Entity-centric Information Management and Integration

Heiko Stoermer
Technical Director, OKKAM EU Project

Information

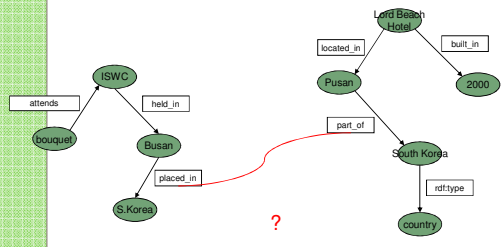
- Data: symbols
- Information: data that are processed to be useful; provides answers to "who", "what", "where", and "when" questions
- Knowledge: application of data and information; answers "how" questions
- Understanding: appreciation of "why"
- Wisdom: evaluated understanding.

- Ackoff, R. L., "From Data to Wisdom", Journal of Applies Systems Analysis, Volume 16, 1989 p 3-9.
- Taken from http://www.systems-thinking.org/dikw/dikw.htm

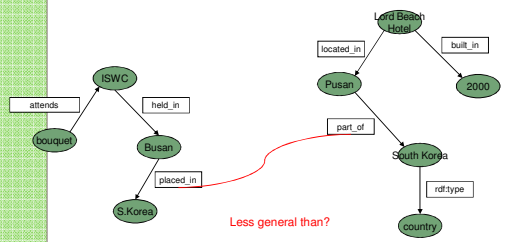
Schemas and Info-Integration

- Shared conceptualization => integrateable information!
- Issue: shared conceptualizations are problematic.

Schema-level Heterogeneity

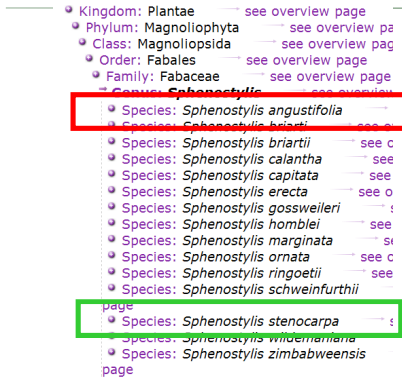


Schema-level Integration



The importance of talking about the same thing

Would classification help?



Schemas?

- Formalization/classification of the world
- ... of part of the world?
- ... of the part of the world I know?
- ... of a part of the world as I see it?
- ... of a part of the world based on my background knowledge?

Sharing Schemas - Issues

- partiality
- granularity
- approximation
- perspective
- gray areas of formalization

InfoSystems and Classifications

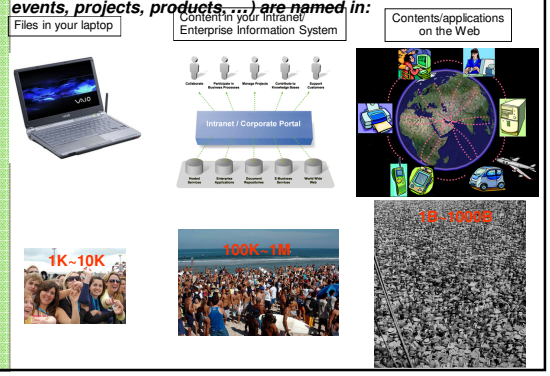
- Purpose and content of Information Systems: classes or "objects"?
- Classification as a vehicle to access an object, its specification, or information about it.

A Different Paradigm

- What we usually want is to find information about something.
- What we currently find are documents or records, attached to schemas or classifications
- We advocate
  - **Entity-centric Information Management**

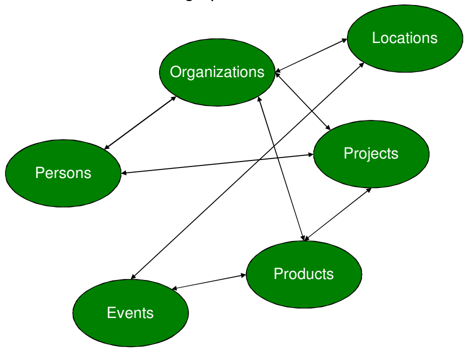
The rational

- ***How many "entities" (persons, locations, organizations, events, projects, products, ...) are named in:***



