Laying out the conceptual foundations for data integration in the humanities

Michele Pasin, Arianna Ciula

Centre for Computing in the Humanities

Kings College, London

michele.pasin@ kcl.ac.uk





Summary

1. An emerging web of data

2. The role of ontologies

3. Creating ontologies *for* the humanities

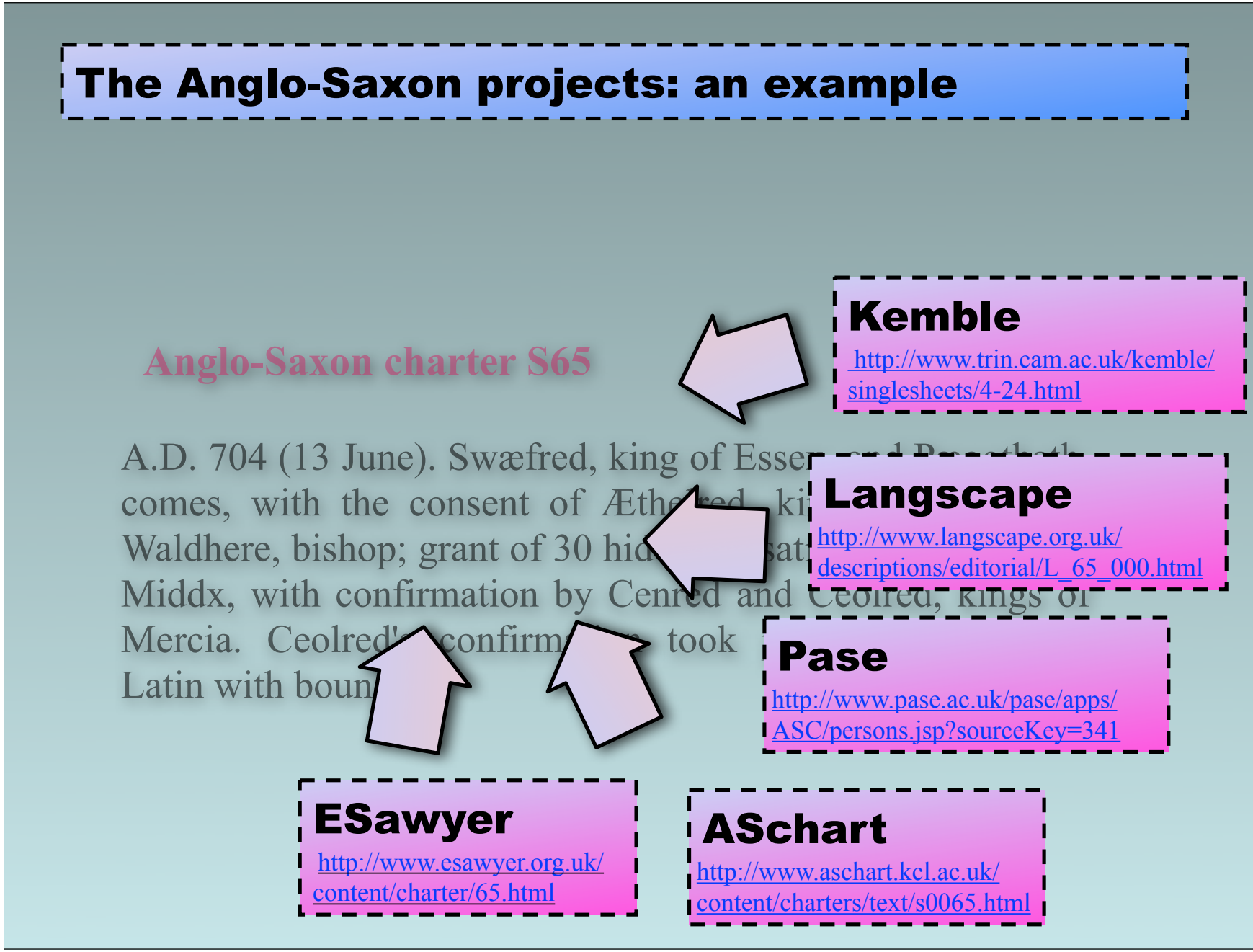
4. Working in a world with many ontologies

The Anglo-Saxon projects: an example

Anglo-Saxon charter S65

A.D. 704 (13 June). Swæfred, king of Essex, and Pæogthath, comes, with the consent of Æthelred, king of Mercia, to Waldhere, bishop; grant of 30 hides (cassati) at Twickenham, Middx, with confirmation by Cenred and Ceolred, kings of Mercia. Ceolred's confirmation took place at Arcencale. Latin with bounds.

The Anglo-Saxon projects: an example



The Anglo-Saxon projects: analysis

Project/Data	ASChart	E-Sawyer	Kemble	LangScape	PASE
Archive	x	x	x	x	x
County				x	x
Estate				x	x
Date	x	x	x	x	x
Summary	x	x	x		x
Manuscript class mark		x	x	x	
Language	x	x			x
Language of part of text				x	
Full text	x				
Part of text				x	
Facsimile			x		
Translation of part of text				x	
Headwords of part of text				x	
Bibliography		x		link	
Partial bibliography			x		
Alternative Identifiers			x		x
Alternative identifiers for part of text				x	
Identification of factoids	link		x		x
Editorial notes			x	x	x

