NUXEO WORLD PARIS 2011

When ECM Meets the Semantic Web

20 Oct 2011 - Olivier Grisel & Stefane Fermigier

nuxeo
Open Source ECM

Business Motivations

Information society

From Wikipedia, the free encyclopedia

For other uses, see Information society (disambiguation).

The aim of the **information society** is to gain competitive advantage internationally through using IT in a creative and productive way. An **information society** is a society in which the creation, distribution, diffusion, use, integration and manipulation of information is a significant economic, political, and cultural activity. The knowledge economy is its economic counterpart whereby wealth is created through the economic exploitation of understanding. People that have the means to partake in this form of society are sometimes called digital citizens. As Beniger[1] shows, this is one of many dozen labels that have been identified to suggest that we are entering a new phase of society.

Source: Wikipedia

Knowledge economy

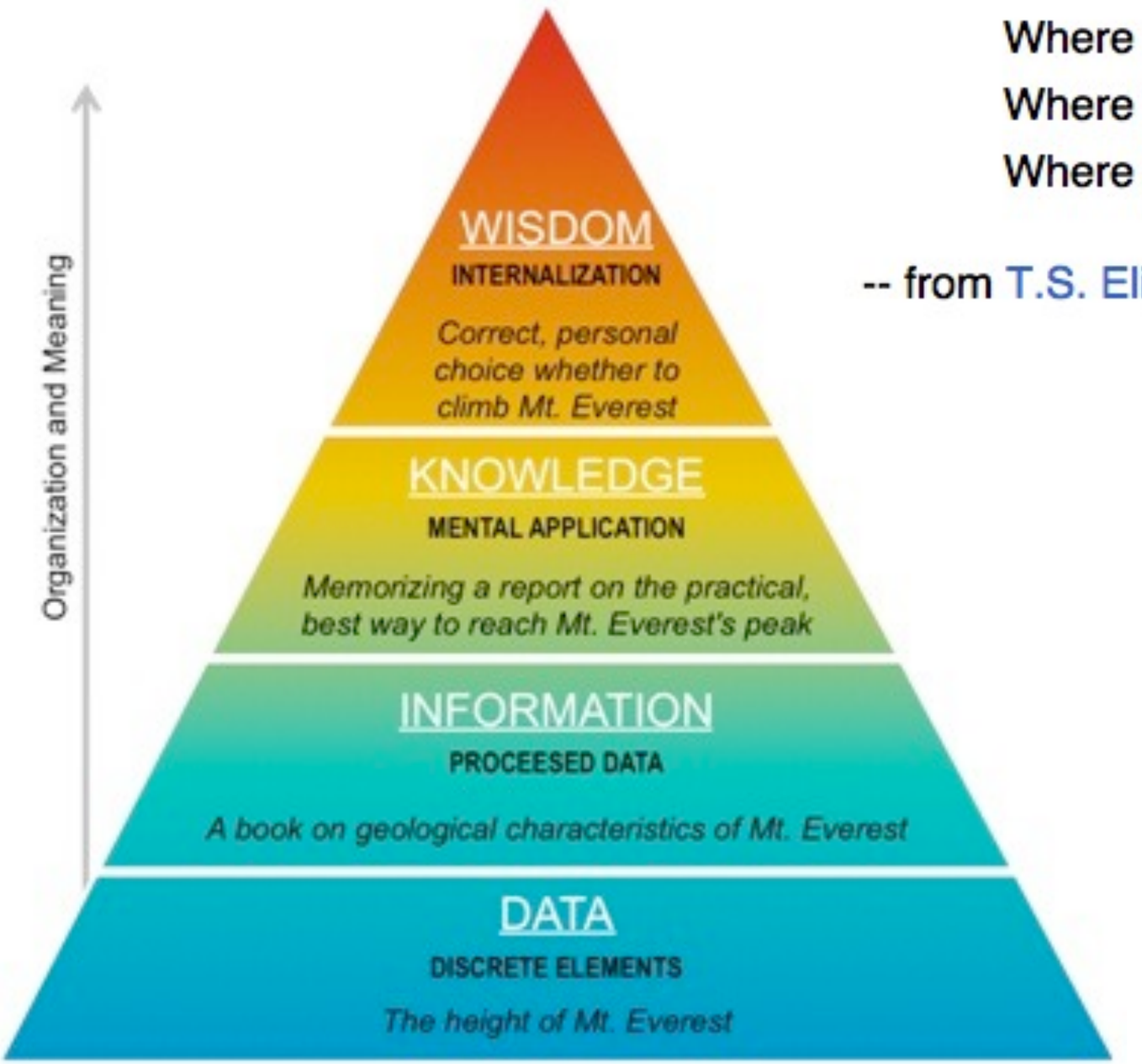
From Wikipedia, the free encyclopedia

The **knowledge economy** is a term that refers either to an **economy of knowledge** focused on the production and management of knowledge in the frame of economic constraints, or to a **knowledge-based economy.** In the second meaning, more frequently used, it refers to the use of knowledge technologies (such as knowledge engineering and knowledge management) to produce economic benefits as well as job creation. The phrase was popularized by Peter Drucker as the title of Chapter 12 in his book *The Age of Discontinuity,* And, with a footnote in the text, Drucker attributes the phrase to economist Fritz Machlup.[1]

The essential difference is that in a *knowledge economy,* knowledge is a product, while in a *knowledge-based economy,* knowledge is a tool. This difference is not yet well distinguished in

