Co-evolving document collections and knowledge structures

CoDAK

Dr. Evgeny Knutov

(MSc Seminar Nov. 11 2013)

The CoDAK project

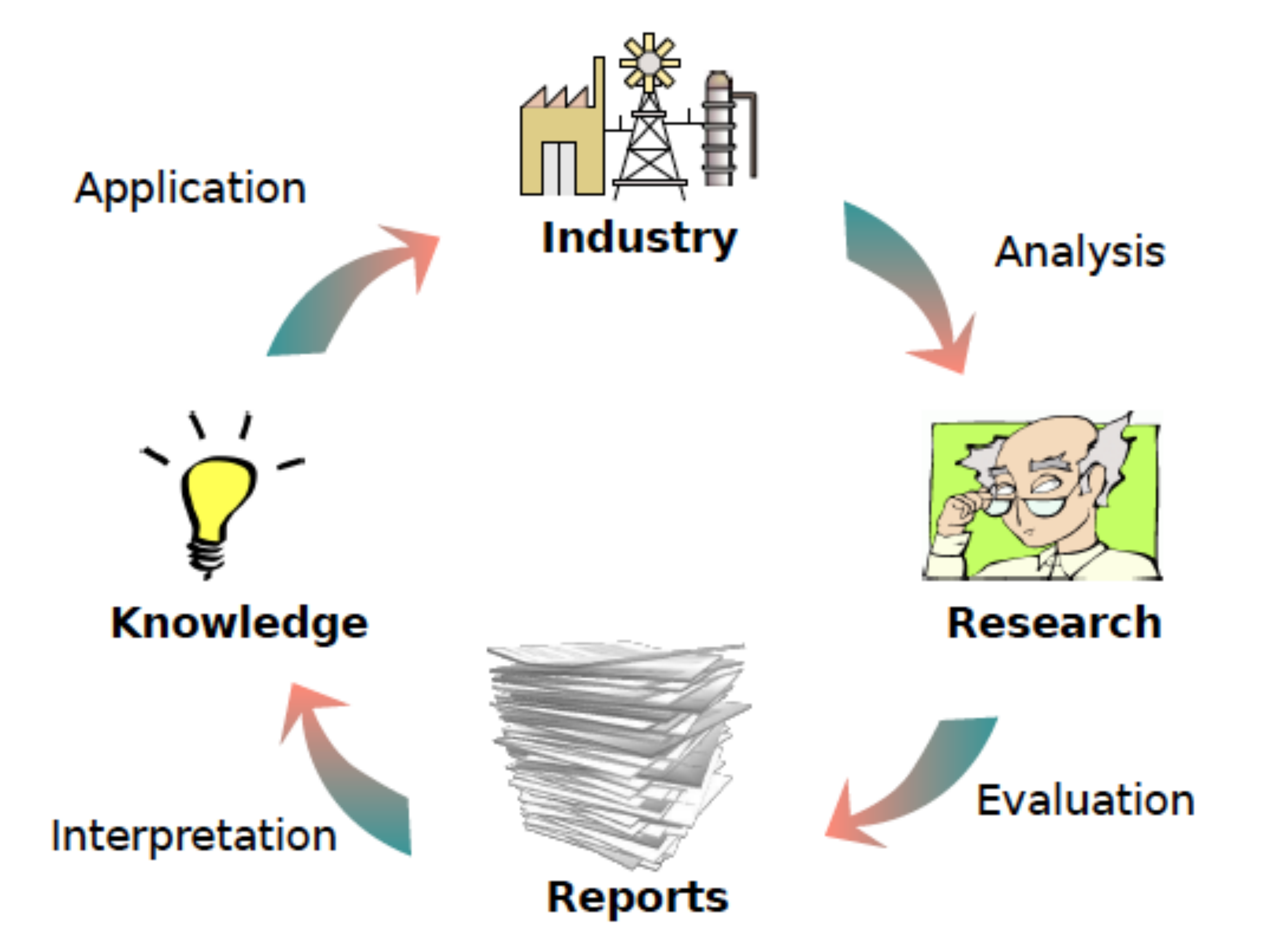
CoDAK: Co-evolving Document Collections and Knowledge Structures

AgentschapNL: dept. Energy & Climate http://www.agentschapnl.nl/

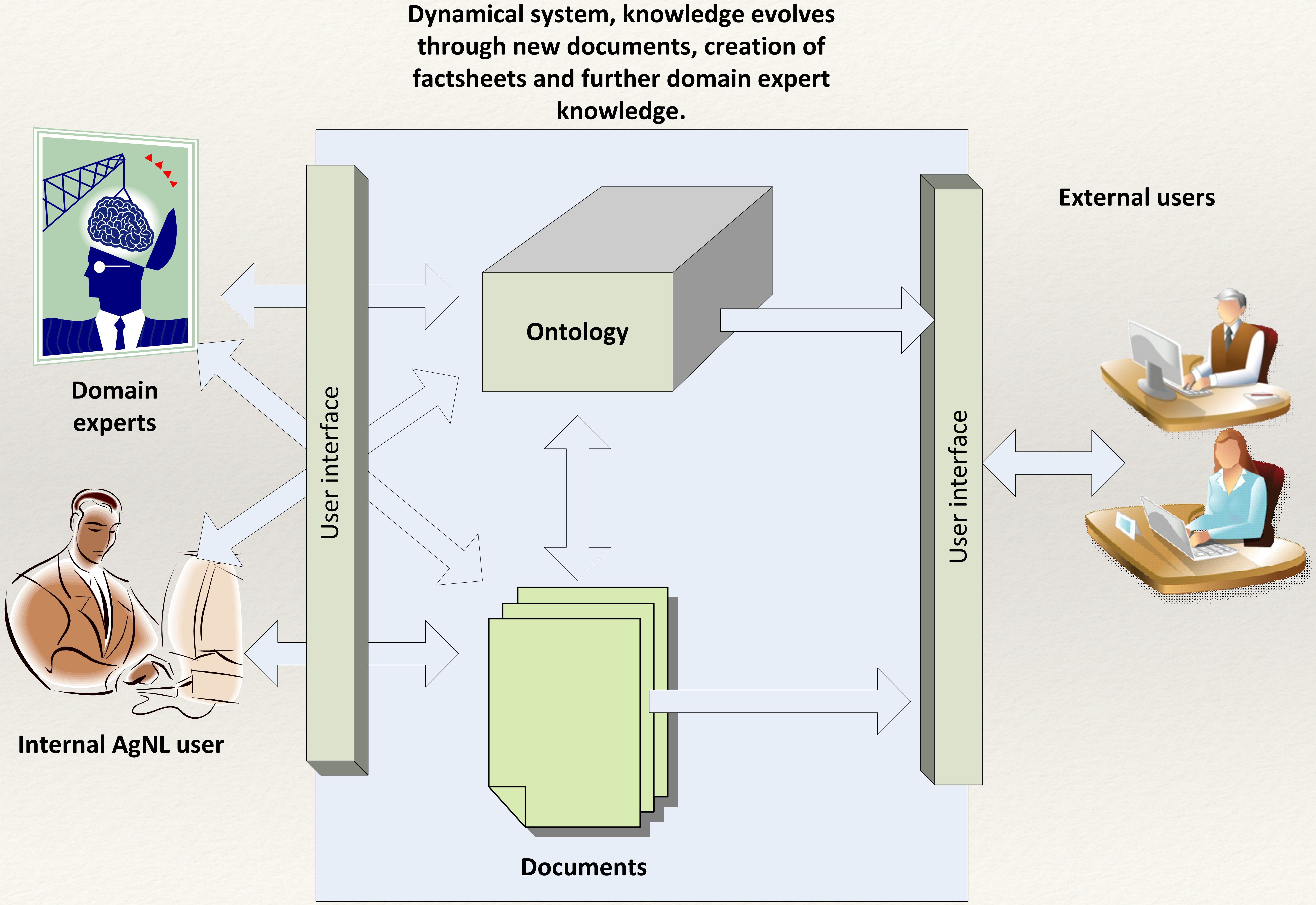
www.semontoweb.com

The CoDAK project (cont.)

