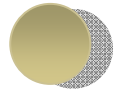
An Entity Name System (ENS) for the Semantic Web

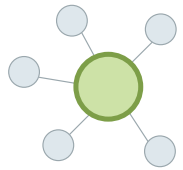
ESWC2008

Paolo Bouquet, **Heiko Stoermer**, Barbara Bazzanella

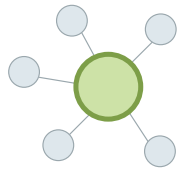
University of Trento, Italy

2008-06-05





- **Introduction and Motivation**
- **The Semantic Web Vision Revisited**
- **The Entity Name System**
- **Issues and Discussion**
- **Outlook**



Introduction and Motivation

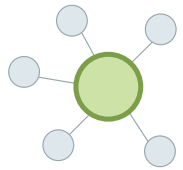
The Semantic Web Vision Revisited

The Entity Name System

Issues and Discussion

Outlook

An ordinary day on the Semantic Web



News about the Social Dinner

REUTERS

Revyu.com reviews on the Sheraton

Metadata about ESWC2008

Pictures and tags about ESWC2008

flickr

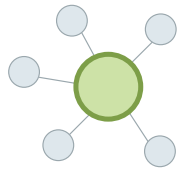
LinkedIn

Updated social networks after ESWC2008

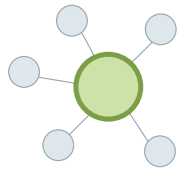
YouTube
Broadcast Yourself

Videos and tags from ESWC2008

Lots of „linked data“ about Tenerife?



- Not quite …
- The reference to Tenerife is somehow “hidden” behind:
  - Different names (e.g Tenerife vs. Teneriffa) in text documents
  - Different URIs are used in different RDF files
  - Different metadata schemas / vocabularies
  - Different keys in databases/XML documents
  - …
- What can be “nice to have” in the Web is a real problem in other contexts.



Introduction and Motivation

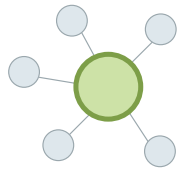
The Semantic Web Vision Revisited

The Entity Name System

Issues and Discussion

Outlook

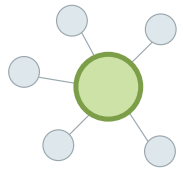
Semantic Web: a long-term vision



The Semantic Web is what we will get if we perform the same globalization process to knowledge representation that the Web initially did to hypertext.

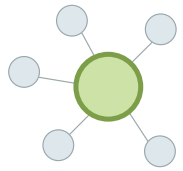
[Tim Berners-Lee, *What the semantic Web isn't but can represent* , 1998]

Semantic Web key ideas: a summary



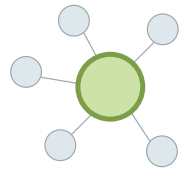
- Names in natural language (like “Tenerife” and “Teneriffa”, “Paolo”, “Paolo Bouquet” and “Bouquet, P.”) can be ambiguous or not unique
- Therefore, when we want to make a statement about a resource, we must use its identifier
- When two nodes in two RDF graphs have the same identifier (URI), they unambiguously refer to the same resource
- The global knowledge space is achieved by applying the operation of merging local graphs into a single (virtual, decentralized) global graph
- Now the virtual global graph can be queried as if it was a single knowledge base

Power to the URI



In our opinion, the concept of the URI to denote entities, and the resulting Global Graph vision, is of of **the** most important distinctions between classic KR and the Semantic Web

The Semantic Web Today



dblp.uni-trier.de

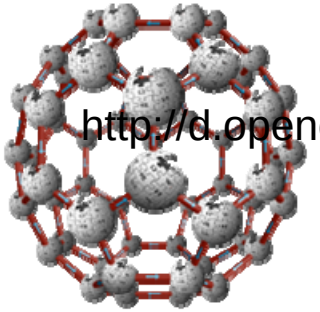
http://dblp.l3s.de/d2r/resource/authors/Frank_van_Harmelen