The Anglo-Saxon projects: desiderata

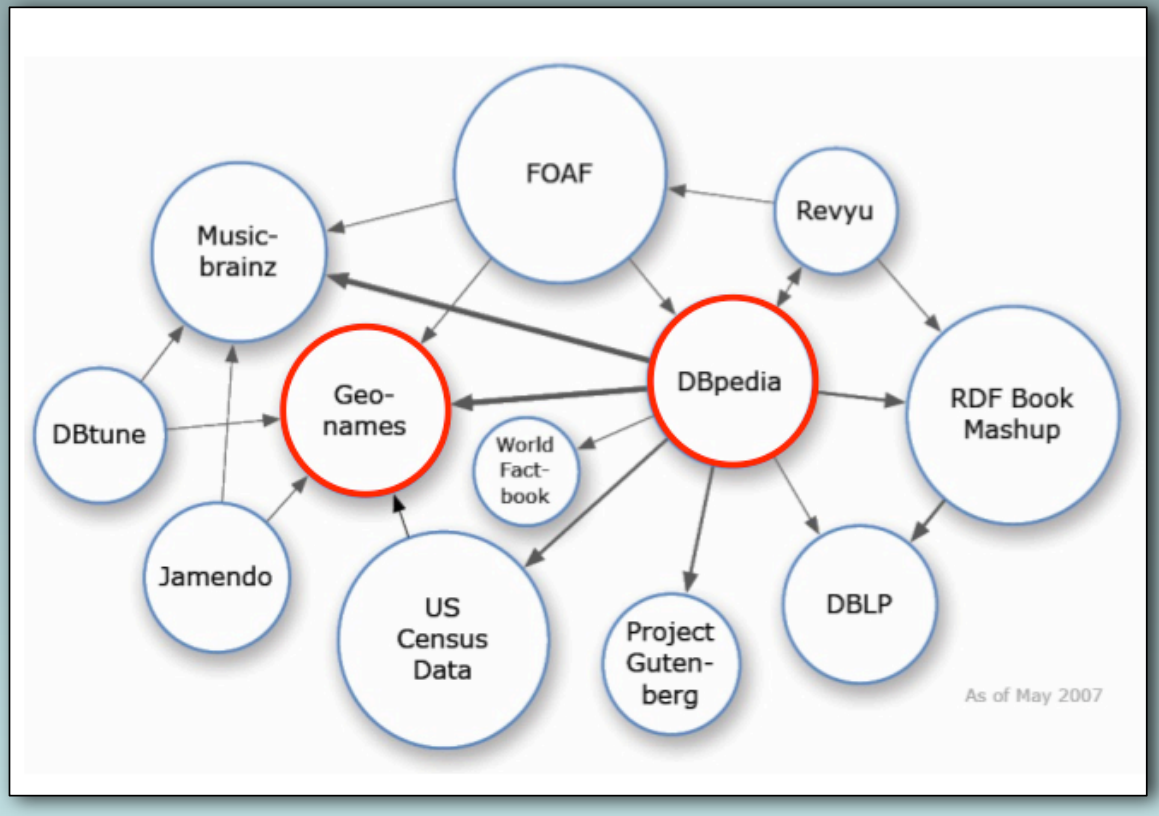
Project/Data	ASChart	E-Sawyer	Kemble	LangScape	PASE
				x	x
				x	x
				x	x
				x	x
					x
				x	
Language	x	x			x
Language of part of text				x	
Full text					
					x
					x
Editorial notes			x	x	x

- **data sharing**
  - maintain provenance and integrity
  - eliminate redundancy
  - allow for comparative perspective (e.g. visualise conflicts of interpretations)

- **models exposure**
  - what is an event (e.g. Anglo-Saxon project: what is a transaction in PASE?), a person, a place?
  - Can a certain consensus be reached? Necessity to establish community of practices around modeling exercises, clusters of consensus around knowledge domains or specific disciplines

