

XML-basierter Workflow in den Wissenschaften

Wolfgang Meier

Institut für Soziologie
Technische Universität Darmstadt

Informatisierung der Arbeit
Workshop: Virtuelle Organisationen und verteilte
Anwendungen

Gliederung

- 1 Wozu XML?
- 2 XML in den Geisteswissenschaften?
 - Überblick
 - Beispiel: SozioNet
- 3 Web Services: Vorteile
- 4 Perspektiven

Wozu XML?

“XML Works, huh? Have you ever used it for anything serious?”

“XML is not well suited to data-interchange, much as a wrench is not well-suited to driving nails.”

Probleme

- Überzogene Marketingversprechen: XML als Patentlösung für Probleme des Datenaustauschs und der heterogenen Datenhaltung
- Als Basis für Web Services ist XML im Vergleich zu älteren Austauschformaten weder schlank, noch effizient, noch besonders einfach (vgl. XML vs. ASN.1, Binary XML)

Wozu dann XML?

- Heterogenität** Informationen müssen nicht in Tabellen und Spalten gepresst werden
- Erweiterbarkeit** Wandel kann modelliert werden, statt ihn zu vermeiden
- Flexibilität** Daten können in Umfang und Struktur variieren
- Nachhaltigkeit** Wechsel der Software bedeutet nicht, daß der Erschließungsaufwand umsonst war

Wozu dann XML?

- Im Gegensatz zu HTML definiert XML eine Meta-Syntax als Basis für domänenspezifische Markup-Sprachen.
- Integration von Text und Daten, von hoch- und schwach-strukturierten Inhalten
- Eine Vielzahl verbundener Standards: XSLT, XQuery, XLink, XInclude ...

XML in den Geisteswissenschaften?

Gerade für die Geisteswissenschaften eröffnen sich neue Möglichkeiten!

- Quellen sind in der Regel heterogen, eine Mischung aus schwach und stark strukturierten Informationen
- Verknüpfung und Rekombination bislang getrennter Datenbestände
- Unterstützung kollaborativer Arbeitsformen durch Nutzung vorhandener Standards, Werkzeuge und Dienste

XML in den Geisteswissenschaften?

- Strukturelle Auszeichnung/Archivierung von historischen und literarischen Quellen oder Transkripten (TEI-Standard: Text Encoding Initiative)
- Austausch bibliographischer Informationen (MODS/METS, RDF, Dublin Core)
- Systematische Erschließung von Quellen
- Spezielle Vokabulare: empirisch-quantitative Untersuchungen, qualitative Textanalyse, HEML (Historical Events Markup Language) ...

Beispiel: SozioNet

- <http://www.sozionet.org>
- Erfassung frei zugänglicher, sozialwissenschaftlicher Ressourcen im Web
- Gefördert vom BMBF
- Bestandteil von Infoconnex in Kooperation mit dem Informationszentrum Sozialwissenschaften
- Beteiligt sind 12 sozialwissenschaftliche Institute und Forschungseinrichtungen

Einige Stichworte

- Definition gemeinsamer Standards für die Publikation von Ressourcen im Web
- Bietet Werkzeuge zur Erstellung qualitativ hochwertiger, bibliographischer Metadaten
- Erschließung über die im Fach anerkannten Werkzeuge: IZ Thesaurus, Klassifikation
- Personalisierte Oberfläche zur Eingabe, Recherche und Verwaltung von Metadaten

Metadaten in SozioNet

- Nutzung weltweit anerkannter Standards (Dublin Core) für die allgemeinen Elemente (Titel, Autor, Abstract etc.)
- Das domänenspezifische Vokabular (z.B. Thesaurus-Verweise) ist davon klar getrennt
- SozioNet schreibt eine einheitliche Grundstruktur vor. Erweiterungen sind zulässig und erwünscht
- Einmal erstellte Metadaten sind in anderen Kontexten nachnutzbar

Verwendete Technologien

- Applikation basiert vollständig auf XML und verwandten Standards
- Oberfläche zur Metadaten-Eingabe in XForms
- Suchfunktionen, Thesaurus-Browser, Nutzeroberfläche sind XQuery-Skripte
- Speicherung von Metadaten und Nutzerdaten in XML Datenbank: kein Mapping auf relationales Datenbankschema
- Thesaurus und Klassifikation liegen ebenfalls als XML vor

Vorteile von Web Services: Z39.50 vs SRW/SRU

- Projekt: **ViBSoz**: Virtuelle Fachbibliothek Sozialwissenschaften
- Integration von Fachdatenbanken (SOLIS) und Bibliothekskatalogen in einem verteilten System
- <http://www.vibsoz.de>
- Basiert auf **Z39.50**: Standardprotokoll für Recherche in bibliographischen Datenbanken

Z39.50: Probleme

- Komplexer Standard: kaum eine Implementierung deckt alle Bereiche ab
- Protokoll ist in ASN.1 definiert: binäres Datenaustauschformat
- Implementierung erfordert komplexe Codebibliotheken
- Unterstützt per default vor allem binäre Austauschformate: MARC, MAB
- Aber: Z39.50 ist gut an den Workflow von Bibliotheken angepasst

Vergleich: SRW/SRU

- Nachfolger von Z39.50 basierend auf Web Services (SRW: SOAP, SRU: REST)
- Deutlich reduzierte Komplexität, deckt aber alle wichtigen Use-Cases ab
- Protokoll ist in allen gängigen Programmiersprachen schnell zu implementieren: keine aufwändigen Codebibliotheken nötig
- Datenaustausch in XML, HTTP als Transportprotokoll
- Setzt auf die langjährigen Erfahrungen mit Z39.50 auf: große Akzeptanz

Perspektiven

- SozioNet ist nur ein Baustein zur Verbesserung des Workflows
- Auf Basis von XML ließen sich vielfältige Quellen in ein solches System integrieren
- Die Standards sind vorhanden, es fehlt aber an der nutzerfreundlichen Integration