http://www.ivan-herman.net/foafExtras.rdf#FrankH

http://dbpedia.org/resource/Frank_van_Harmelen

http://irit.rkbexplorer.com/id/person-4beda57f85d62fab8c6c6cfb7559b7d7

http://irit.rkbexplorer.com/id/person-fedcd2ec9170142953094ba1d46945ae

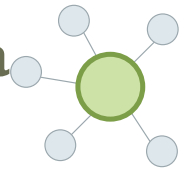


http://revyu.com/people/Frank

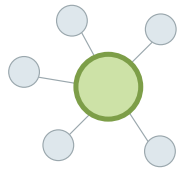
http://d.opencalais.com/pershash-1/5bfcc349-4cf8-3cb3-8259-3681aa40d669

http://ontoworld.org/wiki/Special:ExportRDF/Frank_van_Harmelen

SemWeb Community approach: Linked Data

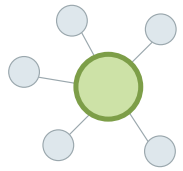


- Main ideas:
  - Proliferation of URIs for entities is unavoidable
  - Let's use the owl:sameAs property to link from one URI to another
  - Create heuristics to find identity between entities
- Issues:
  - Who creates the sameAs statements?
  - Where are the statements stored?
  - What about logical implications of owl:sameAs?
  - Who implements the massive machinery that reasons over the transitive closure of owl:sameAs statements in a globally distributed KB?



Introduction and Motivation
The Semantic Web Vision Revisited
The Entity Name System
Issues and Discussion
Outlook

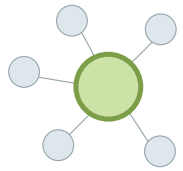
Our proposal: from DNS to ENS



We propose an a-priori approach, an Entity Naming System (ENS):

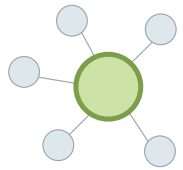
- Basic idea: any description of an entity is "resolved" into its global ID
- Building blocks: ENS servers (repository + "resolution" of names)
- An open, public service which can be invoked by any application in which entities are mentioned

The OKKAM Project



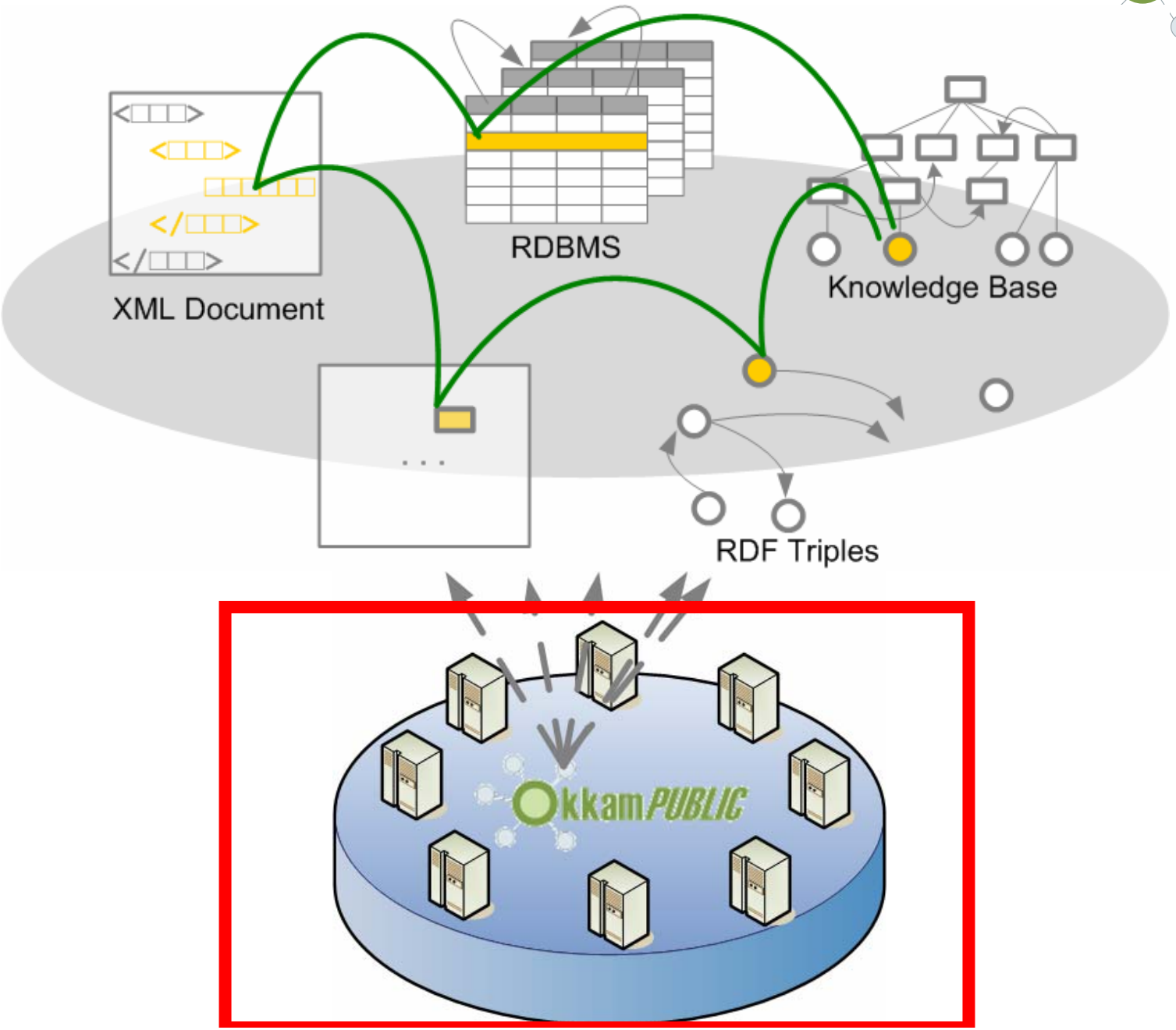
- An architecture and infrastructure to foster the systematic re-use of identifiers for entities.
- Under development in the context of the European Integrated Project „OKKAM“ from 2008 to 2010.
- Approach:
  - issuing globally unique, rigid identifiers for entities
  - enabling you to find and reuse these identifiers, so we can finally talk about the same objects and integrate our information *correctly*
  - indexing external information about entities