An invaluable asset

- "Entities" is what a large part of our information is about:



However …

- ***How many names, descriptions or IDs (URIs) are used for the same "entity"?***

London 런던 ܠܘܢܕܘܢ लंडन लंदन લંડન ለንደን ロンドン
লন্ডন ລອນດອນ இலண்டன் ლონდონი Llundain
Londain Londe Londen Londen Londen Londinium
London Londona Londonas Londoni Londono Londra
Londres Londrez Londyn Lontoo Loundres Luân Đôn
Lunden Lundúnir Lunnainn Lunnon لندن لندن لندن لوندون
לאנדאן לונדון Λονδίνο Лёндан Лондан Лондон
Лондон Лондон Լոնդոն 伦敦 …

The capital of UK, host city of the IV Olympic Games, host city of the XIV Olympic Games, future host of the XXX Olympic Games, the city of the Westminster Abbey, the city of the London Eye, the city described by Charles Dickens in his novels, …

http://sws.geonames.org/2643743/
http://en.wikipedia.org/wiki/London
http://dbpedia.org/resource/Category:London
…

… or …

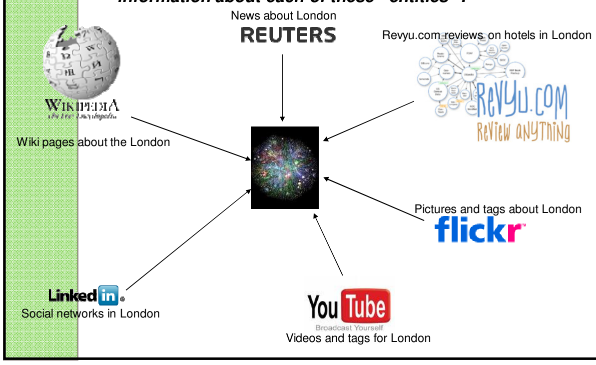
How many "entities" have the same name?

- London, KY
- London, Laurel, KY
- London, OH
- London, Madison, OH
- London, AR
- London, Pope, AR
- London, TX
- London, Kimble, TX
- London, MO
- London, MO
- London, London, MI
- London, London, Monroe, MI
- London, Uninc Conecuh County, AL
- London, Uninc Conecuh County, Conecuh, AL
- London, Uninc Shelby County, IN
- London, Uninc Shelby County, Shelby, IN
- London, Deerfield, WI
- London, Deerfield, Dane, WI

- Or
  - London, Jack
2612 Almes Dr
Montgomery, AL
(334) 272-7005
  - London, Jack R
2511 Winchester Rd
Montgomery, AL 36106-3327
(334) 272-7005
  - London, Jack
1222 Whitetail Trl
Van Buren, AR 72956-7368
(479) 474-4136
  - London, Jack
7400 Vista Del Mar Ave
La Jolla, CA 92037-4954
(858) 456-1850
  - …

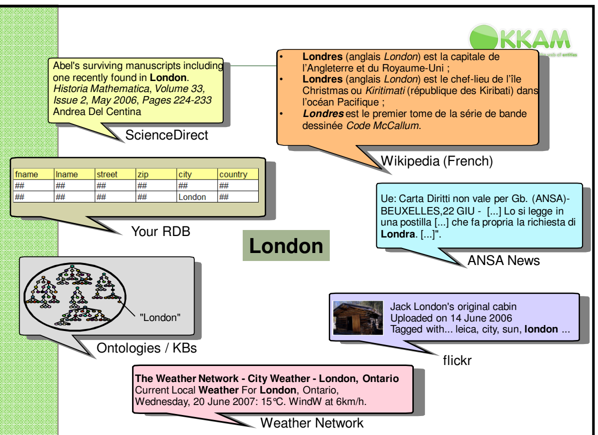
… or …

- ***How many content types / applications provide valuable information about each of these "entities"?***