Source: Wikipedia

The DIKW hierarchy

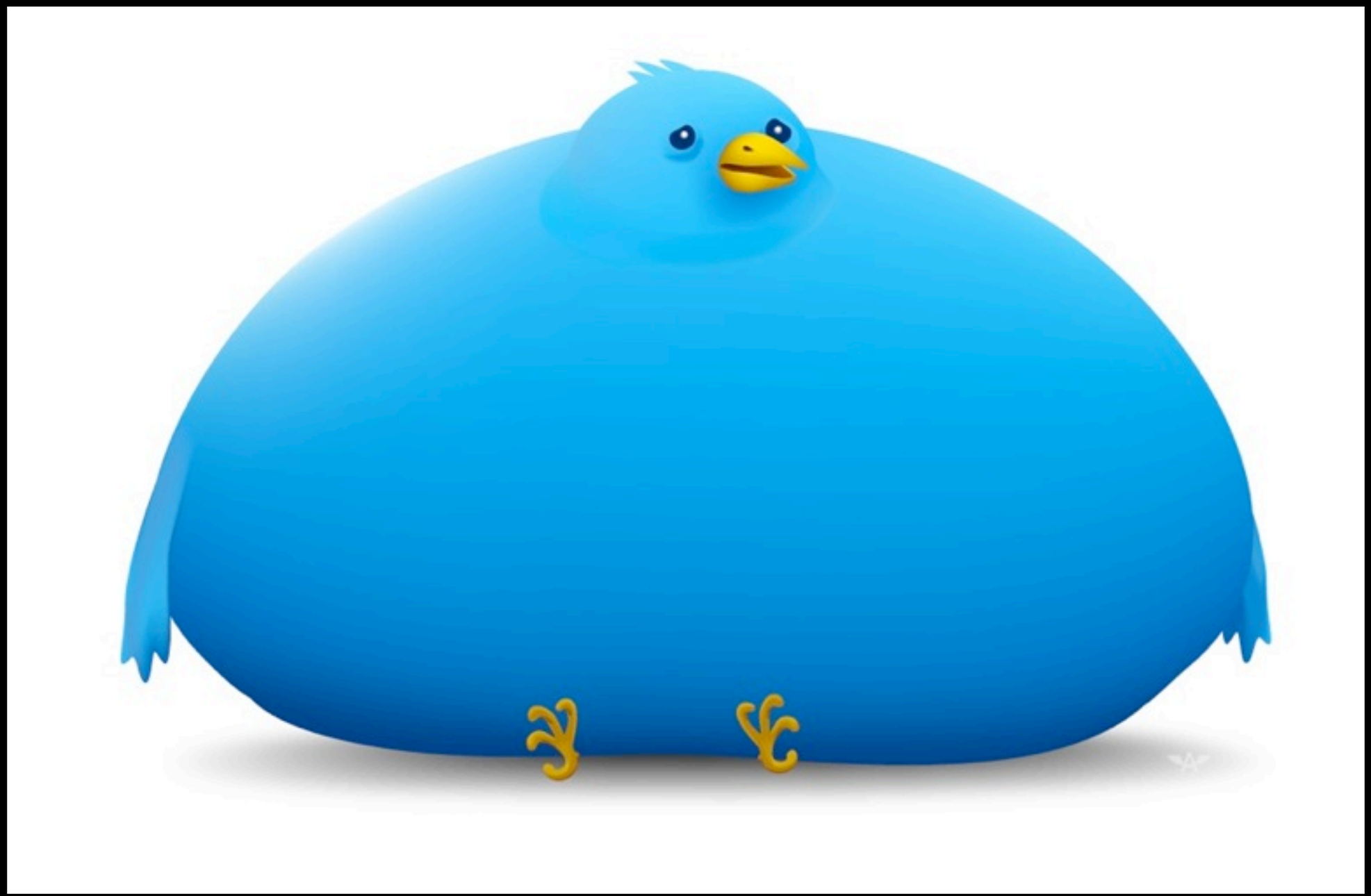


Where is the Life we have lost in living?
Where is the wisdom we have lost in knowledge?
Where is the knowledge we have lost in information?

-- from T.S. Eliot, "Choruses from 'The Rock'"

But every coin has another side

Infobesity!



A few figures

- 50% more data / content / information produced every year
- 1.8 zettabytes of data produced in 2011 (= 1 billion terabytes)
- Employees are drowning in a sea of email, status messages, etc., and spend on average more than 6 hours / weeks unsuccessfully searching for or recreating lost documents

A Solution: the Semantic Web

A Brief History of the Web

- **Web 1.0** (1990-now): web of sites and pages, aka the *World Wide Web*
- **Web 2.0** (2000-now): web of people and of participation, aka the *Social Web* (Blogs, RSS, tags, Facebook, Wikipedia, etc.)
- **Web 3.0** (2010-now): web of data, of meaning and connected knowledge, aka the *Semantic Web*

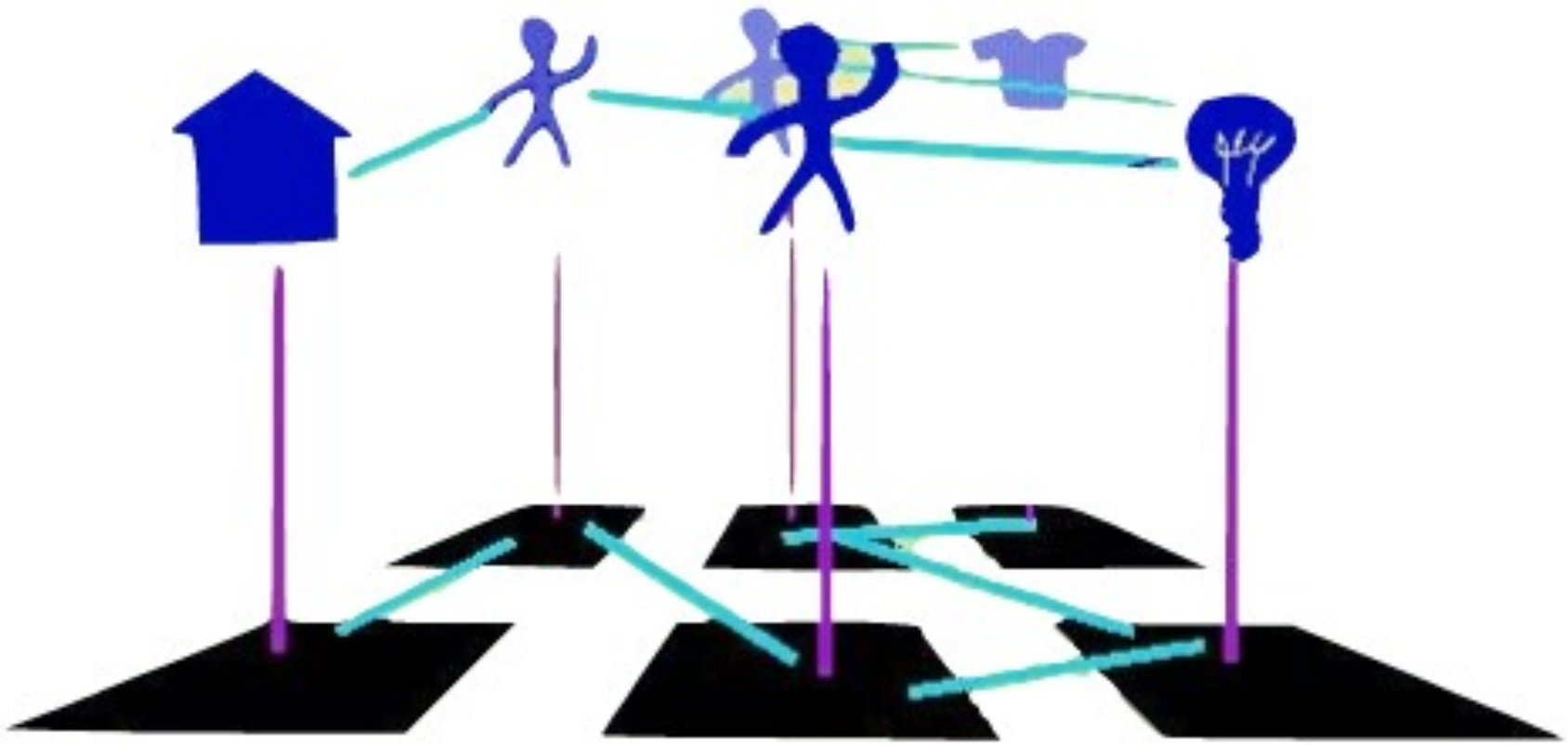

WWW
2009
MADRID
xml 10

"To a computer, then, the web is a **flat**, **boring** world devoid of **meaning**"

