge befassen. Und das Verstehen richtet sein Augenmerk zum großen Teil auf die Eigenart der Sache, auf die Unregelmäßigkeit, auf den Verstoß gegen die Regel. Die Geisteswissenschaft interessiert sich naturgemäß für den *merkmalhaften* Fall. (Und wenn man gelegentlich ausgerechnet das Merkmallose untersuchen will, pflegt man, es durch eine Art Verfremdung zum *Merkmalhaften* zu machen, um es schärfer sehen zu können.)

Die Kunsthistorikerin Jocelyn Penny Small hat einmal bemerkt, wer ein Datenbanksystem einrichtet, kommt in Versuchung, kommt ja gerade unter Druck, die Daten schön regelmäßig einzugeben, alles klar, alles sauber, alles aufgeputzt. So läuft man aber Gefahr, die Daten leicht oder schwerwiegend zu verfälschen, denn est ist sehr leicht, beim Normalisieren die Eigenart der Daten fahren zu lassen. Man kommt sich als Geisteswissenschaftler angesichts der kalten Logik des Datenbanksystems mit seinen Schemata und seinen streng rechteckigen Datentabellen so unordentlich, so desorganisiert, vor, man möchte anläßlich der Digitalisierung aber das wilde Durcheinander in den Daten doch ein bißchen aufräumen.

Aber, wie Small uns ganz zurecht mahnt, est ist gar nicht unsere Sache, die Unordnung der geschichtlichen Überlieferung und der geisteswissenschaftlichen Daten allgemein zu zähmen. Unsre Sache ist es, mittels der kalten Logik der Datenbanksysteme die wirbelnde Unordnung des faktisch Gegebenen möglichst detailliert und wirklichkeitstreu nachzubilden, und das Durcheinander somit zu bewahren. “Preserve the mess,” schreibt Small.

Wie kann man das, wenn man nach Normen und Standards, und nicht nach den Eigenarten der Daten arbeiten muß?

Manche behaupten sogar, das für die geisteswisssenschaftliche Datenverarbeitung Notwendige sei der gängigen kommerziell ausgerichteten Praxis der Informationstechnologie (und der real existierenden Normenorganisationen [standards development organizations]) so fremd, daß wir als Geisteswissenschaftler mehr Schaden als Nutzen daraus ziehen. Manfred Thaller, der als exponierter Befürworter dieses Gedankenganges gelten darf, hat sich dann konsequenterweise schon vor Jahrzehnten angeschickt, eine neue rein geisteswissenschaftliche historische Informatik zu begründen, die mit der herkömmlichen Informatik und Datenbanksystemen kaum etwas mehr als den Namen Informatik gemeinsam haben soll.

Bezeichnet also der Titel des Vortrags also eine Art mythisches Ungeheuer? Hätte er genausogut *Anwendungen des Einhornhaares in der Halbleiteranfertigung* heißen können?

[Pause]

Sie wissen schon genau, daß dem nicht so ist.

Zum einen läßt die Gattung des Abendvortrags in einer wissenschaftlichen Institution wie dem Institut für deutsche Sprache eigentlich keine Dreiminutenvorträge zu. Kein Mensch wird erwartet haben, daß der Vortag etwa so laute:

> Standards in der geisteswissenschaftlichen Textverarbeitung?
>
> Es gibt keine.
>
> Das wärs. Tschüß, Ihr Lieben!

Allein die Tatsache, daß der Empfang nachher noch nicht fertig vorbereitet ist, würde das unmöglich machen.

Den Zuhörern, die die Macht der gattungsbedingten Erwartungen schon gut kennen, wird es also keine Überraschung bereiten, wenn ich sage, Nein, das ist kein Witz. Die Standards haben alles mit der Aufgabe und den Möglichkeiten der geisteswisseschsaftlichen Textverarbeitung zu tun. Aber das Verhältnis bedarf der weiteren Erörterung.

Die Geisteswissenschaften interessieren sich wie gesagt besonders für die Ausnahme, für den merkmalhaften Fall. Aber die Ausnahme kann erst in Hinblick auf eine Regel als Ausnahme erkannt werden. Das Merkmalhafte eines jeden Phänomens wird erst im Gegensatz zu anderen notwendigerweise merkmallose Eigenschaften der Sache überhaupt deutlich. Die Regelmäßigkeit — der Standard — ist der Eigenart der Sache und damit ihrem Verstehen untrennbar verbunden.

Es wäre ein Irrtum, anzunehmen, der einzige Sinn und Zweck der Standards sei, genau vorzuschreiben, wie ein Text, ein Wörterbucheintrag, ein Lexikonartikel auszusehen habe, und daß die Standards Ausdruck der Forderung seien, alles müsse gleichgeschaltet werden. Genau genommen sind alle Standards und Normen nur Hilfsmittel zur Klassifikation der Sachen: es sind eben Definitionen. Sie erlauben es uns, die Sachen als der Norm konform oder nicht konform zu beschreiben. In den meisten kommerziellen Anwendungen verlangt man in der Tat, daß ein Gerät, oder ein Datenstrom, der jeweiligen Norm konform sei; diese Bewertung liegt aber in der Natur der Anwendung, nicht im Wesen der Standards selbst.

Ich möchte heute abend zuerst die besonderen Anforderungen erwähnen, die man an Standards stellen muß, wenn sie für geisteswissenschaftliche Anwendungen taugen sollen. Dann möchte ich einige Anforderungen und Schwierigkeiten beschreiben, die aus dem Bestreben geisteswissenschaftlicher Projekte erwachsen, um Lexika, Wörterbücher, digitalisierte Textausgaben, Korpora, und andere Grundlagenwerke, die wir erstellen, der Nachwelt nutzbar zu machen. Denn was man für die Nachwelt baut, ist eine Art Botschaft oder Nachricht an die Zukunft. Botschaften und Nachrichten kommen aber nicht immer bei dem Empfänger an. Was können und müssen wir machen, um die Wahrscheinlichkeit des Erfolgs zu vergrößern? Zwischendurch werde ich einzelne Standards erwähnen, die besonders relevant sind für die geisteswissenschaftliche Arbeit.

2. Anforderungen an Standards für geisteswissenchaftliche Arbeit

Die geisteswissenschaftliche Arbeit stellt eine Reihe von Anforderungen an Standards, und wenn man für die eigene Arbeit sich einen Standard auswählen muß — eine der schönsten Attribute der heutigen Standards, ist die Tatsache, daß es derer so viele gibt, man hat eine reiche Wahl und darf bzw. muß sich eine eigene Palette davon zusammenstellen — wenn man einen Standard wählen muß, wie beurteilt man, ob er für die geisteswissenschaftliche Arbeit geeignet sei?

Es spielt dabei die Tatsache ein Rolle, daß *normiert wird, was man schon gut versteht*. Das heißt, das Geläufige wird zuerst standardisiert, das Ungewöhnliche erst später oder gar nicht. Die Gegenstände geisteswissenschaftlicher Aufmerksamkeit widerstreben oft die Normierung. Wir untersuchen sie ja eben deshelb, weil man sie nicht vollkommen versteht. Es kann schwierig sein, geeignete Standards zu finden. Was verlangen wir als Geisteswissenschaftler von den Standards?

