# Identity and Reference on the Semantic Web

DEFINITIONS, CHALLENGES, OPPORTUNITIES ….

AND A PARTIAL BIBLIOGRAPHY

Paolo Bouquet (University of Trento, Italy)

*AASSW2007*

Busan (South Korea), 7 November 2007

# The Semantic Web vision: a global knowledge space

"Knowledge representation is a field which currently seems to have the reputation of being initially **interesting, but which did not seem to shake the world** to the extent that some of its proponents hoped.

It made sense but was of limited use on a small scale, but **never made it to the large scale**. <u>This is exactly the state which the hypertext field was in before the Web</u>.

Each field had made certain **centralist assumptions** -- if not in the philosophy, then in the implementations, which prevented them from spreading globally.

But each field was based on fundamentally sound ideas about the representation of knowledge.

**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**. *We remove the centralized concepts of absolute truth, total knowledge, and total provability, and see what we can do with limited knowledge*".

[Tim Berners-Lee, *What the Semantic Web can represent*, 1998]

# Lecture outline

- How the Semantic Web is trying to implement this grand vision
- Two problems in creating the global knowledge space:
  - Schema-level integration
  - Instance-level integration
- The critical role of identity and reference in solving these issues
- Three big issues for identity and reference on the Semantic Web:
  1. The relationship between *identity* and *identifiers*
  2. The idea of using HTTP URIs
  3. An infrastructure for supporting reference through global identifiers (ideas and a proposal)
- Conclusions

# A few basic definitions

**Resource**

- “a resource can be anything that has identity” [RFC2396]
- “the term 'resource' is used in a general sense for whatever might be identified by a URI” [RC3986]
- “... whatever can be identified by a URI, or anything which can be the subject of a discourse, such as cars, people, etc.” [webarch]

**URI (Uniform Resource Identifier)** is “a compact sequence of characters which that identifies a physical or abstract resource” [RFC3986]

**URIs, URLs, URNs:** The **classical view** held that an identifier might specify the location of a resource (a URL) or its name (a URN), independent of location. Thus a URI was either a URL or a URN. The **contemporary view** holds that Web-identifier schemes are, in general, URI schemes, as a given URI scheme may define subspaces.

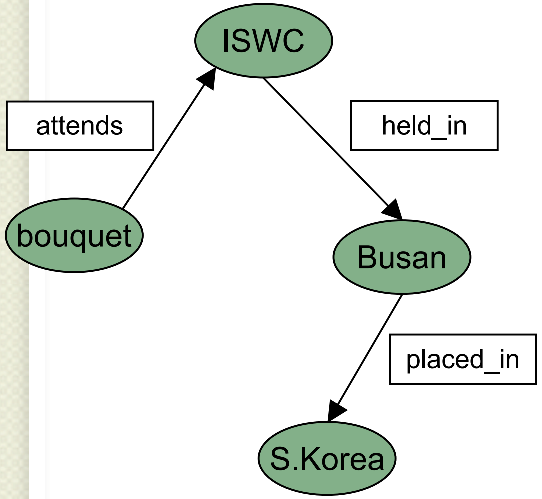
**URI scheme**: the top level of a URI naming structure (e.g. http, file, urn, mailto). They are often associated with a protocol used to access the named resource.

# The Semantic Web: from vision to practice

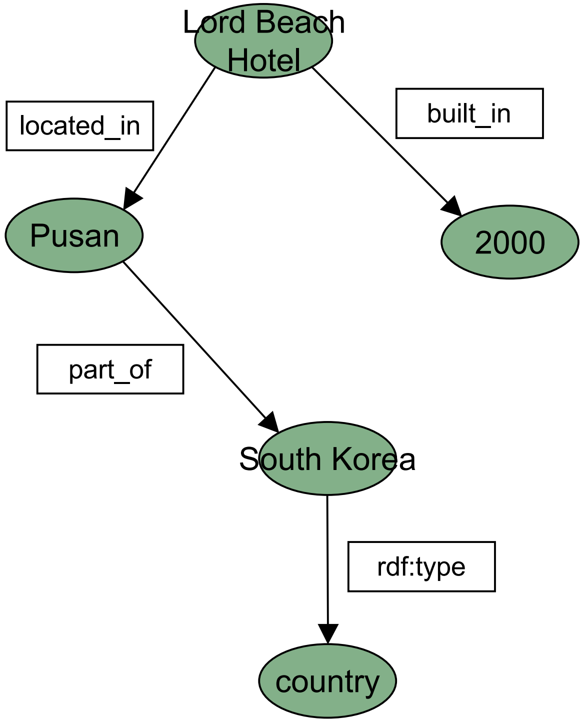
- How the Semantic Web can enable this vision:
  - Anyone can create and publish collections of RDF statements about any collection of resources (documents, people, locations, events, products, topics, …).
  - [If information is contained in databases (for example, in relational form), it can be exported to RDF with some additional work].
  - Any collection of RDF triples defines a “local” graph, whose nodes are resources and arcs are relations between resources.
  - The meaning of relations can be defined in vocabularies/ontologies (using RDFS or OWL), which are themselves specified as graphs
  - Different graphs can be glued together in a virtual global graph of knowledge, which can now be browsed, searched and reasoned about.
- The expected outcome is *the global decentralized space of knowledge* envisaged by Tim Berners-Lee.