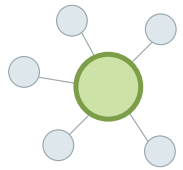
But....



- **Do we need this? Many things can already be identified!**
- Existing Approaches:
  - Entity URIs
  - RFID
  - LSID
  - OpenID
  - DOI/ISBN
  - Wikipedia page
  - ...
  - Problems: Proliferation, verticality, findability (identifiers and systems), non-rigidity, superficiality
- Some "good" approaches exist, and interoperability with them should be pursued

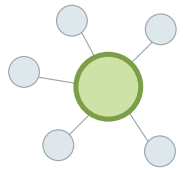
Entity-centric Information Integration





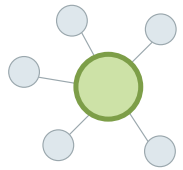
The OKKAM ENS Prototype

ENS Premises



- "Phone Book" vs. Knowlege Base
  - We do not attempt to create a KB about entities
  - We store entity descriptions for only two reasons:
    - distinguishing entities from another
    - finding entities and their identifiers
  - We do not model strong typing

Entity representation in the ENS

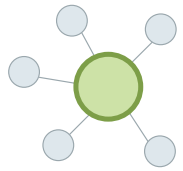


- The ENS repository stores existing URIs + a representation of the corresponding real world entity
  - => Entity Representation Schema (ERS)

- This representation is not meant as a source of information about the entity, it is only used to maximize the chance of getting a match (like a phone directory)

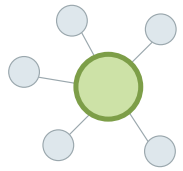
- In OKKAM, an entity representation has 4 main elements:
  - An **ENS URI** for the entity
  - An entity **profile**
  - A collection of **metadata**
  - A list of **alternative URIs**

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

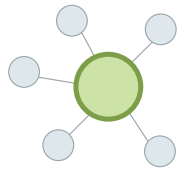


Four main elements:

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]