2.1. Breite / Vollständigkeit

Erstens, die Vollständigkeit, oder wenigstens die Breite.

Der Standard soll für das betreffende Gebiet so vollständig sein, wie möglich. Das gilt für jede Ebene der Datenrepräsentation, vom Zeichensatz bis hin zur Text- bzw. Datenstruktur und zur semantischen Ebene: jede Ebene der Datenrepräsentation muß eine bestimmte Vielfalt an Daten darstellen können.

Jede Norm für Zeichensätze z.B. stellt einen gewissen Vorrat an Zeichen bereit. Braucht man für die Arbeit nur eine Untermenge dieses Vorrats, so ist die Norm für diesen Gebrauch vollständig genug.

In dieser Hinsicht muß der sogenannte Universalzeichensatz (UCS, Universal character set) erwähnt werden. Dieser Universalzeichensatz wird von zwei parallel entwickelten Normen definiert: Unicode (von dem Unicode-Consortium verabschiedet) und der internationalen Norm ISO 10646. Schon die erste Version von diesen Standards hatte einen Zeichensatz für praktisch alle standardisierten Schriftsysteme der Welt bereitgestellt. Inzwischen sind durch weitere Forschung und durch die große Verdienste von vieler Philologen, darunter des Thesaurus Linguae Latinae und des Thesaurus Linguae Graecae, tausende von neuen Zeichen aufgenommen worden, die in historischen Schriftsystemen, und in Dutzenden von Minderheitsschriften, gebraucht werden. Die historische Formen der griechischen und lateinischen Schriften sind inzwischen einigermassen gut vertreten, und in der nächsten Version von Unicode erwartet man, daß auch die notwendigen papyrologischen Zeichen aufgenommen werden.

Für viele Projekte darf die Zeichensatzmisere, an die viele frühere Unternehmen der geisteswissenschaftlichen Datenverarbeitung gelitten haben, und die mancher hier noch gut in Erinnerung hat, als historisches Kuriosum gelten. Man wird abends nach einem Glas Wein vielleicht alte Geschichten erzählen, wie man den Großrechner des Rechenzentrums ausgetrickst hat, um die notwendigen Zeichenformen auszudrucken, aber für viele von uns gehören solche Bemühungen nicht mehr zur Tagesarbeit. Dafür können wir den Philologen dankbar sein, die es möglich gemacht haben, den Universalzeichensatz so zu erweitern.

Ähnlich bietet uns jede Anwendung von SGML oder XML (jeder XML-Auszeichnungssprache, oder XML-Wortschatz wie man in Anlehnung an das Englische “XML vocabulary” sagen könnte) eine bestimmte Anzahl von Terminis an, mit denen man eine bestimmte Anzahl der Textstrukturen oder Textphänomene unterscheiden und auszeichnen kann. Wenn alle Textstrukturen, für die man sich interessiert, mit diesen Terminis ausgezeichnet werden können, so ist die Auszeichnungssprache vollständig genug.

Wer z.B. technische Handbücher mittels DocBook auszeichnen will, wird meistens alles in Docbook finden, was er braucht. Gedichtsammlungen gegenüber weist aber Docbook peinliche Lücken auf: für diesen Zweck kann Docbook nicht als vollständig gelten.

2.2. Erweiterbarkeit / Extensibilität

Zweitens, die Erweiterbarkeit.

Bezeichnend für die geisteswissenschaftliche Arbeit ist, daß die absolute Vollständigkeit ein Chimäre ist, manchmal aus rein praktischen Gründen, aber oft auch aus theoretischen.

Das sogenannte Universal Character Set (UCS), der universeller Zeichensatz von ISO 10646 und Unicode nimmt sich vor, einen Zeichenvorrat für alle Schriftsysteme aller Sprachen bereitzustellen, einschließlich der archaischen Zeichen alter Schriftsysteme. Und wie schon gesagt hat man hier viel geleistet. Es ist aber leicht

einzusehen, daß dies als Zielsetzung bewundernswert, als Beschreibung eines real existierenden und jetzt auf immer geschlossenen Zeichensatzes aber fast undenkbar ist. Denn wir können jederzeit neue Schriftsysteme entdecken oder entschlüsseln. Es ist ja nicht so lange her, daß man durch die Entschlüsseling der Schrift *Linear B* jede Menge neuer Einsichten in die antike Kultur gewonnen hat.

Und es gibt in der Tat eine Reihe von Schriften, die auch in der nächsten Version von Unicode voraussichtlich nicht berücksichtigt werden.

Es gibt auch Zeichen, die man gelegentlich als Zeichen braucht, die aber kaum in einen Standard with Unicode oder ISO 10646 passen. Wilhelm Schlegel (oder was es Friedrich?) benutzt in seinen Tagebüchern oft ein Zeichen, das wie eine Parabol (oder die Hälfte einer Hyperbol) aussieht, mit einem Pünktchen am Fokus. Das Zeichen bezeichnet offensichtlich die Unendlichkeit, oder das Unendliche. Es wäre praktisch undenkbar, Schlegels Tagebücher ohne dieses Zeichen herauszugeben oder zu digitalisieren. Aber es wäre ebenso undenkbar, dieses Zeichen in das Universal Character Set eingliedern zu wollen: es gehört nicht dahin, denn es ist nicht Teil eines kulturell getragenen Schriftsystems, sondern ist eine rein private Abkürzung.

Oder man denke an die Abkürzungen der antiken und mittelalterlichen Handschriften. Wer Handschriften mit paläographischer Genauigkeit nachschreibt, wie etwa in den Handschriftenausgaben der arnamagnaeanischen Institute in Kopenhagen und Reykjavík, wird bestimmen müssen, welche handschriftlichen Unterschiede hier als graphematisch anzusehen sind, und welche als bloß graphetische Unterschiede nicht in die Transkription gehören. Dazu braucht man einen großen Vorrat an Sonderzeichen. Aber will man jede Schreibart mit in Unicode aufnehmen, die in Capellis Katalog der Handschriftenabkürzungen vorkommt?

Es ist also wichtig, daß man den Standard erweitern kann, wenn es unbedingt nötig ist. Ich hätte mich zum Beispiel geweigert, den Universalzeichensatz des Unicodes und der ISO 10646 als den einzigen zulässigen Zeichensatz von XML-Dokumenten zu akzeptieren, wenn den beiden Standards nicht die sogenannte Private Use Area eingegliedert wäre, mittels derer man den Standardzeichensatz erweitern kann.

Die Zeichen der Private Use Area sind nicht standardisiert, und bedürfen der Dokumentation und der Sonderbehandlung, dürfen also nur dann benutzt werden, wenn es dringend notwendig ist. Aber es ist eben manchmal dringend notwendig, einen Standard zu erweitern. Es ist gut, wenn der Standard diese Möglichkeit von vornherein anerkennt und dafür Regel anbietet.