# An example …

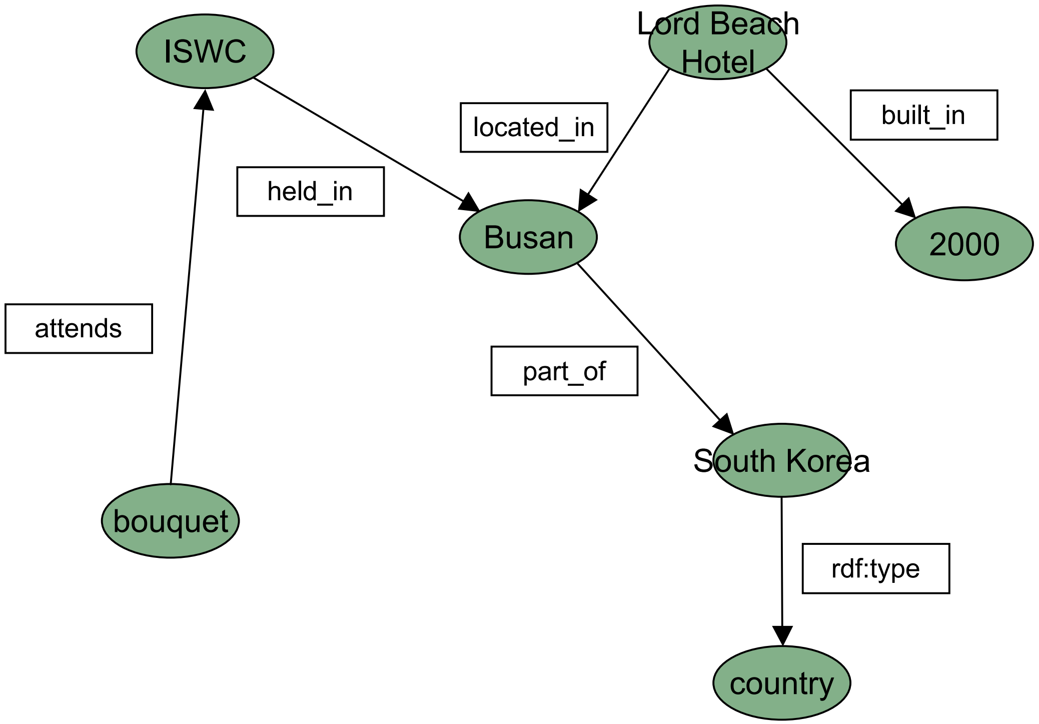
ISWC2007.org



hotel_pusan.org



# An example …



Query: find me an a hotel located where ISWC is held in 2007

# Key ideas: a summary

- Names in natural language (like “Busan” and “Pusan”, “Paolo”, “Paolo Bouquet” and “Bouquet, P.”) can be ambiguous or not unique
- Therefore, when we want to make a RDF statement about a resource,  we must use its URI
- When two nodes in two graphs have the same 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

# An example of RDF statements

```
<rdf:RDF
   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
   xmlns:aassw="http://www.aassw.org/aassw#"
   xmlns:unitn="http://www.unitn.it/unitn-ns#">

   <rdf:Description rdf:about=“bouquet">
     <aassw:name>Paolo Bouquet</aassw:name>
     <unitn:title>Associate Professor</unitn:title>
     <aassw:teaches rdf:Resource=“IRSW”/>
   </rdf:Description>
…
```