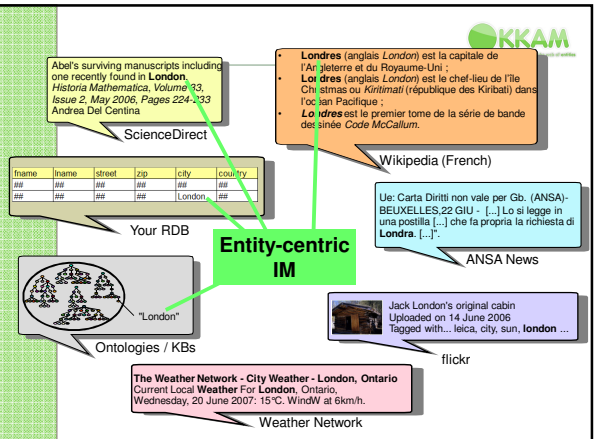
This is an immense loss of value

- Precision and recall of keyword-based **search engines** is deeply affected
- **Semantic search** still mainly relies on keyword search/matching (as very few URIs are consistently reused across RDF datasets)
- **Information integration** (e.g. from heterogeneous data sources) is hard to achieve
- **Mash-ups** from different sources are not easy to create
- **Information extraction** produces poorly integrated results (co-reference is supported mainly in the same document/corpus, not across large-scale distributed environments)

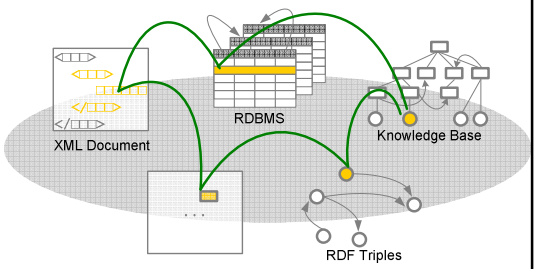


London!?

- Did you mean:
  - London, KY
  - London, Laurel, KY
  - London, OH
  - London, Madison, OH
  - London, AR
  - London, Pope, AR
  - London, TX
  - London, Kimble, TX
  - London, MO
  - London, MO
  - London, London, MI
  - London, London, Monroe, MI
  - London, Uninc Conecuh County, AL
  - London, Uninc Conecuh County, Conecuh, AL
  - London, Uninc Shelby County, IN
  - London, Uninc Shelby County, Shelby, IN
  - London, Deerfield, WI
  - London, Deerfield, Dane, WI
  - London, Uninc Freeborn County,
- Or
  - London, Jack
2612 Almes Dr
Montgomery, AL
(334) 272-7005
  - London, Jack R
2511 Winchester Rd
Montgomery, AL 36106-3327
(334) 272-7005
  - London, Jack
1222 Whitetail Trl
Van Buren, AR 72956-7368
(479) 474-4136
  - London, Jack
7400 Vista Del Mar Ave
La Jolla, CA 92037-4954
(858) 456-1850
  - ...



Entity-centric IM



Identifiers and Integration

- (Semi) Structured Data Sources provide „hooks“:
  - Relational Databases: Primary Keys
  - XML: Attributes
  - RDF/OWL: URIs
- Unstructured Data Sources require annotation:
  - RDFa
  - MicroFormats
  - Metadata

Real Big Problem is ...

Too many identifiers for the same thing out there ... not much used in content production ... and poorly interlinked or not accurate

- How can I avoid **advertising** restaurants in Venice (FL) for a query about Venice (IT)?
- How do we collect sparse information about a specific customer in a complex **Intranet** environment?
- Where do I find **product** XYZ exactly?

- In short: how can we enable mash-ups based on:

```
select * from Web where ID="..."
```

on the Web of Data and/or in an enterprise-wide Intranet ... and/or the Internet of Things?

Identify this!

What is an Entity?

- Possible Answers:
  - Everything that exists.
  - Everything that can be pointed at.
  - Everything that has a name.
  - Everything that can be talked about.
  - Everything that has determinate identity conditions.