In der Praxis kann man oft die Lücken eines Standards wenigstens zum Teil auf höheren Ebenen wieder füllen. Wenn man z.B. wie manche Sachkundige womöglich vermeiden will, Zeichen im Private Use Area von Unicode direkt zu benutzen, kann man die notwendigen Zeichen des Textes durch Auszeichnungen in einer Auszeichnungssprache darstellen, etwa mit einem XML-element namens *Zeichen* oder *char*, wie es in der Auszeichnungssprache MathML oder in den Richtlinien der Text Encoding Initiative beschrieben wird. Diese Ausweichmöglichkeit, die Flucht auf die höhere Ebene, hat ihre Nachteile (die unterschiedliche Darstellung der Zeichen sticht z.B. ins Auge), darf aber nicht vernachlässigt werden.

Bei der Auszeichnung der Textstruktur stößt man etwas schneller an die Grenzen der vorhandenen Standards, die XML-basierte Auszeichnungssprachen definieren. Die Richtlinien der TEI behandeln eine reiche Vielfalt an Textstrukturen, die man in anderen öffentlich zugänglichen Auszeichnungssprachen (wie etwa Docbook oder HTML) vermißt. Ihr Ziel war es ja, die Auszeichnung von Texten zu ermöglichen, die für die geisteswissenschaftliche Forschung interessant sein könnten. Das heißt aber, daß sie beliebige Texte, beliebige Gattungen, in beliebiger Sprache, für beliebige wissenschaftliche Interessen, behandeln müssen. Manches ist da verhältnismäßig gut ausgebaut, aber Lücken gibt es da genug.

Es ist daher Grundprinzip der TEI-Guidelines, daß man die Auszeichnungssprache erweitern und umdefinieren kann, ohne deswegen den Richtlinien nichtkonform zu sein.

- Viele Elemente können mit Hilfe des Attributs *type* ohne großen Aufwand spezialisiert werden.
- Neue Elemente können der Sprache hinzugefügt werden.
- Fast alle vordefinierte Elemente der Auszeichnungssprache dürfen weggelassen werden.

- Die Grundstruktur existierender Elemente darf geändert werden.

Mit diesen Mitteln, versucht man in der TEI die Ausnahmefreundlichkeit der Richtlinien und die Toleranz für Spezialfälle zu erhöhen.

Man kann natürlich noch mehr machen.

Der Sinn des Namens *Extensible Markup Language* (‘erweiterbare Auszeichnungssprache’) besteht darin, daß man mit XML beliebige Textstrukturen auszeichnen kann, eben weil man eigene XML-Tags, und damit ganze eigene Auszeichnungssprachen, definieren kann.

Bei der Entwicklung von XML und vordem von SGML hat man die Erweiterbarkeit der Auszeichnungssprache dadurch garantiert, daß man gar keine Auszeichnungssprache definiert hat (der Name ist historisch begründet, ist aber irreführend), sondern nur eine Metasprache zur Definition von Auszeichnungssprachen definiert hat. Hier sieht man eine wichtige Methode der Erweiterbarkeit, die ich die Flucht in die Metasprache, oder die Flucht in die Abstraktion nenne. Eben weil man über die richtige Auszeichnungssprache uneinig ist, einigt man sich darauf, daß ein jeder eine eigene Auszeichnungssprache muß definieren können.

2.3. Toleranz für unvollständigkeit

Es ist auch wichtig, daß der Standard es dem Wissenschaftler freiläßt, Informationen unvollständig anzugeben, denn manchmal weiß man eben nicht alles über die Gegenstände der Untersuchung.

Dies kann eine heikle Sache werden, denn die automatische Nachprüfung der Daten um Vollständigkeit ist ein wichtiger Hilfsmittel, um Fehler bei der Dateneingabe und bei der Verarbeitung und Wiederspeicherung der Daten zu entdecken.

3. Botschaften an die Zukunft

Soweit zu den Anforderungen, die wir als Geisteswissenschaftler an die Standards stellen müssen.

Es gibt auch einige Anforderungen, die an die geisteswissenschaftlichen Projekte gestellt werden, eben weil unsere Projekte als Ziel haben, Werkzeuge für die Nachwelt vorzubereiten.

3.1. Lebensdauer der Technik und der Daten

Die Computertechnik entwickelt sich nach wie vor im rasenden Tempo. Viele Organisationen rechnen damit, daß alle Hardwares jede zwei bis drei Jahre ersetzt werden; manche versuchen, die Maschinen im Durchschnitt fünf Jahre lange in Betrieb zu halten — Maschinen im Alter von fünf Jahren neigen aber zu unerwarteten und katastrophalen Pannen. Wer denkt schon daran, Rechner im Dreißig- oder Zwanzigjahrenzyklus, oder selbst im Zehnjahrenzyklus, zu erneuern?

Die Softwares bleiben oft etwas länger leistungsfähig. Die verschiedenen Versionen einer Software ersetzen sich vielleicht regelmäßig, aber es gibt durchaus Softwares, die jahrzehntelang zugänglich sind, oder sogar jahrzehntelang den Markt beherrschen. Jahrzehntelang, aber man kann noch nicht sagen: über viele Jahrzehnte hinweg. (Tustep, jetzt bald im Alter von dreißig Jahren, ist ja in der Softwarewelt ein Greis.)

Aber die Daten, um die wir uns kümmern, leben viel länger. Selbst kommerzielle Daten bleibe viel länger aktuell, als Hardware und Software. Der Vertrag mit dem Telefondienst läuft fünf, oder zehn, oder fünfzig Jahre. Die ärztlichen Unterlagen sollten idealerweise uns das ganze Leben lang zugreifbar bleiben, oder noch länger, denn die [medical history] unserer Eltern können durchaus bei der Diagnose von Belang sein. Für Steuerzwecke haben oft Immobilien eine Abschreibungsdauer von etwa dreißig Jahren, werden aber viel länger in Stand gehalten.

Und das sind nur die einfachsten rein kommerziellen Beispiele. Wer die menschlichen Sprachen,

Literaturen, und Kulturen als Forschungsgegenstand hat, pflegt, Daten im Alter von fünf bis zwanzig Jahrhunderte zu verarbeiten. Selbst die Vorbereitung eines Wörterbuches nimmt in manchen Fällen mehr Zeit in Anspruch, als die Geschichte der elektronischen Rechentechnik aufzuweisen hat.

Wenn wir bei jeder neuen Maschine, bei jeder Aneignung einer neuen Software, alle Daten wieder neu erstellen müßten, so kämen wir nie über die Anfangsstadien unserer Projekte hinaus.

Wenn wir unsere Daten und Forschungsresultate, und die Werkzeuge, die wir uns bauen — Korpora, Ausgaben, Wörterbücher, Lexika — nicht nur fü den eigenen Gebrauch erhalten wollen, aber noch über das eigene Leben hinaus der Nachwelt zur Verfügung stellen wollen, sollten wir uns ernsthaft darüber Gedanken machen, wie wir die Daten in einer dauerhaften und nachhaltbaren Form [Format?] speichern können, aus der wir wenns notwendig ist auch anwendungsspezifischen Formen ableiten können. Wie macht man das? Wie sichert man die Daten für die Zukunft?

Man kann diese Fragen vielleicht am besten beantworten, wenn man hier die Botschaften an die Zukunft im Licht eines allgemeinen Kommunikationsmodells betrachtet, das vom dem großen Strukturalisten Roman Jakobson 1960 vorgeschlagen wurde.