Tim Berners Lee, http://www.w3.org/Talks/WWW94Tim/

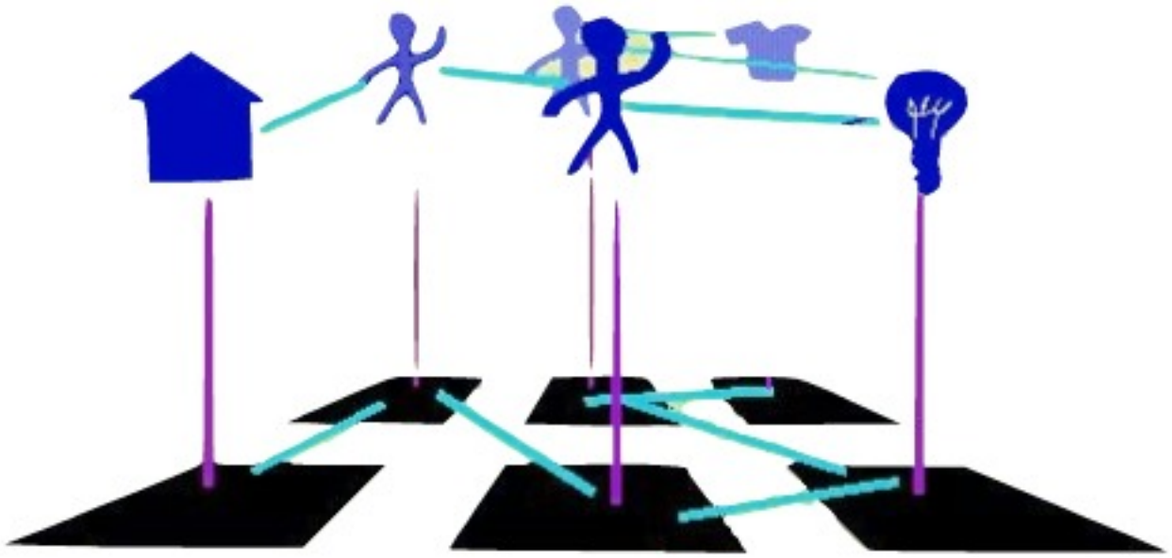
"This is a pity, as in fact **documents** on the web describe **real objects** and imaginary **concepts**, and give particular **relationships** between them"

Tim Berners Lee, http://www.w3.org/Talks/WWW94Tim/



"Adding semantics to the web involves two things: allowing **documents** which have information in **machine-readable** forms, and allowing **links** to be created with **relationship values**."

Tim Berners Lee, http://www.w3.org/Talks/WWW94Tim/



"The Semantic Web is not a separate Web but an **extension** of the current one, in which information is given well-defined **meaning**, better enabling **computers and people** to work in cooperation."

Tim Berners Lee, http://www.w3.org/Talks/WWW94Tim/

Means and Tools

4 stages

- **Extract** meaning from raw data / content
- **Connect** information to form knowledge
- **Reason** about this knowledge
- **Present** this knowledge in actionable form

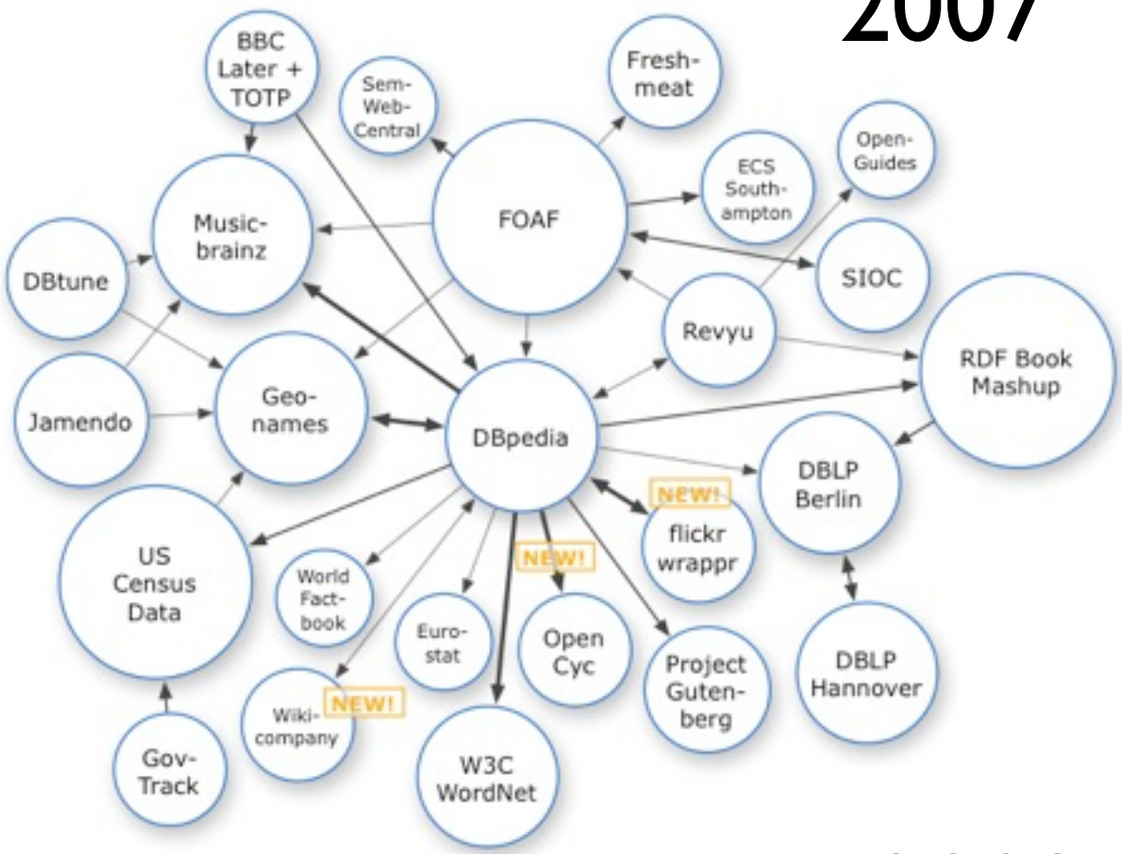
Extracting

- Leverage metadata embedded in or associated with documents (when they exist)
- Or use machine learning, NLP (Natural Language Processing) and image processing algorithms to **extract meaning** from text / images
- Examples include: named entities extraction, automatic categorization / tagging, sentiment analysis, etc.