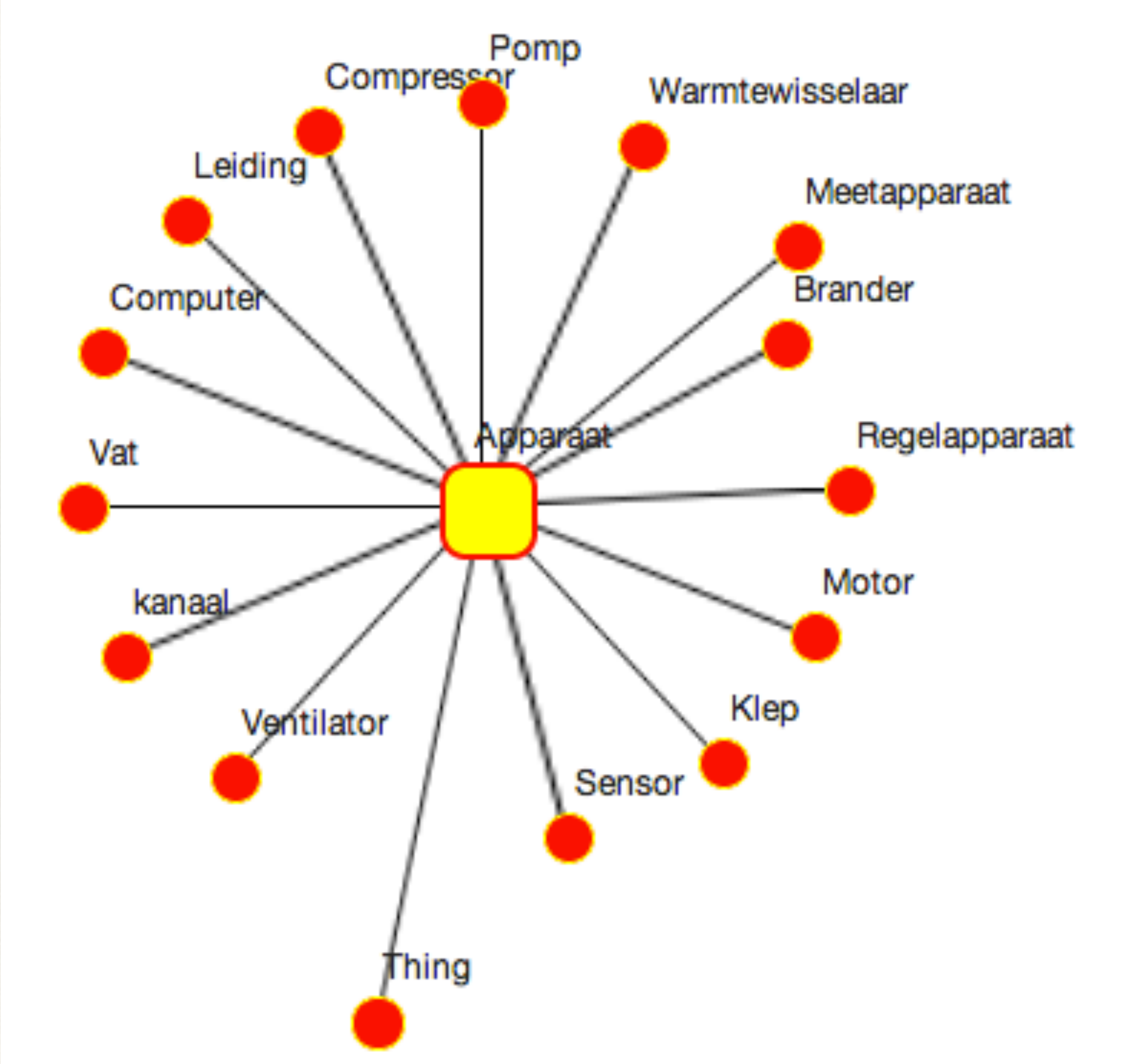
- Problem statement
- Research
- Evolution of CoDAK
- Demos
- CoDAK language
- New research objectives

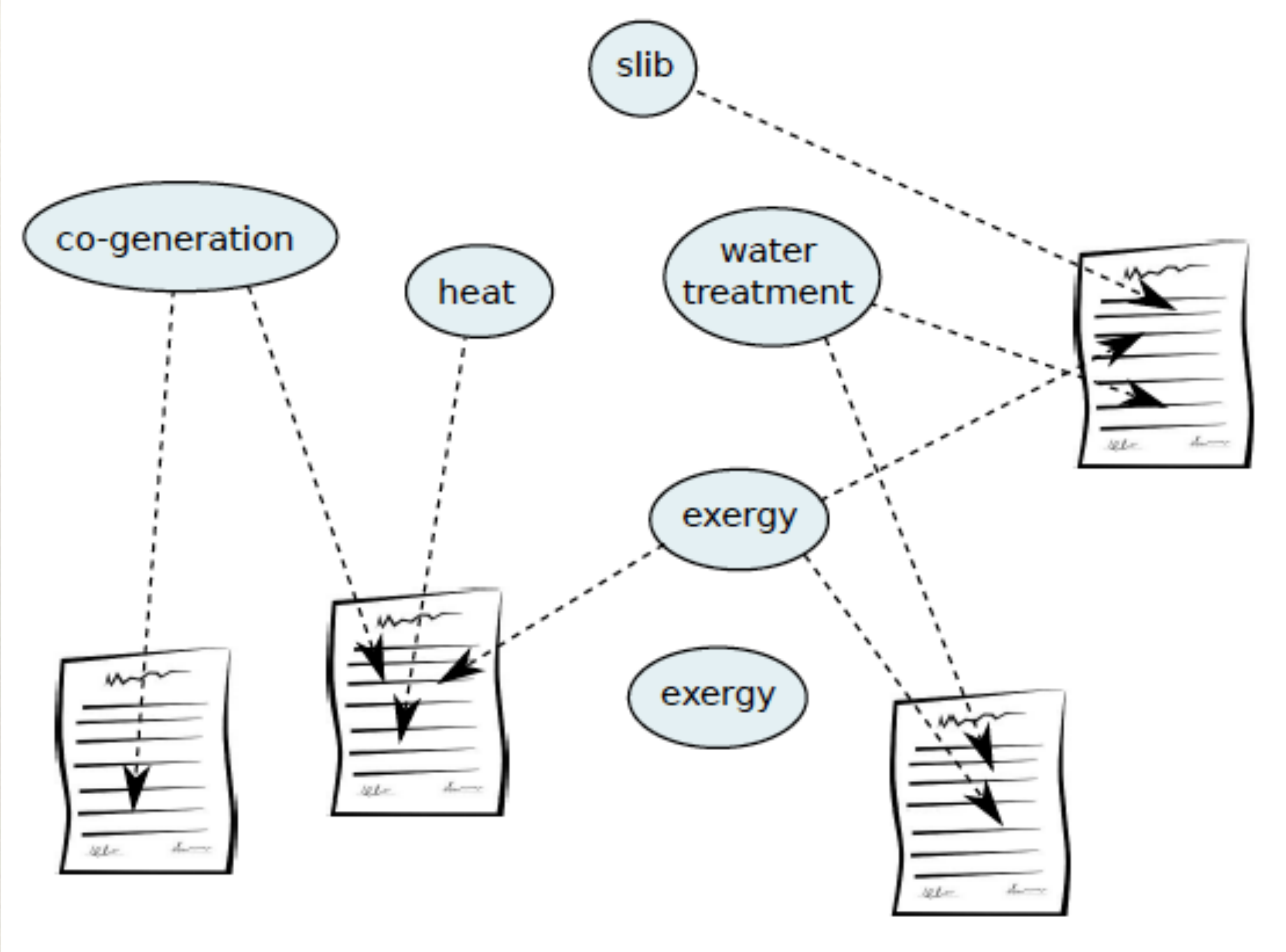


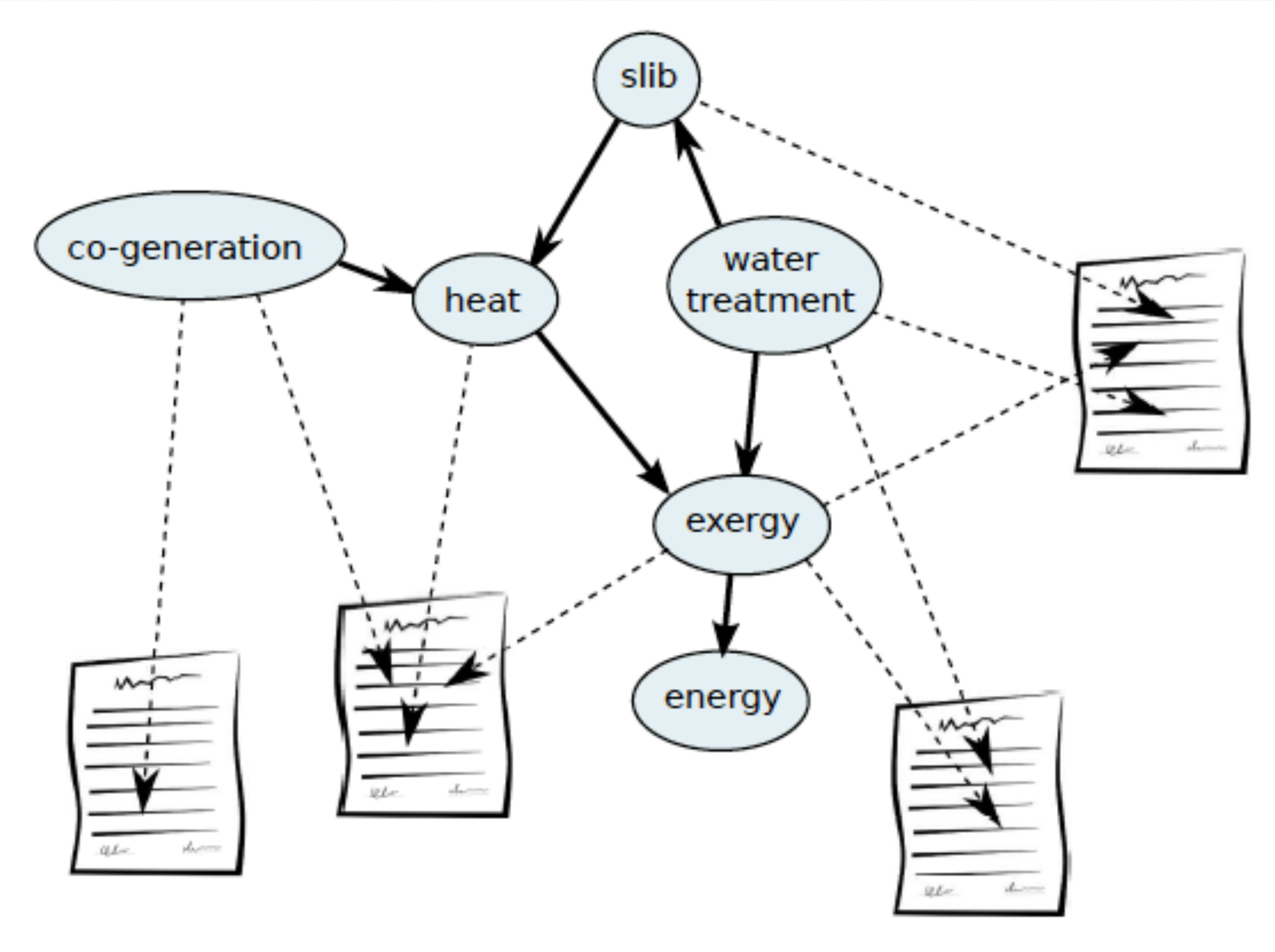
How people at AgNL see the system


Dynamical system, knowledge evolves through new documents, creation of factsheets and further domain expert knowledge.
Domain experts
Internal AgNL user
User interface
Ontology
Documents
User interface
External users