3.2. Erfolg und Fehlschlag

Als Zukunftsicherung der Daten bezeichne ich das Bemühen, sicherzustellen (oder wenigstens die Wahrscheinlichkeit zu erhöhen, daß die Daten, die wir mühevoll und teuer erstellen, für die Nachwelt nutzbar seien. Im Grunde genommen gleicht dies dem Problem des Datenaustausches mit anderen Projekten oder Organisationen: Die Zunkunft ist ein fremdes Land, man macht dort vieles anders. In mancher Hinsicht unterscheiden sich diese zwei Probleme: der Empfänger der Daten z.B. sollen wir oft selbst sein, und nicht andere, aber im wesentlich kennen wir auch dann den Empfänger auf eine sehr unvollkomme Weise. Wir werden bis dahin einiges gelernt, oder vergessen, unsere Institutionen werden vielleicht neue Richtungen eingeschlagen haben. Die Lage kann sich drastisch geändert haben. (Das kommt nicht ganz selten vor, wenn man neue oder neuartige Daten und Softwares bereitstellt: lange, bevor man den Originalplan zu Ende gefürht hat, können die ersten Teillieferungen die Lage grundliegend ändern.)

Und selbst wenn der Empfänger die selbe Ziele hat, die wir erwarten, kennen wir doch nicht die technische Umgebung, in der er arbeitet, wir wissen nicht, was für Hardware und Software zur Verfügung stehen wird, unsere Daten auszunutzen. Wir wissen es nicht und können es nicht erfahren, denn die Kommunikation mit der Nachwelt ist eine Art Einbahnstraße, ein Write-Only Datenträger. Der Empfänger kann nicht unsere Botschaft empfangen und inspizieren, und dann uns eine Rückmeldung schicken “Botschaft nicht verstanden. Bitte nochmals senden.” Wer Botschaften an die Zukunft sendet, bekommt keine Rückmeldungen. Es ist eine Art Flaschenpost, oder als ob man Botschaften an Spionen senden würde, die so geheim arbeiten müssen daß sie sich keine Rückmeldung leisten können.

In dieser Lage müssen wir alle Fehlermöglichkeiten des Vorgangs voraussehen und vermeiden. Alle Bruchstellen der Verbindung zwischen Sender und Empfänger müssen untersucht werden, um mögliche Pannen zu vermeiden. Dazu diene uns das Kommunikationsmodell von Roman Jakobson.

Ein Mitteilung, so Jakobson, hat offensichtlich einen Sender, und einen Empfänger

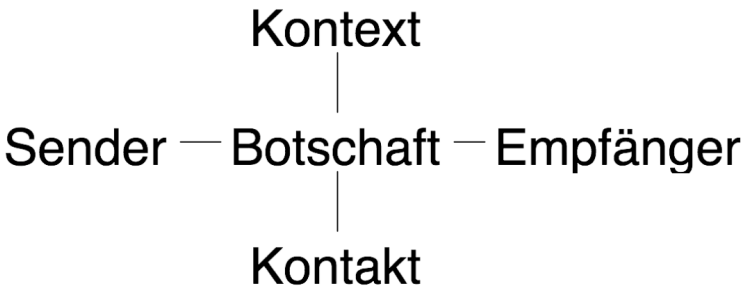
Sender — Botschaft — Empfänger

Die Mitteilung kann vor allem dazu bestimmt sein, über den Sender Auskunft zu geben. Jakobson schreibt demgemäß eine *Ausdrucksfunktion* oder emotive Funktion der Sprache und der Mitteilung zu. Oder sie kann dem Empfänger einen Auftrag oder einen Befehl geben: das ist die Aufforderungsfunktion oder die konative Funktion.

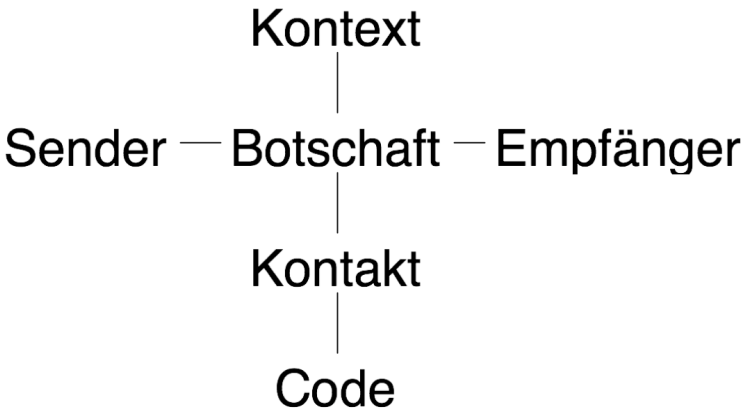
Meistens aber handelt es sich um eine Mitteilung, die sich auf eine Sachlage in der Welt (oder im *Kontext*)

bezieht; eine so ausgerichte Mitteilung übt die referentielle Funktion der Sprache aus.

Die Kommunikation findet aber nur dann statt, wenn der Sender und der Empfänger körperlich in Kontakt stehen. Die Mitteilung muß ja physikalisch vom Sender zum Empfänger kommen. Im Fall von gesprochenen Mitteilungen heißt das, daß der Sprecher (der Sender) und der Hörer (der Empfänger) in unmittelbarer Nähe zueinander sind, es sei denn, Lautsprecher oder Funkgeräte kommen ins Spiel. Im schriftlichen Fall muß der Schriftträger vom Sender zum Empfänger kommen, entweder direkt oder durch eine Art Staffellauf, wo mehrere Abschriften einen Teil der Strecke zurücklegen können. Werke der Antike oder des Mittelalters kann man heute nur dann lesen, wenn wenigstens *eine* Hs die Feinde der Information überwunden hat und bis in unsere Zeit überlebt hat. Selbst in der Neuzeit sind viele Werke dem Krieg, der Zensur, oder den Kaminen überlebender Verwandter des Autors zum Opfer gefallen. Eine Mitteilung oder Botschaft kann auch als Zweck haben, einfach sicherzustellen, daß dieser Kontakt richtig funktioniert. (Hallo? Hallo? Hören Sie? Funktioniert diese Lautsprecheranlage?) Das ist die *phatische* Funktion der Sprache. Wenn wir den Kontaktweg hinzufügen, sieht die Zeichnung so aus:



Der Kontakt aber genügt nicht. Die Kommunikation findet nur dann statt, wenn der Sender und der Empfänger beide das selbe sprachliche System (denselben Code) beherrschen. Die *metalinguistische* Function der Sprache dient dazu, die Gemeinsamkeit des sprachlichen Systems herzustellen oder wieder in Gleichgewicht zu bringen.



Jakobson wollte die Funktionen der Sprache erläutern, aber sein Modell kann uns dazu dienen, die verschieden Ausfallarten der Kommunikation zu verstehen.

3.2.1. Ausfall beim Sender

Die erste mögliche Bruchstelle in der Verbindung von Sender und Empfänger liegt beim Sender. Wenn wir Botschaften an die Zukunft senden, dann sind wir das. Wir können aus Absicht oder Versehen die Daten überhaupt nicht senden; wir könnten sie verlieren oder löschen, statt sie aufzubewahren und wiederzubenutzen.