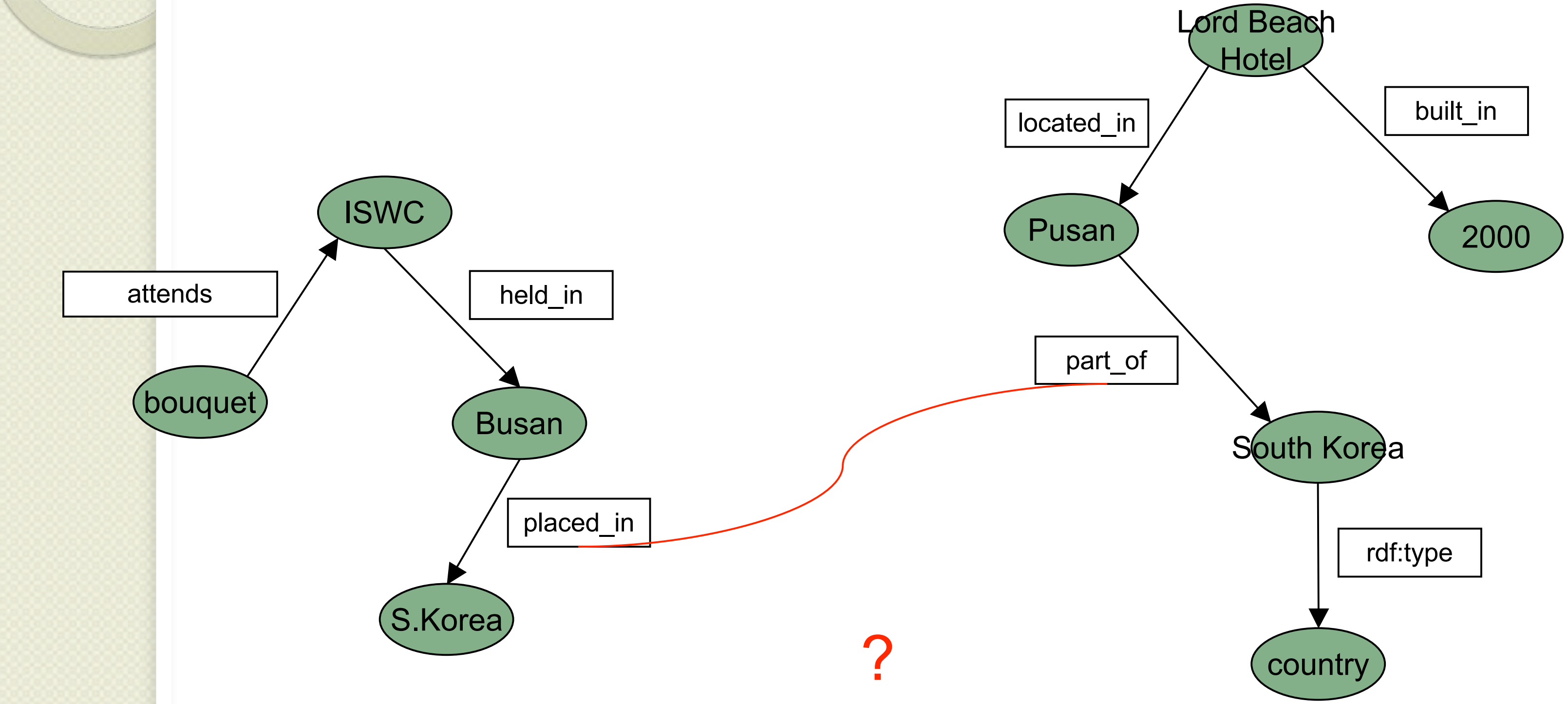
# Knowlegde integration in the global space

To implement this scenario, however, we need to address and solve two serious integration problems:

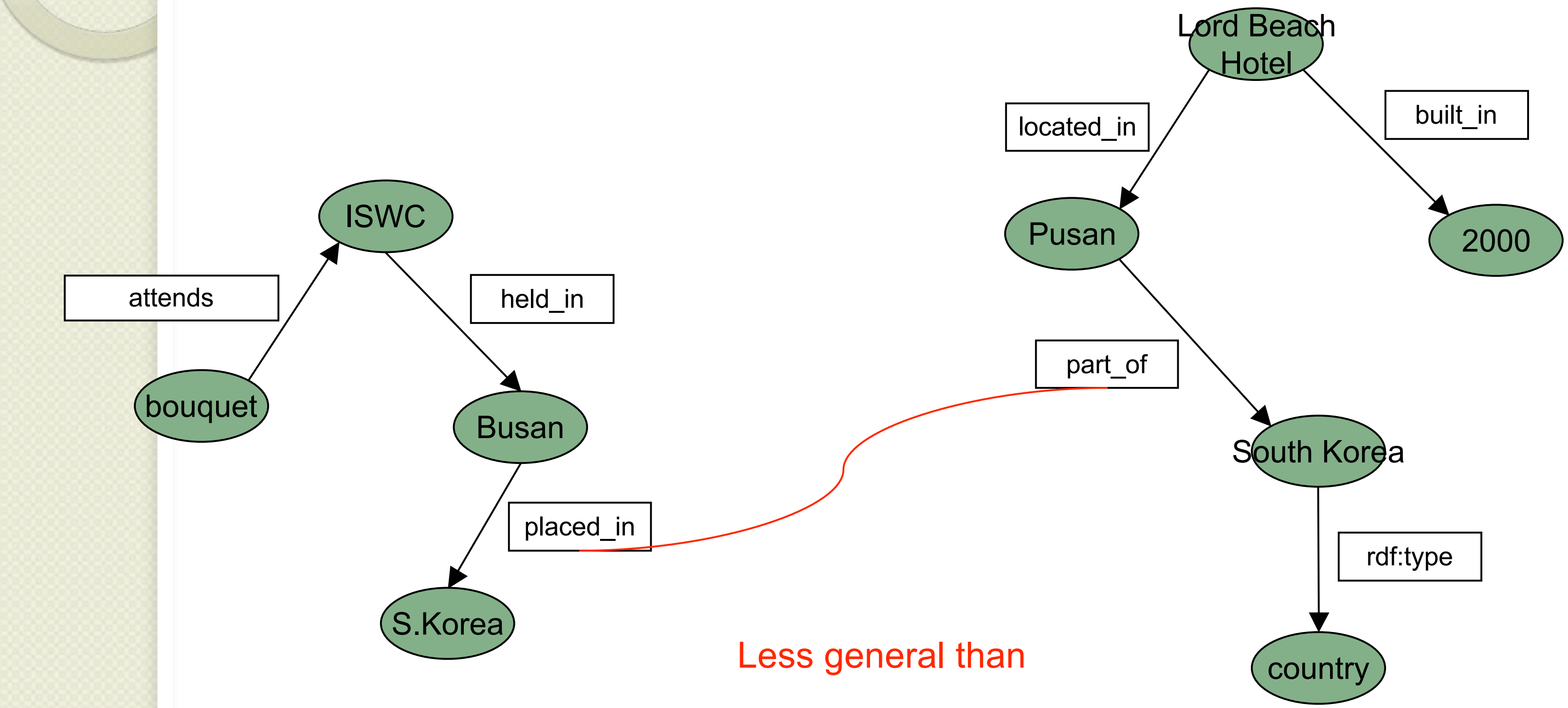
- **Schema-level heterogeneity**: matching different vocabularies/ontologies so that equivalent (or related) classes and properties can be put in relation or collapsed

- **Instance-level mismatch**: resolving identities between instances, so that all statements about it can be correctly integrated

