Die Rolle von Metadaten und XML in einem multilingualen Multimedia-System über Gesundheitsfragen

Dietmar RÖSNER

Zusammenfassung

Im EU-Projekt CATCH sollen Konzepte und Werkzeuge entwickelt werden für das Management multilingualer und multimedialer Informationsobjekte. Eine zentrale Rolle kommt dabei Metadaten und Dokumentstrukturen zu. Diese werden mit den von der Auszeichnungssprache XML bereitgestellten Sprachmitteln kodiert und sollen die flexible Nutzung des Pools an Informationsressourcen ermöglichen.

Einleitung

Das Projekt CATCH (Langtitel: ***C**itizen **A**dvisory System based on **T**elematics for **C**ommunication and **H**ealth*) wird im Rahmen des Europäischen *Telematics Application Program* seit Januar 1998 bis voraussichtlich Juni 2000 gefördert. Zu den Hauptzielen der Aktivitäten im Bereich Gesundheit dieses Förderprogramms gehört es, solche Anwendungen zu entwickeln, die dazu beitragen, daß das europäische Gesundheitssystem die Erwartungen der Bürger erfüllt.

Die Arbeiten von CATCH haben das Ziel, ein Rahmensystem zu schaffen, mit dem Informationen und Dienste zu gesundheitsrelevanten Fragen für europäische Bürger entsprechend deren Bedürfnissen bereitgestellt werden können. Der so verbesserte Zugang zu Gesundheitsinformationen soll dazu beitragen, die Selbstverantwortung der Bürger zu stärken. Informationen über gesunden Lebensstil, über Vermeidung von Risikofaktoren und andere Möglichkeiten zur Prävention sollen Gesundheitsrisiken verringern helfen. Eine wichtige Auswirkung der besseren Information soll letztendlich auch sein, daß sich die Ausgaben im Gesundheitsbereich verringern.

Beim Ansatz von CATCH liegt der Schwerpunkt auf folgenden Aspekten:

- Die Endnutzer der angebotenen Informationen und Dienste sind europäische Bürger.
- Den Fragen der Multilingualität, aber auch der soziokulturellen Unterschiede in Europa wird besondere Aufmerksamkeit gewidmet.
- Die Möglichkeiten der neuen Medien zu erhöhter Interaktivität und auch Adaptivität an den jeweiligen Nutzer sollen erprobt und entwikkelt werden.
- Die Entwicklung und Implementation des Systems erfolgt aus einer benutzerorientierten Perspektive. Dabei werden insbesondere zwei Gruppen von Nutzern betrachtet:
  — Endnutzer als „Informationskonsumenten",
  — Autoren als „Informationsproduzenten".
- Informationsanbieter können verschiedene der im Gesundheitsbereich handelnden Organisationen sein (Krankenkassen, Fachverbände, Selbsthilfegruppen, Kliniken, niedergelassene Ärzte usw.).
- Neben der Bereitstellung von Information ist auch die Integration von Diensten ein wichtiges Anliegen.

Von der Einzellösung zum flexiblen Rahmensystem

Ziel der derzeitigen Arbeiten von CATCH ist es, anhand einiger ausgewählter Themengebiete (Hautkrebs, kardiovaskuläre Erkrankungen) exemplarisch aufzuzeigen, wie medizinische Informationssysteme für Laien zu gestalten sind, mit Prototypen und realen Benutzern Erfahrungen zu sammeln, die Erfahrungen zu abstrahieren und so schließlich ein *Rahmensystem* zu erreichen, daß es anderen Informationsanbietern erleichtern wird, ihre Inhalte für Internet- oder auch Kioskanwendungen (s.u.) aufzubereiten.

Zu den wesentlichen Ergebnissen der laufenden Projektphase sollen gehören:

- eine Methodik zur Entwicklung und zum Management von multilingualen und multimedialen medizinischen Informationssystemen,
- ein Pool von textuellen und nichttextuellen Informationsobjekten, anhand deren die Methodik entwickelt, erprobt und evaluiert wurde,

- prototypische Software, mit der die Erstellung der Informationsobjekte und ihre automatisierte Nutzung erleichtert werden soll.