Oder, und das ist eine zweite Ausfallart, es könnte vorkommen, daß wir nicht sagen, was wir meinen. Im Bereich XML heißt das, wir könnten dem Tagmißbrauch, dem schlechten Modellieren, oder anderen semantischen Übeln unterliegen. Die Semantik soll später diskutiert werden. Im Moment genügt es, zu sagen: wenn mann bestimmte Informationen in Zukunft wiederbenutzen will, so ist es wichtig, im Klaren über die Natur dieser Information zu sein. In einem literarischen Werk wird dieses Anliegen dazu führen, daß man am liebsten die eine Stelle as Personennamen, die andere als terminus technicus, die dritte als Fremdwort, auszeichnet, auch wenn in der Stilvorlage alle drei in der gleichen Schriftart (etwa: schräg) gesetzt werden sollen. Wenn man mal die Stilvorlage ändert (und das kommt doch vor), wird sich die Schriftart der einen oder der anderen Stelle ändern, doch nicht die Tatsache, daß es um Personnenname, terminus technicus, oder Fremdwort handelt. Wenn man sich auf die sachliche Auszeichnung konzentriert, so vermeidet man viele unnötige Änderungen.

Die Fähigkeit, das zu sagen was man sagen will, statt die Aussage einem vordefinierten Schema von semantischen Primitivfunktionen anzupassen, läßt den Gebrauch von SGML und XML fast wie eine Befreiung erscheinen, wenn man an andere Methoden der Textdarstellung gewohnt ist.

Damit verbunden ist ein ernüchternde Verantwortung, denn wenn man genau das sagen kann, was man sagen will, so muß man sich eben entscheiden, was man eigentlich sagen will.

3.2.2. Ausfall beim Empfänger

Eine zweite Bruchstelle stellt der Empfänger dar. Es kann sein, daß der zukunftiger Empfänger unserer Botschaft gar nicht auf diese Botschaft achtet, nicht zuhört, fängt damit nichts an. Dagegen kann man nicht viel unternehmen, außer daß wir es dem Empfänger leicht machen, zu wissen, worum es sich bei dieser Botschaft dreht. So kann man vielleicht verhindern, daß unsere Arbeit aus Versehen weggeschmissen wird, weil der Empfänger (und hier bitte die Erinnerung daran wach halten, daß es sich hier sehr oft um uns selbst handelt) nicht mehr die Bedeutung oder den Ursprung der Daten durchschaut.

Eine zweite Ausfallart beim Empfänger besteht darin, daß wir vergessen, *daß wir nicht wissen, wer der Empfänger ist*. Wir wissen vor allem nicht, was der Empfänger *kann*, was seine Fähigkeiten sind. Es ist folglich meistens sinnlos, ihm per Flaschenpost zu bestimmten Tätigkeiten anzuregen, ihm Befehle zu erteilen, ihm eine Botschaft mit *imperativer Semantik* zukommen zu lassen. Eine deklarative Semantik hat viel größere Chancen, auch in zukünftiger Zeit relevant zu bleiben, genauso wie die deklarative Semantik heute eine Schlüsselposition hat, wenn man die Wiederverwendbarkeit, die Geräteunabhängigkeit, und die Anwendungsunabhängigkeit der Daten gewährleisten will.

3.2.3. Ausfall beim Kontakt

Die dritte mögliche Bruchstelle liegt darin, daß man den Kontakt zwischen Sender und Empfänger verliert.

Diese Ausfallart tritt dann ein, wenn der Datenträger verloren geht, aus internen Gründen nicht mehr zu lesen ist, oder mit neu erworbenen Maschinen nicht mehr zu lesen ist. In den 80er Jahre haben gewissenhafte Benutzer ihre Dateien regelmäßig auf Disketten gespeichert, um sie zu archivieren. Jetzt sitzen dies Benutzer auf einem großen Haufen Disketten in der Größe von 5 Zoll, die keine Maschine mehr lesen kann. Wenn man noch 3-Zoll Disketten hat, soll man sie schnell auf neue Datenträger überspielen, bevor die letzte zugängliche Maschine im Haus, die ein Diskettentreibwerk noch hat, spurlos verschwindet.

Manche Bibliotheke und Rechenzentren versuchen, diese Ausfallart dadurch auszuweichen, indem sie alle

Datenträger regelmäßig kopieren. Viele setzen dafür Softwares für die Verwaltung von digitalen Bibliotheken ein, die das Kopieren der Daten und der dazugehörigen Metadaten bewerkstelligen. Solche Softwares sind dazu konzipiert, sehr große Datenmassen zu bewältigen, aber die Verbindung zu dem ursprünglichen Kontext geht in solche Massensystem leicht verloren. Es rät sich, dabei möglichst alles aufzuschreiben, was der zukünftiger Empfänger vielleicht wissen muß, wenn er die Daten innerhalb dieser Digitalbibliothekssoftware eines Tages herumliegen findet.

3.2.4. Ausfall im Code

Eine Ausfallart, die in der Vergangenheit den geisteswissenschaftlichen Projekten große Schwierigkeiten bereitet hat, ist die Möglichkeit, daß der Sender und der Empfänger verschieden Zeichensätze benutzen. Eben weil der Zeichensatz von so grundlegender Bedeutung für die Textdatenverarbeitung ist, und von allen Teilsystemen unterstützt werden muß, wird die Wahl des Systemzeichensatzes vielen Benutzern völlig unsichtbar. Stillschweigend setzen alle Softwares im System den gleichen Zeichensatz voraus. Wer nicht gegen diesen Systemzeichensatz wegen seiner Unvollständigkeit ständig kämpfen muß, fragt sich gar nicht, wie der Zeichensatz des Systems überhaupt heißt, bis es beim Datenaustausch mit einem fremden System zu einer Panne kommt.

Hier hat die Entwicklung vom Universalzeichensatz unheimlich viel geholfen. Auch wenn man auf die Private Use Area zurückgreifen muß, um Sonderzeichen zu kodieren, hat man mit dem Universalsatz einen gemeinsamen Anhaltspunkt.

Wenn der Empfänger einmal die Zeichenkodierung verstanden hat, beginnt die schwierige Arbeit, das Datenformat zu verstehen. Es sind dem menschlichen Geist beim Erstellen von Datenformaten praktisch keine Grenzen gesetzt, und der menschliche Geist hat sich dankbar auf diesem Gebiet energisch und reichlich entfaltet.

Wer eine Botschaft an die Zukunft senden will, hat drei Arten von Datenformatten zu erwägen:

- proprietäre (geschlossene) Formate
- eigene, selbstdefinierte Formatte, den eigenen Daten und dem eigenen Bedarf nach Belieben angepast
- öffentlich zugängliche, öffentlich dokumentierte Formatte

Proprietäre Formate bieten sich an und sind sehr bequem, solange die dazugehörige Software weit verbreitet ist und sowohl dem Sender wie auch dem Empfänger zugänglich ist. Für den Datenaustauch über geographischen und organisatorischen Grenzen hinweg werden proprietäre Formate oft mit Erfolg eingesetzt. Aber das meist recht kurze Leben solcher Formate macht sie für eine Botschaft an die Zukunft eher untauglich.

