

Paragon - semantic web software under development

Author: Christoph Pingel

Introduction/Purpose

Purpose

An open source implementation of an ontology-based, interoperable knowledge management for the arts and related fields, including a 'personal workspace' to store one's own 'facts'.

Includes modules for integration with Personal Information Management (PIM) and Web Content Management (CMS) as well as intermediate layers for visualisation data and database mappers.

Motivation

In the art world, there is no agreed-upon standard for semantically sound data exchange that is easy to implement and use.

Rich ontologies with thousands of entries are usually not available as open, shareable knowledge.

There is also little support for curators or artists who wish to do their content management (for artistic, publishing or management purposes) based on an ontology. While wikis and weblogs and simple syndication standards are just beginning to be used by some early adopters, these tools, promising as they may be, still lack the necessary interoperability on the semantic level, that is, beyond exchange of URLs, keywords, and tracebacks.

From the standpoint of open source/open access it's a good thing to have an implementation of state-of-the-art, standard compliant tools (OWL/reasoning engine), for several reasons:

The semantic web is not left to large companies to define and explore.

Sharing content and/or structure on a 'win-win' basis could provide a model for other communities as well.

In the long run, open source ontology tools could provide a basis for 'accumulative' knowledge that remains in the public domain (like, say, citation history graphs that are very useful for students, but usually only available as a commercial service)

Good semantical data could provide a playing field for people who want to experiment with new visualisation techniques, editing interfaces, data mining or knowledge discovery technologies etc.

It can be seen as an invitation to the arts sphere for experimentation with the data.

There are also research interests in such a system, e.g in the fields of IT, interface design methodology,

cognitive science, culture theory etc. Some examples:

How do we use advanced visualisation technologies for displaying large data sets without overwhelming the user? (Trees, semi-lattices, constellations, Kohonen maps, etc.)

How 'intuitive' can an interface for semantic data be?

How can users be supported in building ontologies by data mining and similar approaches?

What is the relationship between mnemotechnical visualizations that for example the scholastics used (cf. Ivan Illich: 'Im Weinberg des Textes') and 3- or more-dimensional representations of 'ontological' spaces? Are there ways to 'exploit' the intricate relationships between media formats and their modes of use? What about interface 'metaphors' or conceptual blends? How helpful and/or misleading can they be given an ontology as basis?

What is there to learn for user interface design if we assume a 'fluent' surface where browsing, exploring, and editing (cut/paste, rearrange, align, set up new relations, etc.) happen on the same screen? Is it to add just another layer to what Manovich has called a 'mixture' of more tradition paradigms (cinema, algorithms, book)? Or is there a useful and sensible 'meta' approach to all of these activities?

What we will see: We will see blurring boundaries between Personal Information Management, Project Management and Web Content Management. A blend of social computing and semantic web. The project wants to prefigure the outcome of this trend according to the values of open source/open access.

Strategy

- 1 - Acquire some data to play with,
- 2 - build the database,
- 3 - make the semantic links in the data explicit by
- 4 - building an ontology
- 5 - explore data mining, knowledge discovery etc. to help with ontology building/filling database
- 6 - find partners for implementation/use/requirements
- 7 - develop methodology to link semantic data spaces
- 8 - implement modules for communication (XML-RPC, OWL, XSLT, .dot)
- 9 - build better visualization tools
- 10 - start all over, now doing everything at once

Outline of the System

Core system

The core system is the so-called 'Service Provider', a server including a database with a certain amount of collected 'knowledge', reasoning engine, some business logic, web services and a portal web site. Typically, such a system could be hosted by an at least moderately large institution or initiative with a certain interest to entertain a portal.

The core system can also be run as a so called 'Data Provider' but without portal functions or providing services to users from outside. Institutions could use it as a general purpose database system with translators e.g. for the web content management system. Yet still, from the outside it could be used as a 'data provider', that is, it is always possible to obtain a OWL representation of say the current exhibitions and event schedules for further use (e.g. by the local information systems to incorporate the dates into their information systems).

Extensions

Generally, there are two ways to extend the system:

Using interfaces to existing web services (could be set up on the same machine for that purpose) like ConceptNet, Amazon/Alexa, Google and several others that will appear in the coming years.

Incorporating modules into the 'Service Provider' software that support knowledge discovery (e.g. crawling and association rule learning), relevance evaluation (giving the relations a weight) or visualisation services (or whatever seems useful)

Architecture Overview

Database

The database is organized according to principles outlined by Polygon and similar initiatives: Objects and attributes are stored in different, very narrow tables that contain only the barebones information and some administrative data. Objects are connected with other objects and attributes by means of special link tables. Where appropriate, the possible kinds of relations and attribute names are restricted by a thesaurus.

Business Logic

We consider everything to be 'business logic' that serves to format database output for a certain purpose. Since the database is not in any way restricted as far as linking between the items is concerned, it is necessary

to restrict output if it is to be used for a certain purpose.

The business logic for an exhibition for example would restrict db output to the exhibition, institution, curators, topic etc, but *exclude* any links from the artists to things that are not themselves linked to the exhibition or the artist. That is, for example, we include exhibitions that appear in the artist's exhibition history, but we exclude other artists that appeared in the group exhibition our artist participated in.

The business logic of the documentation of a curatorial process excludes everything besides the curator's entries and the artists and topics they deal with.

Typically, it is business logic's task to take a certain definition and provide an OWL representation of it.

The business logic also deals with exports to other formats and data import via database mappers.