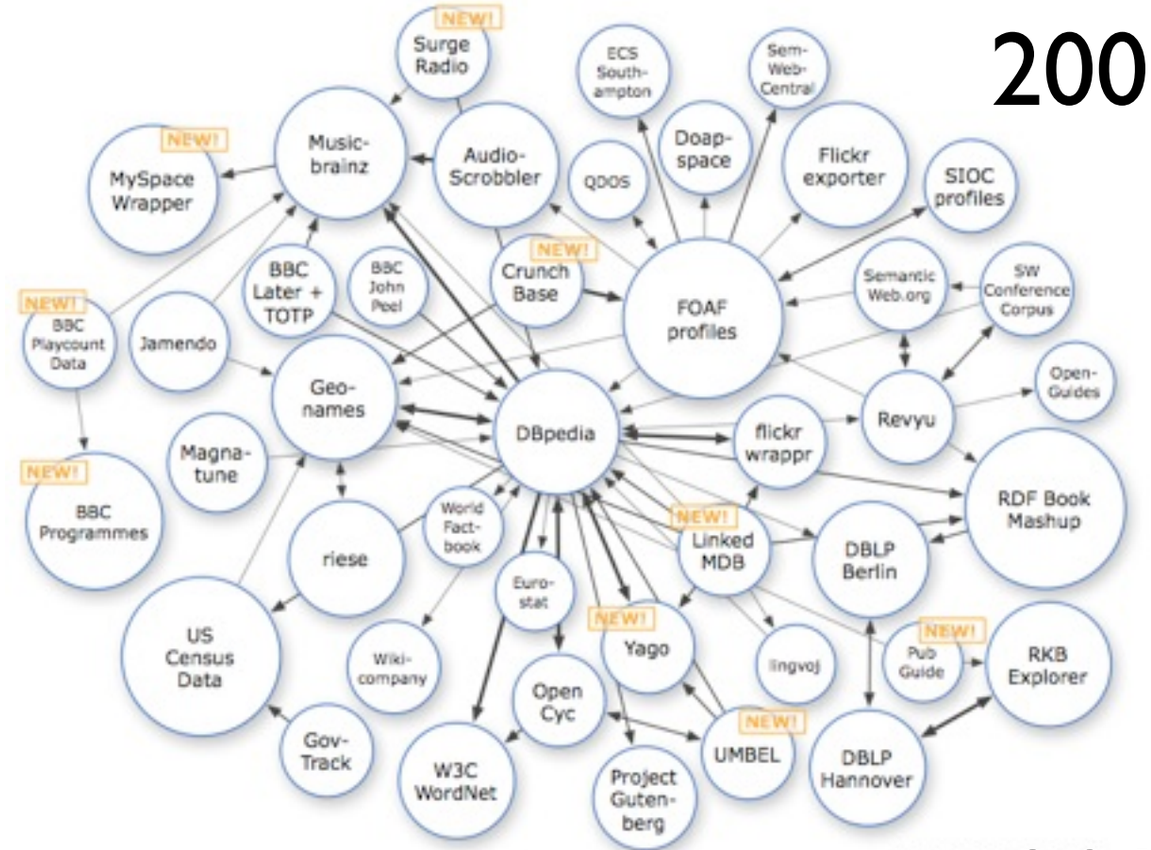
Interlude: Linked Open Data

nuxeo
Open Source ECM

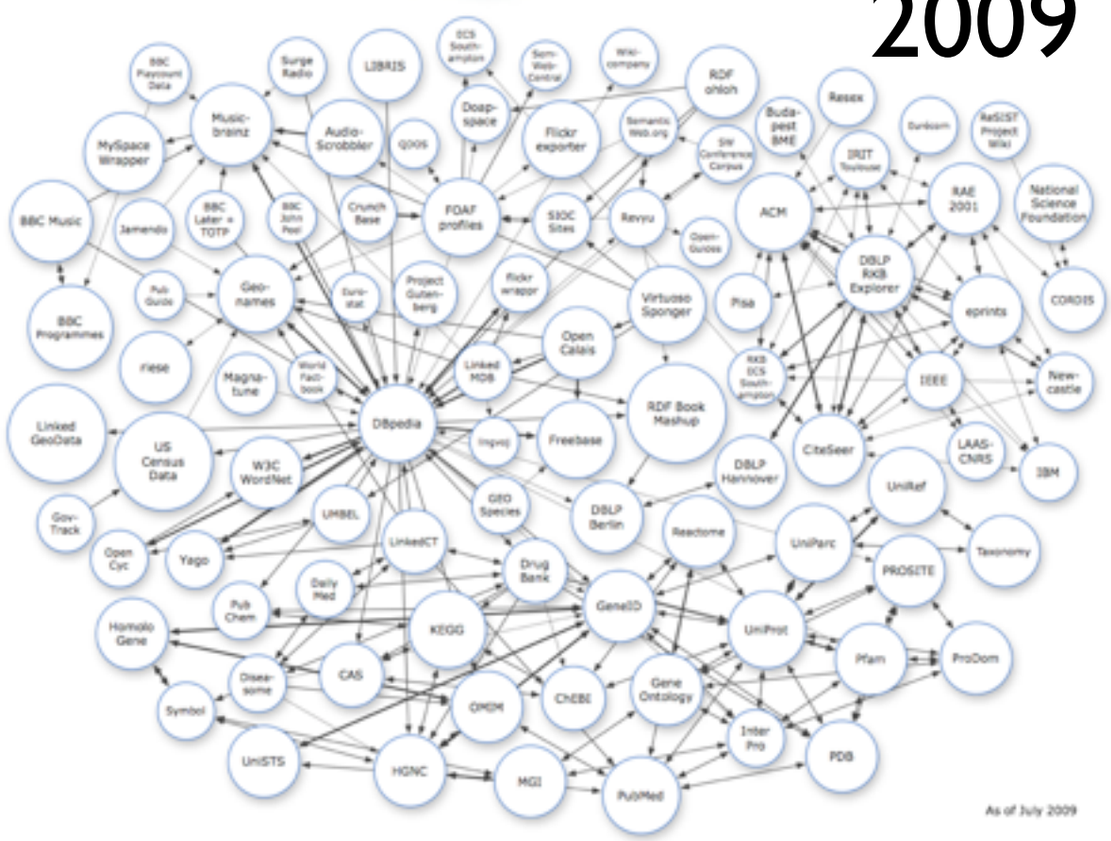
2007



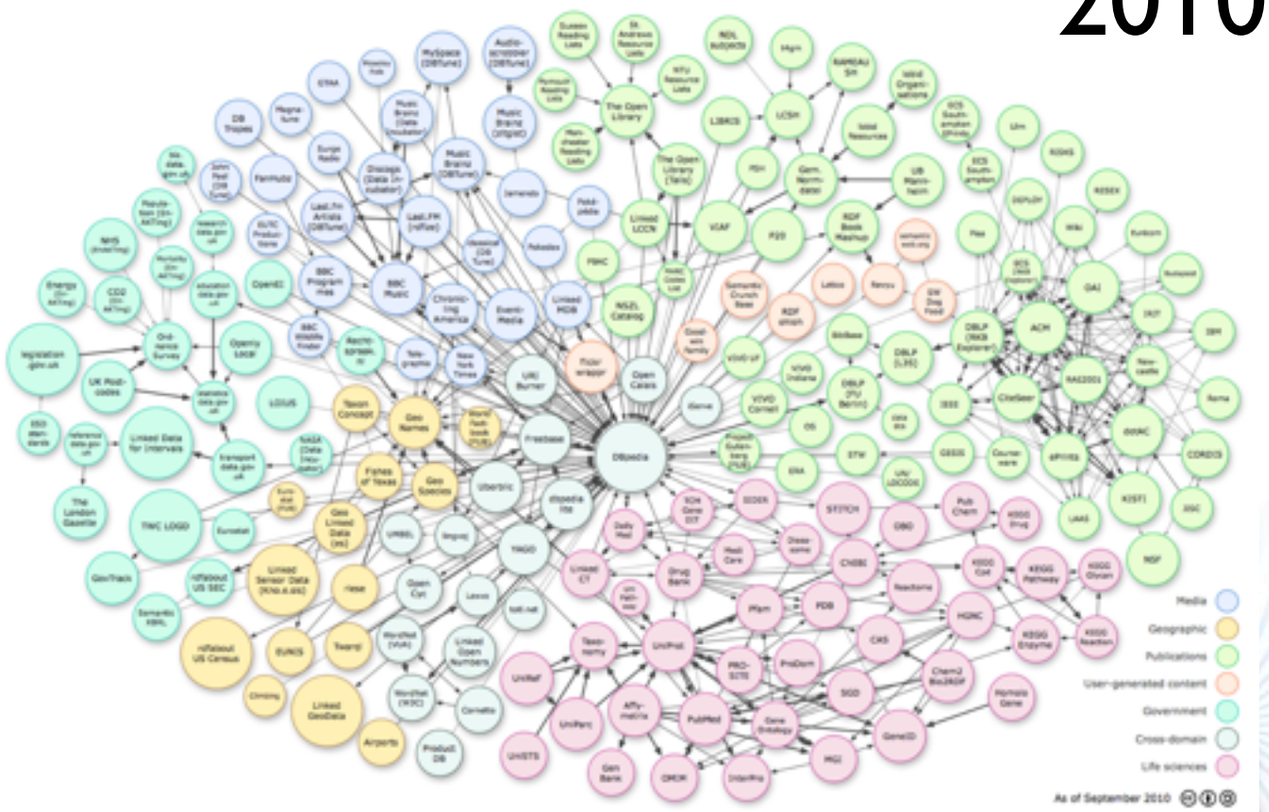
2008



2009

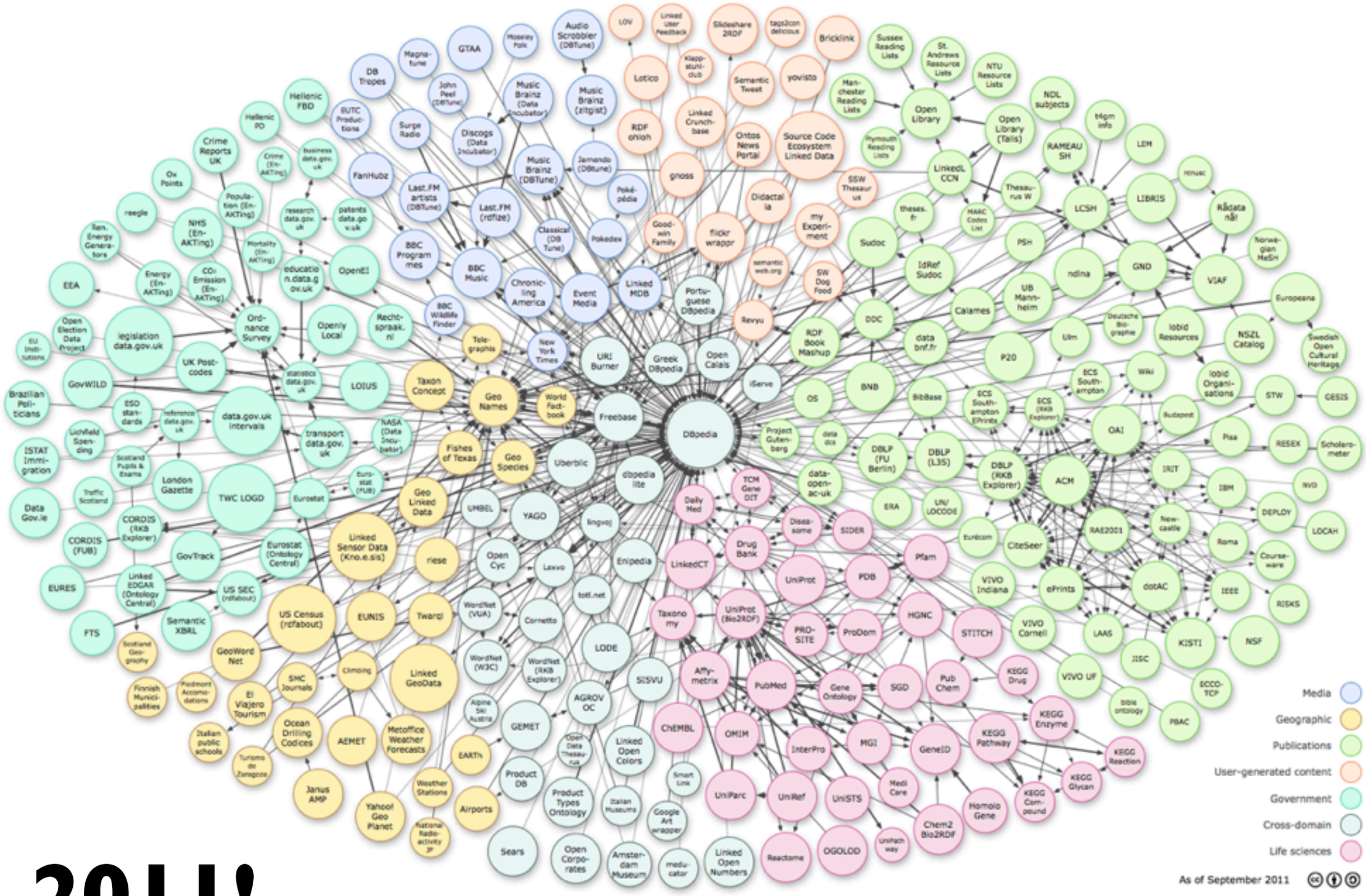


2010



nuxeo

Open Source ECM



2011!

NUXEO WORLD PARIS 2011

Linking

- Many Linked Open Data repositories have been made available over the last 10 years
- RDF and graph database systems are now available to manage this huge mass of information (billions of triples)
- **Match** information extracted from content with these public (or internal) data/knowledge bases

Reasoning

- When you are working on reliable metadata (ex: RDFa embedded in web pages), you can use rule / inference engines to infer **actionable knowledge** from your content (ex: shopping recommendation engine)
- Rules can also be used to **clean up** / flag errors when working with unreliable (e.g. automatically extracted) information

Presenting

- Allow the users of your system to interact with the knowledge thus extracted or produced, in a way that allows them to do their jobs better
- A smart presentation system solves the information overload issue by **contextualizing** the information, i.e. presenting only information relevant to what the user is currently doing

R&D Projects Involving Nuxeo

IKS project

nuxeo
Open Source ECM

- European R&D project under the FP7, with 13 partners (6 SMEs) and a 8.5M EUR budget
- Goal: create a semantic software "stack" that will be used by CMS vendors to add semantic features to their products
- Started in Jan. 2009, will last until Dec. 2012
- First tangible result: **Apache Stanbol** (more about this later)

SAMAR project

- French collaborative R&D project with 10 partners, and a 4.5M EUR budget
- Goal: create a platform for managing multimedia content in arabic, for news agencies such as AFP
- Will include: automated translation, named entities extraction, content classification
- First results: integration between **Nuxeo and Temis** (more later)

State of the Art
Semantic ECM at Nuxeo

The Semantic Engine

- From unstructured content to Knowledge
- Language guessing
- Topic classification (Business, Sports, Media, ...)
- Named Entities extraction and linking
- Relationships and properties extraction

Demo time!


apache
stanbol™

http://stanbol.demo.nuxeo.com/engines

apache stanbol™ The RESTful Semantic Engine

/engines

Web View REST API

Enhancement Engines

There are currently **3** active engines.