Ein wichtiges Teilthema ist in diesem Zusammenhang die *Unterstützung des Autorenprozesses*. Ein anderes ist die *Wiederverwendbarkeit* von multimedialen Ressourcen (d.h. von Texten, Bildern, Videos, ggf. Geräuschdateien u.ä.). Schließlich sollen Konzepte entwickelt und erprobt werden, die helfen, *multilinguale Informationsbestände* konsistent zu halten und den Prozeß der Übertragung von einem Quelldokument in Dokumente verschiedener Zielsprachen zu unterstützen (für das Projekt wurde Englisch als Sprache für die 'Masterversionen' der in den verschiedenen Zielsprachen – derzeit: Englisch, Deutsch, Italienisch, Portugiesisch – anzubietenden Texte verabredet).

Zentral für alle diese Anliegen ist, daß bereits bei der Erstellung von beispielsweise Texteinheiten für diese geeignete *Meta-Information* (BACA 1998) erfaßt wird. Meta-Information kann sich dabei sowohl auf den Erstellungs-/ Überarbeitungsprozeß beziehen oder auf den Inhalt oder auf die Art der intendierten Verwendung und dadurch bestimmte Aspekte der Inhaltspräsentation. Für die einheitliche Kodierung von Meta-Information wird XML (*extensible markup language*, BRAY et al. 1998, BOSAK 1996) eingesetzt.

Das CATCH-Konsortium

Das CATCH-Konsortium besteht – wie bei europäischen Projekten üblich – aus sehr unterschiedlichen Partnern, die in der Arbeitsteilung des Projekts unterschiedliche Rollen haben:

- Medizinische Experten aus Universitätskliniken und Institutionen der Gesundheitsprävention entwickeln die medizinischen Inhalte und erproben das System an ihren Einrichtungen.
- Die AOK Sachsen-Anhalt ist Anwendungspartner und erprobt das System als Kiosk in seinen Geschäftsräumen und in der Internet-Version als Teil seines Informationsangebots im WWW.
- Die für medizinische Systeme zuständige Abteilung von Hewlett-Packard Italien (HPI) zeichnet für die Fragen des zukünftigen Einsatzes solcher Systeme und ihrer Vermarktung verantwortlich.

- Die Entwicklung der informatischen Konzeption und die implementatorischen Aufgaben bei der Entwicklung der Prototypen teilen sich die universitären Partner aus Ulster und Magdeburg.

Erstellung der Texte

Die medizinischen Experten, die derzeit Texte für das Projekt schreiben, kommen aus Deutschland (Hautkrebs), Italien und Portugal (Herz/ Kreislauf). Im Projekt wurde beschlossen, mit folgender 'Sprachpolitik' zu arbeiten:

Die 'Masterversion' aller textuellen Dokumente ist in Englisch zu verfassen. Dies kann der Experte/die Expertin selbst tun, auch wenn er/sie nicht Muttersprachler/in ist, oder ein Übersetzer kann eingeschaltet werden.

Auf der Basis der englischen 'Masterversion' werden die Versionen für die verschiedenen Zielsprachen erstellt. Dies kann durch Übersetzer oder durch medizinische Experten geschehen, die Muttersprachler der Zielsprache sind. Im ersteren Fall werden die Ergebnisse der Übersetzung durch medizinische Experten geprüft und – ggf. nach Korrekturen – freigegeben.

Arten von erfaßten Metadaten

Teilt man die in CATCH erfaßten und weiterverarbeiteten Metadaten grob ein, so ergeben sich einerseits technische und bibliographische, auf den Autoren-und Publikationsprozess bezogene Daten und solche, die zum Inhalt des Textes und zu seiner Verwendung wichtige Hinweise geben.

Die bibliographischen Metadaten umfassen in Anlehnung an den sog. Dublin Core (DCMI 1999) den Autor, die Quelle der Information, das Datum der Erstellung, aber auch, ob der medizinische Inhalt zertifiziert ist, und wenn ja, von welcher Organisation.