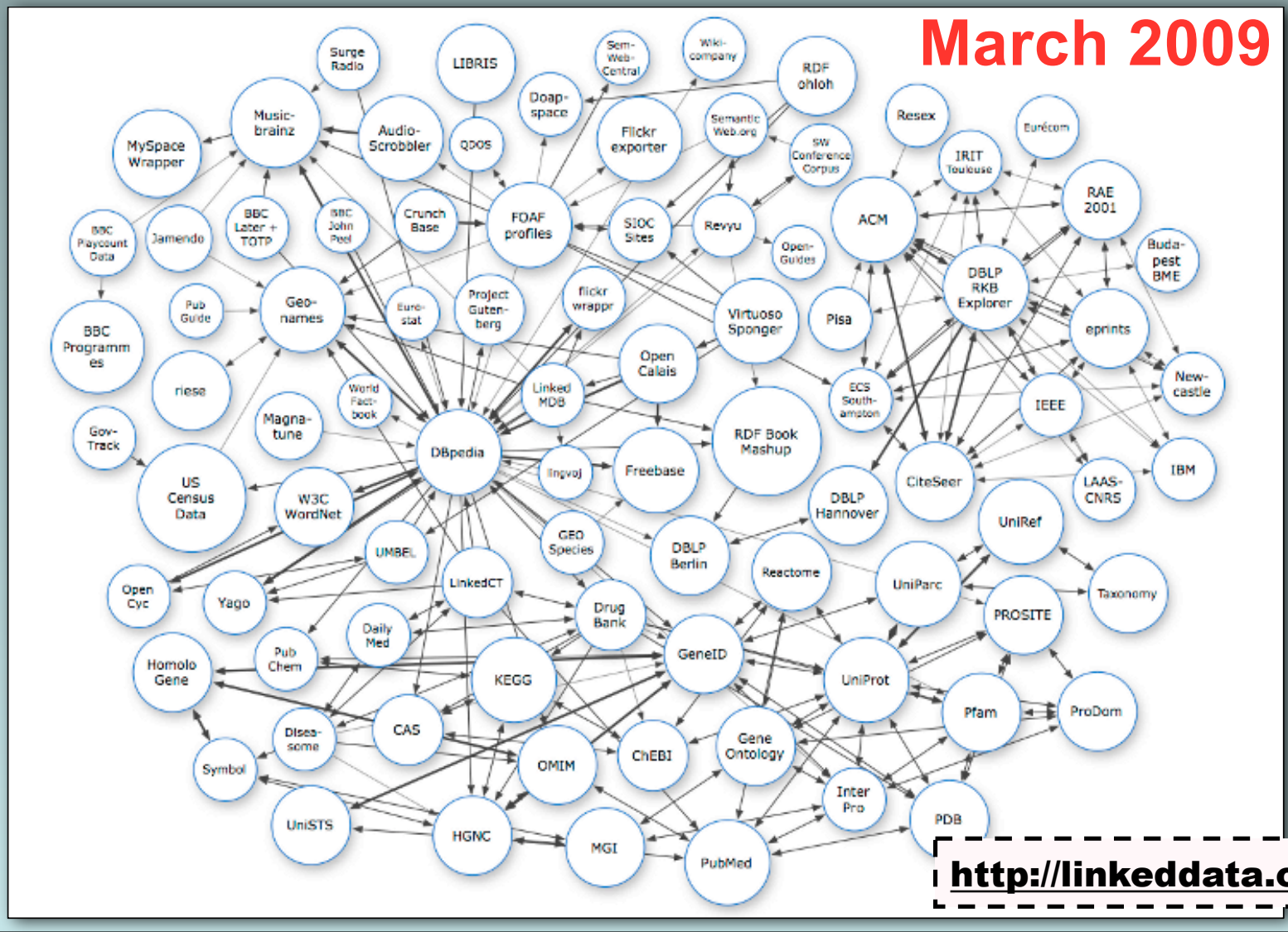
The LinkedData initiative

May 2007

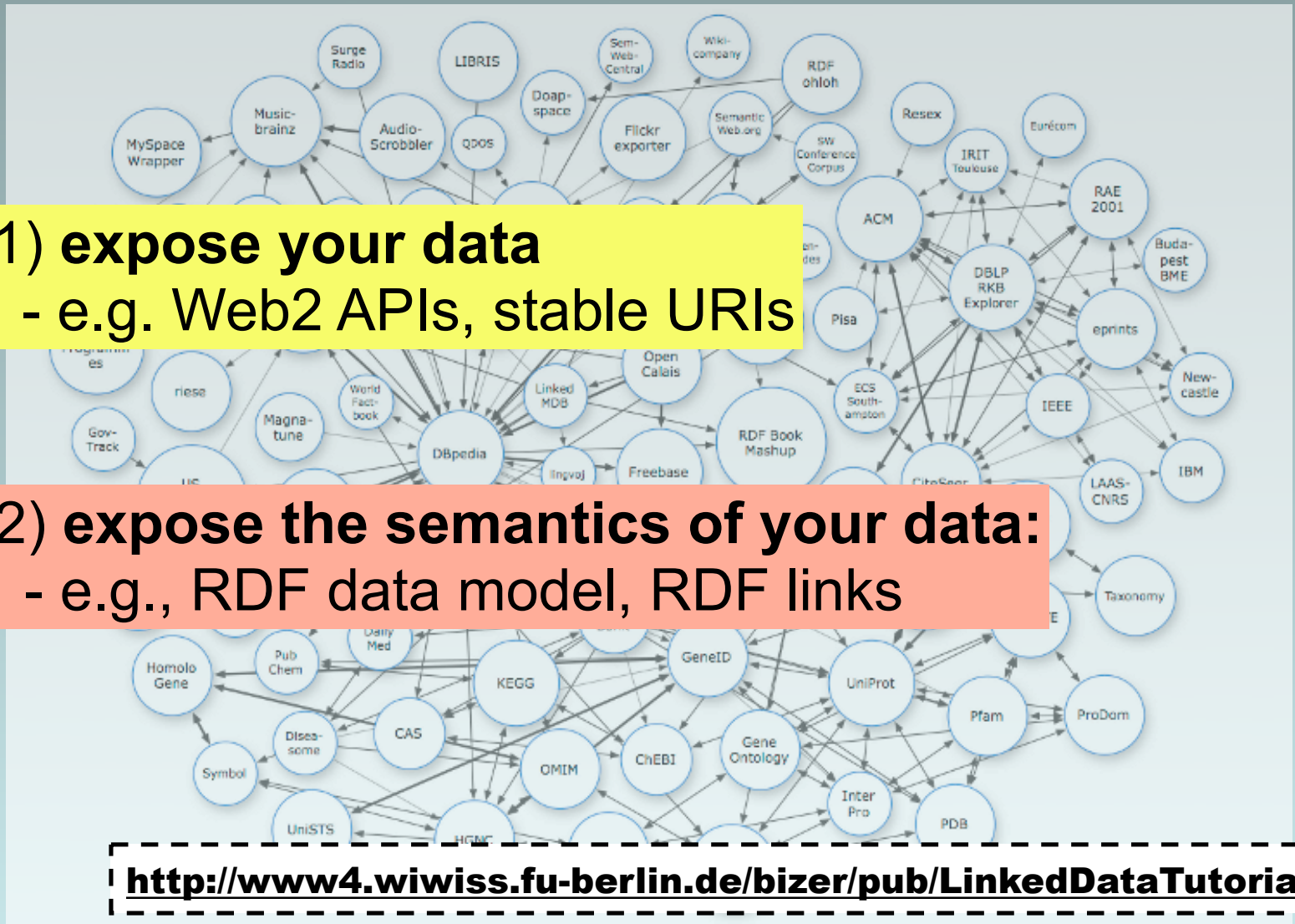


http://linkeddata.org/

An emerging web of data



An emerging web of data: in a nutshell

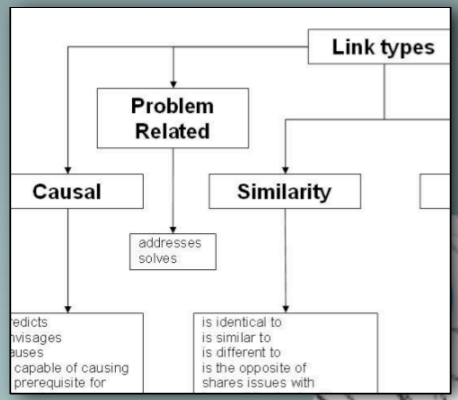


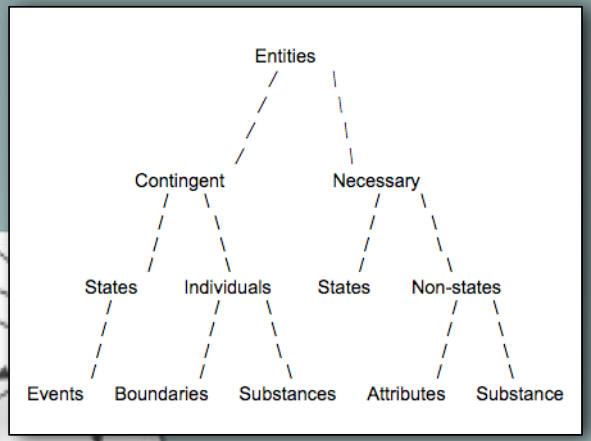
1) **expose your data**
- e.g. Web2 APIs, stable URIs

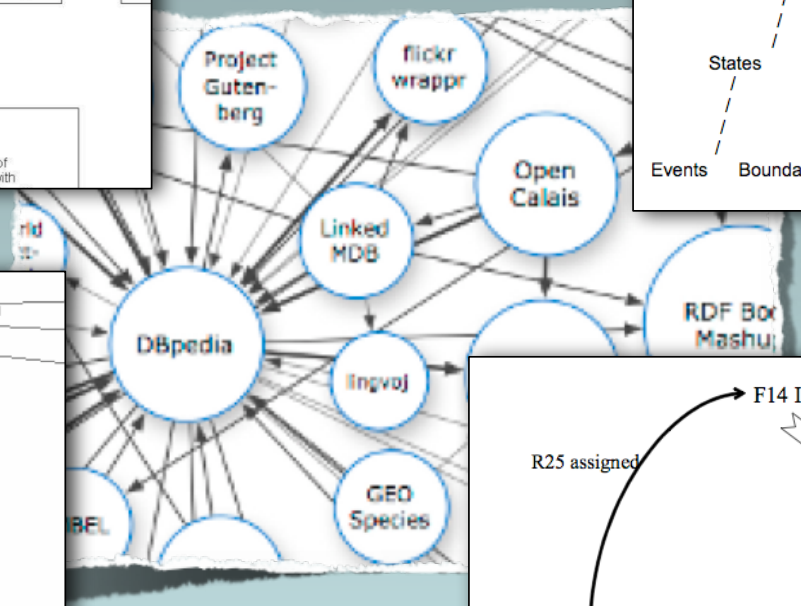
2) **expose the semantics of your data:**
- e.g., RDF data model, RDF links