- org.apache.stanbol.enhancer.engines.opennlp.impl.NamedEntityExtractionEnhancementEngine
- org.apache.stanbol.enhancer.engines.autotagging.impl.EntityMentionEnhancementEngine
- org.apache.stanbol.enhancer.jersey.cache.CachingDereferencerEngine

You can enable, disable and deploy new engines using the OSGi console.

Paste some text below and submit the form to let the active engines enhance it:

```
John Smith works at Smith Consulting in London, United Kingdom.
```

Output format: JSON-LD Run engines

Extracted entities

People	Organizations	Places
John Smith	Smith Consulting	London
		United Kingdom

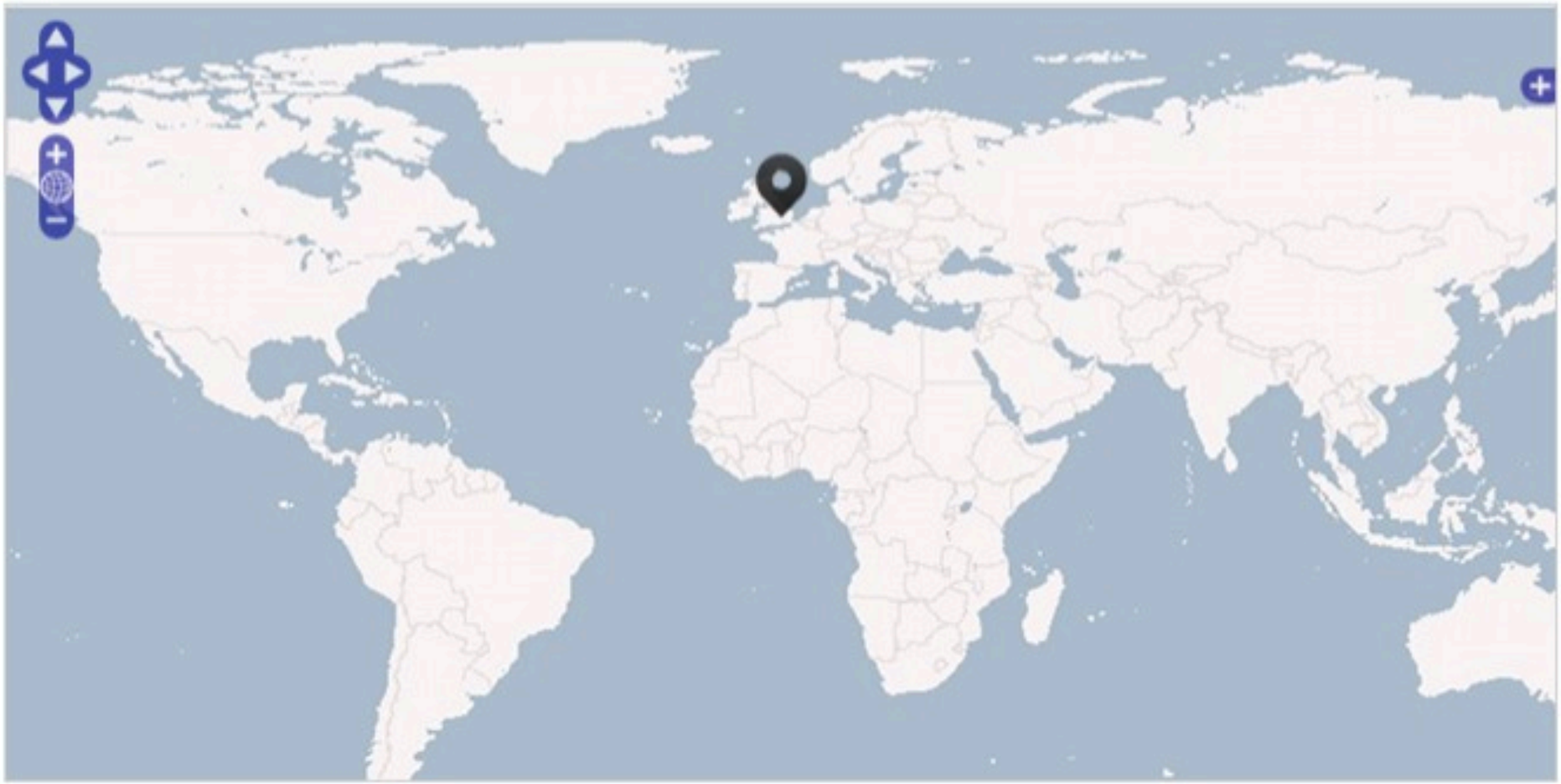
http://stanbol.demo.nuxeo.com/engines

John Smith

Smith Consulting

London

United Kingdom



Raw RDF output


```
{
  "@": [
    {
      "@": "<urn:enhancement-4cf4870a-fc4e-4418-c966-ae4f0a873854>",
      "a": [
```

http://stanbol.demo.nuxeo.com/engines#

apache stanbol™ The RESTful Semantic Engine

/engines

Web View REST API

Enhancement Engines

Stateless REST analysis

This stateless interface allows the caller to submit content to the Stanbol enhancer engines and get the resulting enhancements formatted as RDF at once without storing anything on the server-side.

The content to analyze should be sent in a POST request with the mimetype specified in the `Content-type` header. The response will hold the RDF enhancement serialized in the format specified in the `Accept` header:

```
curl -X POST -H "Accept: text/turtle" -H "Content-type: text/plain" \
     --data "John Smith was born in London." http://stanbol.demo.nuxeo.com
/engines
```

The list of mimetypes accepted as inputs depends on the deployed engines. By default only `text/plain` content will be analyzed

Stanbol enhancer is able to serialize the response in the following RDF formats:

- `application/json` (JSON-LD)
- `application/rdf+xml` (RDF/XML)
- `application/rdf+json` (RDF/JSON)
- `text/turtle` (Turtle)
- `text/rdf+nt` (N-TRIPLES)

By default the URI of the content item being enhanced is a local, non de-referencable URI automatically built out of a hash digest of the binary content. Sometimes it might be helpful to

RESTful is Beautiful

localhost:8080/nuxeo/nxpath/default/default-domain/workspaces/News articles.1319058853893/Libya Rebel

Muammar Gaddafi

nuxeo • DM | Home | Document Management | Admin Center | Studio | Administrator | SeamReload | Search | Advanced search

> default-domain > Workspaces > News articles > Libya Rebels edge closer to Tripoli.html

- default-domain
  - Entities
  - Sections
  - Templates
  - Workspaces
    - AFP Arabic
    - Another Workspace
    - News articles

WorkList | Clipboard

No document in clipboard

Libya Rebels edge closer to Tripoli.html

Summary | Edit | Files | Publish | Relations | Workflow | Alerts | Comments | History | Manage

Libyan rebels edged closer to the capital city of Tripoli on Sunday to help fellow mutineers inside the city who declared a final clash with leader Muammar Gaddafi.
Following a night marred with gunfire, the rebels said that they controlled a handful of Tripoli's localities. With the rebels within about 25 km of Tripoli, Gaddafi's hold on power looks fragile. He labelled the rebels, who had been fighting for the past six months, as "rats" and said that he would not yield to their demands.
A coordinated revolt that rebels
instantly after Muslim clerics ca
advancing to the capital's peripl
uprising, which is in its sixth-mo
"The rebels may have risen too
Oliver Miles, a former British an
quite the extent they think it ha
The rebels' advance toward the
them and the capital. Governm
a burned-out tank, and some to
filled some walls with graffiti, or
In Benghazi, the rebels' main st
everything was going according
and others are coming in from c
Ghoga, vice chairman of the reb
Gaddafi — in hiding since the N
late yesterday that he had no i
Moussa Ibrahim, in a briefing fo
armed units defending Tripoli from the rebels wholeheartedly believe that if this city is captured, the blood will run everywhere; so they may as well fight to the end."