Andere Metadaten dienen dazu, den Hauptgegenstand und die Intention der jeweiligen Informationseinheit zu charakterisieren. Beispiele können sein:

- der Text soll Information über einen Gegenstandsbereich vermitteln,
- er soll eine Definition geben,

- er soll instruieren, eine bestimmte Aktion auszuführen oder Aufgabe zu erfüllen,
- er soll den Benutzer überzeugen, ein bestimmtes Verhalten anzunehmen,
- er soll den Benutzer nach Informationen fragen (als Teil einer Interaktionssequenz oder in einem Fragebogen).

Schließlich kann der Autor mit Metadaten charakterisieren, ob und wie er das Zielpublikum berücksichtigt und welcher Mittel er sich dafür bedient:

- Wer soll angesprochen werden durch den Text? (z.B. allgemein interessierte Bürger, Patienten, Angehörige, Kinder, ... ?)
- Berücksichtigt die präsentierte Information Unterschiede zwischen den Benutzern?
- Wie ist der Sprachstil: ist er wissenschaftlich, neutral, alltäglich, emotional, instruktiv usw.?

Nutzung von Metadaten

Wenn multimediale Ressourcen (in erster Linie Texte, aber auch Bilder usw.) mit Meta-Informationen versehen sind, die auch den Inhalt charakterisieren oder typisieren, dann lassen sich bestimmte Prozesse automatisieren, die bei Dokumenten, die nur als maschinell nicht interpretierbarer Text (oder entsprechend als unanalysierte Bilder) vorliegen, nur schwierig zu bewerkstelligen sind. So kann z.B. im Datenbestand festgestellt werden, ob Texte mit einer bestimmten inhaltlichen Charakteristik vorliegen, in welchen Sprachausprägungen sie vorliegen, welche Aktualität sie haben. Wird ein Text verändert, so lassen sich die ihm entsprechenden parallelen Texte in den anderen Sprachen leicht identifizieren und die Verantwortlichen für den Inhalt können entscheiden, ob die Änderung des einen Textes entsprechende Änderungen in den anderen Sprachversionen nach sich ziehen muß oder nicht.

Zur Nutzung von CATCH: Internet vs. Kiosk

Das CATCH-System wird im Internet angeboten werden. Obwohl die private Nutzung des Internet zunimmt, wird damit derzeit nur ein kleiner Teil der möglichen Nutzer erreicht.

Geplant ist daher, das System an verschiedenen Orten auch über öffentlich zugängliche sog. Informationskioske anzubieten.

Die Kioskversion des Systems erfordert, da sie nicht mit Maus und Tastatur, sondern über einen berührungssensitiven Bildschirm bedient wird, eine modifizierte Benutzerführung. Auch inhaltlich kann sich das Informationsangebot von der Internet-Version unterscheiden. So kann lokale Information hinzugefügt werden (z.B. Information über die Klinik, in der der Kiosk steht: Ärzte, Personal, Räume, Zeiten, ...) oder das thematische Angebot kann auf bestimmte Fachgebiete konzentriert, dort aber mit detaillierterer Information angereichert werden (z.B. Konzentration auf das Gebiet Hautkrankheiten und detaillierte Information über die praktizierten Behandlungsmethoden im Kiosk der dermatologischen Universitätsklinik).

Weil die Erstellung der Informationsobjekte teuer und zeitaufwendig ist, muß versucht werden, einen hohen Grad an Wiederverwendbarkeit zu erreichen. Metadaten können helfen, die Auswahl der für eine spezielle Ausprägung des Systems gewünschten Informationsobjekte aus dem gesamten Pool verfügbarer Informationsressourcen zu automatisieren. Bis zu welchem Grad eine solche Unterstützung bei der Konfiguration spezieller Systemausprägungen gehen kann, ist Gegenstand der weiteren Untersuchungen im Projekt.

Ein ausführliches Beispiel

Die Tags, die in den textuellen Informationsobjekten von CATCH verwendet werden, gehören insbesondere den folgenden Kategorien an:

- Tags zum Erfassen bibliographischer Metadaten,
- Tags zur Markierung struktureller Einheiten eines Texts,
- sog. ‘in-line tags’, mit denen Information zur semantischen Klassifikation von Termen eines Textes gegeben wird.