Selbstgemachte Formate sind oft eine gute Wahl, weil sie so gut der Eigenart der Daten und den Bedürfnissen des Senders angepaßt werden können. Aber wer ein solches Eigenformat definiert, muß damit rechnen, daß er das Format auch gründlich dokumentieren muß. Denn ohne Dokumentation wird der Empfänger wenig mit den Daten anzufangen wissen. Es waren keine dreißig Jahre seit der Marslandung von Viking (1975 gelauncht, 1976 gelandet), als man die Meßdaten des Landers durchsuchen wollte, um mögliche Zeichen von Leben auf Mars zu finden. Das Magnetbandformat aber, in dem die Daten elektronisch erhalten sind, wurde leider nie dokumentiert, bzw. es wurde die Dokumentation nicht gefunden, und alle Daten wurden neu von Papierausdrucken mit der Hand eingegeben.

Für eine Botschaft an die Zukunft scheinen aus solchen Gründen sich die offene Formate wie XML besonders gut (oder wenigstens weniger schlecht) zu eignen. Solche Formate sind gut dokumentiert, die Dokumentation läßt sich ohne große Mühe finden (wenigstens heute - wir wollen hoffen, das sei auch in Zukunft der Fall), und es scheint unwahrscheinlich, daß das Wissen um XML und andere offene Formate jemals gänzlich aus der Welt verschwindet. Das XMLformat weist viel Redundanz auf, so daß es Datenverfall verhältnismäßig gut widersteht — wenigstens wird es weniger Wahrscheinlich, daß die Daten korrumpiert werden, ohne daß man es merkt. Auch wenn XML so aus der Mode fiele, daß es keine XML-softwares mehr gäbe, ist das Format im Grunde so einfach, daß man selbst einen Parser dafür schreiben

könnte, um die Umformatierung in ein neues Format zu erleichtern.

Zusammenfassend kann man sagen, daß gegen Ausfälle beim Kontakt und beim Code es brauchbare technische Mittel gibt, wenn man diese Mittel konsequent und diszipliniert einsetzt. Probleme beim Sender unde beim Empfänger dagegen, verlangen nicht technische sondern institutionelle Lösungen.

3.2.5. Ausfall in der Semantik

Die letzte Ausfallart ist die der Semantik.

Die Kommunikation kann selbst dann spektakulär versagen, wenn der Sender etwas mitteilen will, der Empfänger zuhören will, und die in dem gemeinsamen Code verfaßte Mitteilung erfolgreich beim Empfänger ankommt. Drei verschiedene Ausfallarten gibt es hier, die alle mit der Erfassung der Bedeutung der Botschaft zu tun haben.

Die erste Ausfallart scheint ein hoffnungsloser Fall zu sein. Der Empfänger empfängt, entschlüsselt, und versteht die Botschaft, und entdeckt dann erst, daß die Botschaft für den Empfänger weder interessant noch nützlich ist. Der Empfänger will vielleicht etwas über die Weissagung in der Antike erfahren, und schaut sich die Daten an, weil sie angeblich u.a. auch von Orakeln handeln. Er findet darin aber nur Information zu einem gewissen Datenbanksystem, das ihn leider nicht interessiert. Ganz verhindern kann man wohl diese Art des Ausfalls nicht, aber wir können und sollen es dem Empfänger so leicht wie möglich machen, zu sehen, welches Oracle wir eigentlich meinen.

Die zweite Ausfallart besteht darin, daß der Empfänger die Botschaft erfolgreich entziffert, alle Daten richtig den betreffenden Objekten in der Anwendungsdomäne zuweist, begreift die volle Bedeutung der Botschaft aber nicht. Dagegen ist auch kein Kraut gewachsen: daß man gelegentlich die volle Bedeutung einer Tatsachenmenge nicht begreift, gehört weniger zu der Problematik der Kommunikation, als zu der Problematik des Lebens.

Die dritte semantische Ausfallart ist ganz einfach. Man bekommt ein XMLdokument, versteht also mühelos die Elementstruktur der Daten, kennt aber die vorliegende Auszeichnungssprache nicht, versteht also nicht, welche Bedeutung den Elementen und Attributen des Dokuments zuzuweisen ist. Diese Ausfallart dürfte eine der am öftesten auftretenden sein, wenn es um den Austausch von XMLdokumenten geht. Sie kann wenigsteins teilweise vermieden werden, aber nicht ohne Arbeit.

Zu diesem Thema gibt es viel zu sagen — zuviel, vielleicht, denn ich vermute, der Empfang ist inzwischen doch fertig vorbereitet, und Sie haben vielleicht schon Durst. Ich versuche mich also kurz zu fassen.

3.3. Nachhaltige Semantik

Die eigene Auszeichnungssprach dem Empfänger verständlich zu machen, erfordert eine gewisse menschliche Intelligenz, und es ist schwierig, dafür ein Regelwerk zu erstellen, das objektiv oder intersubjektiv nachprüfbar wäre, und das uns den Erfolg garantieren würde. Einige allgemeine Ratschläge kann man allerdings geben.

Regel 1: Man denke darüber systematisch nach, was man sagen will.

Man braucht nicht unbedingt, eine formale Ontologie mit Definitionen in der Web Ontology Language (OWL) oder mit Topics in Topic-map-format zu formulieren, aber es lohnt sich zu fragen: worüber, über welche Arten von Wesen, wollen wir Aussagen machen? Wörter? Sprachen? Texten? Werken? Belegstellen? Wenn man eine formal definierte Ontologie erstellen würde, was für Dinge würde man voraussetzen? Welche Eigenschaften würde man ihnen zuweisen? Zu den Methoden für solche systematische Überlegungen gibt es eine kleine, weit verstreute Literatur, die die Modellierung und die Erstellung von Auszeichnungssprachen behandelt. Ich empfehle allen u.a. das Buch von Eve Maler und Jeanne El Andaloussi, *Developing SGML DTDs: From Text to model to markup*.

Regel 2: die Auszeichnungssprache sorgfältig entwerfen, mit dem Ziel, daß die Einzeldokumente, die mit dieser Sprache ausgezeichnet werden, so gemeinverständlich wie nur möglich sein sollen.

Rein mechanisch produzierten Auszeichnungssprachen können beliebig schwerverständlich werden.

Regel 3: die Auszeichnungssprache, und Ihren Gebrauch dieser Sprache, dokumentieren!

Große Bibliotheke haben oft ein Hauptexemplar des bibliothekarischen Regelwerks, nach dem sie Bücher katalogisieren. Dieses Hauptexemplar wird oft mit unbeschriebenem Papier durchschossen, damit die lokal adoptierten Zusatzregeln, die lokale Auslegung schwieriger Fälle, usw. festgehalten werden können. Manche geisteswissenschaftliche Projekte pflegen auch eine solche lokale Erweiterung ihres Regelwerks. Bei einer allgemein gehaltenen Auszeichnungssprache wie den Richtlinien der TEI sind solche lokale Erweiterungen durchaus notwending, und müssen dokumentiert werden, wenn die Daten dem Empfänger verständlich sein sollen.