Muammar Muhammad al-Gaddafi (born 7 June 1942) is a Libyan revolutionary and the country's head of state from 1969 to the present day. Gaddafi became head of state after removing King Idris from power in a 1969 bloodless coup, after which he established the Libyan Arab Republic. His almost 42 years in power make him one of the longest-serving non-royal rulers in history.

Link to another entity

Mentions in current document

(...) the capital city of Tripoli on Sunday to help fellow mutineers inside the city who declared a final clash with leader **Muammar Gaddafi** . Following a night marred with gunfire, the rebels said that they controlled a handful of Tripoli's localities. With the rebels (...)

State	Project	Version	0.0

This document is **unlocked** | Lock

People

- Abdul Hafiz Ghoga — Abdul Hafiz Ghoga is a Libyan human rights lawyer, who rose to prominence as the spokesman for the National Transitional (...)
- Moussa Ibrahim — Moussa Ibrahim is a Libyan political figure, serving as Libyan Minister of Information and the official spokesman for Muammar Gaddafi (...)
- Muammar Gaddafi — Muammar Muhammad al-Gaddafi (born 7 June 1942) is a Libyan revolutionary and the country's head of state from 1969 to (...)
- Oliver Miles

Organizations

- Libya Rebels
- National Transition Council

localhost:8080/nuxeo/nxpath/default/default-domain/entities/Muammar Gaddafi@view_documents?tabIds=%

Muammar Gaddafi

nuxeo • DM | Home | Document Management | Admin Center | Studio | Administrator | SeamReload | Search | Advanced search

> default-domain > Entities > Muammar Gaddafi

- default-domain
  - Entities
  - Sections
  - Templates
  - Workspaces

WorkList | Clipboard

No document in clipboard

Muammar Gaddafi

Summary | Edit | Relations | Alerts | Comments | History | Manage

Occurrences in documents

	Title	Modified	Author	State
	Libya Rebels edge closer to Tripoli.html	10/20/2011 12:40 AM	Administrator	Project
	afp.com-20110615T020009Z-TX-PAR-QFU53.xml	10/20/2011 12:24 AM	Administrator	Project
	afp.com-20110615T001023Z-TX-PAR-QFS22.xml	10/20/2011 12:24 AM	Administrator	Project
	afp.com-20110615T071432Z-TX-PAR-QGG11.xml	10/20/2011 12:24 AM	Administrator	Project
	afp.com-20110615T052303Z-TX-PAR-QGA07.xml	10/20/2011 12:24 AM	Administrator	Project
	afp.com-20110615T044702Z-TX-PAR-QFY87.xml	10/20/2011 12:24 AM	Administrator	Project

Add link to new document

Muammar Muhammad al-Gaddafi (born 7 June 1942) is a Libyan revolutionary and the country's head of state from 1969 to the present day. Gaddafi became head of state after removing King Idris from power in a 1969 bloodless coup, after which he established the Libyan Arab Republic. His almost 42 years in power make him one of the longest-serving non-royal rulers in history.

Also known as ムアンマル・アル=カッザーフィー, Muammar Gaddafi, معمر القذافي, Muammer Kaddafi, Muammar al-Gaddafi, Mu'ammar Gheddafi, Muamar el Gadafi, מועמר קדאפי, Mouammar Kadhafi, 穆阿迈尔·卡扎菲, Каддафи, Муаммар

Entity Type Person

Remote knowledge base

Muammar Gaddafi

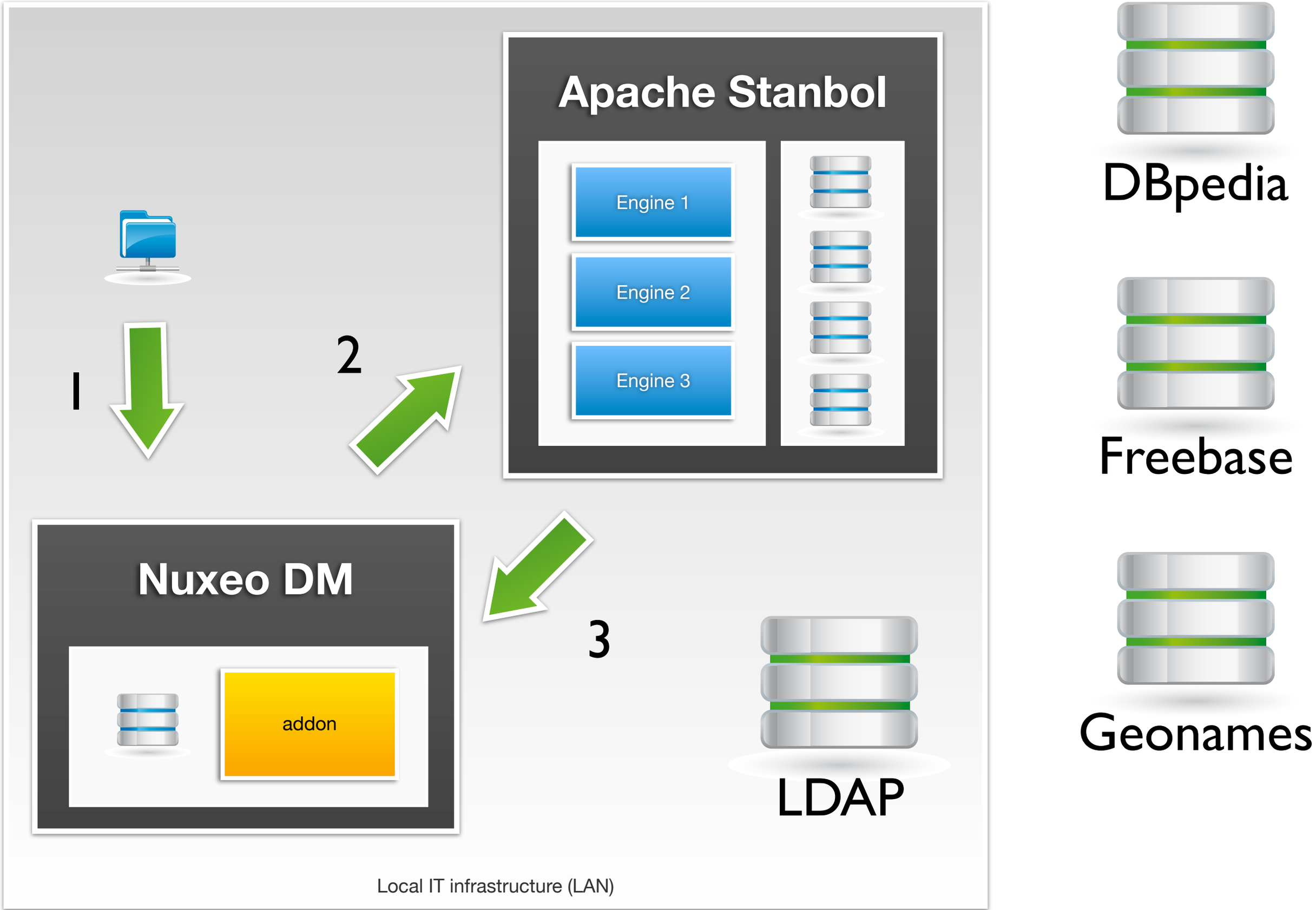
This Nuxeo DM instance is not registered to Nuxeo Connect: you won't be able to get the most recent fixes and stay up to date.
Nuxeo DM version: 5.4.3-SNAPSHOT. Register and enable Nuxeo Connect to benefit from automatic maintenance.