Ontology and Documents

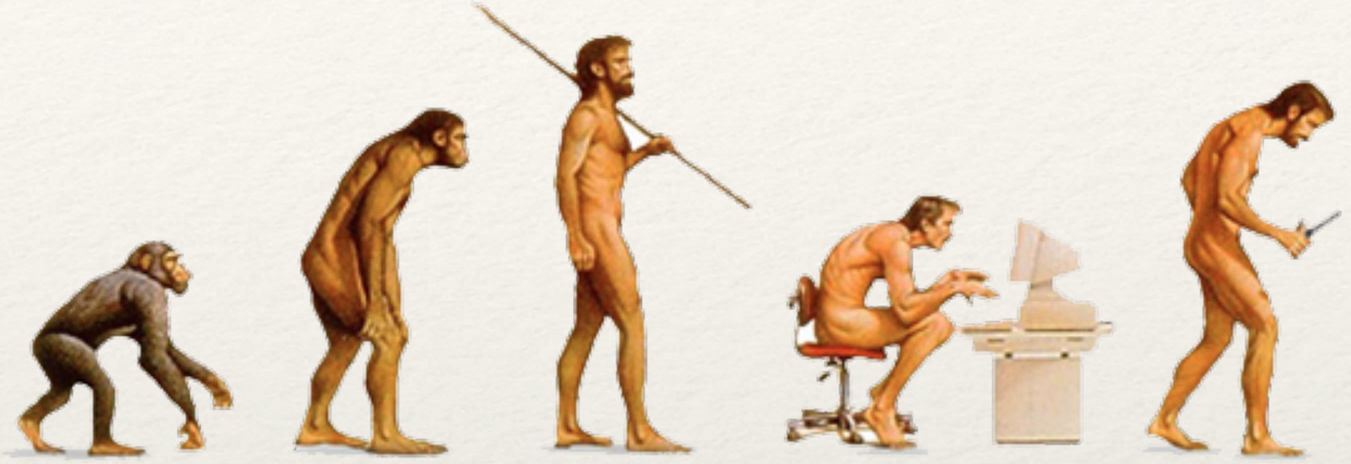






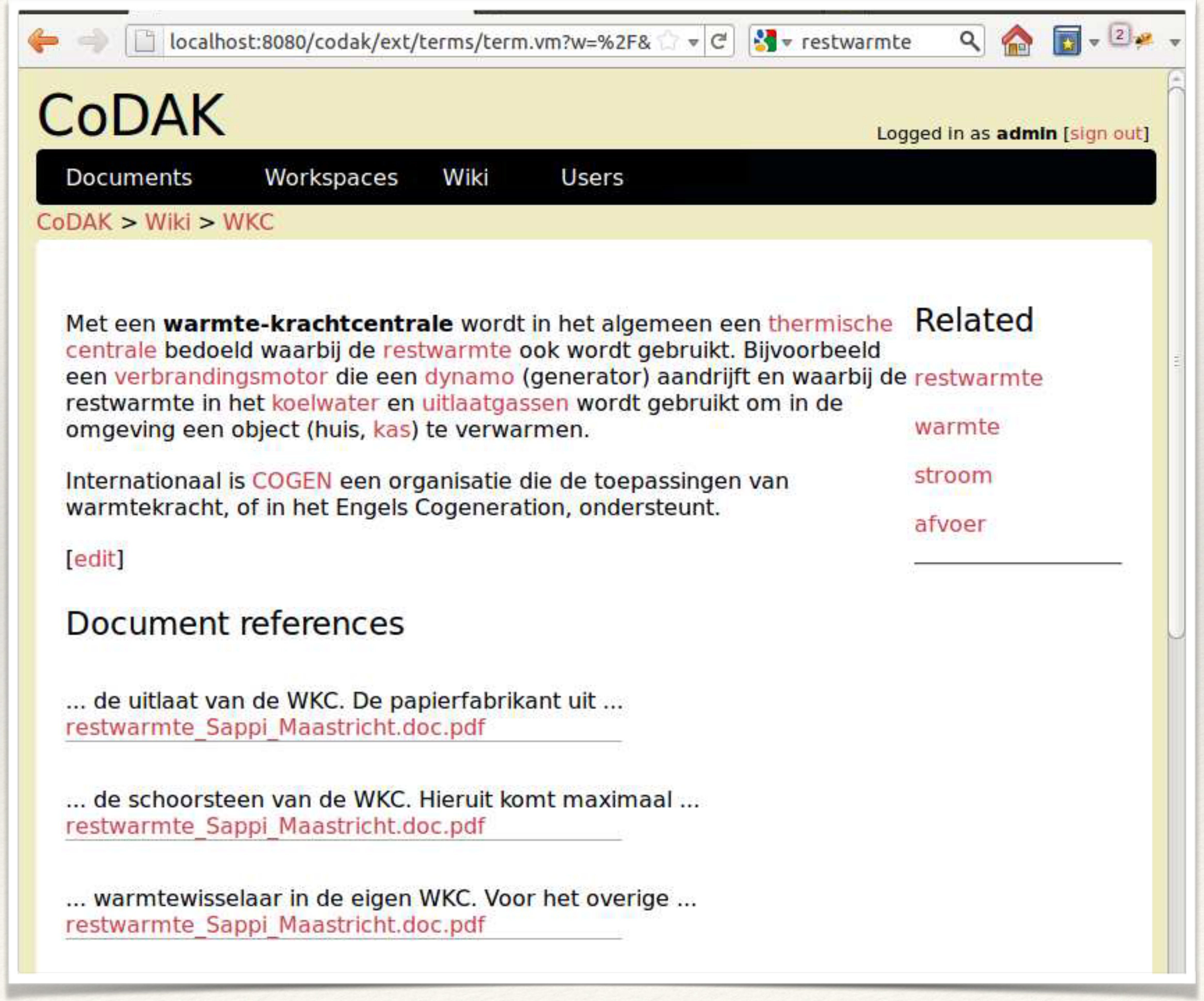
Evolution of CoDAK

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



First versions

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



CoDAK

Logged in as **admin** [sign out]

CoDAK > Wiki > WKC

Met een **warmte-krachtcentrale** wordt in het algemeen een thermische centrale bedoeld waarbij de restwarmte ook wordt gebruikt. Bijvoorbeeld een verbrandingsmotor die een dynamo (generator) aandrijft en waarbij de restwarmte in het koelwater en uitlaatgassen wordt gebruikt om in de omgeving een object (huis, kas) te verwarmen.

Internationaal is COGEN een organisatie die de toepassingen van warmtekracht, of in het Engels Cogeneration, ondersteunt.

[edit]

Related

restwarmte

warmte

stroom

afvoer

Document references

... de uitlaat van de WKC. De papierfabrikant uit ...
restwarmte_Sappi_Maastricht.doc.pdf

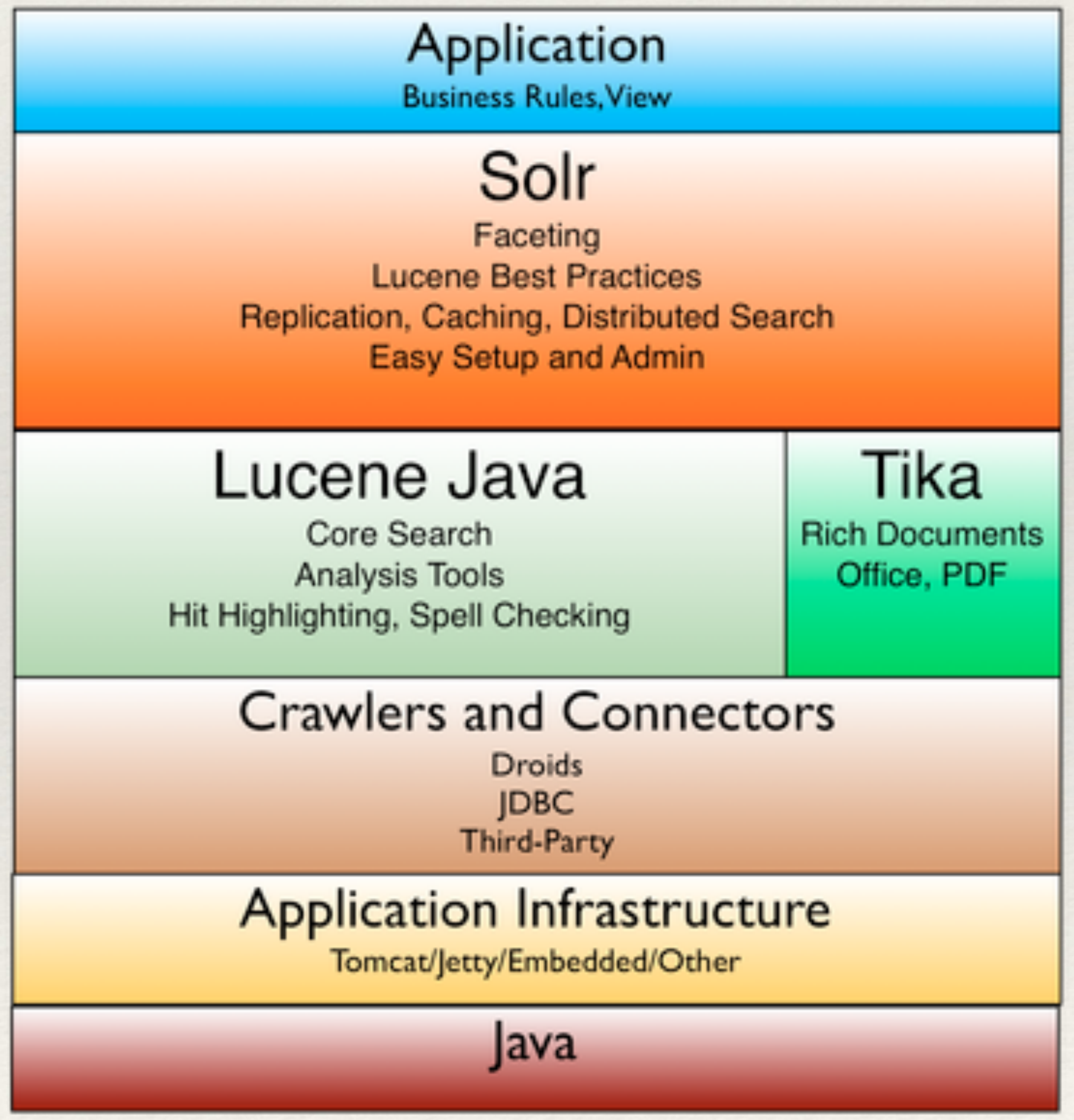
... de schoorsteen van de WKC. Hieruit komt maximaal ...
restwarmte_Sappi_Maastricht.doc.pdf