Zusamenfassend sollte das Markup Vokabular (oder genereller gesagt das verwendete Datenformat) in allen für Langlebigkeit angelegten Daten auf verschiedene Arten dokumentiert werden:

1. Generelle Dokumentation auf hoher Ebene
2. Referenzinformationen für jedes Element und Attribut
3. Anmerkungen zu lokaler Anwendung, wenn die lokale Anwendung eine konsistente Variante eines weit verbreiteten Vokabulars ausmacht.
4. Beschreibungen der Bedeutung des Markups mittels einer 'Übersetzung der Bedeutung des Markups oder von Markup-Konstrukten in eine oder mehrere formale Notationen: Prädikatenlogik erster Stufe, RDF, Prolog etc.

Regel 4. den Tagmißbrauch vermeiden!

Der Tagmißbrauch (engl. Tag abuse) schadet der Nutzbarkeit von Dokumentation, denn wenn Tagmißbrauch begangen wird, dann beschreibt die Dokumentation nicht mehr die Sprache, in der die Daten ausgedrückt werden. Wenn Elemente oder Attribute nicht angemessen hinsichtlich ihrer definierten Semantik benutzt werden, sind die Daten weniger einfach wiederverwendbar, weil sie nicht so verlässlich verarbeitet werden können.

Der Tagmißbrauch definiert man als Unverträglichkeit zwischen der beabsichtigten Semantik und der tatsächlichen Verwendung eines Tags Es ist schwer, ihn mit automatischen Methoden aufzuspüren. Aber es gibt Methoden, welche das notwendige menschliche Eingreifen einfacher und effizienter machen. So genannte false-color Fassungen von Dokumenten können vorbereitet werden. Sie stellen in auffälligen Farben Markierungen von spezifischen Passagen bereit, welche ein Mensch überprüfen sollte (z.B. alles in Rot was als ein Personenname ausgezeichnet ist, oder alle Ortnamen mit blauem Hintergrund). Die Semantik des Markups kann in natursprachige Sätze übersetzt werden, so daß sie hinsichtlich Inkonsistenzen und Irrelevanz überprüft werden kann. Vergleiche [Marcoux 2006] für weitere Diskussionen.

Regel 5. Ergänzende Dokumente sollen bereitgestellt und dokumentiert sein.

Soviel relevanter Kontext wie möglich muß man an den Empfänger weiterleiten. Wichtige Metadaten, die spezifisch sind für ein bestimmtes Dokument, sollten wahrscheinlich eher innerhalb des Dokuments gespeichert werden als extern, so daß es weniger wahrscheinlich wird, daß sie verloren gehen. Die Verfügbarkeit solchen zusätzlichen Materials kann weitreichend zum Verständnis des angemessenen Kontextes für die Interpretation der Daten beitragen, und hilft somit Mißverständnisse oder ein Unverständnis der Daten zu verhindern.

Regel 6. Früh und oft validieren und verifizieren!

Man kann viele Problem dadurch verhindern, indem man regelmäßige Validierung und Verifikation durchfürht. Im allgemeinen Fall ist die Semantik formaler Sprachen nur für wohlgeformte Äußerungen wohldefiniert. Nicht valide Dokumente haben nicht notwendigerweise eine feste Interpretation. Es muß

deshalb früh und oft validiert werden.

Das selbe trifft für semantische Validierungs- und Verifikationsprozeduren zu. (Der Leser sollte sich bewußt sein daß Forscher und Praktiker aus dem Bereich der Programmverifikation *Verifikation* als mechanischen Prozeß bezeichnen, und *Validierung* als zugehörigen nicht mechanischen Prozeß. Die Markup Community folgt der Tradition der formalen Logik, indem sie den Ausdruck *Validität* als mechanisch überprüfbare Eigenschaft auffaßt; nicht selten wird der Ausdruck *Verifikation* benutzt um einen zugehörigen nicht mechanischen Prozeß zu bezeichnen. Wer sich mit Anderen unterhält, die Interesse an dem Thema haben, tut gut, sicherzustellen, daß man man sich u.U. die Terminologie sich gegenseitig erklärt.)

Zusammenfassend:

1. Überlegen, was Sie überhaupt sagen wollen!
2. Die Auszeichnungssprache mit Sorgfalt entwerfen, und die Elementnamen, die Attributnamen, und die Verschachtelungsstruktur mit Bedacht wählen!
3. Die Ausz.spr. dokumentieren, vorzugsweise auf verschiedene Arten:
    a. Dokumentation in Prosa auf hoher Ebene
    b. detaillierte Beschreibung jedes Elements und jedes Attributes
    c. Dokumentation zur lokalen Interpretationen und Verwendung
    d. Beschreibung, in Prosa und als ausführbarer Programmcode, zumindest eines Teils der Bedeutung einer Dokumentinstanz in einer radikal anderen Notation wie Prädikatenlogik erster Stufe oder RDF.
4. Den Tagmißbrauch vermeiden!
5. Zusätzliche Dokumente (Dokumentation, Schemata, Stylesheets etc.) für den Empfänger bereitstellen!
6. Sowohl die Syntax als auch die Semantik der Dokumente systematisch validieren!

4. Schlußwort

Die Daten, mit denen wir arbeiten und die die für uns wichtig sind, sind oft beständiger als die Anwendungen und Werkzeuge, mit denen sie erzeugt werden. Die Kosten würden unerschwinglich sein, wenn wir alle für uns wichtigen Daten jedesmal neu erzeugen müssten, wenn wir Hardware oder Software auswechseln. Die Kosten werden immer noch hoch sein, wenn wir unsere Daten *transformieren* müssen, indem wir sie in einem oft verlustbehafteten Prozeß in ein neues System importieren. Es ist viel besser, unsere Daten in einer Form zu haben, die unverändert bleiben kann, so lange die Daten bestehen, die als Format zur Langzeitarchivierung verwendet werden kann, und die eine einfache Transformation in anwendungsspezifische Formate erlaubt. Der hier präsentierte Entwurf ist ein erster Versuch, den Kontext einer Antwort für dieses Problem zu erkunden.

Ein abschließender Gedanke.

Wie bereits bemerkt — standardisiert werden kann nur das, was man versteht.

Wir werden als Gesellschaft Standards, die der Komplexität, Variabilität und Vielfältigkeit menschlicher Kultur und unseres kulturellen Erbes gerecht werden, nur dann erzeugen, wenn Personen mit dem notwendigen Wissen und der notwendigen Erfahrung aktiv an der Entwicklung der Standards teilnehmen.

Aus der Sicht von Standardisierungsorganisationen ist dieses Ziel nur erreichbar, wenn ausreichende Vorraussetzungen für die öffentliche Teilnahme gegeben sind. Im Falle des W3C, welches ich am besten kenne, wurden mit diesem Umstand im Hinterkopf verschiedene Eigenschaften der Organisation und des W3C-Prozeßes definiert.

- Die W3C-Mitgliedsgebühren für gemeinnützige Organisationen sind gegenüber denjenigen für die Vollmitgliedschaft stark reduziert.
- Unter geeigneten Umständen dürfen im W3C Sachkundige an der Arbeit von Arbeitsgruppen teilnehmen, auch wenn deren Organisation kein Mitglied des W3C ist. Meine eigene Teilnahme beim W3C begann als *invited expert* in der Arbeitsgruppe, in der die XML-Spezifikation erstellt wurde.