> How to enable Nuxeo Connect
> Subscribe to Nuxeo Connect

The Apache Software Foundation
http://www.apache.org/

=

Semantic Engines
(Apache OpenNLP)
+
Fast Linked Data local index
(Apache Solr)
+
Semantic Rule Engine
(Apache Jena)


Apache Stanbol
Engine 1
Engine 2
Engine 3
1
2
3
Nuxeo DM
addon
LDAP
DBpedia
Freebase
Geonames
Local IT infrastructure (LAN)

How to build engines?

Training statistical models for NER with Wikipedia and DBpedia

- Extract **sentences with link positions** in Wikipedia articles
- DBPedia to the **find type of the target entity** (Person, Location, Organization)
- **Apache Pig scripts** to compute the **join + format** the result as training files for OpenNLP
- **Apache OpenNLP** to build and evaluate the models
- **Apache Hadoop** for distributed processing
- **Apache Whirr** for deployment and management on Amazon EC2 cluster

```
-- Register the project jar to use the custom loaders and UDFs
REGISTER $PIGNLPROC_JAR

parsed = LOAD '$INPUT'
  USING pignlproc.storage.ParsingWikipediaLoader('$LANG')
  AS (title, wikiuri, text, redirect, links, headers, paragraphs);

-- filter and project as early as possible
noredirect = FILTER parsed by redirect IS NULL;
projected = FOREACH noredirect GENERATE title, text, links, paragraphs;

-- Extract the sentence contexts of the links respecting the paragraph
-- boundaries
sentences = FOREACH projected
  GENERATE title, flatten(pignlproc.evaluation.SentencesWithLink(
    text, links, paragraphs));

stored = FOREACH sentences
  GENERATE title, sentenceOrder, linkTarget, linkBegin, linkEnd, sentence;

-- Ensure ordering for fast merge with type info later
ordered = ORDER stored BY linkTarget ASC, title ASC, sentenceOrder ASC;
STORE ordered INTO '$OUTPUT/$LANG/sentences_with_links';
```

```
-- Load wikipedia, instance types and redirects from DBpedia dumps
wikipedia_links = LOAD '$INPUT/wikipedia_links_$LANG.nt'
 USING pignlproc.storage.UriUriNTriplesLoader(
    'http://xmlns.com/foaf/0.1/primaryTopic')
  AS (wikiuri: chararray, dburi: chararray);

wikipedia_links2 =  FILTER wikipedia_links BY wikiuri IS NOT NULL;

-- Load DBpedia type data and filter out the overly generic owl:Thing type
instance_types =
  LOAD '$INPUT/instance_types_en.nt'
  USING pignlproc.storage.UriUriNTriplesLoader(
    'http://www.w3.org/1999/02/22-rdf-syntax-ns#type')
  AS (dburi: chararray, type: chararray);

instance_types_no_thing =  FILTER instance_types BY type NEQ 'http://www.w3.org/2002/07/owl#Thi
joined = JOIN instance_types_no_thing BY dburi, wikipedia_links2 BY dburi;
projected = FOREACH joined GENERATE wikiuri, type;

-- Ensure ordering for fast merge with sentence links
ordered = ORDER projected BY wikiuri ASC, type ASC;
STORE ordered INTO '$OUTPUT/$LANG/wikiuri_to_types';
```

```
sentences = LOAD '$INPUT/$LANG/sentences_with_links'
  AS (title: chararray, sentenceOrder: int, linkTarget: chararray,
      linkBegin: int, linkEnd: int, sentence: chararray);

wikiuri_types = LOAD '$INPUT/$LANG/wikiuri_to_types'
  AS (wikiuri: chararray, typeuri: chararray);

-- load the type mapping from DBpedia type URI to OpenNLP type name
type_names = LOAD '$TYPE_NAMES' AS (typeuri: chararray, typename: chararray);

-- Perform successive joins to find the OpenNLP typename of the linkTarget
joined = JOIN wikiuri_types BY typeuri, type_names BY typeuri USING 'replicated';
joined_projected = FOREACH joined GENERATE wikiuri, typename;
joined2 = JOIN joined_projected BY wikiuri, sentences BY linkTarget;

result = FOREACH joined2
 GENERATE title, sentenceOrder, typename, linkBegin, linkEnd, sentence;

-- Reorder and group by article title and sentence order
ordered = ORDER result BY title ASC, sentenceOrder ASC;
grouped = GROUP ordered BY (title, sentenceOrder);

-- Convert to the OpenNLP training format
opennlp_corpus =
FOREACH grouped
GENERATE opennlp_merge(
   ordered.sentence, ordered.linkBegin, ordered.linkEnd, ordered.typename);
```

```
$ opennlp TokenNameFinderEvaluator -encoding utf-8 \
   -model fr-ner-location \
   -data ~/data/fr/opennlp_location/test
```

Performance evaluation for NER on a French extraction with 100k sentences

class	precision	recall	f1-score
location	0.87	0.74	0.80
person	0.80	0.68	0.74
organization	0.80	0.65	0.72

Performance evaluation for NER on a English extraction with 100k sentences

class	precision	recall	f1-score
location	0.77	0.67	0.71
person	0.80	0.70	0.75
organization	0.79	0.64	0.70

Training statistical models for topic classification from Wikipedia and DBpedia

- Filter category tree from **DBpedia SKOS entries** (~500k)
- **Pig scripts** to compute the **joins with articles abstracts** for all the articles categorized in Wikipedia
- Export as 2.8GB TSV file to be indexed in **Apache Solr**
- Use Solr **MoreLikeThisHandler** to find the top 3 most related Wikipedia category for any kind of text
- **Apache Whirr & Hadoop** for deployment and management on Amazon EC2 cluster

Wrap Up on Recent Work

- Full offline mode: Stanbol EntityHub
- Multi-lingual Indexes
- New UI for occurrences reviews
- Temis Luxid Annotation Factory integration

What's next?

- Stanbol and Temis connection in Admin Center
- Embedded Stanbol mode for easy deployment
- More OpenNLP models for more languages
- Finalize topic classification - handle hierarchy
- Tight integration with Nuxeo DM search features

Thank you for your attention!