... warmtewisselaar in de eigen WKC. Voor het overige ...
restwarmte_Sappi_Maastricht.doc.pdf

Java tryouts

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

The Lucene Stack



Application
Business Rules,View

Solr
Faceting
Lucene Best Practices
Replication, Caching, Distributed Search
Easy Setup and Admin

Lucene Java
Core Search
Analysis Tools
Hit Highlighting, Spell Checking

Tika
Rich Documents
Office, PDF

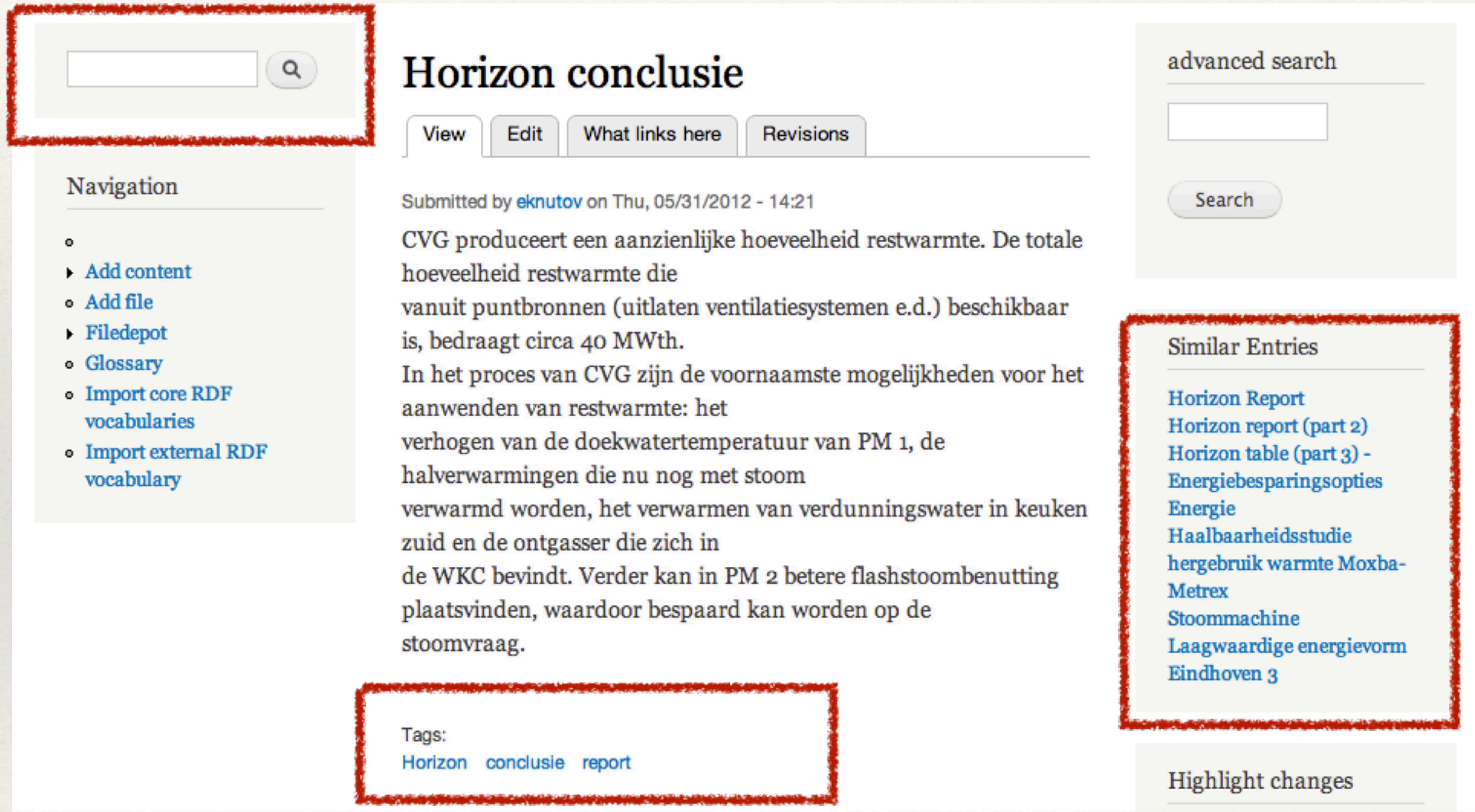
Crawlers and Connectors
Droids
JDBC
Third-Party

Application Infrastructure
Tomcat/Jetty/Embedded/Other

Java

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**



Horizon conclusie

Submitted by eknutov on Thu, 05/31/2012 - 14:21

CVG produceert een aanzienlijke hoeveelheid restwarmte. De totale hoeveelheid restwarmte die
vanuit puntbronnen (uitlaten ventilatiesystemen e.d.) beschikbaar is, bedraagt circa 40 MWth.
In het proces van CVG zijn de voornaamste mogelijkheden voor het aanwenden van restwarmte: het
verhogen van de doekwatertemperatuur van PM 1, de halverwarmingen die nu nog met stoom
verwarmd worden, het verwarmen van verdunningswater in keuken zuid en de ontgasser die zich in
de WKC bevindt. Verder kan in PM 2 betere flashstoombenutting plaatsvinden, waardoor bespaard kan worden op de stoomvraag.

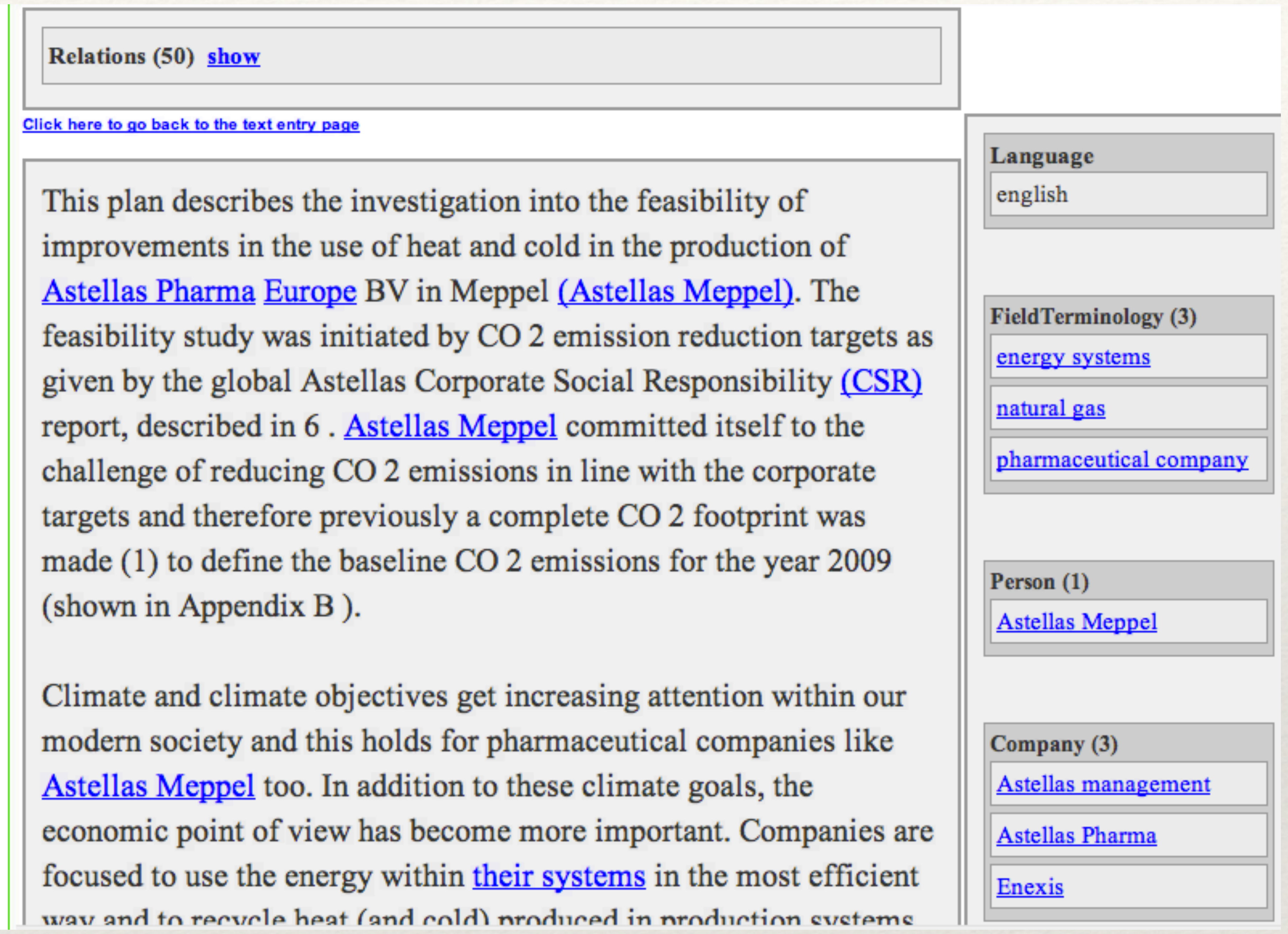
Tags:
Horizon conclusie report

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



Relations (50) show

Click here to go back to the text entry page

This plan describes the investigation into the feasibility of improvements in the use of heat and cold in the production of Astellas Pharma Europe BV in Meppel (Astellas Meppel). The feasibility study was initiated by CO 2 emission reduction targets as given by the global Astellas Corporate Social Responsibility (CSR) report, described in 6 . Astellas Meppel committed itself to the challenge of reducing CO 2 emissions in line with the corporate targets and therefore previously a complete CO 2 footprint was made (1) to define the baseline CO 2 emissions for the year 2009 (shown in Appendix B).