Ein Beispiel soll die verschiedenen Kategorien verdeutlichen. Wir verwenden dazu Ausschnitte aus einem Text aus einer Kollektion von Fragen und Antworten zum Themenbereich Prävention und Früherkennung von Hautkrebs.

```
<?xml version="1.0"?>
<CATCH-INFO-ELEMENT>
<META authors="Dr. Schramm, Luckert" supervisor="Prof. Gollnick" copy-
right="UDV, 1999"/>
<META translated-by="DR"translation-date="March-15-99"
time-to-translate="50min"/>

<RHETORICAL-QUESTION-ANSWER-PAIR>
<RHETORICAL-QUESTION>
7. How to perform <SELFHELP>self examination</SELFHELP> and
<SELFHELP>self diagnosis</SELFHELP>!
</RHETORICAL-QUESTION>

<ANSWER>
In prevention (of skin cancer) we distinguish between <MEDTERM>primary
prevention</MEDTERM>  through  information  and  early  detection  as
<MEDTERM>secondary prevention</MEDTERM>.

<APPEAL>There is no doubt: In addition to primary prevention early detec-
tion plays the most significant role in the fight against cancer. Don't
forget: You are the most important factor in early detection!
</APPEAL>

It is a big advantage that the <ORGAN>skin</ORGAN> –– in contrast to many
other organs –– is visible and can be examined without technical devises
and without <MEDTERM>invasive examination methods</MEDTERM>. We thus have
the basis for an examination method applicable by everybody.

For all those <DISEASE>malignant diseases of the skin</DISEASE> that
develop visibly regular self examination offers a big chance to detect
<DISEASE>cancer</DISEASE> already in an early stage.

...
```

Bibliographische Metadaten

Bibliographische Metadaten erfassen Informationen über den Lebenszyklus von Informationsobjekten, z.B. zu ihrer Erstellung, Übersetzung und Aktualisierung. In CATCH wird zu diesem Zweck der sog. Dublin Core als Ausgangspunkt verwendet, aber um Angaben erweitert, die für das Projekt zusätzlich von Belang sind.

Im obigen Beispieltext sind die Metadaten über die Autoren, den Betreuer, das Copyright und über die Übersetzung bibliographische Metadaten enthalten.

Strukturen

Um die Flexibilität im Umgang mit Informationsressourcen und ihre Wiederverwendbarkeit zu maximieren, werden in CATCH konsequent die Vorteile ausgenutzt, die eine Architektur bietet, bei der logische Dokumentstrukturen von allen Fragen des Layouts oder der Bildschirmausgabe vollständig separiert werden.

Textuelle Informationsobjekte im Informationspool von CATCH enthalten nur Tags mit Bezug zur logischen Struktur der Dokumente. Die Abbildung von logischen Dokumentstrukturen auf Ausgabeformate oder das dynamische Erzeugen von Layout sind ein von der Speicherung und Verwaltung der Informationsobjekte getrennter Prozess.

Durch die Auszeichnung mit strukturellen Tags können die Autoren den Prozessen, die auf den Informationsobjekten arbeiten, verdeutlichen, um welche Teile es sich handelt und was ihre jeweilige Funktion ist.

Der Beispieltext wird als Instanz eines Strukturelements mit der Bezeichnung `<RHETORICAL-QUESTION-ANSWER-PAIR>` gesehen, das sich zusammensetzt aus den Elementen `<RHETORICAL-QUESTION>` und `<ANSWER>`.

Die Begründung für diese Strukturierung ist, daß der einleitende Satz des Dokuments (`How to perform ...`) eine rhetorische Frage aufwirft, die durch den restlichen Text beantwortet wird (im Element `<ANSWER>`). Bei dieser Art von Analyse ist der zugrundeliegende Sprechakt wichtiger als die syntaktische Form der sprachlichen Oberfläche. Der einleitende Satz hätte genausogut Formen annehmen können wie *'How can I perform self examination and self diagnosis?'* oder *'Self examination and self diagnosis'*.

