

WebStand Project

Theme

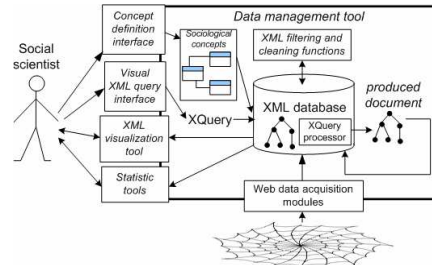
Create a customizable application platform to simplify the work of sociologists working with Web Data (web pages and emails)

Specific sociological application

Analysis of the W3C Standardization process, through the analysis of mailing lists over several years

Features

- Modular, customizable application
- Development of pluggable services libraries (Email acquisition and analysis, CV extraction, etc.)
- Persistent, temporal, P2P source centric XML store
- Simple graphical XQuery language
- Compatible with various formats used in sociology, thanks to XML

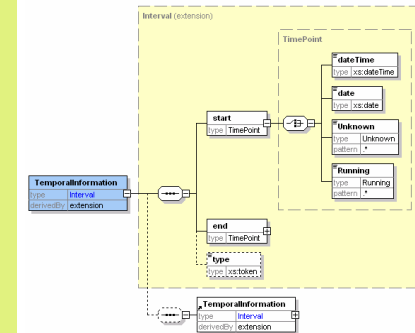


Context

Any XML element in the database can be temporalized, through the use of temporal annotations (Allen intervals), defined according to an XSD.

These annotations can themselves be annotated.

Information stored is monotonous. We never update facts, we simply add new facts with different temporal intervals, to guarantee we are able to reconstruct at any point in time the knowledge situation at any previous date.



Example facts :

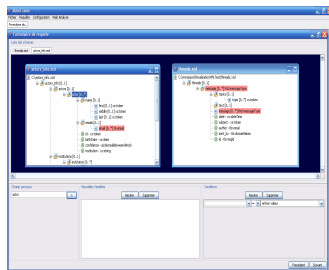
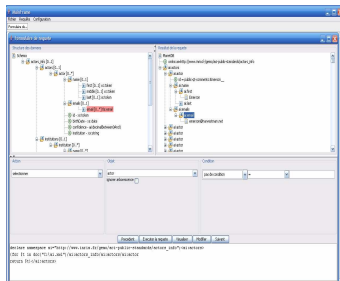
- In 2002 we learn that John Doe joined XML Corp. as CEO on 1/1/2001
- In 2007 we learn that John Doe left XML Corp. on 31/1/2003

Example Queries

- Who did John Doe's work for in 2004 ?
- In 2005, who did we think John Doe worked for in 2004 ?

<WEB

stand />



Integrated platform

- XML and XSD editor
- Easy MonetDB-XQuery database management
- Two forms QBE-like designed to meet sociologists' needs: Simple form to build primary queries, advanced form to construct FLOWR
- Collaborative work through Web service interface



Project Funding: ANR JCJC 05

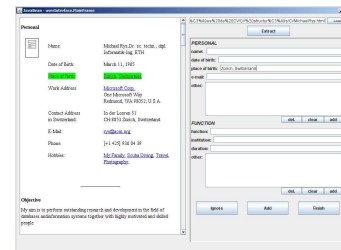
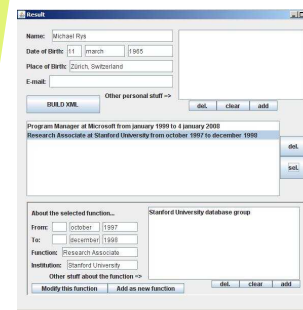
Participants : INRIA-Futurs, U. Versailles, U. Paris XI, U. Aix-Marseille, U. Paris- Dauphine, Ecole des Mines de Nancy, Telecom Paris



Sample Extraction Module : CV Crawler



- Automatically crawls the Web, powered by eXalead search engine
- Semi automatically fills in predefined form using crawled information
- Generates XML documents following a a given schema
- Results are integrated into the database



User Interface

CV Crawler