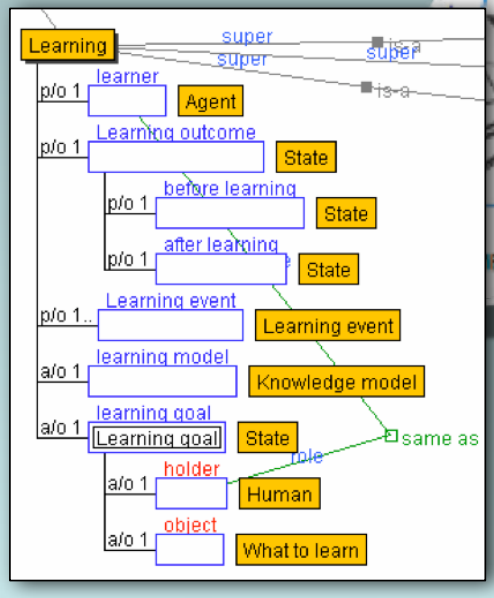
http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/

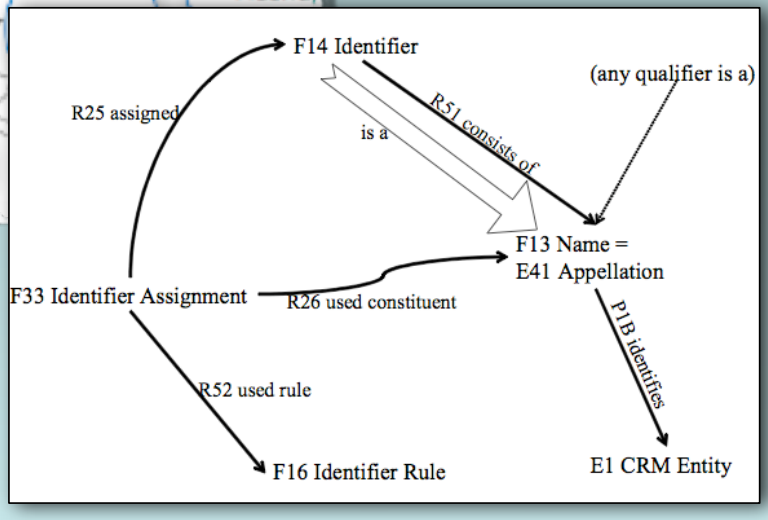
Creating semantic models: ontologies



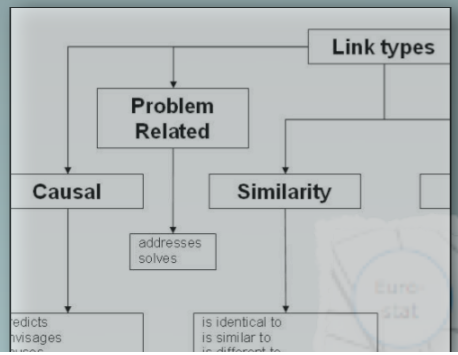


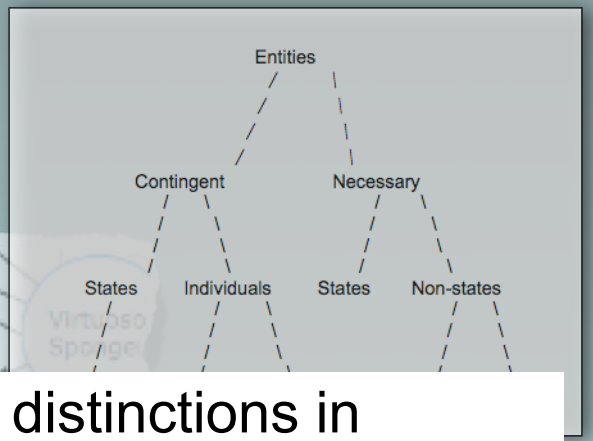






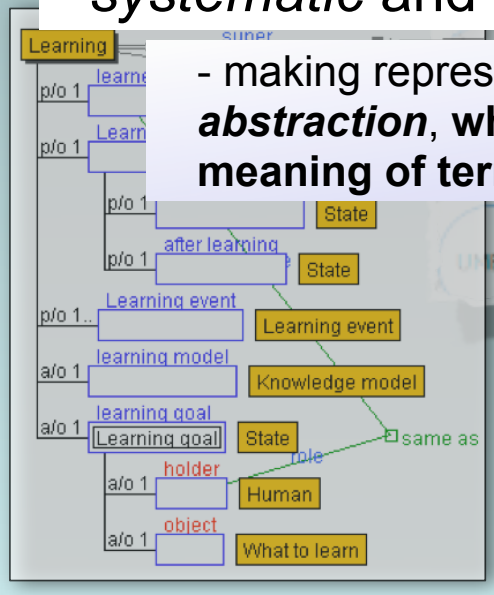
Creating semantic models: ontologies

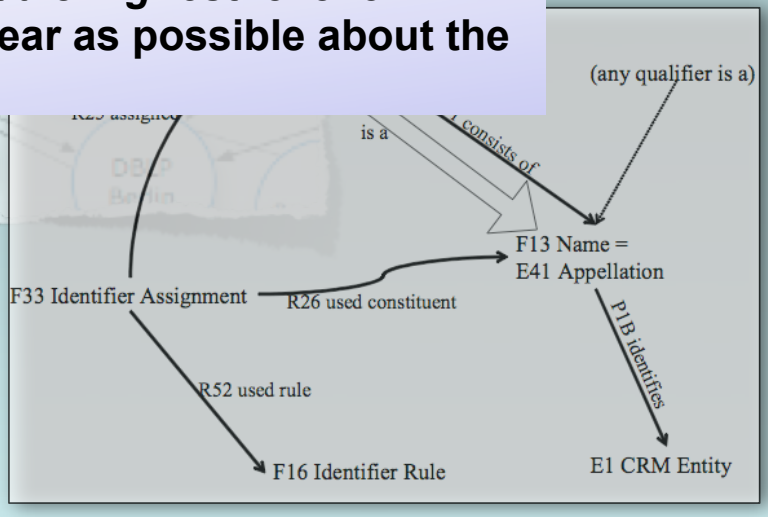




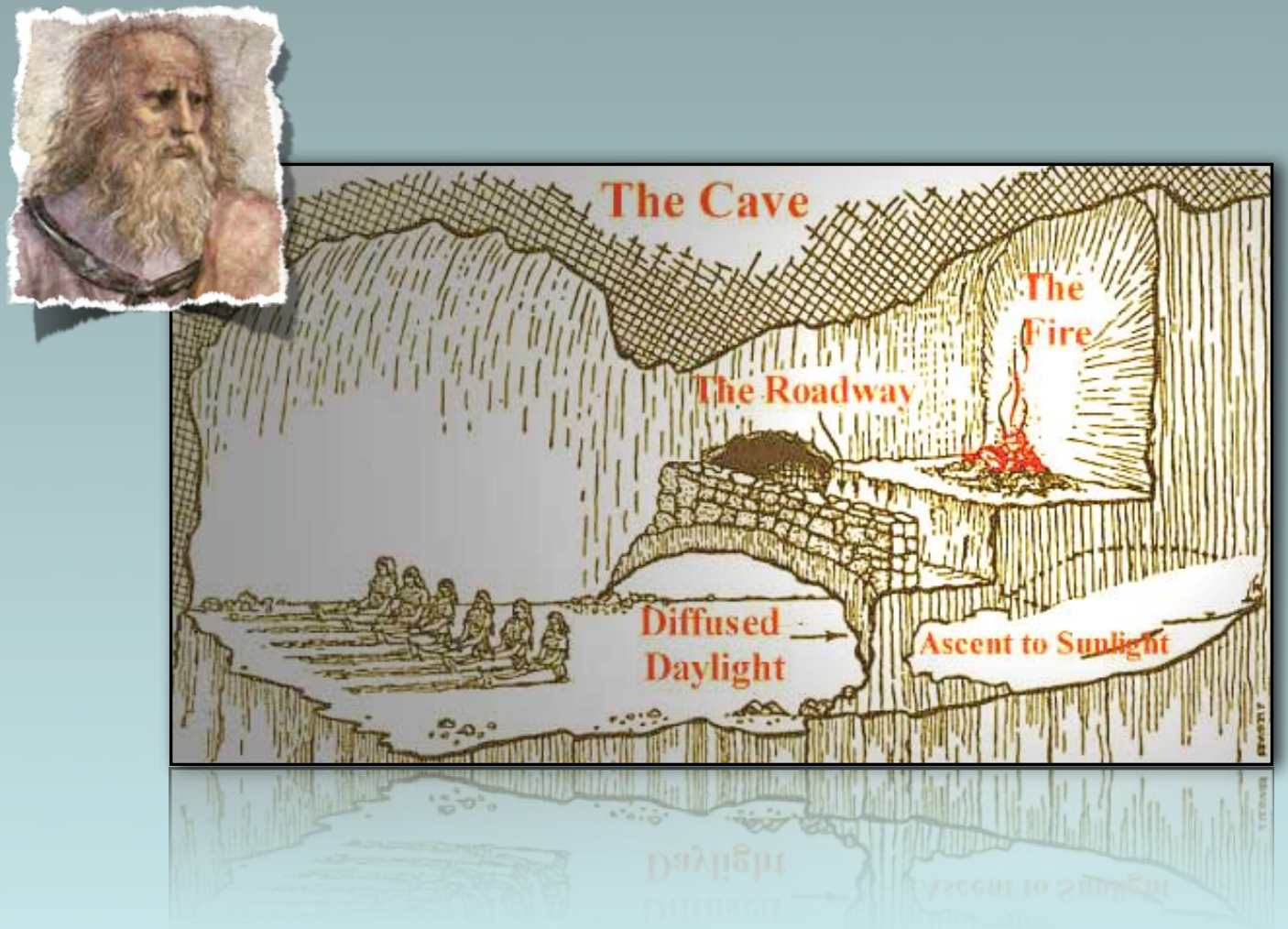
- a *theory* of how to make ontological distinctions in *systematic* and *coherent* manner

- making representational choices at the ***highest* level of *abstraction***, **while still being as clear as possible about the meaning of terms**



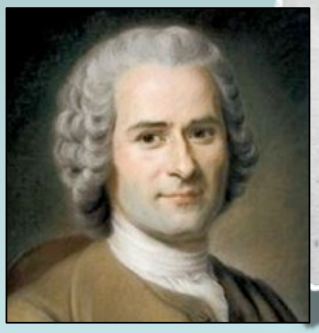


The role of ontologies: the ‘realist’ position



The role of ontologies: the ‘pragmatic’ position

software applications





DU

CONTRACT SOCIAL;

OU,

PRINCIPES

DU

DROIT POLITIQUE.

PAR J. J. ROUSSEAU,

CITOYEN DE GENEVE.

— *fœderis æquas*

Dicamus leges. Æneid. XI.

A AMSTERDAM,

Chez MARC MICHEL REY.

MDCCLXII.



research communities

The ontological approach: a few principles

- determine an **essential** property for each concept and instance

  - Proper use of is-a relation should inherit the “Essential” property of its super classes (= identity criteria checking)