Examples please…

- We think more about instances (individuals, objects), not so much about universal resources (e.g. classes or properties)
- Why?
  - Because schemas embody viewpoints, entities don't [let's say...]
  - Because classes in Semantic Web ontologies are *already* uniquely identified by their URI
- For example:
  - Sonia Bergamaschi, WWW2007, UNITN, London, UK are all entities
  - Pegasus, π, √2, ... are entities
  - "ComputerScience" as a topic *may* be an entity (borderline ...)
  - "Pizza Margherita" in a food ontology ... ??
  - "La Divina Commedia", "MS Word", "VW Golf" are tricky entities ...
  - The class "Person" and the property "Has_email_address" are not entities
- Types of entities to start with: people, locations, organizations, events, products, ...

Entity Identification: HOWTO

- Describe
- Identify
- Reference
- Integrate

Describing an Entity

- Specification vs. Conceptualization
  - no unique, globally valid classification of an entity
  - classification of an entity not always sufficient for differentiation
  - no one necessary and sufficient set of attributes to describe an entity
- => schemaless approach

Identifying an Entity

- **Give an identifier**!
- Required: A globally unique, findable, rigid identifier for all types of entities.
- Semantic Web standard: fully qualified URIs.

Referencing an Entity

- With a good, findable identifier, reference is as easy as creating a hyperlink in HTML.

Integrating Information

- With a good, a-priory approach for the reuse of identifiers, integration is straight-forward and syntactical.

The OKKAM Vision

- **Building The Entity Name System**
- An architecture and infrastructure in development to enable **entity-centric information management**.
- Approach:
  - issuing globally unique, rigid identifiers for entities
  - enabling you to find and reuse my entities, so we can finally talk about the same objects and integrate our information *correctly*
  - referencing external information about entities

The OKKAM Project

Project Title	**Enabling the Web of Entities**. A scalable and sustainable solution for systematic and global identifier reuse in decentralized information environments
Acronym	**OKKAM**
Starting date	01/01/2008
End date	30/06/2010
Duration	30 months
No. of partners	12 (5 universities, 4 SMEs, 3 large companies)
Effort	763 person months (more that 63 person years)
Total cost	7.352.931,34 Euro
EU contribution	5.125.000,00 Euro
Type	Large-scale Integrating Project

The OKKAM consortium

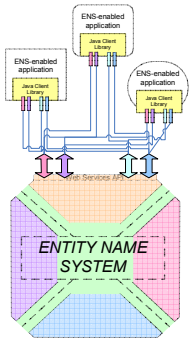
1. University of Trento, IT
2. L3S Hannover, DE
3. SAP Research, DE
4. Elsevier, NL
5. Expert System, IT
6. Europe Unlimited, BE
7. MAC, IE
8. EPFL, CH
9. DERI Galway, IE
10. University of Malaga, SP
11. INMARK, SP
12. ANSA, IT