Climate and climate objectives get increasing attention within our modern society and this holds for pharmaceutical companies like Astellas Meppel too. In addition to these climate goals, the economic point of view has become more important. Companies are focused to use the energy within their systems in the most efficient way and to recycle heat (and cold) produced in production systems

Language
english

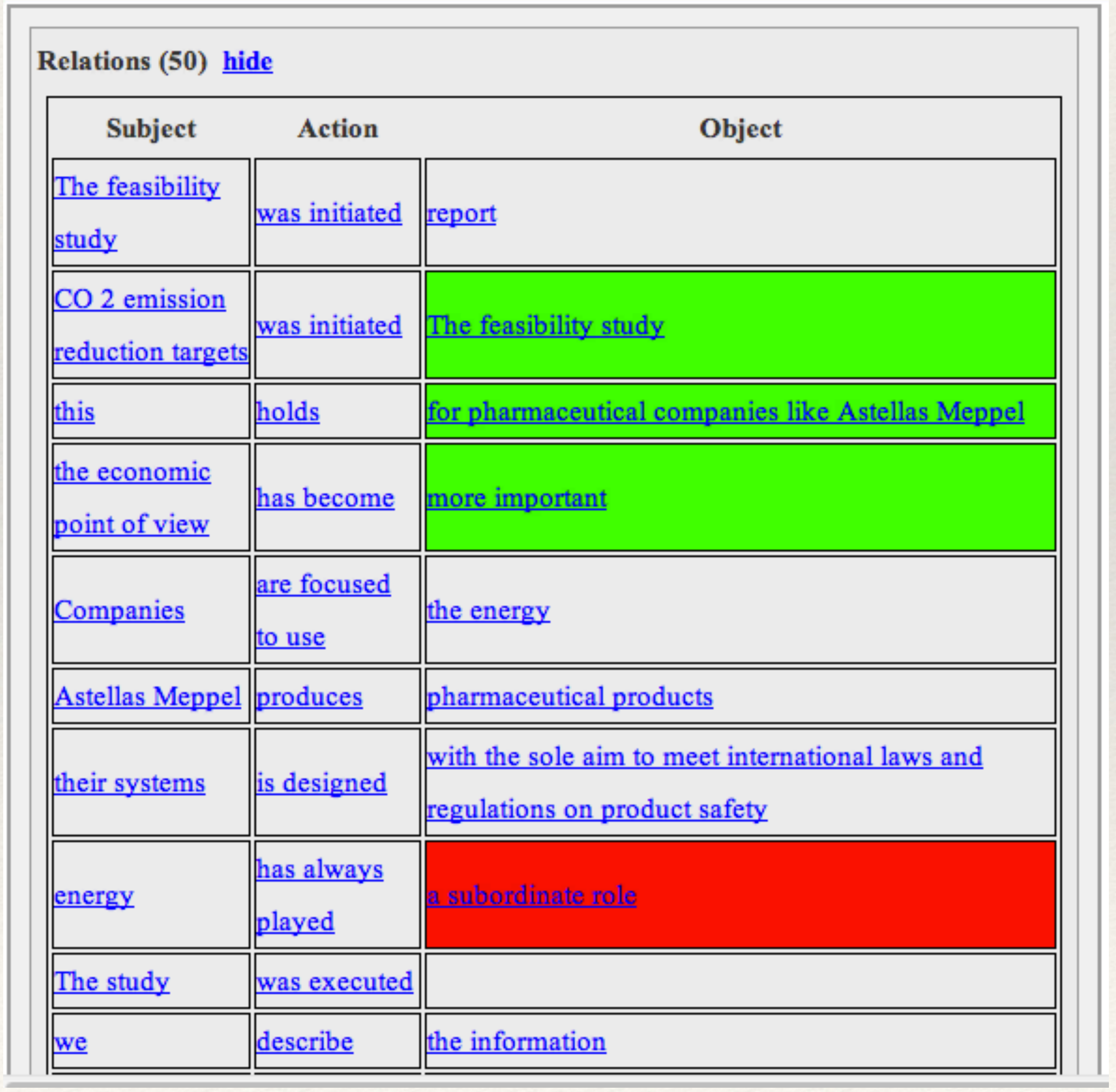
FieldTerminology (3)
energy systems
natural gas
pharmaceutical company

Person (1)
Astellas Meppel

Company (3)
Astellas management
Astellas Pharma
Enexis

extracted entities

extracted relationships



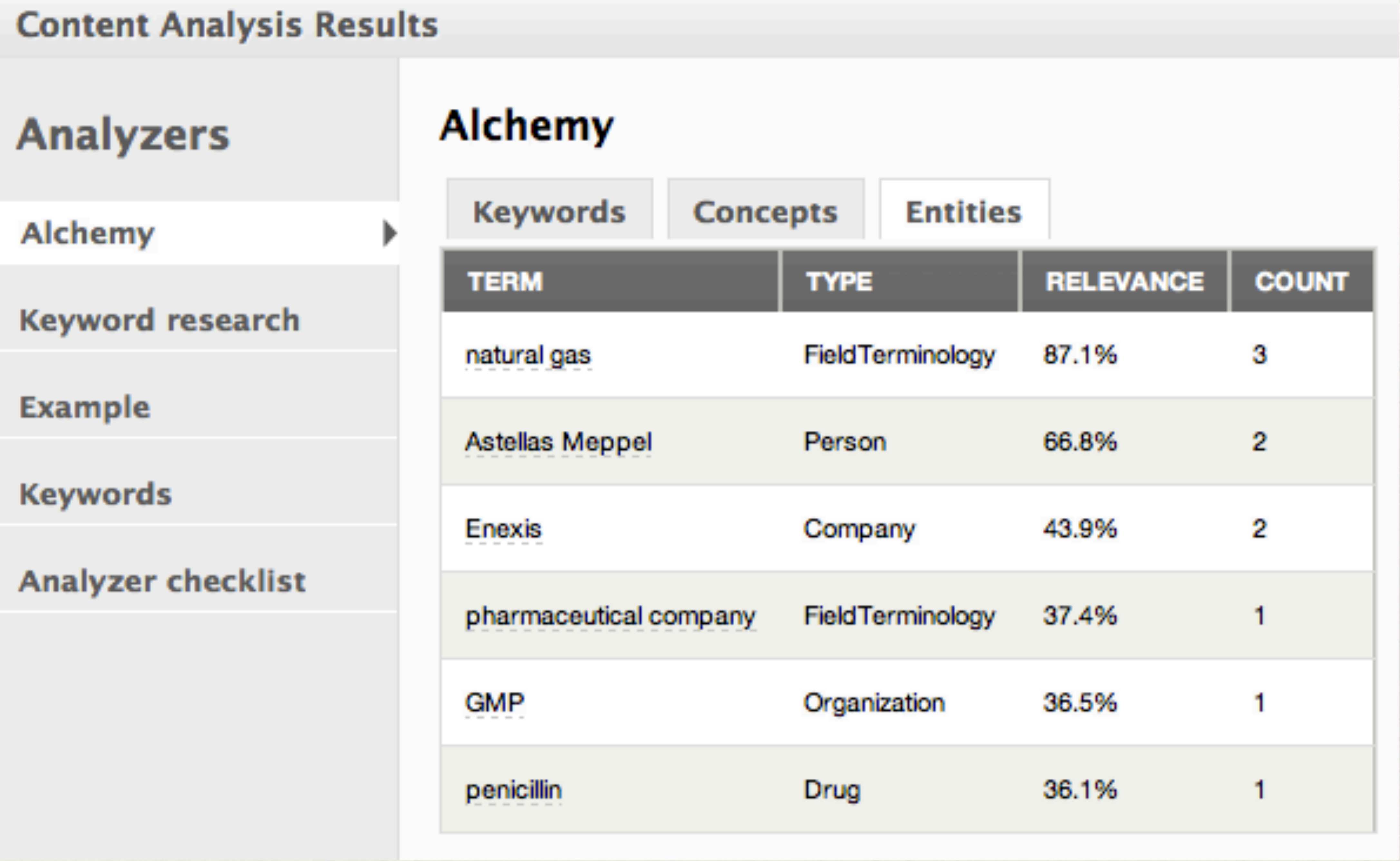
Relations (50) hide

Subject	Action	Object
The feasibility study	was initiated	report
CO 2 emission reduction targets	was initiated	The feasibility study
this	holds	for pharmaceutical companies like Astellas Meppel
the economic point of view	has become	more important
Companies	are focused to use	the energy
Astellas Meppel	produces	pharmaceutical products
their systems	is designed	with the sole aim to meet international laws and regulations on product safety
energy	has always played	a subordinate role
The study	was executed	
we	describe	the information