Das Layout könnte verdeutlichen, daß die `<RHETORICAL-QUESTION>` auch als Titel der Informationseinheit fungiert. Der Beispieltext gehört zu einer Serie aus mehreren Elementen vom Typ `<RHETORICAL-QUESTION-ANSWER-PAIR>`. In einer dynamisch kreierten Übersichtsseite zu dieser Kollektion könnte jeweils das Element `<RHETORICAL-QUESTION>` als Verweis zum zugehörigen Text vom Typ `<ANSWER>` verwendet werden.

Im Beispieltext sind weitere Einheiten entsprechend ihrer Funktion ausgezeichnet:

- explizite Appelle:

```
<APPEAL>There is no doubt: In addition to prima-
ry prevention early detection plays the most
significant role in the fight against cancer.
Don´t forget: You are the most important factor
in early detection!
</APPEAL>
```

Mit einem Appell will ein Autor die grundsätzliche Einstellung der Leser zu einem Thema in seinem Sinne beeinflussen (hier: Wichtigkeit regelmäßiger Untersuchung der Haut).

- Empfehlungen:

```
<RECOMMENDATION>The examination should be car-
ried out <DETAIL>with a completely nude body
</DETAIL>.</RECOMMENDATION>
```

Empfehlungen beziehen sich auf Aspekte des Verhaltens des Lesers. Der Autor schlägt vor, bestimmte Handlungen auszuführen oder dies in einer bestimmten Weise zu tun. In einer negativen Variante kann auch empfohlen werden, bestimmte Handlungen zu unterlassen.

Auch hier gilt wieder, daß das Layout verschiedene Struktureinheiten jeweils unterschiedlich behandeln und ihre Funktion hervorheben kann (z.B. mit unterschiedlichen Farben, Zeichensätzen, Hervorhebungen, ...). Zusammen mit den Autoren aus dem CATCH-Konsortium wird derzeit die Liste von Strukturelementen weiterentwickelt und vervollständigt. Zu ihr werden Tags gehören für:

- `<DEFINITION>`

Eine `<DEFINITION>` besteht aus dem zu definierenden Term (`<DEFINIENDUM>`) und dem Definitionstext (`<DEFINIENS>`).

- `<WARNING>`

Eine Warnung soll den Leser, möglicherweise zum wiederholten Male, auf mögliche Risiken oder Gefahren aufmerksam machen.

Semantische Klassifikation

Autoren können natürlichsprachliche Terme – das können einzelne Wörter oder Wortgruppen sein –, die sie in ihren Texten für wichtig

erachten, mithilfe sog. in-line tags semantisch klassifizieren. So finden wir im Beispieltext u.a. die folgenden Auszeichnungen: ‘skin’ wird als Körperteil (<ORGAN>), ‘cancer’ als Bezeichnung einer Krankheit (<DISEASE>) und ‘invasive examination methods’ als (hier zunächst nicht weiter spezifizierter) medizinischer Term (<MEDTERM>) gekennzeichnet. Die vorgesehenen Auszeichnungen und ihre Bedeutung sind als Ontologie organisiert. Eine erste Version dieser vor allem aus der Perspektive von medizinischen Laien organisierten Strukturierung des Sachgebiets ist an die medizinischen Experten zur Evaluation und Erweiterung übergeben worden. Derzeit wird untersucht, inwieweit der Vorschlag zu einer weltweit einheitlichen medizinischen Terminologie – das sog. Unified Medical Language System (UMLS 1999) – sich dafür eignet, direkt in CATCH integriert zu werden.

Diskussion

Andere Systeme

Von anderen Systemen, die im Internet medizinische Informationen für Laien anbieten, unterscheidet sich CATCH insbesondere durch die europäische Dimension der Mehrsprachigkeit und die für das Management multilingualer Informationsobjekte entwickelte Methodik.

Viele der Angebote im Internet sind nur auf englisch verfügbar. Ein deutschsprachiges Angebot bietet das kommerziell betriebene System *lifeline* (lifeline 1999).