Interoperability / Web Services

Using the business logic as an abstraction layer, it will not be a problem to implement whatever web services will be required, be it XML-RPC (preferred), SOAP, REST or even WAP.

Although we can, in a sense, 'afford' net connection problems since every service provider is a 'stand alone' system, we need a model of how to deal with contradictory facts during import. According to the rules, some facts may be forbidden if contradictory.

Semantics

An evolving art and humanities ontology will be at the same time guideline and outcome of the work. By using semantic grounding, several things become possible or easier to do:

- map different databases/models onto a common model
- keep the intellectual work that usually goes into any structured model of a domain, store it and make it available for later use
- makes it easier and quicker to define and implement output translators
- makes it easier and quicker to set up search systems (browser or API based)

Data Mining and Knowledge Discovery

Besides interoperability, that is perhaps the technically most challenging aspect of the project. In recent years, extensive research has been done, and some of the results are available as open source or free for non-commercial use. It may depend on the application domain which among the many models/methods (data-mining, knowledge discovery, self-organizing maps etc.) is to be applied.

Tools

Apache: open source implementation of HTTP

Python: widely used programming language, easy to learn and use, necessary library support is there: xmlrpc, sax, orange, mod_python (among others that may be useful)

MySQL: Relational Database Management System

Frontier: formerly closed source, now open source cross platform scripting environment, will soon run Python additionally to Usertalk and AppleScript. It has a hierarchical object database and makes use of outlines extensively. Since there is a visual interface for the object database, it's a very good prototyping platform, esp. since it will run Python soon.

ConceptNet: free (for non-commercial use) natural language processing library written in Python, supports emotion sensing, concept extraction and other very useful stuff. ConceptNet is based on a large natural language vocabulary and predicate database. Can also be run as stand-alone XML-RPC server.

orange: open source data mining and knowledge discovery library for Python written in C/C++

Implementations

iconoclash exhibition browser

Flash applet reads XML-files

caching (xml once read is kept in local storage)

server based Flash->QTVR and QTVR->Flash linking mechanism

Preliminary 'core system' implementation

- classes for web access, (simple) datamining, and import from other databases are there
- ontology is evolving
- database is set up, currently containing about 60.000 things and 13.000 persons and groups.
- will be accessible through a minimal user interface (like del.icio.us) from January 2005 on.

Implementation Ideas

As a further 'proof of concept', I'd like to do some projects with a focus on non-mainstream topics and a narrow, but very fine-grained set of data from the domain (which I know quite well).

The Electric Miles Davis

Paul Tingen wrote a wonderful book about the 'electric' Miles Davis from 1967-1992. In the 'data' part of the book, all recording sessions, participating musicians, released albums etc. are listed in detail. It would be

wonderful to have an 'electric Miles Davis ontology' that incorporates this information and additional stuff (like influences, private events, producers, etc.) and an intelligent graphical display thereof.

The Mental Spaces / Conceptual Integration Portal

One of the very promising fields in cognitive science is defined by Gilles Fauconnier and Mark Turner, with a little help from their friends like Lakoff/Johnson, Eve Sweetser, etc.

It would be interesting to have a relatively fine-grained, evolving ontology of their research that could be used at the same time as a 'graphical research database' and an entry point for the interested public.

It would contain seminal case studies, Lakoffs database of conceptual metaphor, political subjects, links to related projects, an ever evolving graph of the different fields of cognitive science, cognitivism, AI, AL, bottom-up robotics, etc.

Vision

- Seamless integration of knowledge bases from heterogenous sources
- An (open source) licensing system for binary content (pictures, movies, audio) (Creative Commons?)
- Blurring the boundaries between the spaces of personal annotation, knowledge management and web content management
- Adequate toolset for 'visual knowledge management' beyond trees and tables
- Establishment of a 'cognition theory of the networked workspace', probably following Hutchins' paradigm of 'material anchors for conceptual blends'

Related Projects and Standards

AIFB/ontoprise (ontologies, knowledge discovery / knowledge management)

The AIFB at Karlsruhe university and its spin-off 'ontoprise' currently concentrate on business applications. Closed source, commercial projects. No art applications, no cognitive science (metaphors, conceptual blending etc.), no 'playground'

Chandler (PIM)

Chandler by Mitch Kapor's 'Open Software Foundation' is an attempt, according to Kapor, to fuse the positive aspects of commercial desktop application development (usability, pro design) with the positive aspects of open source. The idea is to build a Personal Information Manager (PIM) that is built around a calendar, IMAP email client and contact list but will eventually include other data sources like RSS etc.

Medienkunstnetz

Medienkunstnetz is an educational project by ZKM and Hochschule für Grafik und Buchkunst Leipzig that

aims to cover a large set of work, artists and topics.

OWL (modeling, semantics)

OWL, the Web Ontology Language is the W3C standard for the semantic web that is supported by more and more ontology tools and initiatives

plumbdesign (e.g. visual thesaurus)

Plumbdesign is a US company that makes excellent visualization software for large data sets. While some design issues remain, by and large the set a high standard for appealing visualisations of random tree oder lattice structures.

Polygon (database architecture)

Polygenesys is a company from Munich that developed the database aided relation management tool Polygon

Protegé

Protegé is a set of ontology development tools (with possible database backend) developed at Stanford University written in Java.

v2_archive

The v2_archive is an ontology-based archive/ web portal of the 20-year history of the Rotterdam 'medialab for unstable media'.

Christoph Pingel, December 1, 2004