

Knowledge acquisition, organization and maintenance for heterogeneous information resources



Nguyen G., Laclavík M., Babík M., Gatíal E., Ciglan M., Balogh Z., Oravec V., Hluchy L.

giang.ui@savba.sk, laclavik.ui@savba.sk, <http://ikt.ui.sav.sk/>, <http://nazou.fiit.stuba.sk/>

Institute of Informatics, Slovak Academy of Science, Dúbravská cesta 9, 845 07 Bratislava, Slovakia

Motivation

acquire information from public sources and transform the information to knowledge structures

Methodology

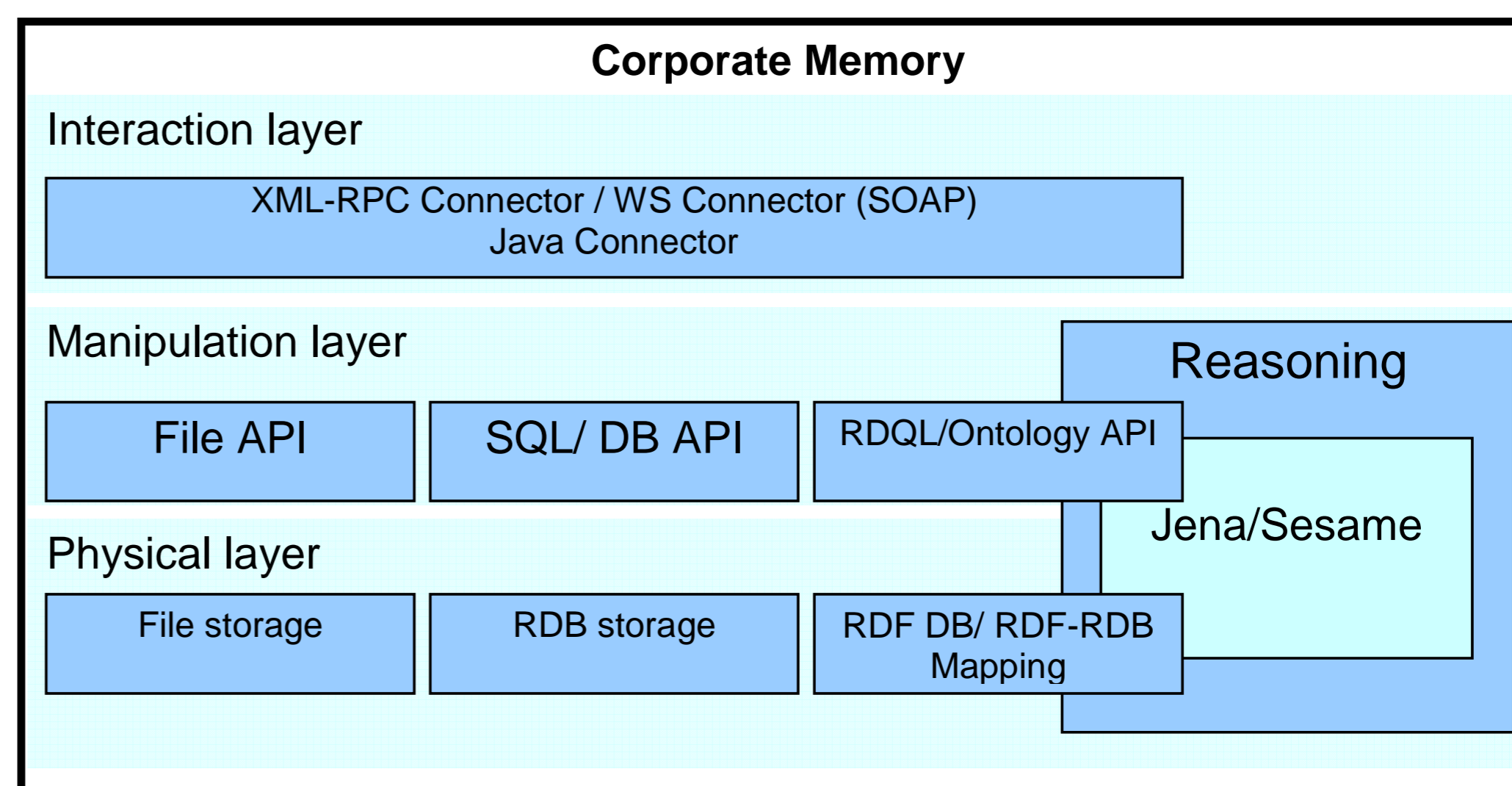
Tools for knowledge acquisition, organization and maintenance; different tools require data in different form

Corporate Memory

framework for integration of independent tools at the data level to ensure interoperability

