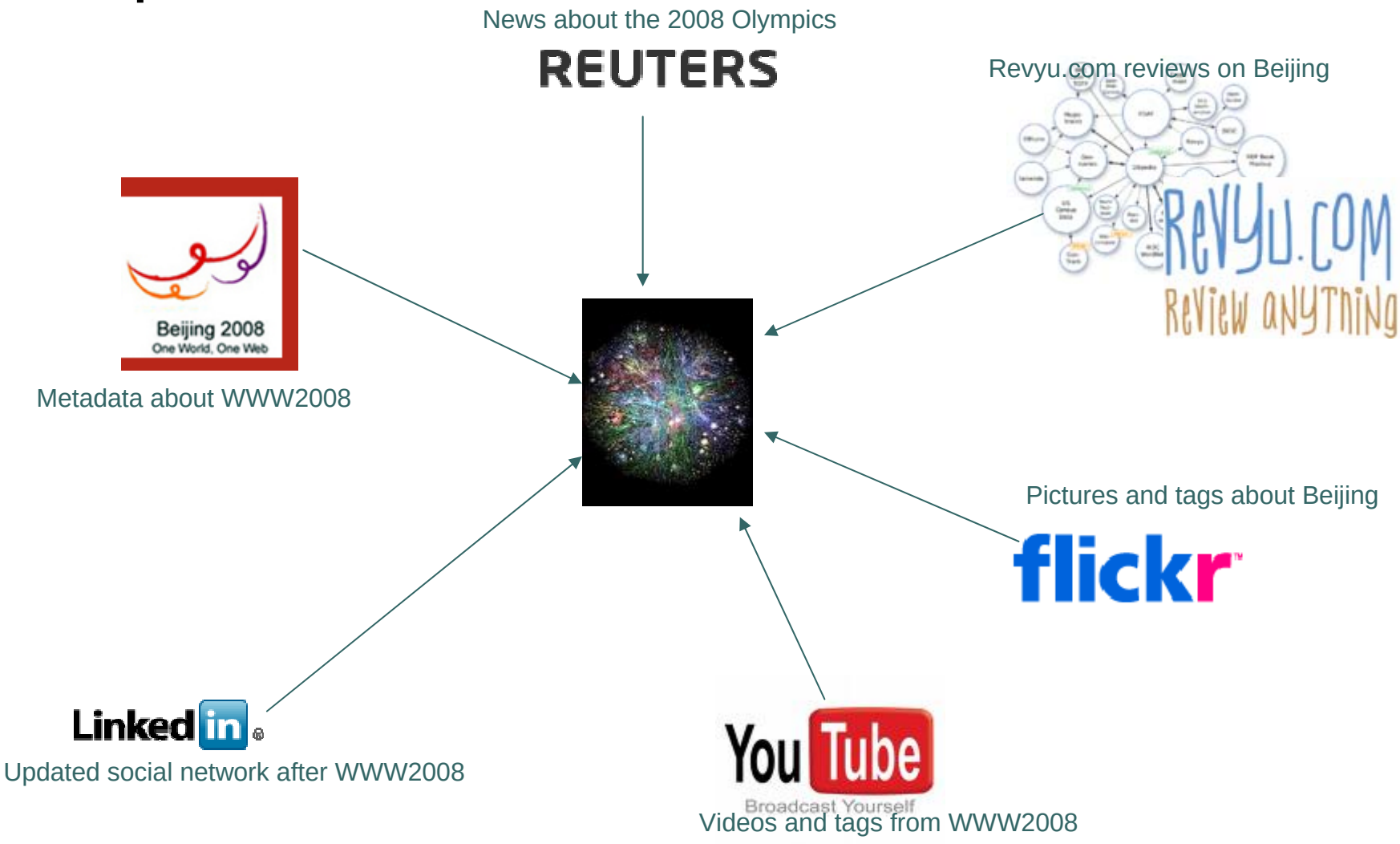
# An Entity Name Systems (ENS) for the [Semantic] Web