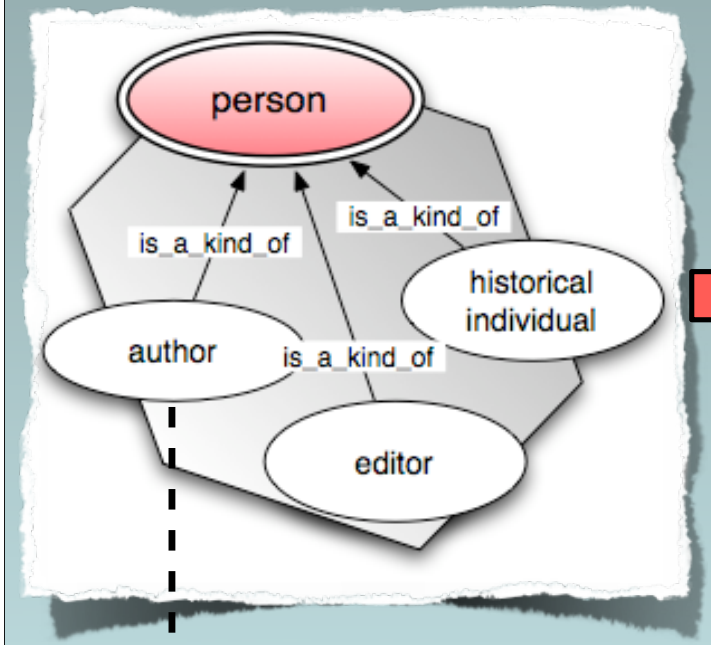
- **concepts** rather than **terms**

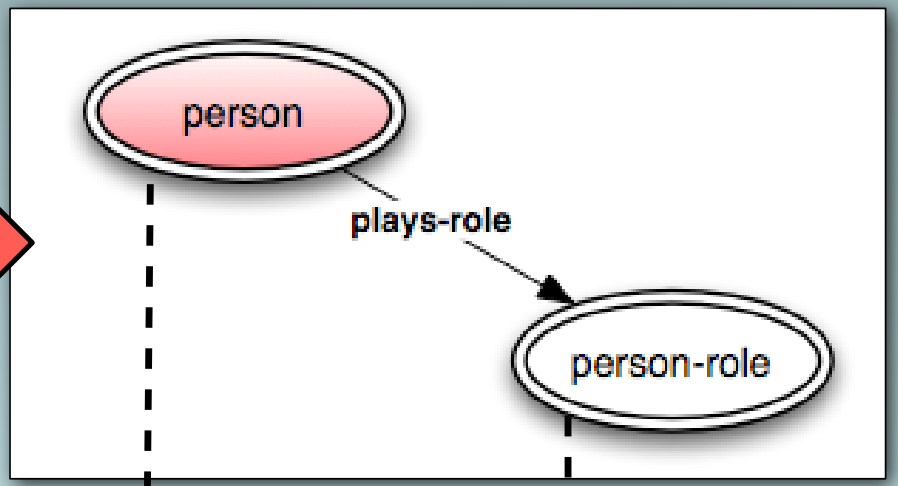
  - people are easily trapped by the endless terminological discussion departing from the underlying conceptual structure of the target domain

- **role** concepts vs **basic** concepts

  - Clear and consistent differentiation between basic concepts (man, rice, oil, etc.) and role concepts(teacher, food, fuel, etc.).

Example: looking for essential properties... #1



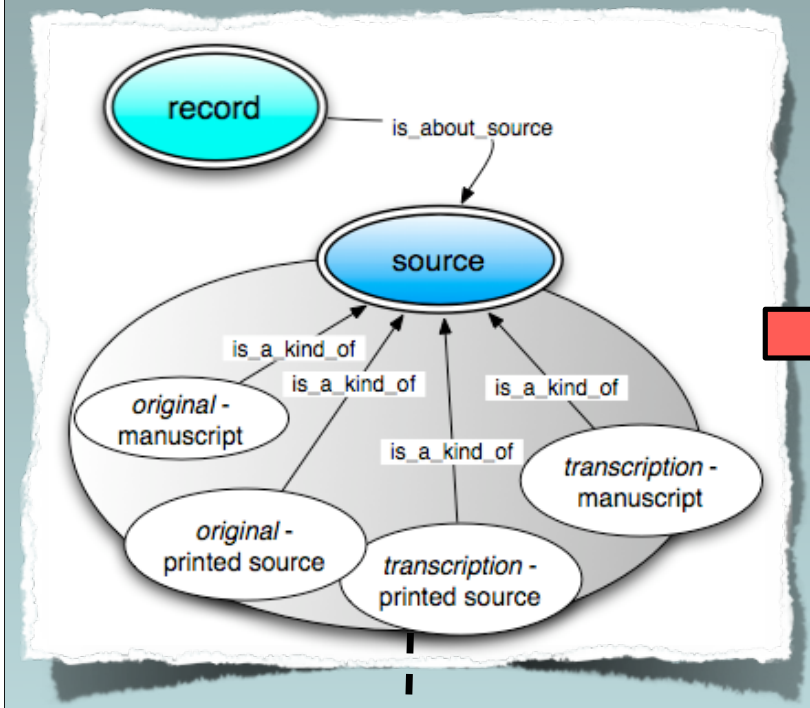


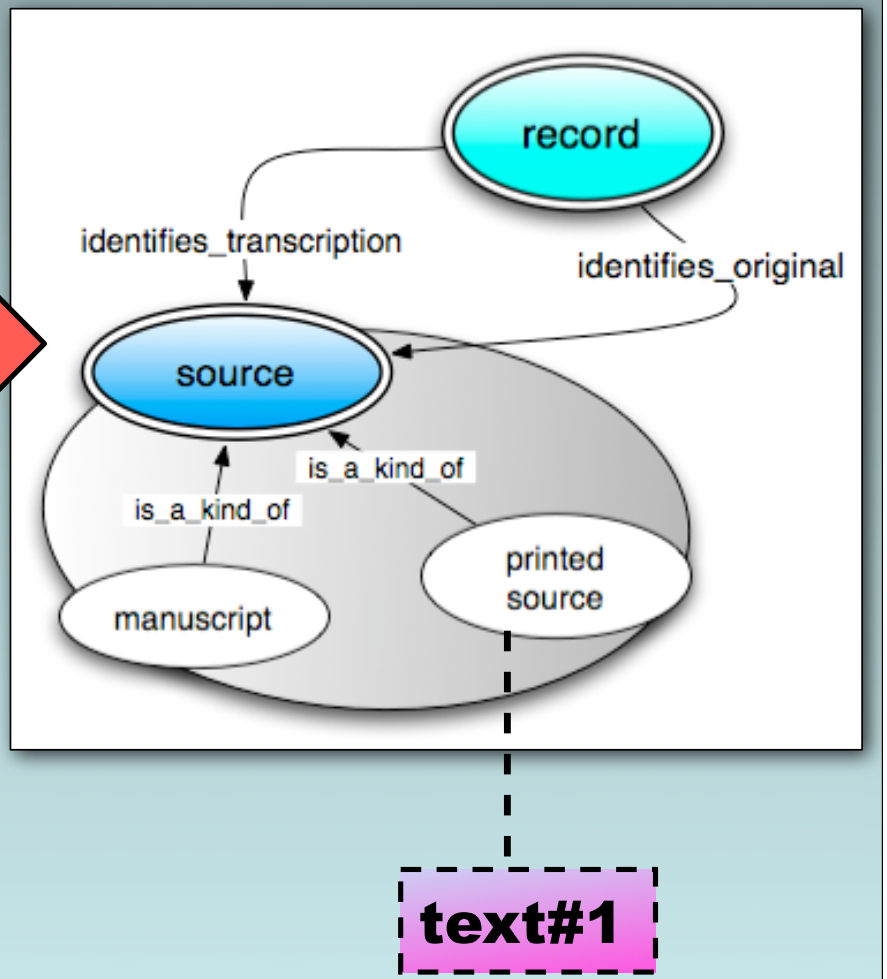
Mr. Jones

Mr. Jones

author, editor, common person...

