Co-evolving document collections and knowledge structures

CoDAK

Dr. Evgeny Knutov

(MSc Seminar Nov. 11 2013)
The CoDAK project

CoDAK: Co-evolving Document Collections and Knowledge Structures


www.semontoweb.com
The CoDAK project (cont.)

❖ Problem statement
❖ Research
❖ Evolution of CoDAK
❖ Demos
❖ CoDAK language
❖ New research objectives
Dynamical system, knowledge evolves through new documents, creation of factsheets and further domain expert knowledge.
Ontology and Documents
Evolution of CoDAK

- multiple versions
- scrum approach
- deployment
- user studies
- new endeavours
First versions

- custom development
- uses Frog NLP module
- PDFBox for PDF documents processing
Java tryouts

- pure Java-based implementation
- includes:
  - **Lucene** indexed PDF files (with PDFbox)
  - **Sail** ontology index
  - using Lucene Multi-index search
Drupal-based

- need to manage and create documents (friendly)
- Drupal CMS
- working towards “WorkingSheets”
- uses both internal search index (only for specific nodes) and external Solr
Drupal?

- open-source content management system (framework)
- consists of the core and additional modules (primarily PHP)
- data construction (types, views, versions)
- https://drupal.org/

- not?-out-of-the-box
- want something else to work together - do it yourself
- bunch of CGI
- new research challenges
  - FactSheet
  - UI
Extracting for Drupal

- Alchemy entity extraction
- Drupal taxonomies
- Automatic tagging

- nothing works with Drupal as expected;
- most of the Drupal modules are tailored to fit in specific implementation;
- lots of Ajax was failing in IE
- external API
Analyzing content in Drupal

- automatically get the most relevant information of your report, table, etc. connected to many other documents and people to help you explore the domain;
- identifies people, companies, geo features
- automatically tags analyzed content
- identifies facts and subject-object-action relationships
- topic categorization
Sustainable Water Fund: First call results

Number of applications for the Sustainable Water Fund exceeds all expectations.

On May 7th, the first stage for the tender for the Sustainable Water Fund closed. A total of 61 applications was received. This exceeded even the most optimistic estimates. To the Ministry of Foreign Affairs and NL Agency this shows that there is great momentum for PPPs in the water sector and that the Sustainable Water Fund has the potential to really contribute to water safety and water security in developing countries.

The total requested indicative subsidy from the Fund is EUR 200 million, more than 4 times the available budget for this tender. This means that the Fund has managed to leverage roughly EUR 150 million from water sector actors in the Netherlands and abroad. The financial limitations of the tender, however, will determine how many proposals can be funded. The Ministry of Foreign Affairs and NL Agency will keep the sector informed about this process.

Applications have been received for 27 target countries.

Read [here](#) more about the first call of result.
“Factsheet” generation tryouts

- generate simple reports
  - on a certain pre-defined topic
  - term-based
- using different report templates
- add reports to the document base
“Factsheet” generation tryouts (cont.)

**Report term**

**Template**

**Onderneming**

Een bedrijf is een organisatie van arbeid en kapitaal. Een bedrijf dat gericht is op het maken van winst wordt veelal een onderneming genoemd. Een bedrijf dat tastbare producten genoemt.

Bij commerciële organisaties is er sprake van een onderneming waar risico's worden gelopen en waar het streven naar winst noodzakelijk is voor het voortbestaan. Om hun klanten en aan te spreken met hun diensten, wordt er nauwgezet gelet op de vier -P's: prijs, plaats, product en promotie. Tegenwoordig handelen bedrijven ook steeds vaker natuurlijk Profit. In deze ruimere benadering wordt er ook rekening gehouden met de meerwaarde of kosten voor mens, maatschappij en de omgeving (m.a.w. duurzaam ondrep).
HTML5 + Graphics

http://localhost/codak01/codak_10o.html
HTML5 + Graphics (cont)

- badly scalable search engine (good UI though)
- lots of documents pre-processing (into HTML)
- buggy jOWL package
- IE5,6,7 visualization limitations
- limited set-up
And again..., robust, scalable, simple HTML

- from scratch
- only search, no UI at first
- incrementally adding features
- textual representation for the ontology
- tracking users’ behaviour
First deployed versions based on

- **Solr**, starting from ver. 3.6
- current **Solr** ver. 4.5.1
- **Sesame** - Data Store and SPARQL backend
- using only XMLHttpRequest
- XML data only
- PHP for interacting with the internal user and searches DataBases
- independent of anything
● one of the deployed version
● Demo time
CoDAK language

- Expanded keyword search (use Sesame to construct Solr query)
- locatie*, alles*, etc.
- Ex.
  - gas+Limburg (9 docs)
  - gas locatie*limburg (11)
- Easily extendable, based on keyword-query table, no parsing required
Other version of CoDAK

- Heat and Energy-related
- Separation technology-related
- Map integration
- Plain version (sandbox)
Research opportunities

❖ major:
❖ CoDAK language
❖ Deep search map integration
❖ “tamper” with Google
❖ cluster documents/enhance ontology
❖ minor:
❖ Ontology visualisation/interaction
❖ Analyze user queries to enhance the ontology
❖ Apply in a field of your own