Paolo Bouquet
University of Trento (Italy)
Coordinator of the FP7 OKKAM IP

LDOW @ WWW2008 – Beijing, 22 April 2008

# An ordinary day on the [Semantic] Web



# Lots of new “linked data” about Beijing?

- Not quite … [see the idea of “information islands from Falcons]
- The reference to Beijing is somehow “hidden” behind:
  - Different names (e.g Beijing vs. Peking) in text documents
  - Different URIs are used in different RDF files
  - Different metadata schemas / vocabularies
  - Different keys in databases
  - …

# So what can’t we (easily) do?

- Straight integration of RDF content via simple graph merging
- Reasoning requires mapping beforehand
- Linking multimedia (and Web2.0) content to RDF content
- Getting the best from business intelligence / Web mining apps
- Multimedia search
- …

# What can we do about it?

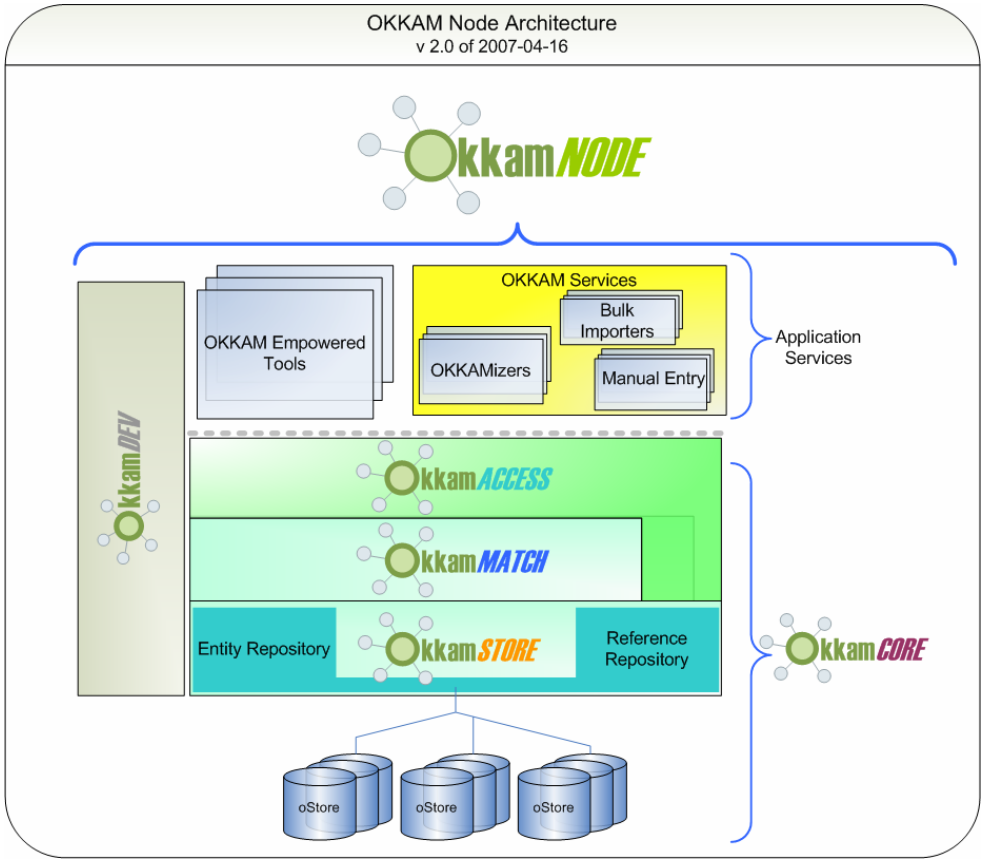
OKKAM aims at developing an **Entity Name System** for the [Semantic] Web which can make sure that the same entity (individuals, like person, location, organization, event, product, …) is referred to through the same URI across:

- any type of content / format
- any application
- any domain

all over the Web … and beyond

# How does it work?

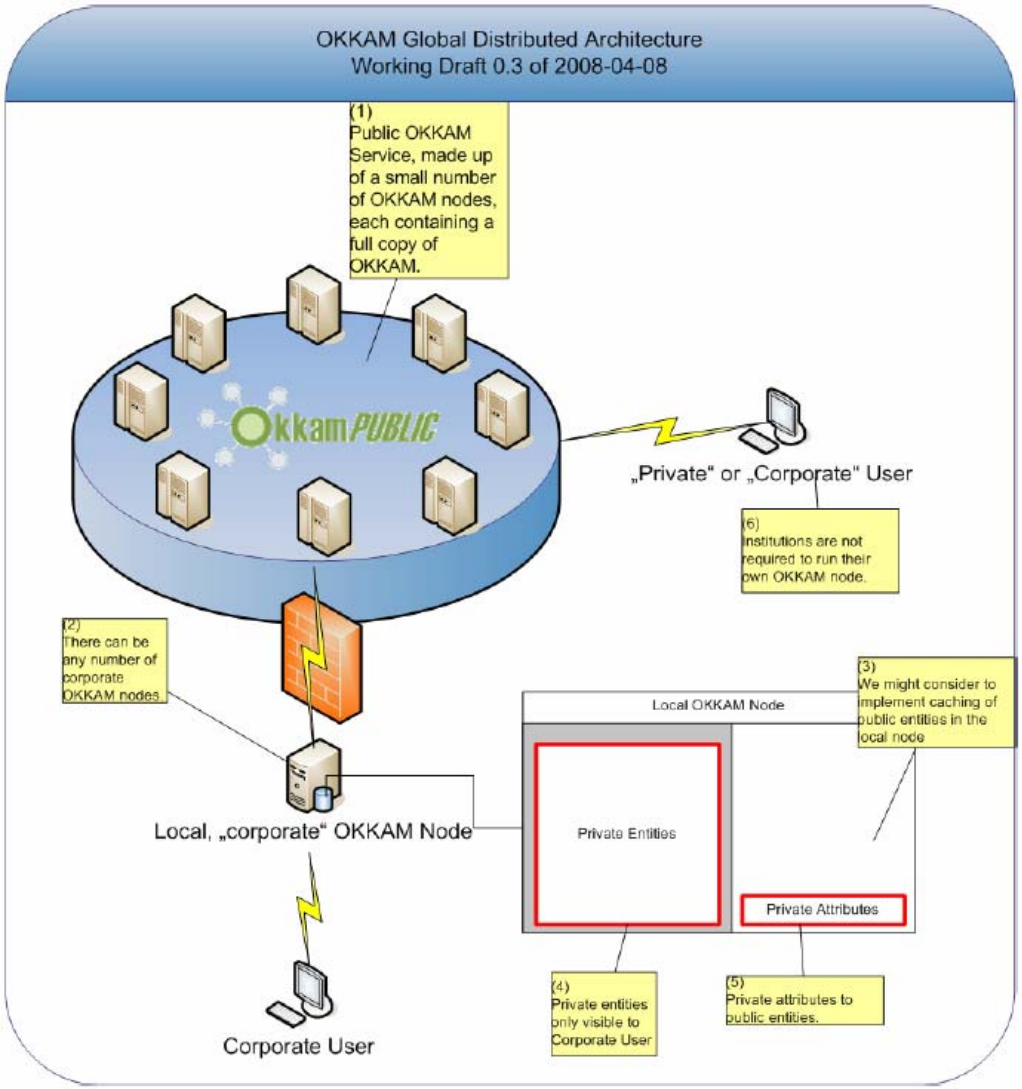
[How we are trying to do it in the EU-funded project OKKAM]




http://fp7.okkam.org

- Global distributed storage of URIs of billions of URIs
- Supports entity matching for finding an entity's URI (based on simple profiles + links to external resources)
- Provides simple APIs for any human/application which needs to find the URI of an entity
- Makes available services for the automatic annotation of different content types with global URIs
- Offers secure and trustworthy methods for access control

# Global and decentralized




- Replicated public nodes for the Web
- Local “corporate” nodes for non public data (+ cache)

# Entity representation schema (ERS): the key concepts

- The ENS repository stores existing URIs + a representation of the corresponding real world entity
- This representation is not meant as a source of info about the entity, it is only used to maximize the chance of getting matching right (like a phone directory)
- In OKKAM, an entity representation has 4 main elements:
  - A OKKAM URI for the entity
  - An entity **profile**
  - A collection of **metadata**
  - A list of **alternative URIs** (including the **preferred URI**, if any)