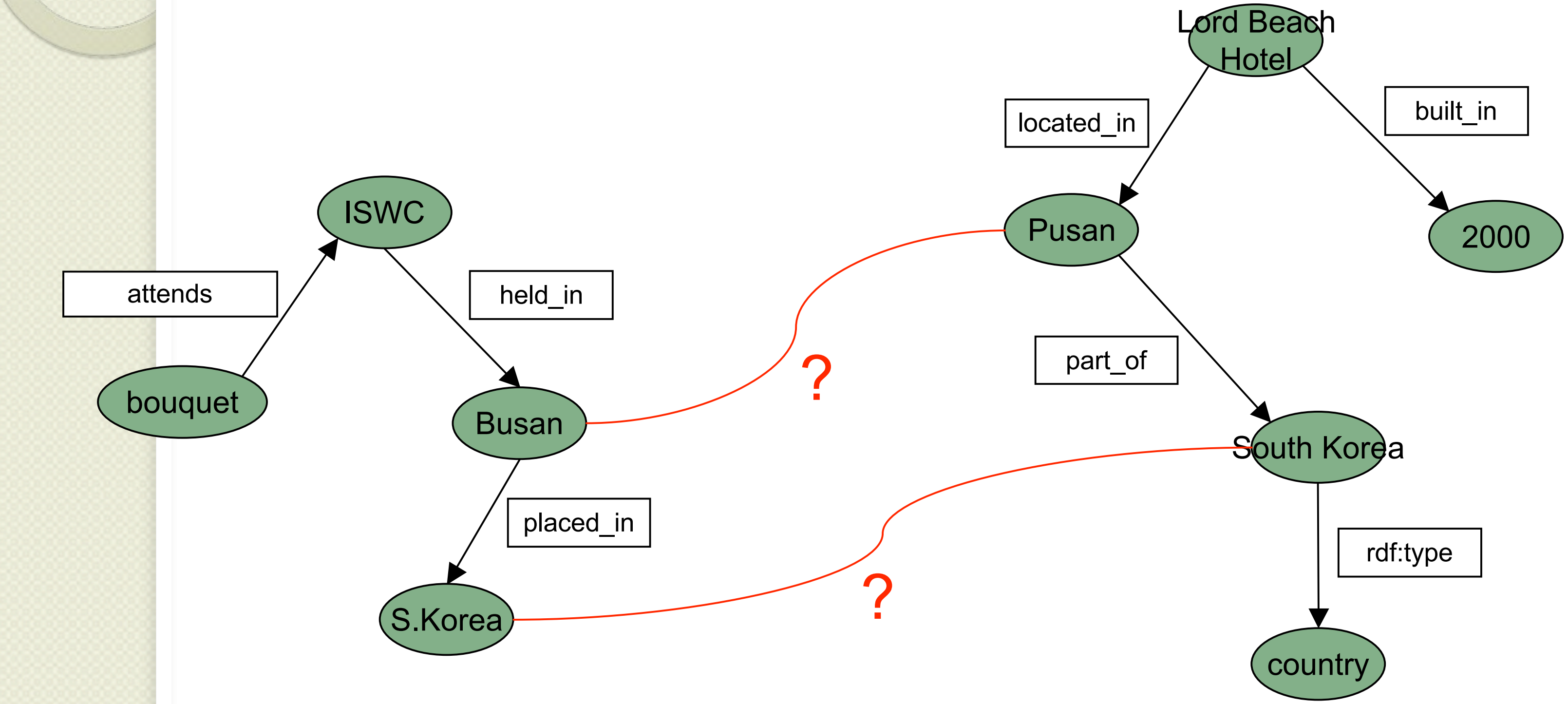
# Issue #1: schema-level heterogeneity



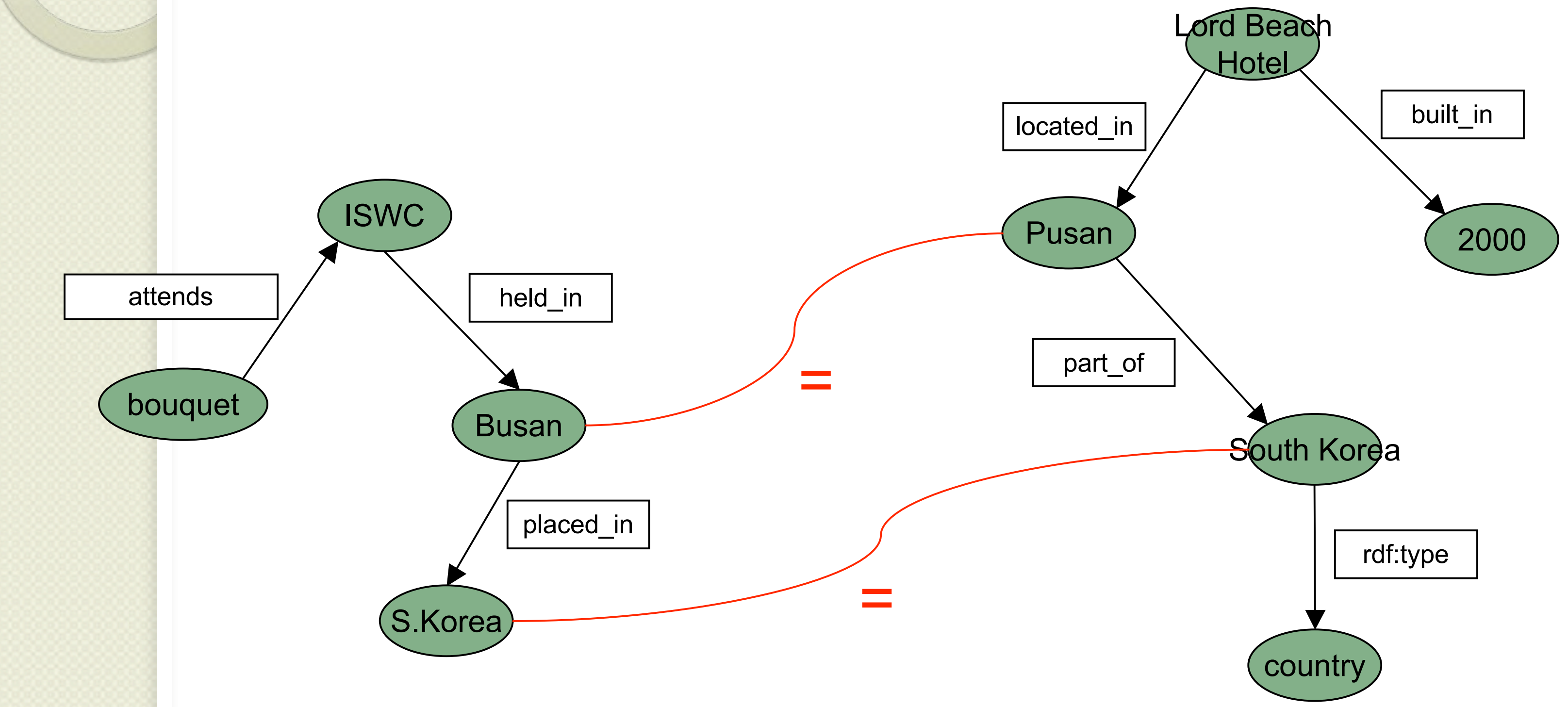
# Issue #1: schema-level integration



# Isssue #2: entity-level mismatch



# Isssue #2: entity-level integration



# State-of-the-art

- Schema-level integration is well-studied and supported (see ontology matching methods and tools)

- Entity-level integration is mostly neglected, perhaps because most people believe it is an easy (or at least easier) problem, mainly technological and not worth investigating for researchers.

[May this be a partial explanation of why the Semantic Web is not taking off as fast as hoped??]

# However ….

the situation of instance-level integration is currently as follows:

- The RDF metadata of some of the most important WWW and Semantic Web conference are poorly integrated at the instance level
- FOAF profiles rely on *ad hoc* methods for identifying people (and don't identify much else)
- The most common available ontology editors (e.g. Protégé) or metadata management systems (e.g. in digital libraries) generate "local" URIs for any newly created instance
- Some efforts exist to create reusable URIs (e.g. LSID, DOI), but are very vertical or commercial
- URI retrieval and reuse in general is not well supported

This way, the Semantic Web will never happen …

# Identity and Reference

- Identity and Reference is not much a “topic” which can be studied, it is all about creating awareness and establishing a good practice

- This practice has two aspects:
  - Agreeing on a syntax for URI (and this is done, see in particular [RFC3986])
  - Implementing methods for reusing URIs once they’ve been introduced
  - Creating applications (e.g. search engines, browsers, social software, etc.) which exploit global URIs

# Three critical issues

- Identity vs. Identifiers: the difference between abstract resources and individuals

- What kind of URIs?

- What infrastructure for managing reference on the Semantic Web?

# 1. Identities vs. identifiers: abstract resources

- Concepts and properties are defined within a vocabulary/ontology (they are context dependent)
- The meaning of a concept or property is therefore referred to by its local URI

- For example, imagine that

  `http://www.my_ont.org/things#whale`

  is defined as a subclass of mammal, and instead

  `http://www.my_world.net/objects#whale`

  as a subclass of fish: two **different** concepts!

# 1. Identities vs. Identifiers: individuals

- Individuals are not defined by a vocabulary
- They are not equivalent to their attributes
- In a sense, they are independent from the context in which they are introduced
- For example,

  `http://www.my_ont.org/university#Bob`

  is of type a professor, and

  `http://www.my_world.net/soccer#Bob`

  is a soccer player. But in the real world he is **the same** Bob with two different identities.