- Alchemy entity extraction
- Drupal taxonomies
- Automatic tagging

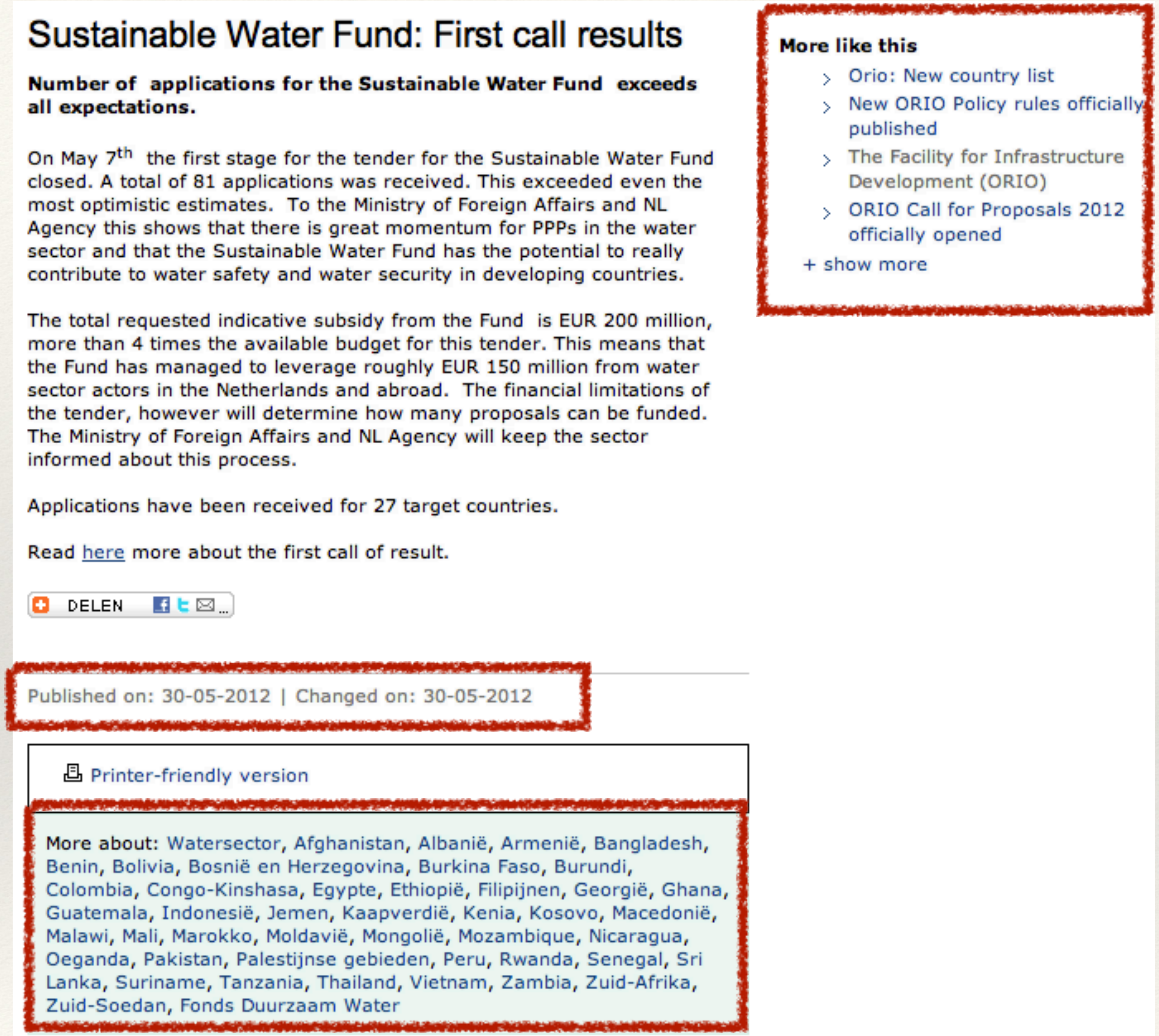
- nothing works with Drupal as expected;
- most of the Drupal modules are tailored to fit in specific implementation;
- lots of Ajax was failing in IE
- external API

Analyzing content in Drupal



TERM	TYPE	RELEVANCE	COUNT
natural gas	FieldTerminology	87.1%	3
Astellas Meppel	Person	66.8%	2
Enexis	Company	43.9%	2
pharmaceutical company	FieldTerminology	37.4%	1
GMP	Organization	36.5%	1
penicillin	Drug	36.1%	1

- 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 81 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.

DELEN

Published on: 30-05-2012 | Changed on: 30-05-2012

Printer-friendly version

More about: Watersector, Afghanistan, Albanië, Armenië, Bangladesh, Benin, Bolivia, Bosnië en Herzegovina, Burkina Faso, Burundi, Colombia, Congo-Kinshasa, Egypte, Ethiopië, Filipijnen, Georgië, Ghana, Guatemala, Indonesië, Jemen, Kaapverdië, Kenia, Kosovo, Macedonië, Malawi, Mali, Marokko, Moldavië, Mongolië, Mozambique, Nicaragua, Oeganda, Pakistan, Palestijnse gebieden, Peru, Rwanda, Senegal, Sri Lanka, Suriname, Tanzania, Thailand, Vietnam, Zambia, Zuid-Afrika, Zuid-Soedan, Fonds Duurzaam Water

More like this

- Orio: New country list
- New ORIO Policy rules officially published
- The Facility for Infrastructure Development (ORIO)
- ORIO Call for Proposals 2012 officially opened

+ show more

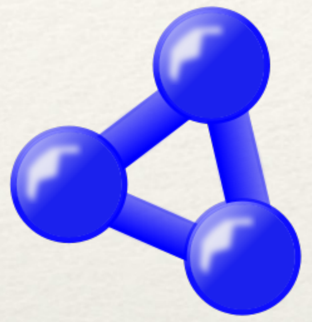
Agentschap NL
Ministerie van Economische Zaken, Landbouw en Innovatie

Actual AgentschapNL web site

“Factsheet” generation tryouts

- generate simple reports
  - on a certain pre-defined topic
  - term-based
- using different report templates
- add reports to the document base

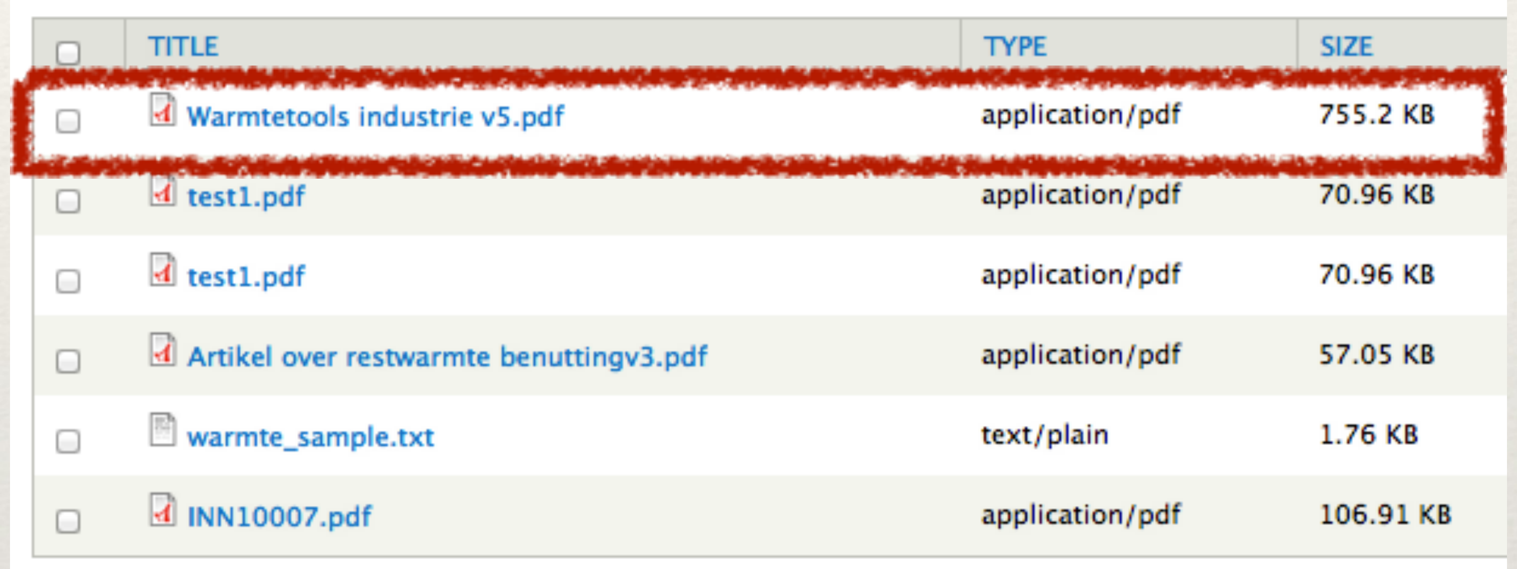
onderneming



“**warmte**”
OWL/RDF

query ontology

CODAK



□	TITLE	TYPE	SIZE
□	Warmtetools industrie v5.pdf	application/pdf	755.2 KB
□	test1.pdf	application/pdf	70.96 KB
□	test1.pdf	application/pdf	70.96 KB
□	Artikel over restwarmte benuttingv3.pdf	application/pdf	57.05 KB
□	warmte_sample.txt	text/plain	1.76 KB
□	INN10007.pdf	application/pdf	106.91 KB