Selbst wenn man also nicht bei einer Organisation arbeitet, die dem W3C beitritt, kann man möglicherweise in einer Arbeitsgruppe als *invited expert* mitarbeiten.

- Jede W3C Empfehlung (Recommendation) wird mindestens drei Mal für öffentliche Kommentare veröffentlicht, und manche Spezifikation noch öfter: als *Last Call Working Draft*, als *Candidate Recommendation*, und als *Proposed Recommendation*. In jeder Phase hat die verantwortliche Arbeitsgruppe die absolute Verpflichtung zu versuchen, allen Kommentaren und Einwänden gerecht zu werden und den Kommentierer zufrieden zu stellen, unabhängig davon, ob er ein W3C-Mitglied vertritt, oder nicht. Aus meiner eigenen Erfahrung kann ich bestätigen, daß die Komentare von Nicht-Mitgliedern genauso ernst genommen werden und genauso viele Umarbeitungen der Spezifikation hervorrufen können, wie die Kommentare von Mitgliedern. Das heißt, selbt wenn eine Organisation nicht dem W3C beitritt und selbst wenn man kein *invited expert* ist, hat man die Möglichkeit, Kommentare abzugeben, und die Arbeitsgruppe hat die Verpflichtung sich ernsthaft mit diesen Kommentaren zu beschäftigen. (Natürlich kommen Kommentare je frühre desto besser an. Wenn man erst in der Schlußphase massenhafte Abänderungen vorschlägt, ist kaum zu erwarten, daß die Arbeitsgruppe die Vorschläge freudig annimmt.)

Viel kann aber auch unternommen werden von Seite der Organisationen, welche sich mit der Studie und dem Erhalt des menschlichen kulturellen Erbes beschäftigen. Es gibt im Verhältnis zu Softwareherstellern viel zu wenig W3C-Mitgliedsorganisationen, die die Forschung und die Nutzer von Standards repräsentieren. Die Boeing Company z.B. vertritt im W3C auf hervorragende Weise die Interessen der Benutzer des Webs, aber auch ihr sind als Einzelmitglied Grenzen gesetzt!

Die Arbeit des W3C, und des W3C als ganzes, würde sehr von der Teilnahme von Universitäten und anderen kulturellen Institutionen profitieren. Sind die Einrichtungen, die hier im Workshop vertreten sind, Mitglieder des W3C? Ich habe mir die Liste unserer Mitglieder angeschaut und muß leider sagen: nein.

Bitte treten Sie dem W3C bei! Wenn Sie das nicht können, dann nehmen Sie bitte auf andere Weise teil, indem Sie Spezifikationen kommentieren und sich mit der Arbeit des W3C vertraut machen! Und tun Sie auch Ähnliches bei anderen Standardisierungsorganisationen, die öffentliche Kommentare erlauben.

So können wir erreichen, daß die Standards, die vom W3C oder andere Normierungsorganisationsen verabschiedet werden, und die wissenschaftlicher Projekte, wie sie hier vertreten sind, sich durch ihren Erfahrungsaustausch gegenseitig bereichern können.

[Pause]

Und jetzt kann ich im Ernst sagen: Das wär's. Ich danke für die Aufmerksamkeit.

A.

Berners-Lee, Tim, Dan Connolly, and Ralph R. Swick. 1999. Web Architecture: Describing and Exchanging Data, W3C Note 7 June 1999. [Cambridge, Sophia-Antipolis, and Tokyo]: World Wide Web Consortium. On the Web at http://www.w3.org/1999/04/WebData

Capelli, Adriano. 1899. *Dizionari di Abbreviature latine ed italiane usate nelle carte e codici specialment del medio-evo* Milano: Ulrico Hoepli. [Und viele spätere Ausgaben.]

Hazaël-Massieux, Dominique, and Dan Connolly. 2005. Gleaning Resource Descriptions from Dialects of Languages (GRDDL), W3C Team Submission 16 May 2005. [Cambridge, Sophia-Antipolis, and Tokyo]: World Wide Web Consortium. On the Web at http://www.w3.org/TeamSubmission/2005/SUBM-grddl-20050516/

Jakobson, Roman. 1960. "Closing Statement: Linguistics and Poetics", in *Style In Language*, ed. Thomas A. Sebeok (Cambridge: MIT Press, 1960), pp. 350-377.

[LOCKSS Project.] 2006. LOCKSS: Lots of Copies Keep Stuff Safe. Project home page at http://www.lockss.org/lockss/Home.

Marcoux, Yves. 2006. "A natural-language approach to modeling (extended draft) Why is some XML so difficult to write?" Extreme Markup Languages 2006 (forthcoming).

Maler, Eve, and Jeanne El Andaloussi. 1996. *Developing SGML DTDs: From text to model to markup*. Upper Saddle River, NJ: Prentice Hall PTR, 1996. xxiv, 532 pp; index.

Small, Jocelyn Penny. “Retrieving Images Verbally: No More Key Words and Other Heresies”. *Library Hi Tech* 9.1 (1991): 51-60.

Sperberg-McQueen, C. M., Claus Huitfeldt, and Allen Renear. Meaning and interpretation of markup. *Markup Languages: Theory & Practice* 2.3 (2001): 215-234. On the Web at <URL:http://www.w3.org/People/cmsmcq/2000/mim.html>

Sperberg-McQueen, C. M., and Eric Miller. 2004. On mapping from colloquial XML to RDF using XSLT, Extreme Markup Languages 2004. On the Web at http://www.mulberrytech.com/Extreme/Proceedings/html/2004/Sperberg-McQueen01/EML2004Sperberg-McQueen01.html

Vorthmann, Scott, Jonathan Robie, and Lee Buck. 2000. Schema adjunct framework. Draft Specification 30 November 2000. [Chapel Hill]: Extensibility. http://www.extensibility.com/saf/spec/ (no longer available?) http://xml.coverpages.org/SchemaAdjunctFramework200011.html

Vorthmann, Scott, and Jonathan Robie. 2001. Beyond schemas: Schema adjuncts and the outside world. *Markup Languages: Theory & Practice* 2.3 (2001): 281-294.

Wrightson, Ann. 2001. "Some Semantics for Structured Documents, Topic Maps and Topic Map Queries." Extreme Markup Languages 2001. On the Web at http://www.mulberrytech.com/Extreme/Proceedings/html/2001/Wrightson01/EML2001Wrightson01.html

Wrightson, Ann. 2005. "Semantics of Well Formed XML as a Human and Machine Readable Language." Extreme Markup Languages 2005. On the Web at http://www.mulberrytech.com/Extreme/Proceedings/html/2005/Wrightson01/EML2005Wrightson01.html

Wrightson, Ann. 2006. "Conveying Meaning through Space and Time using XML." Extreme Markup Languages 2006. Forthcoming.

Notes

[1] Für ihre Hilfe bei der deutschen Formulierung dieser Gedanken habe ich Herrn Prof. Dr. Kurt Gärtner und Herrn Dr. Felix Sasaki herzlich zu danken. Der Zuhörer ahnt gar nicht, wie viel er ihnen schuldet.