Das gleichnamige Unternehmen plant zwar, seine Dienste längerfristig europaweit anzubieten; auf welcher methodischen Basis dies geschehen soll, scheint aber noch nicht ausgearbeitet zu sein.

Seiteneffekt: multilinguales Korpus

Die Arbeiten in CATCH werden – als Nebeneffekt – zu einem Korpus mit einer großen Zahl paralleler Texte führen. Durch die Auszeichnungen von Diskurselementen und durch die semantischen Tags für

wichtige Terme sollte dieses Korpus sich besonders für vergleichende Studien eignen.

Die in CATCH gewählte Vorgehensweise erzwingt die strukturelle und inhaltliche Parallelität der Dokumente in den verschiedenen Sprachen.

Eine vergleichbare Situation liegt bei mehrsprachiger technischer Dokumentation vor. Auch hier gibt es meist ein Masterdokument, das Struktur und Inhalt der Dokumente in den anderen Zielsprachen bestimmt.

Lohnt der Aufwand?

Im Projekt CATCH wird der Ansatz verfolgt, daß die Autoren der Texte Metadaten erfassen. Eine immer wieder gestellte Frage ist: Wie wird die in CATCH erarbeitete Methodik von den Autoren angenommen?

Wir haben darauf noch keine abschließenden Antworten. Der derzeitige Stand ist, daß die im Projekt als Autoren beteiligten medizinischen Experten eine erste Version eines „Autorenhandbuch" (CATCH author's guide) erhalten haben, in der die Methodik beschrieben wird. Die Reaktionen darauf sind grundsätzlich positiv, diskutiert wird vor allem, wieviel Auszeichnungen sinnvoll sind und wie der Aufwand für die Autoren so gering wie möglich gehalten werden kann.

Unterstützende Softwarewerkzeuge werden derzeit als Erweiterungen für Texteditoren implementiert und demnächst an die Autoren ausgeliefert. Sie werden die vordefinierten Tags in Auswahlmenus anbieten. Dabei wird für das 'in-line tagging' zum Kategorisieren von Termen auf eine Ontologie der Domäne und für das Auszeichnen der Dokumentstrukturen auf DTDs für häufig wiederkehrende Informationseinheiten zurückgegriffen. So wird nicht nur der Aufwand für das Eintippen von Tags, sondern auch die kognitive Belastung für das Memorieren der Bezeichner vermieden.

In der Evaluationsphase des Projekts im Frühjahr 2000 sollen die Erfahrungen der Autoren sowohl mit der Methodik als auch mit den unterstützenden Werkzeugen ausgewertet werden. Dabei wird die zentrale Arbeitshypothese des Projekts kritisch zu überprüfen sein, nämlich, daß der Nutzen für die Wiederverwendbarkeit, Wartbarkeit, Anpaßbarkeit an Benutzerbedürfnisse usw., der sich mit verfügbarer Meta-Information

realisieren läßt, den Aufwand für die Bereitstellung der Meta-Information aufzuwiegen vermag.

Informationen über CATCH

Aktuelle Demonstratorversionen des Systems finden sich im Internet unter:

- `http://paris.cs.uni-magdeburg.de/aok/`
- `http://catch.infc.ulst.ac.uk/catchii/main.htm`

Literatur

BACA, Murtha (1998): Introduction to Metadata – Pathways to Digital Information. Getty Information Institute.

BOSAK, Jon (1996): XML, JAVA, and the future of the web. `http://sunsite.unc.edu/-pub/sun-info/standards/xml/why/xmlapps.htm` .

BRAY et al. 1998): Tim B., Jean PAOLI, C.M. SPERBERG-MCQUEEN, Extensible Markup Language (XML) 1.0 . `http://www.w3.org/TR/1998/REC-xml-19980210` .

DCMI (1999): Dublin Core Metadata Initiative. homepage of the *dublin core metadata initiative*. `http://purl.oclc.org/dc/` .

lifeline (1999): Startseite des Informationsdienstes *lifeline*. `http://www.lifeline.de` .

UMLS (1999): National Library of Medicine. homepage of the Unified *Medical Language System* (UMLS). `http://www.nlm.nih.gov/research/umls/` .