search PDF repository

table:
types of companies

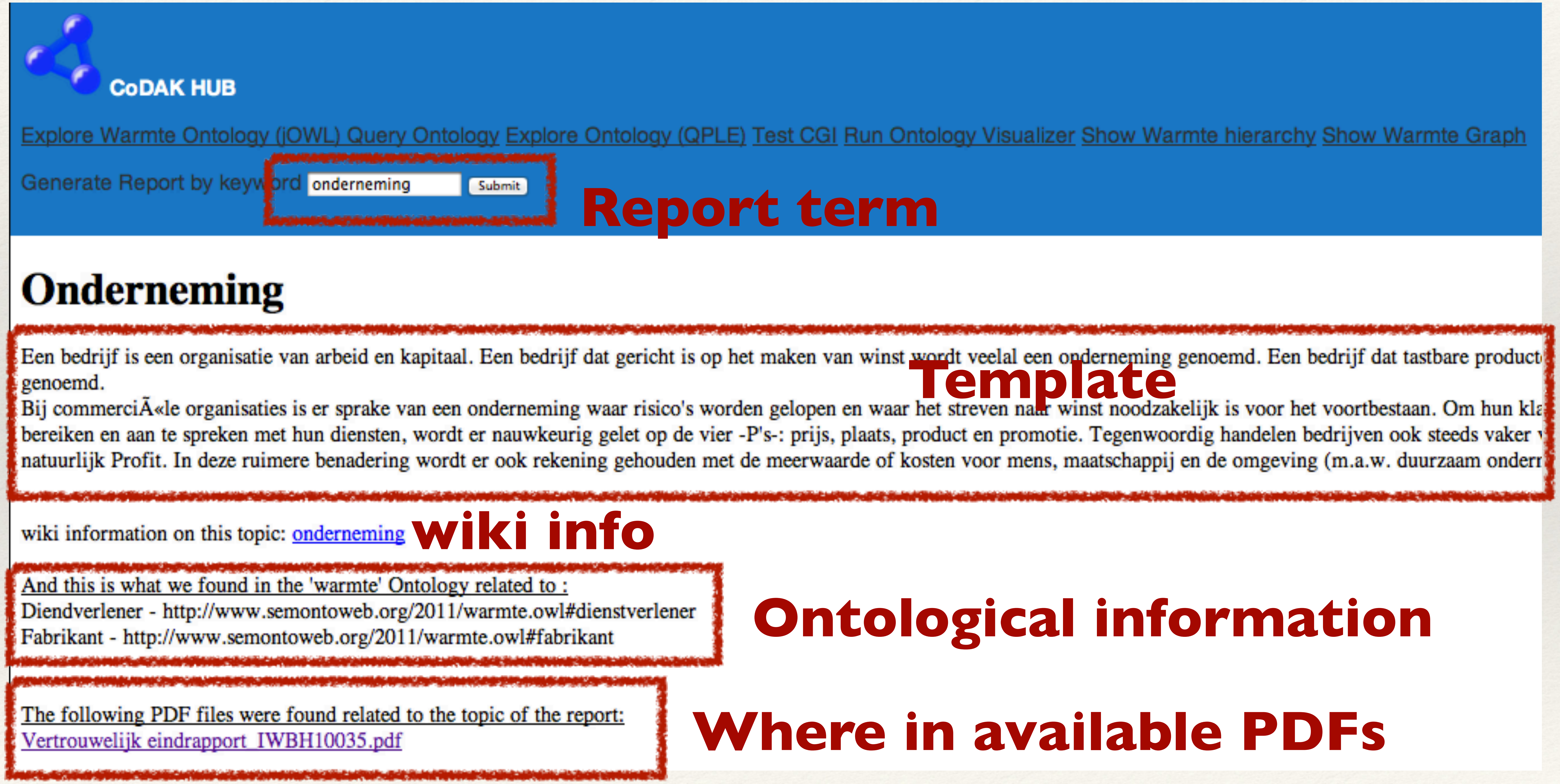
generate page with a report

table:
companies mentioned

templates →

→ can be saved to PDF, etc.

“Factsheet” generation tryouts (cont.)



CoDAK HUB

Explore Warmte Ontology (jOWL) Query Ontology Explore Ontology (QPLE) Test CGI Run Ontology Visualizer Show Warmte hierarchy Show Warmte Graph

Generate Report by keyword onderneming Submit

Onderneming

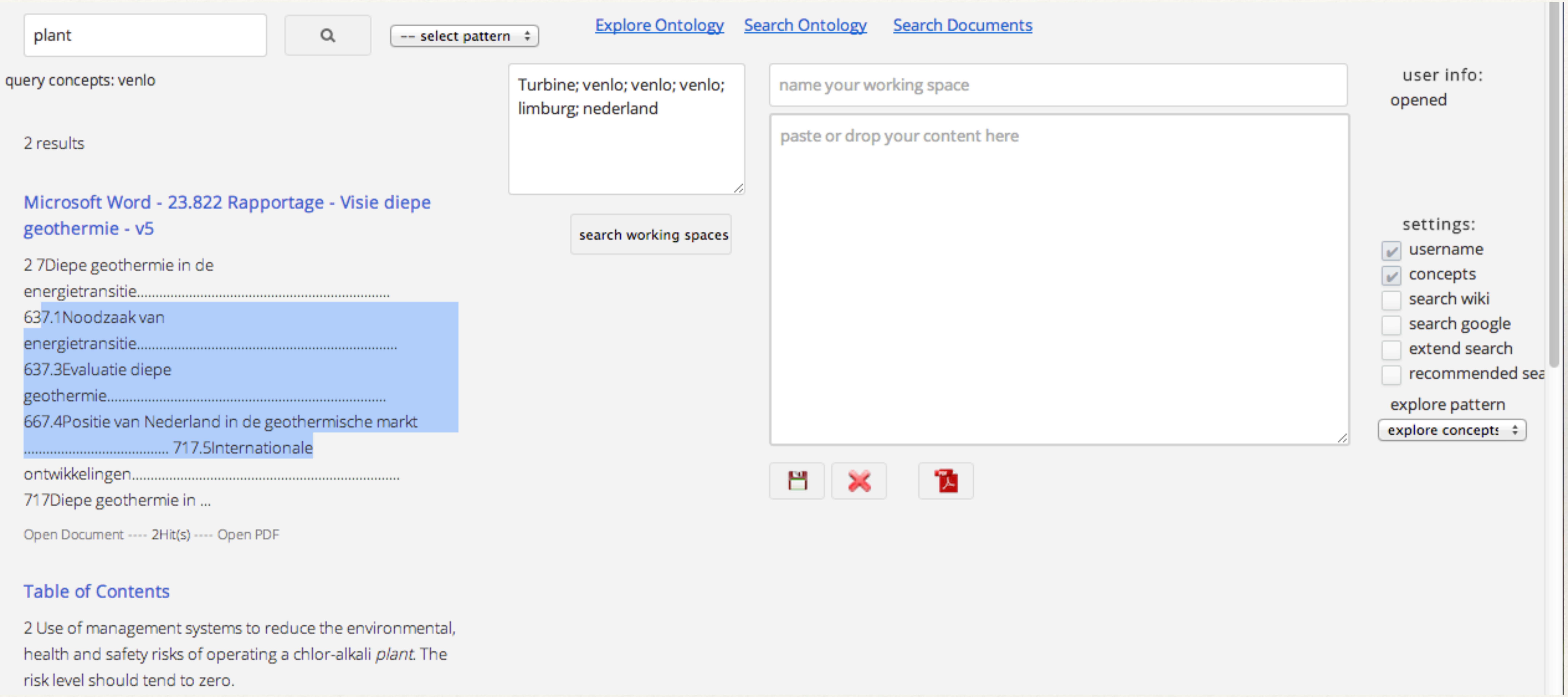
Een bedrijf is een organisatie van arbeid en kapitaal. Een bedrijf dat gericht is op het maken van winst wordt veelal een onderneming genoemd. Een bedrijf dat tastbare product
genoemd.
Bij commerciÃ«le organisaties is er sprake van een onderneming waar risico's worden gelopen en waar het streven naar winst noodzakelijk is voor het voortbestaan. Om hun kla
bereiken en aan te spreken met hun diensten, wordt er nauwkeurig gelet op de vier -P's-: prijs, plaats, product en promotie. Tegenwoordig handelen bedrijven ook steeds vaker v
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 onderr

wiki information on this topic: onderneming

And this is what we found in the 'warmte' Ontology related to :
Diendverlener - http://www.semontoweb.org/2011/warmte.owl#dienstverlener
Fabrikant - http://www.semontoweb.org/2011/warmte.owl#fabrikant

The following PDF files were found related to the topic of the report:
Vertrouwelijk eindrapport_IWBH10035.pdf

HTML5 + Graphics



plant

-- select pattern

Explore Ontology Search Ontology Search Documents

query concepts: venlo

2 results

Microsoft Word - 23.822 Rapportage - Visie diepe geothermie - v5

2 7Diepe geothermie in de energietransitie.. 637.1Noodzaak van energietransitie.. 637.3Evaluatie diepe geothermie.. 667.4Positie van Nederland in de geothermische markt .... 717.5Internationale ontwikkelingen.. 717Diepe geothermie in ...

Open Document ---- 2Hit(s) ---- Open PDF

Table of Contents

2 Use of management systems to reduce the environmental, health and safety risks of operating a chlor-alkali *plant*. The risk level should tend to zero.

Turbine; venlo; venlo; venlo; limburg; nederland

search working spaces

name your working space

paste or drop your content here

user info:
opened

settings:

- ☑ username
- ☑ concepts
- □ search wiki
- □ search google
- □ extend search
- □ recommended sea