# 1. Identities vs. identifiers

As a consequence:

- The URI of a concept/property may include its identity (what it is)

- The URI of an individual/instance should not, and indeed it is not desirable (think of some simple examples)

# 2. What kind of URIs?

Tim Berners-Lee proposed (see [LinkedData]) to use HTTP URIs for referring to resources on the Semantic Web:

- They have some advantages (e.g. DNS look up)
- However, they also have some disadvantages:
  - Tend to encode an identity
  - Tend to confuse a real world entity with its “proxy” (see [IRE]), e.g. a person with her personal web page
- May change through time or disappear

# 3. Managing reference on the SemWeb

- The globalization of hypertexts was grounded on:
  - URIs (URLs) for addressing [digital] resources
  - URIs (URLs) as **global** identifiers for specifying hyperlinks between [digital] resources
- By analogy, the globalization of knowledge representation should be grounded on:
  - URIs for identifying [digital, as well as non digital] resources (people, products, devices, locations, events, …)
  - URIs as **global** identifiers for specifying "links" (properties, relations) between [digital, as well as non digital] resources

# However …

- The use of URIs in the Web was grounded on a solid infrastructure:
  - URLs are resolved to unambiguously locate resources in the Internet
  - hyperlinks are truly universal

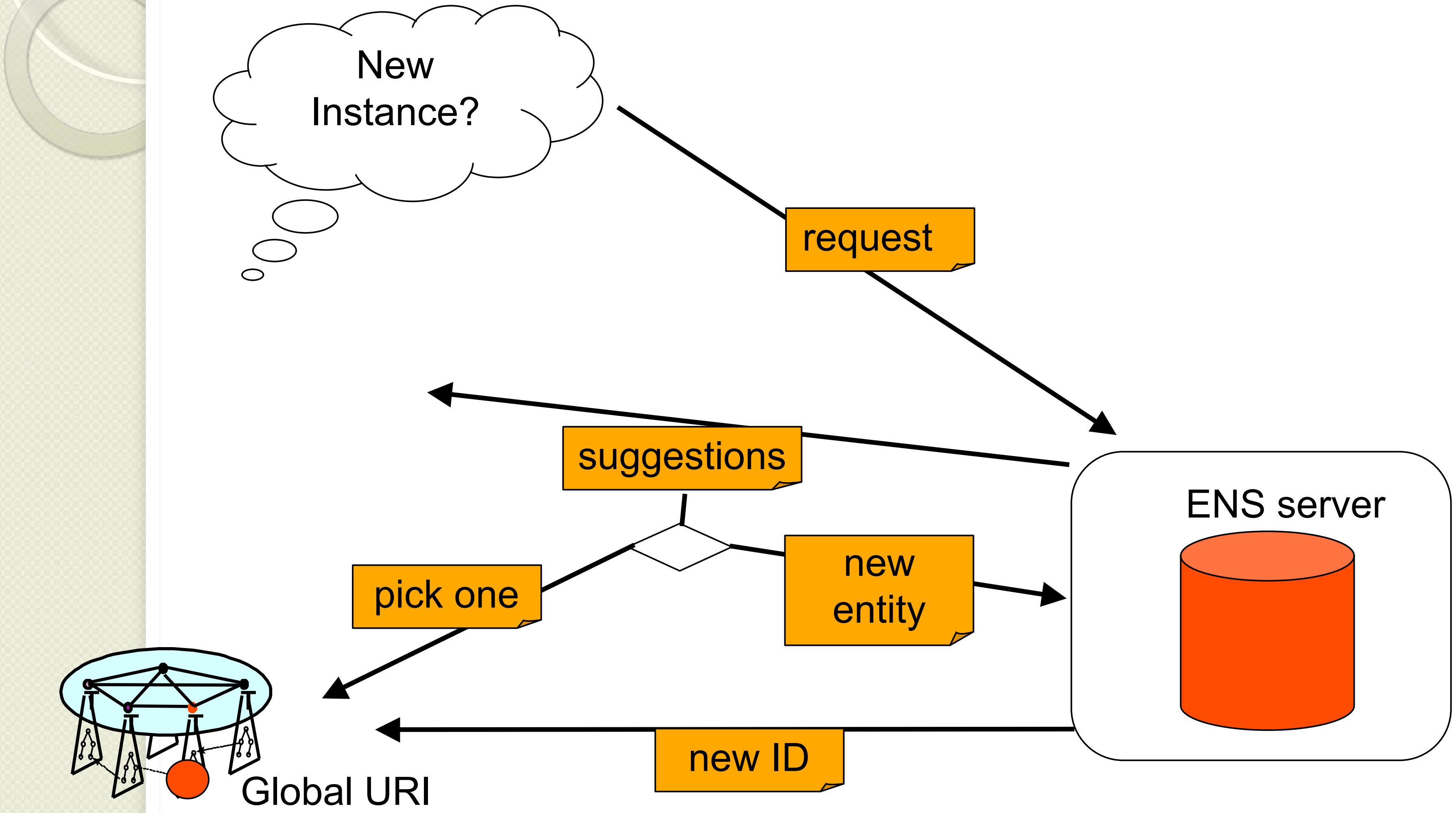
- But what about the Semantic Web?
  - the idea of “resolving” URIs for non digital resources is at least problematic (see e.g. the “identitiy crisis” discussion)
  - “links” between resources are not truly *universal*