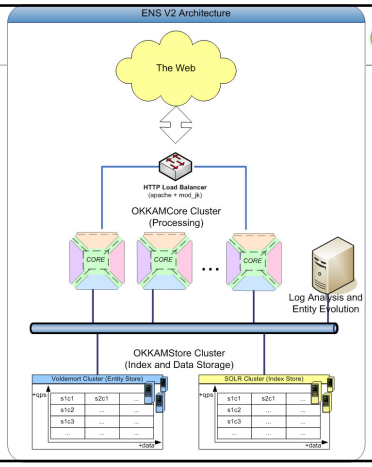
5 universities

4 SMEs

3 corporations

Overview



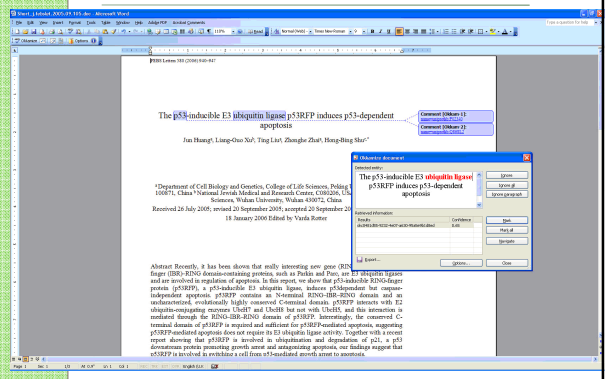


Three OKKAM applications

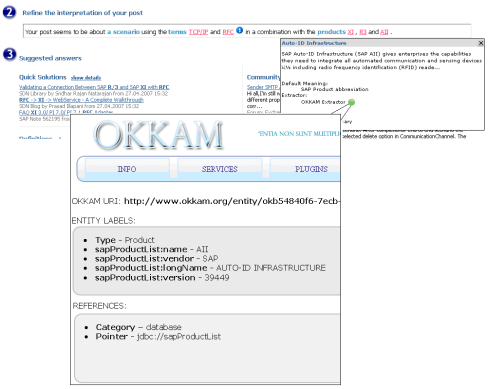
In the project, we have three application scenarios:

- Entity-centric **semantic search engine** (DERI Galway)
- Entitity-centric **organizational knowledge management systems**: entity-centric product lifecycle management within SAP
- Multimedia **authoring environments**: supporting the creation of news (ANSA) and scientific papers (Elsevier) in an ENS-empowered environment

Entity-centric Authoring



Entity-centric Enterprise IKM



Entity-centric Search

ENS-empowered tools

- Any application which is used to create/edit content can be extended with "plugins" which enable the application to get from the ENS the right ID for a named entity
- Tools we have already empowered:

Annotating texts created with MS Word with IDs automatically obtained from the ENS

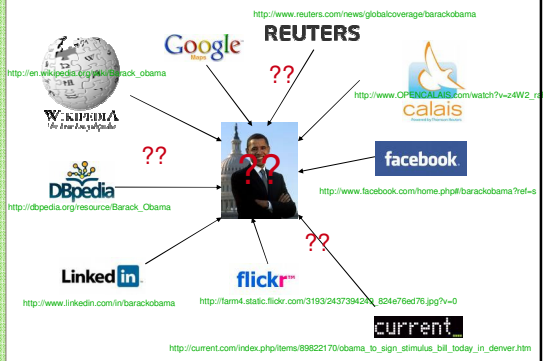
Reusing URIs in RDF/OWL knowledge bases built with Protégé with IDs automatically obtained from the ENS

Introducing global IDs to content created through web-based authoring interfaces (e.g. FOAF content)

Entity-centric Browsing 1

Entity-centric Browsing 2

Today…

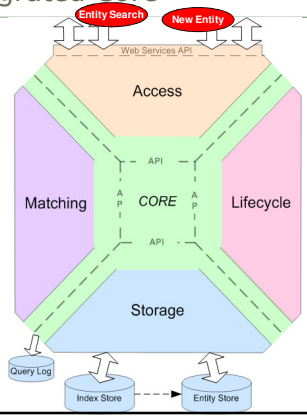


Tomorrow…

The OKKAM Core

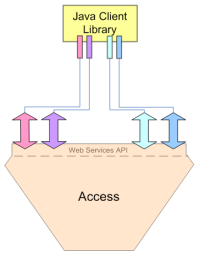
- Integrates results and contributions from
  - Storage Layer
  - Entity Matching
  - Entity Import
  - OkkamCore Infrastructure
  - Lifecycle Management
- About 10 individual direct contributors from 5 partner

Integrated Core



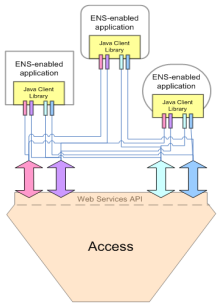
Stable API and Services

- Internal Interfaces
- Query language
- Web Services
  - Entity Search
  - Entity Creation
  - Entity Retrieval
  - Alternative ID Lookup
- Client Library for Java



Running Tools

- Proof-of-concept tools that are running today:
  - OkkamWebSearch
  - Foaf-O-Matic
  - Plugin for MS Word



How we got there...

- Traditional split of Development vs. Production environments
- Production node:
  - UNITN (temporary)
- Development nodes:
  - EPFL (Storage)
  - L3S (Matching)
  - MAC (Integration)
  - UNITN (Lifecycle & Import)