# ERS: Entity profiles

- Three main elements:
  1. A semantic type (but we support only a small number – 8 to 10 – very high level categories, the rest must be found out there on the Web …)
  2. A collection of name/value pairs (but very few, those which are most likely – or most used – to make sure that we got the right URI)
     - [We don't assume any predefined vocabulary for attributes (though we may suggest a few ones for improving matching)]
  3. A collection of typed links to external resources (RDF stores, HTML pages, PDF files, multimedia resources, …) which refer to that entity

# ERS: Entity metadata

Four main elements:

1. **General metadata** (e.g. creation time)
2. **Statistics metadata** (e.g. last modified, # of time retrieved, # of time selected, time last selected)
3. **Provenance metadata** (e.g. source, agent)
4. **Access control metadata** (e.g. owner, authority, subordination)

[Metadata are available also for every single name/value pair of an entity profile]

# ERS: alternative URIs

- A collection of alternative URIs (aliases, synonyms) for the same real world entity
- One of them can be marked as preferred and can be always returned to users/application instead of the internal ENS URI

Dereferincing alternative URIs may provide background knowledge for advanced entity matching methods

# Entity matching

Obviously related to well-known problems: record linkage, deduplication, entity resolution, disambiguation, …

The ENS basic use case is as simple as follows:

- An application needs to find the URI for an entity
- From local information a look-up query is composed (mainly simple keywords or name/value pairs)
- The ENS tries to find the entities in the ENS repository which better matches the query
- A ranked list of results is returned (ranking is based both on similarity measures and statistical information on social use of the ENS)

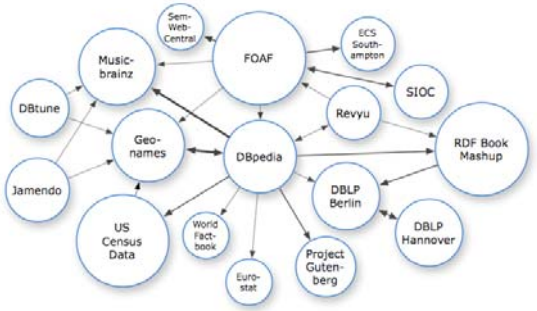
# ENS-enabled tools

Content creation tools which are extended to interact with the ENS

Our first prototypes:

1. Foaf-O-matic: create your FOAF profile with pre-existing URIs
2. Okkam4P: a Protégé 3.3.x plugin for creating individuals/instances with pre-existing URIs
3. O4MSW: a MS Word plugin for annotating MS Word files with pre-existing URIs

# ENS and Linked Data




- Complementary, in principles not competitors (though the ENS is mainly about reusing URIs in content creation, Linked Data is about linking data about a resource)
- The Linked Data content is a fantastic source of entities and name/value pairs for building entity profiles in the ENS
- Lots of methods and tools used for URI disambiguation can be shared and reused
- The ENS can be used by Linked Data tools to look-up for URIs in a single simple service (through APIs)
- The extension to non-RDF content may allow linking RDF data with unstructured data on the Web

# However …

- The ENS is based on the idea that, in general, having multiple URIs for the same thing is a bug, not a feature (is good for browsing, not for information integration and reasoning)
- *Ex post* vs. *Ex ante* approach: using billions of distributed `owl:sameAs` statements will become impractical. Hopefully, the use of a single URI for the same entity may simplify the global graph of the Semantic Web
- The practice of using `owl:sameAs` for interlinking heterogeneous URIs is semantically disputable
- URIs should not encode an identity, so it should make no difference which URI is used for an entity (provided it is unique and standard)
- The Linked Data methods do not support well the creation of new URIs

# An *extra*ordinary day on the [Semantic] Web

http://www.okkam.org/entity/
ok78dfda18-2c96-45a5-a7e5-9093ed919424

REUTERS

http://www.okkam.org/entity/
ok78dfda18-2c96-45a5-a7e5-9093ed919424

Beijing 2008
One World, One Web

http://www.okkam.org/entity/
ok78dfda18-2c96-45a5-a7e5-9093ed919424

http://www.okkam.org/entity/
ok78dfda18-2c96-45a5-a7e5-9093ed919424

flickr

LinkedIn

http://www.okkam.org/entity/
ok78dfda18-2c96-45a5-a7e5-9093ed919424

YouTube
Broadcast Yourself

http://www.okkam.org/entity/
ok78dfda18-2c96-45a5-a7e5-9093ed919424