# 3. ENS: a “DNS” for the Semantic Web

We need something like an Entity Naming System (ENS):

- Cornerstone:  ENS server (repository + “resolution” of names)
- Architecture: distributed, supports federation of local ENS servers, replicated (no single point of failure)
- ENS server vs. Entity Base (or Entity Knowledge Base): supporting reuse vs. collecting and providing knowledge about entities
- Basic schema: set of attribute/value pairs (called “labels”) with no predefined semantics – the minimum required for distinguishing entities from one another

# An example: creating instances in an ontology



# Conclusions: Expected advantages

- Improving instance-level integration of information:
  - merge of RDF graphs
  - ontology integration (instance level)
  - entity resolution in databases
  - enabling distributed queries on autonomous information sources
- Integrating text-based with multimedia resources
  - Tagging text, pictures, audio & video files with global IDs
- Enabling new entity-centric methods for navigating/searching:
  - entity-centric search engines
  - entity-centric web navigation
- Applications:
  - business intelligence
  - publishing & news
  - knowledge management
  - ...

# Conclusions: A bootstrap for the Semantic Web?

- So far, very little incentive to publish RDF information on the Web: what for?

- Introducing the ENS can be the basis for simple but powerful integration of semantic-based information on the web

- The hope: RDF content will grow as people will perceive the return on publishing information in this form

- The ENS may provide a bridge between the current Web and the Semantic Web

# References

Addressing resources on the (Semantic) Web:

- [webarch] I. Jacobs and N. Walsh. Architecture of the World Wide Web. TR W3C, 2004. http://www.w3.org/TR/webarch.
- http://www.w3.org/Addressing . In particular:
  - [RFC2396] T. Berners-Lee et al., Uniform Resource Identifier (URI): Generic Syntax, August 1998
  - [RFC3986] T. Berners-Lee et al., Uniform Resource Identifier (URI): Generic Syntax, January 2005 (obsoletes 2396)
- http://www.w3.org/DesignIssues/ . In particular:
  - T. Berners-Lee, What HTTP URIs identify (2005/6)
  - T. Berners-Lee, What the semantic Web isn't but can represent (1998)
  - [LinkedData] T. Berners-Lee, Linked Data (2006/7)

# References - II

- A foundational approach:
  - A. Gangemi, V. Presutti, The bourne identity of a web resource. Architecture and Philosophy of the Web Identity, Reference, and the Web (IRW2006), WWW2006 Workshop, Edinburgh, Scotland, May, 23rd 2006.
- The “identity crisis”
  - S. Pepper and S. Schwab. Curing the Web’s Identity Crisis: Subject Indicators for RDF. Technical report, Ontopia, 2003. http://www.ontopia.net/topicmaps/-materials/identitycrisis.html (2003)
- Ideas towards a ENS:
  - P. Bouquet, H. Stoermer, M. Mancioppi, and D. Giacomuzzi. OkkaM: Towards a Solution to the “Identity Crisis” on the Semantic Web. In In Proceedings of SWAP 2006, the 3rd Italian Semantic Web Workshop, Pisa, Italy, 2006. CEUR Workshop Proceedings.

# References

- Some philosophical background:
  - S. Kripke, *Naming and Necessity*. Cambridge, Mass.: Harvard University Press, 1980