Example: looking for essential properties... #2





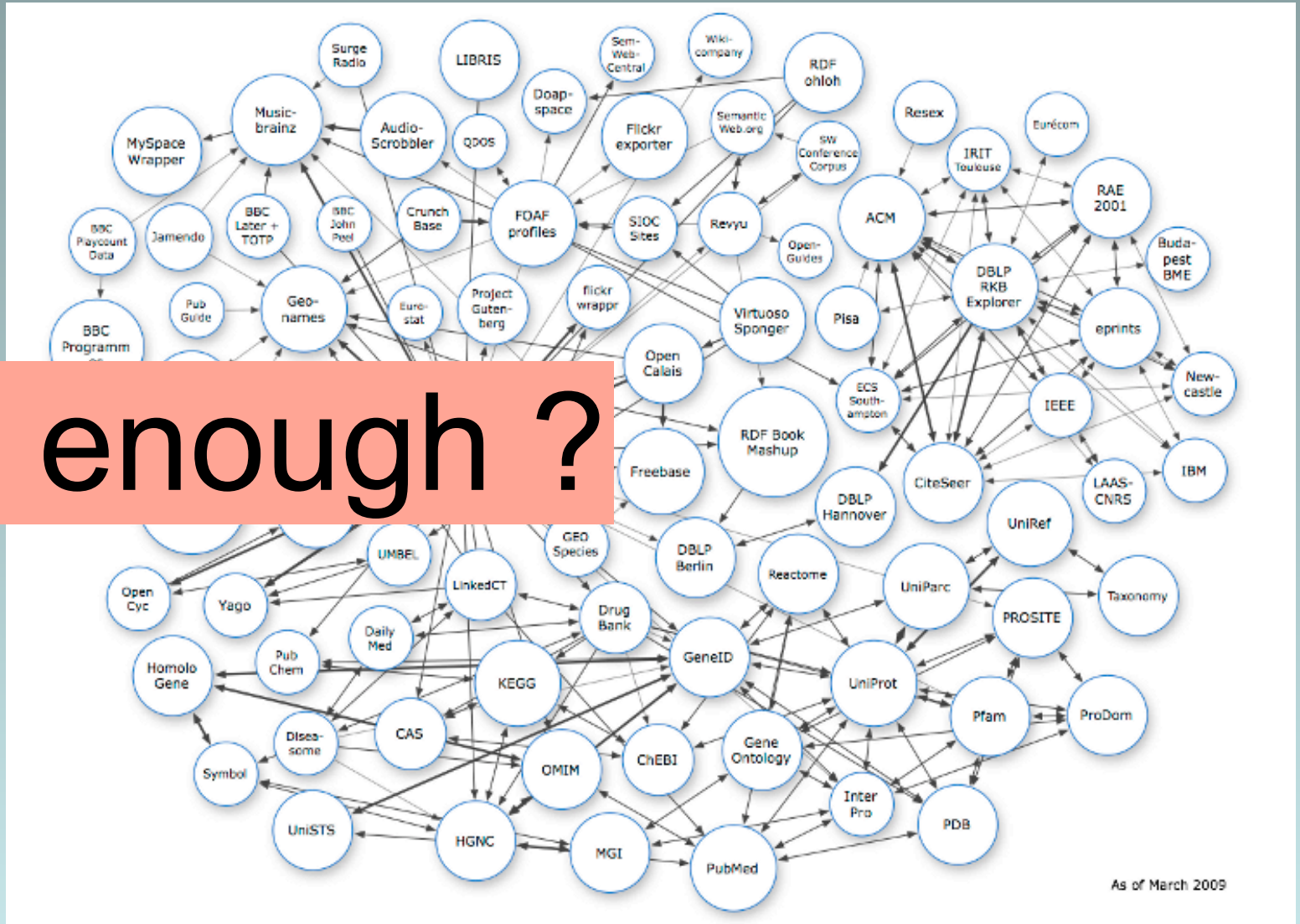
Common ‘things’ we mention in our contracts:

- information objects
  - key characteristics of entities that can carry information, that can be seen as (or part of) a *representation*
- physical features of information objects
  - e.g., materials, conditions, preservation ...
- abstract features of information objects
  - e.g., the *contents* of an information object, the Hamlet as a work
  - e.g., the *linguistic* features of an information object (latin, english, etc.)
  - e.g., aspects of the *discourse* used to communicate the contents of an information object (e.g., proem, dispositive word, bound, curse etc.). These aspects will vary with different projects!

Common ‘things’ we mention in our contracts:

- people & places
  - prosopographic and topographic information
- time & events
  - the temporal aspects are omnipresent!
  - event-types must be specialized depending on the domain of investigation
- abstract ideas
  - e.g. theories, viewpoints, concepts [what we talk about in philosophy]

A network of ontologies....