explore pattern

explore concepts

- 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

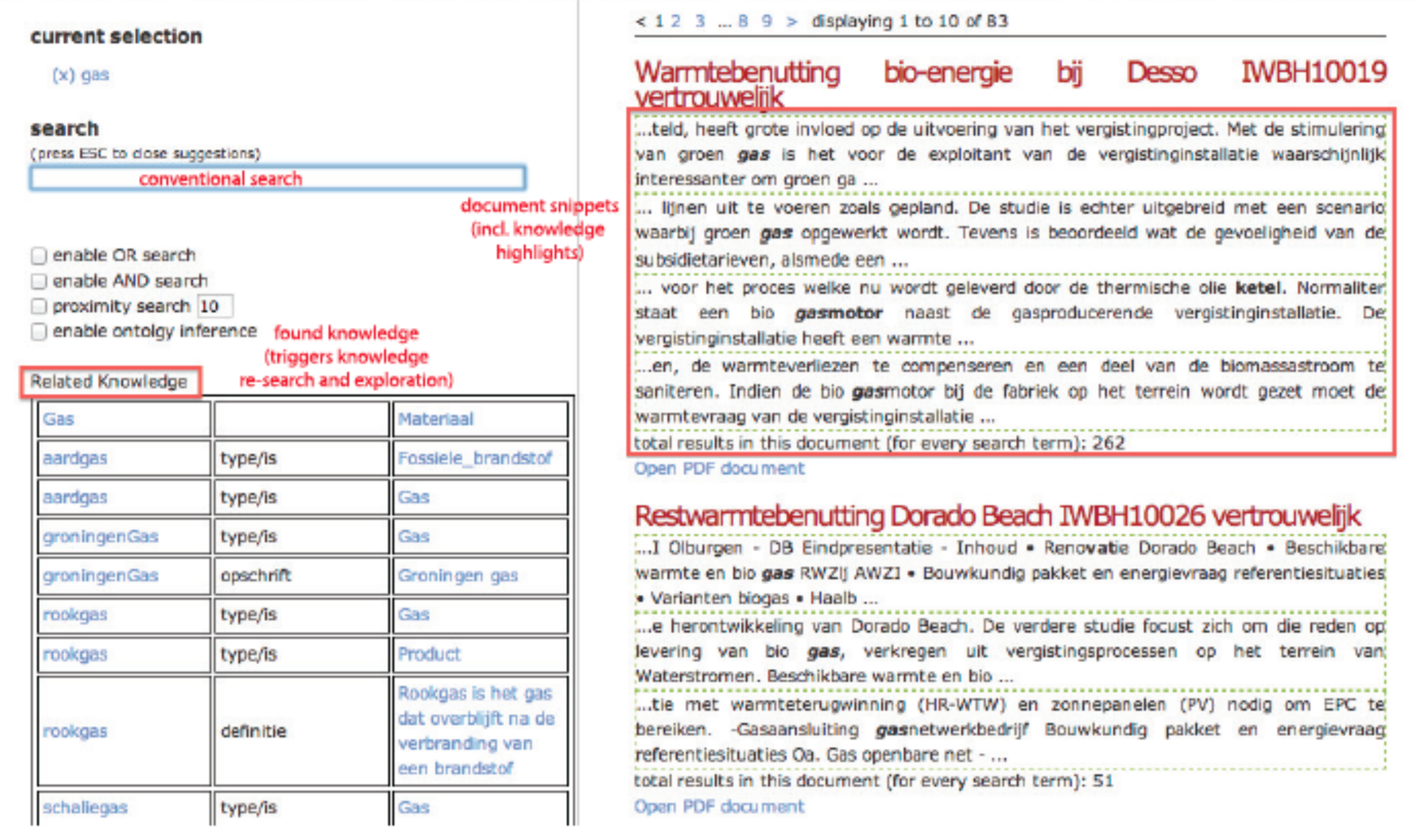
- from scratch
- only search, no UI at first
- incrementally adding features
- textual representation for the ontology
- tracking users’ behaviour

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

Deployed version

- one of the deployed version
- Demo time



current selection

(x) gas

search

(press ESC to close suggestions)

conventional search

document snippets (incl. knowledge highlights)

□ enable OR search

□ enable AND search

□ proximity search 10

□ enable ontolgy inference

found knowledge (triggers knowledge re-search and exploration)

Related Knowledge

Gas		Materiaal
aardgas	type/is	Fossiele_brandstof
aardgas	type/is	Gas
groningenGas	type/is	Gas
groningenGas	opschrift	Groningen gas
rookgas	type/is	Gas
rookgas	type/is	Product
rookgas	definitie	Rookgas is het gas dat overblijft na de verbranding van een brandstof
schaliegas	type/is	Gas

< 1 2 3 ... 8 9 > displaying 1 to 10 of 83

Warmtebenutting bio-energie bij Desso IWBH10019 vertrouwelijk

...teld, heeft grote invloed op de uitvoering van het vergistingproject. Met de stimulering van groen ***gas*** is het voor de exploitant van de vergistinginstallatie waarschijnlijk interessanter om groen ga ...

... lijnen uit te voeren zoals gepland. De studie is echter uitgebreid met een scenario waarbij groen ***gas*** opgewerkt wordt. Tevens is beoordeeld wat de gevoeligheid van de subsidietarieven, alsmede een ...

... voor het proces welke nu wordt geleverd door de thermische olie **ketel**. Normaliter staat een bio ***gasmotor*** naast de gasproducerende vergistinginstallatie. De vergistinginstallatie heeft een warmte ...

...en, de warmteverliezen te compenseren en een deel van de biomassastroom te saniteren. Indien de bio ***gas***motor bij de fabriek op het terrein wordt gezet moet de warmtevraag van de vergistinginstallatie ...

total results in this document (for every search term): 262

Open PDF document

Restwarmtebenutting Dorado Beach IWBH10026 vertrouwelijk

...I Olburgen - DB Eindpresentatie - Inhoud • Renovatie Dorado Beach • Beschikbare warmte en bio ***gas*** RWZI AWZI • Bouwkundig pakket en energievraag referentiesituaties • Varianten biogas • Haalb ...

...e herontwikkeling van Dorado Beach. De verdere studie focust zich om die reden op levering van bio ***gas***, verkregen uit vergistingsprocessen op het terrein van Waterstromen. Beschikbare warmte en bio ...

...tie met warmteterugwinning (HR-WTW) en zonnepanelen (PV) nodig om EPC te bereiken. -Gasaansluiting ***gas***netwerkbedrijf Bouwkundig pakket en energievraag referentiesituaties Oa. Gas openbare net - ...

total results in this document (for every search term): 51

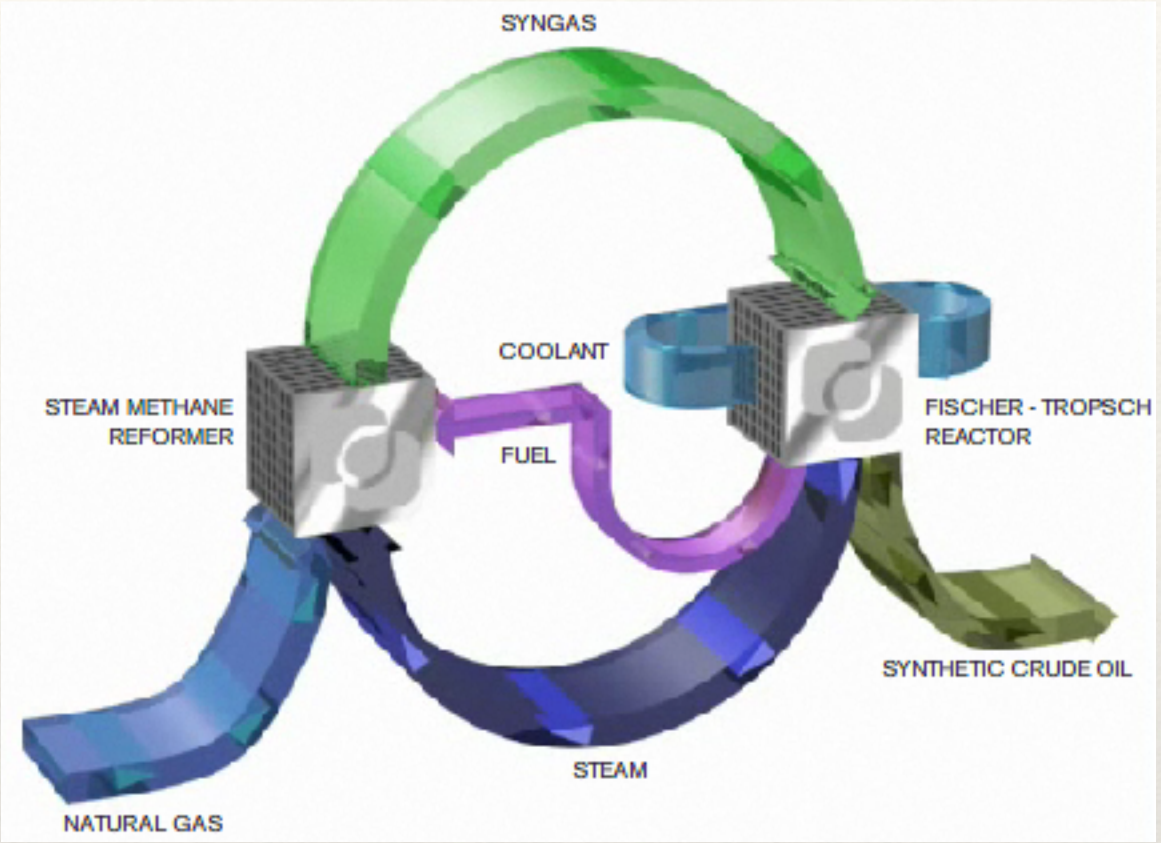
Open PDF document

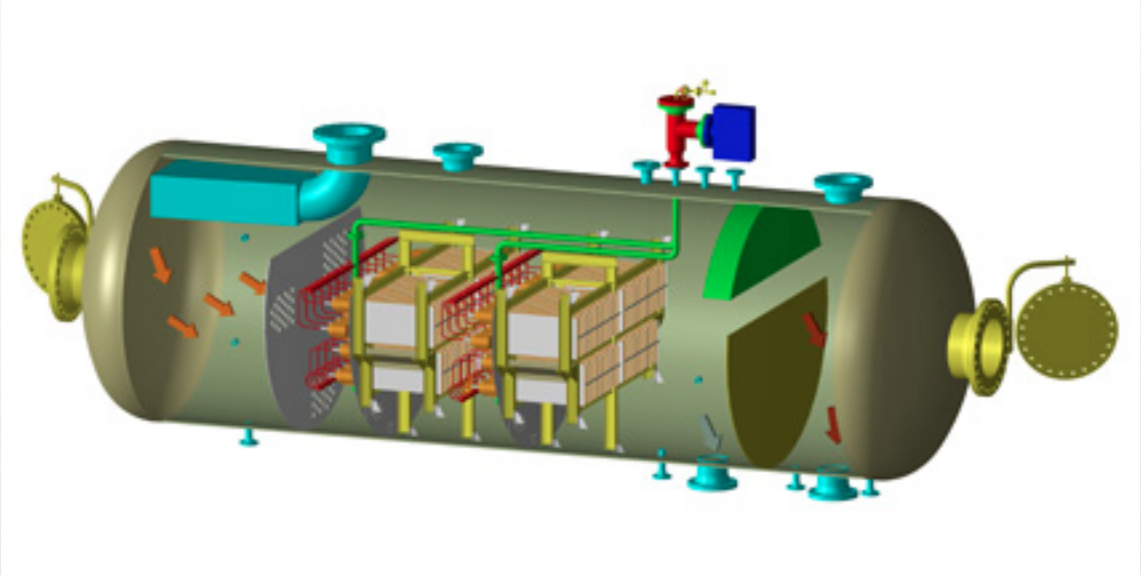
CoDAK language

- 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