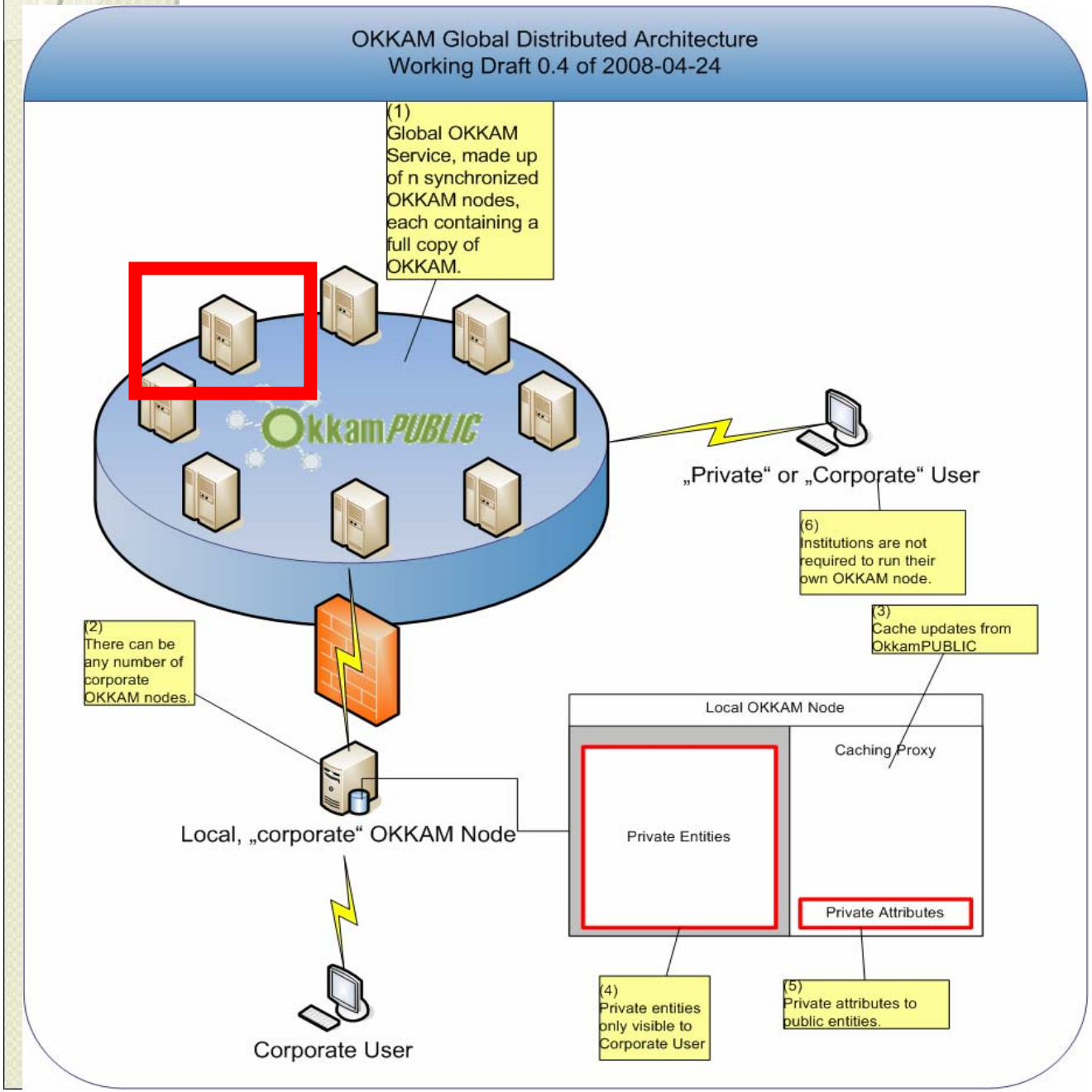
ERS: alternative URIs



- 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

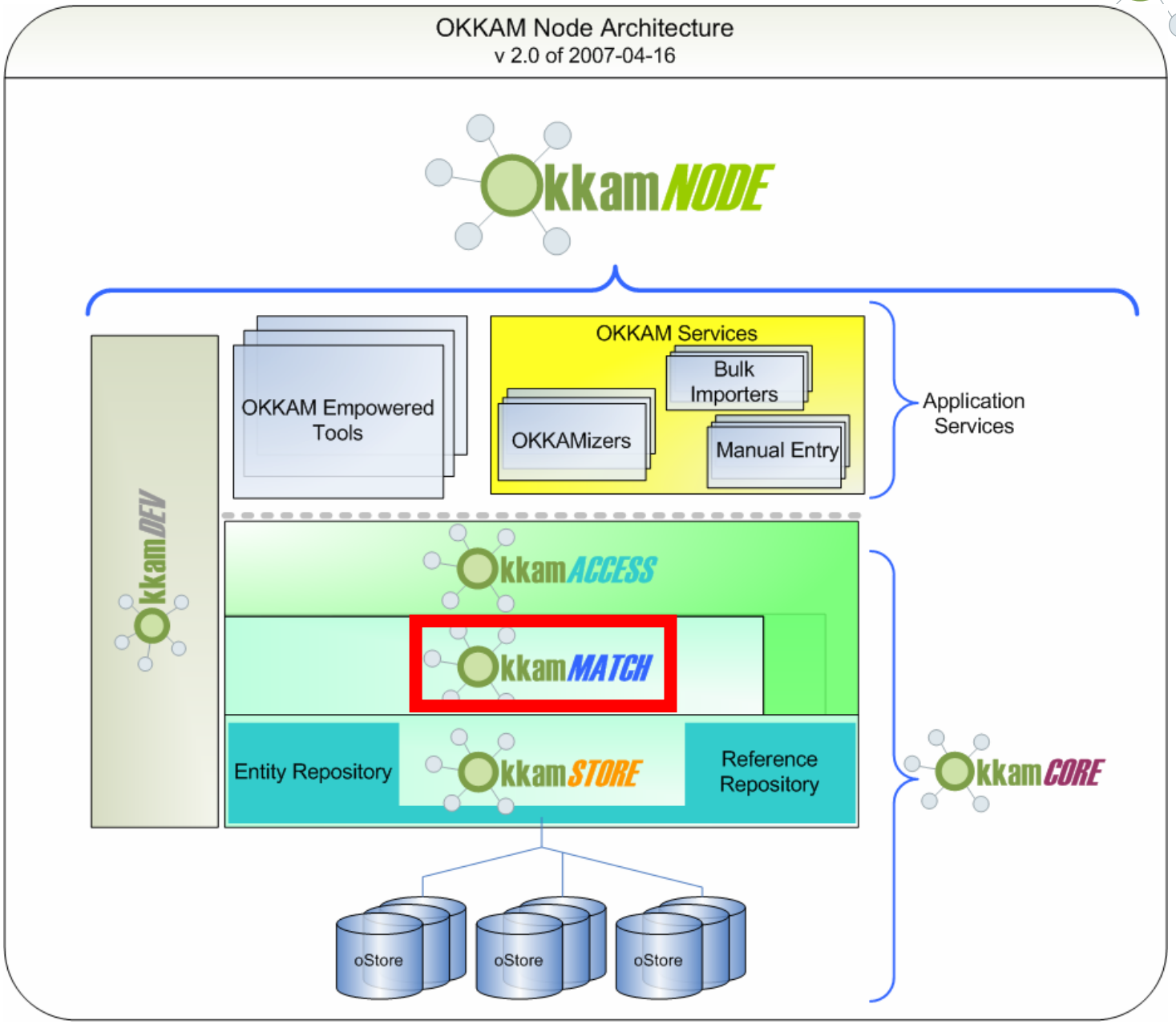
Dereferencing alternative URIs may provide background knowledge for advanced entity matching methods

OKKAM ENS – Global and Decentralized

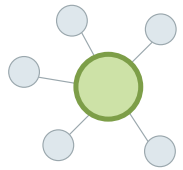


- Replicated public nodes for the Web
- Local „corporate“ nodes for non-public data (and cache)

One OKKAM Node

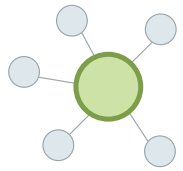


Okkam*MATCH*: Motivations



- Begin with a baseline algorithm that is generic, i.e. independent of
  - representation/formalization
  - existance of certain data
  - typing
  - special heuristics
- Create a benchmark for future developments
- Provide architecture that allows for new algorithms to be plugged and evaluated against the baseline

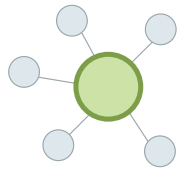
Okkam*MATCH*: Ranking



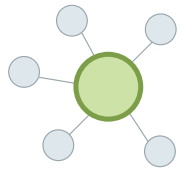
IR-based approach:

- input query and entity profile can be seen as "documents"
- IR knows distance measures
- We use "Monge-Elkan" field matching to compute the similarity between query and candidate profiles on the fly.
- This allows us to return a ranked list instead of just a result set from the data store.