Data in Corporate Memory

- Files
- Relational Database
- Semantic Metadata, Ontologies



Used Technology

- RDF, OWL, SPARQL
- XML, XML-RPC
- Java

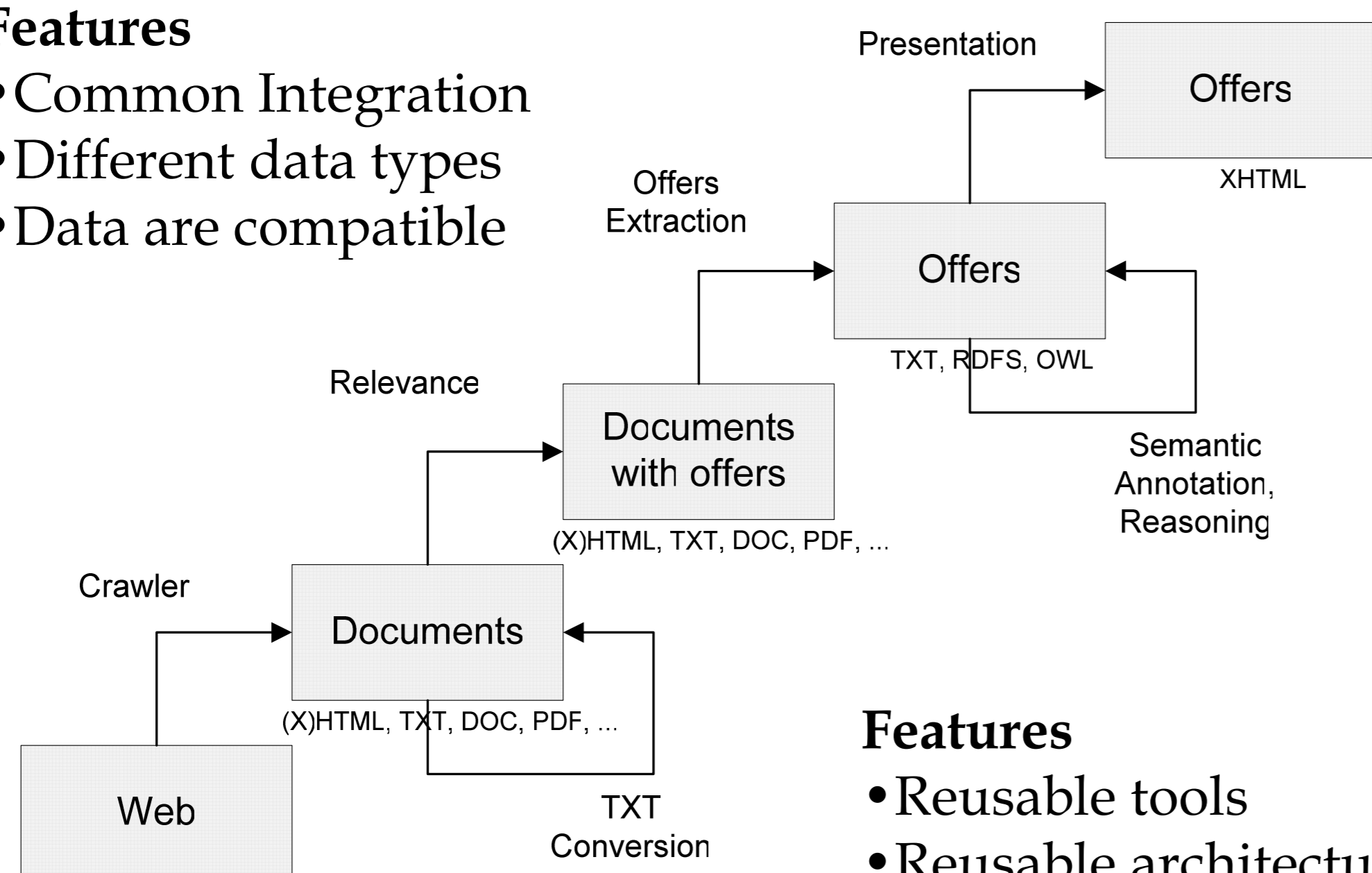
Used Tools

- Sesame
- Jena
- OGSA-DAI
- MySQL



Features

- Common Integration
- Different data types
- Data are compatible



Features

- Reusable tools
- Reusable architecture

Data Transformation Chain

- **RIDAR** (Relevant Internet Data Resource Identification) connects to existing search engines and identify relevant web resources
- **WebCrawler** and **ERID** (Estimate Relevance of Internet Documents) recursively explore web resources and store
- **DocConverter** transforms documents to TXT format.
- **ExPoS** (Offer Extraction) and **OSID** (Offer Separation for Internet Documents) extract offers (e.g. job offers) from document. If there is more offers on one document, or if there is only one it select offer without header, footer, menu or banners.
- **DaiDocIndexing**, **DaiDocSearch**, **JDBSearch** index text documents and offers; this allow other tools (searching, clustering) to use indexes for further processing.
- **Ontea** (Ontology based text annotation) annotates text version of offers by ontology individuals which are detected via regular expressions as relevant semantic properties of the offer. Ontea thus create offers ontology metadata from offer document version according to domain ontology.
- Tools **Prescott** and **faceted browser** support presentation, which transforms ontological data to XML and XML is further transformed to HTML via XSL. Indexes or found clusters are also used by presentation to search, categorize and navigate in offers accumulated in CM.

