SDV Vol.32.1-2008

SDV. Sprache und Datenverarbeitung
International Journal for Language Data Processing
Heft 1/2008

Inhalt

Abstracts

Vorwort

Autor:
Schneider, Roman & Schröder, Bernhard

Aufsatztitel:
Vorwort

Jahrgang:
32
Heft:
01 (2008)
Seiten:
5-6

Abstract:
Mit der rasanten Entwicklung einer ganzen Familie von XML-basierten texttechnologischen Standards seit mehr als zehn Jahren ist die Hoffnung verbunden, dass dadurch die technische Handhabung von linguistischen Daten enorm erleichtert werde und computerlinguistische Projekte davon in der Praxis stark profitierten. Dies um so mehr, als auch die meisten etablierten Datenbankmanagementsysteme (DBMS), Open Source-Lösungen sowie fachspezifische und/oder akademische Datenbankprodukte das Potential strukturorientierter Auszeichnungssprachen erkannt und Schritt für Schritt integriert haben. XMLStandards für die Modellierung der Mikrostruktur informationeller Angebote sowie für die Kommunikation mit externen Applikationen, kombiniert mit den bekannten Vorzügen einer datenbankgetützten Verwaltung und Weiterverarbeitung - aus dieser Perspektive heraus betrachtet sollte sich die computerlinguistische Forschung zukünftig vermehrt um das „Was?“ und immer weniger um das oft leidige „Wie?“ kümmern dürfen. In der Tat gibt es für viele Probleme, die vor einigen Jahren noch umfangreiche Spezialentwicklungen erforderten, inzwischen sehr hilfreiche Standard-Werkzeuge. Doch gilt dies durchgängig? Und welche Überlegungen sind vor Implementierungsentscheidungen nach wie vor unerlässlich? Hier soll der interdisziplinäre GLDV-Workshop „Datenbanktechnologien für hypermediale linguistische Anwendungen“, der im Rahmen der Konvens 2008 in Berlin stattfindet, für mehr Klarheit sorgen, indem er Erfahrungen mit hypermedialen Datenbanken aus unterschiedlichen texttechnologischen Projekten zusammenbringt. Die Vorträge und Systemdemonstrationen des Workshops liegen den Artikeln dieses Heftes zugrunde. Die Beiträge zeigen deutlich, dass nach wie vor nicht alle Werkzeuge alle naheliegenden Erwartungen erfüllen und jedes Projekt eine sorgfältige Bestandsaufnahme der verfügbaren technischen Lösungen erfordert. Die Berichte kommen aus höchst unterschiedlichen Anwendungsbereichen. Das Spektrum umfasst beispielsweise sowohl die Verwaltung unterschiedlicher Annotationsformen für geschrieben - wie für gesprochensprachliche Korpora; weiterhin werden Schnittstellen zu Online-Wörtebüchern und Wikis sowie die Verwaltung fachterminologischer Ontologien thematisiert.

Choosing an XML database for linguistically annotated corpora

Autor:
Eckart, Richard

Aufsatztitel:
Choosing an XML database for linguistically annotated corpora

Jahrgang:
32
Heft:
01 (2008)
Seiten:
7-22

Abstract:
XML has become the de-facto standard for representing linguistically annotated corpora. It seems safe to assume that storing and querying an XML-encoded, annotated corpus in an XML database is a straightforward procedure. In reality, however, it is not. This article aims to provide guidelines for deciding whether to use an XML database and how to choose a suitable product. To this end we examine the following questions: Which aspects should be considered before choosing to store an XML-encoded annotated corpus in an XML database? Which facilities does a database need to provide in order to be suitable for storing and querying annotated corpora? Do current XML databases offer these facilities, and, if not, can they be added?

XML Schema driven Database Management of Speech Corpus Metadata

Autor:
Gasch, Joachim

Aufsatztitel:
XML Schema driven Database Management of Speech Corpus Metadata

Jahrgang:
32
Heft:
01 (2008)
Seiten:
23-33

Abstract:
Electronic speech corpora need to bring together several heterogeneous data formats like audio and video data, corpus-, event- and speaker documentation and time aligned media annotations. The metadata management system has to drive data capture, XML native database storage, dynamic publishing and information retrieval processes. This article describes an XML schema based standardization approach where metadata (documentation and annotation information) of different speech corpora is centrally validated and natively stored within an object-relational XML database.

E-VALBU: Advanced SQL/XML processing of dictionary data using an object-relational XML database

Autor:
Schneider, Roman

Aufsatztitel:
E-VALBU: Advanced SQL/XML processing of dictionary data using an object-relational XML database

Jahrgang:
32
Heft:
01 (2008)
Seiten:
33-44

Abstract:
Contemporary practical lexicography uses a wide range of advanced technological aids, most prominently database systems for the administration of dictionary content. Since XML has become a de facto standard for the coding of lexicographic articles, integrated markup functionality – such as query, update, or transformation of instances – is of particular importance. Even the multi-channel distribution of dictionary data benefits from powerful XML database services. Exemplified by E-VALBU, the most comprehensive electronic dictionary on German verb valency, we outline an integrated approach for advanced XML storing and processing within an object-relational database, and for a public retrieval frontend using Web Services and AJAX technology.

WikiDB: Building Interoperable Wiki-Based Knowledge Resources for Semantic Databases

Autor:
Mehler, Alexander & Gleim, Rüdiger & Ernst, Alexandra & Waltinger, Ulli

Aufsatztitel:
WikiDB: Building Interoperable Wiki-Based Knowledge Resources for Semantic Databases

Jahrgang:
32
Heft:
01 (2008)
Seiten:
47-70

Abstract:
This article describes an API for exploring the logical document and the logical network structure of wikis. It introduces an algorithm for the semantic preprocessing, filtering and typing of these building blocks. Further, this article models the process of wiki generation based on a unified format of syntactic, semantic and pragmatic representations. This three-level approach to make accessible syntactic, semantic and pragmatic aspects of wiki-based structure formation is complemented by a corresponding database model – called WikiDB – and an API operating thereon. Finally, the article provides an empirical study of using the three-fold representation format in conjunction with WikiDB.

Database-driven Access to Heterogeneous XML-Contents Using Domain Ontology of German Grammar

Autor:
Sejane, Ineta

Aufsatztitel:
Database-driven Access to Heterogeneous XML-Contents Using Domain Ontology of German Grammar

Jahrgang:
32
Heft:
01 (2008)
Seiten:
71-87

Abstract:
The present article features the Ontology of German Grammar, one of six information modules in the Web Information System on German Grammar known as GRAMMIS. The ontology module was developed to enhance information retrieval and to inform the user about terminological issues in German grammar. It consists of a knowledge base and an interface with special functions that facilitate both information search and retrieval. The knowledge base itself consolidates – as much as possible – the heterogeneous, nonstandardized

grammatical terminologies used in texts on German grammar and gives general information about grammatical concepts. The interface provides means (1) to navigate the ontology and learn more about the categories and vocabulary of grammar; (2) to retrieve texts from all GRAMMIS modules at once; and (3) either to expand or narrow the query according to user needs and the amount of information found in GRAMMIS. The ontology is implemented in an object-relational database management system, thus ensuring its optimal interaction with other GRAMMIS resources mostly encoded in XML. On the other hand, its current structure allows transformation to conventional ontology as well as thesaurus formats if needed.