A value-based ranking algorithm

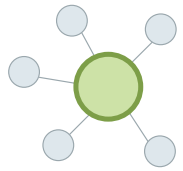


```
q = concatenate(valuesOf(query))
forall candidates
  p = concatenate(valuesOf(profile))
  s = computeSimilarity(p,q)
  rankedResult.store(s)
rankedResult.sort()
```



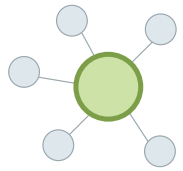
Experimental results

Okkam*MATCH*: Experimental Results



- Experiment:
  - align two populated ontologies (ISWC2006 & ISWC2007) with the help of the ENS
  - merge ontologies
  - compare entity overlap with manually established standard
  - performed on "person" entities

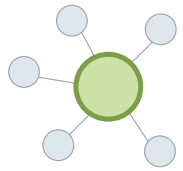
Okkam*MATCH*: Integration Experiment



	Okkam*MATCH*	Gold standard
Total Positives	68	48
True Positive	45	48
True Negative	385	403
False Positive	20	0
False Negative	1	0
Precision	0.69	1
Recall	0.98	1
F-Measure	**0.81**	**1**

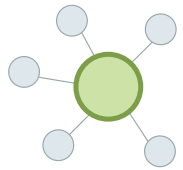
- Results*
  - high recall
  - moderate precision

*results for similarity threshold of 0.90 which has found to be "optimal"

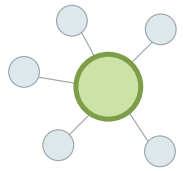


Introduction and Motivation
The Semantic Web Vision Revisited
The Entity Name System
Issues and Discussion
Outlook

Identity and Reference on the SemWeb

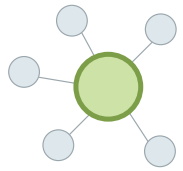


- Outcomes of the IRSW2008 Workshop @ ESWC
  - Controversy: what's in a URI?
  - Proliferation vs. Convergence
  - Centralized vs. Decentralized Mgmt
  - Browsing vs. Reasoning



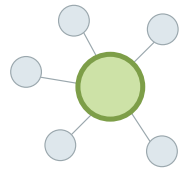
Introduction and Motivation
The Semantic Web Vision Revisited
The Entity Name System
Issues and Discussion
Outlook

Improvements for 2008



- Move from naive relational data store to a combination of HBase distributed storage backend and Lucene indexing
  - (=> first „serious“ population of entities)
- Move from generic, naive entity matching to new matching architecture
  - (=> better performance ;-))
- More OKKAM-empowered tools
  - MSWord plugin for entity annotation
  - New version of Foaf-O-Matic
  - NeOn plugin
  - Firefox plugin
  - ...

An *extra*ordinary day on the Semantic Web



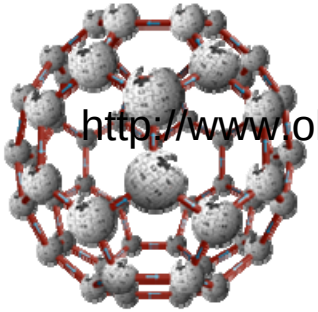
dblp.uni-trier.de

http://www.okkam.org/entity/ok20070630118580279 7287

http://www.okkam.org/entity/ok20070630118580279 7287

http://www.okkam.org/entity/ok2007063011858027

http://www.okkam.org/entity/ok20070630118580279 7287



http://www.okkam.org/entity/ok20070630118580279 7287

http://www.okkam.org/entity/ok200

http://www.okkam.org/entity/ok20070630118580279 7287

Please participate in our experiment!

Win an iPod!


fp7.
kkam
.org