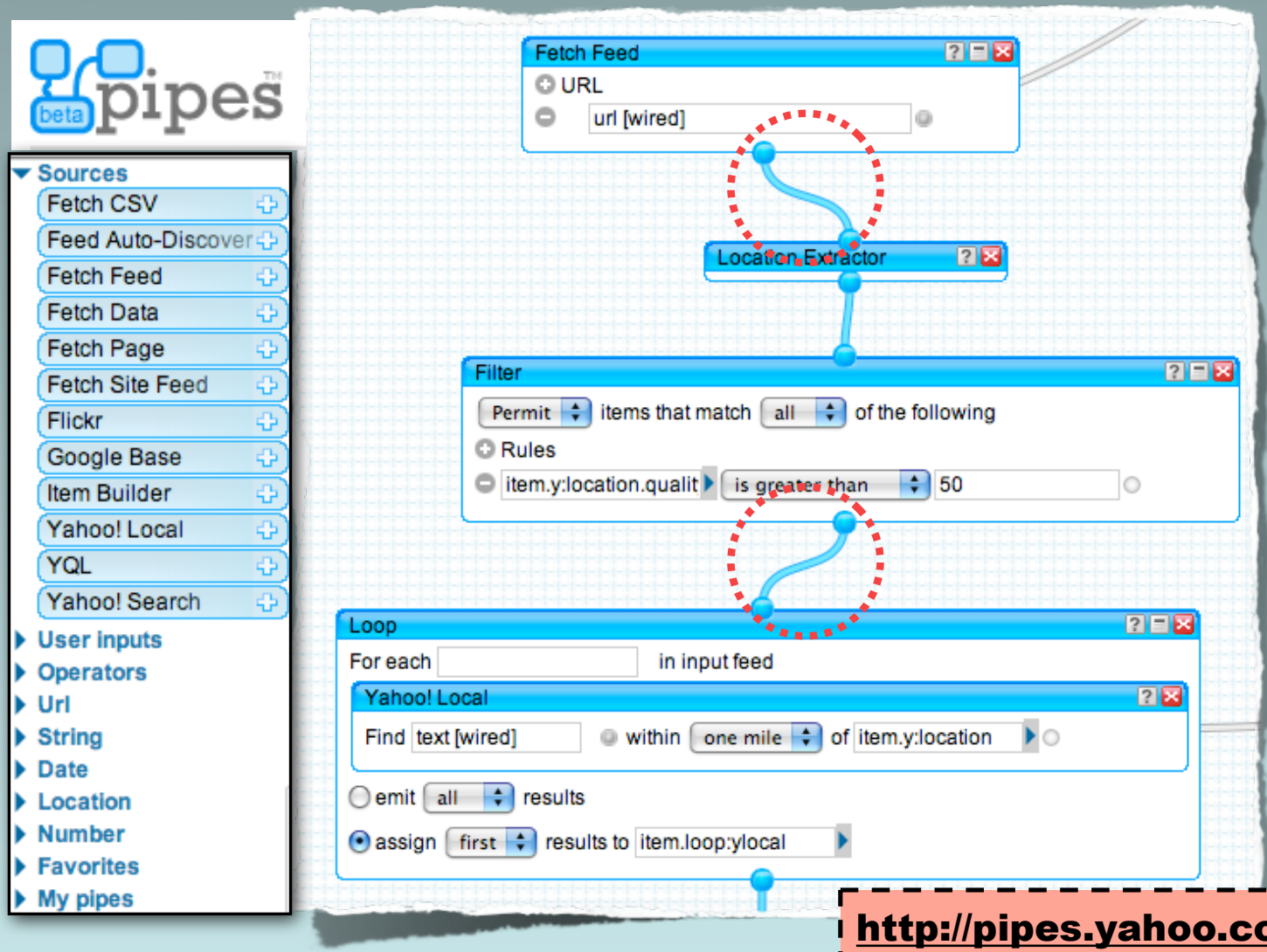


enough ?

Common ‘ways of talking’ about these things:

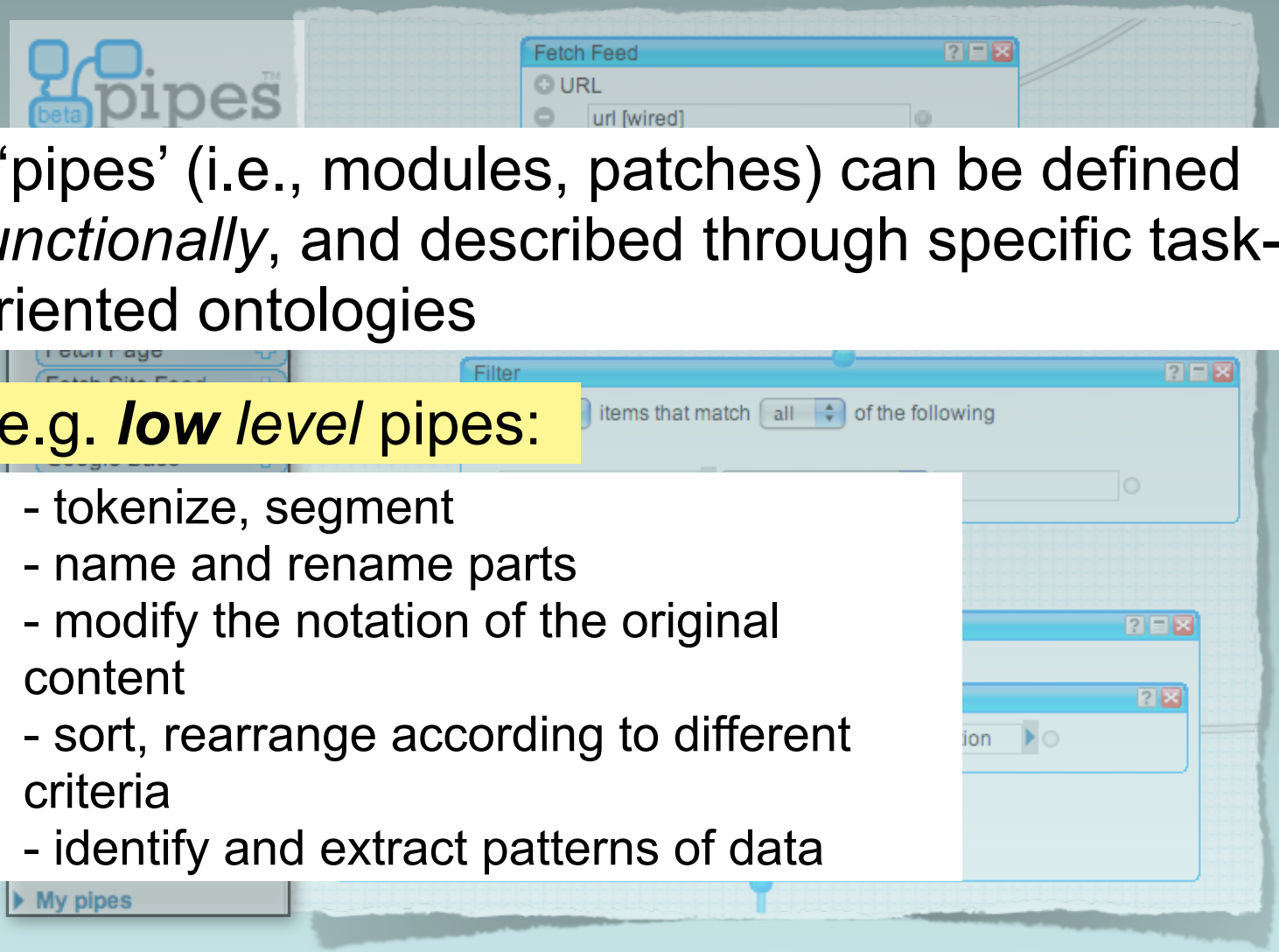
- uncertainty
  - information is missing or contradictory
  - dates are incomplete, or just unknown
- interpretations
  - what we say is not what the text says
  - for keeping track of *who* says what
  - for allowing *contradictory* views on the same subject
- debate
  - being able to represent the *arguments* supporting a view
  - being able to represent the *arguments* challenging a view

Let's find the right pipes (for the hum.scholar)



http://pipes.yahoo.com/

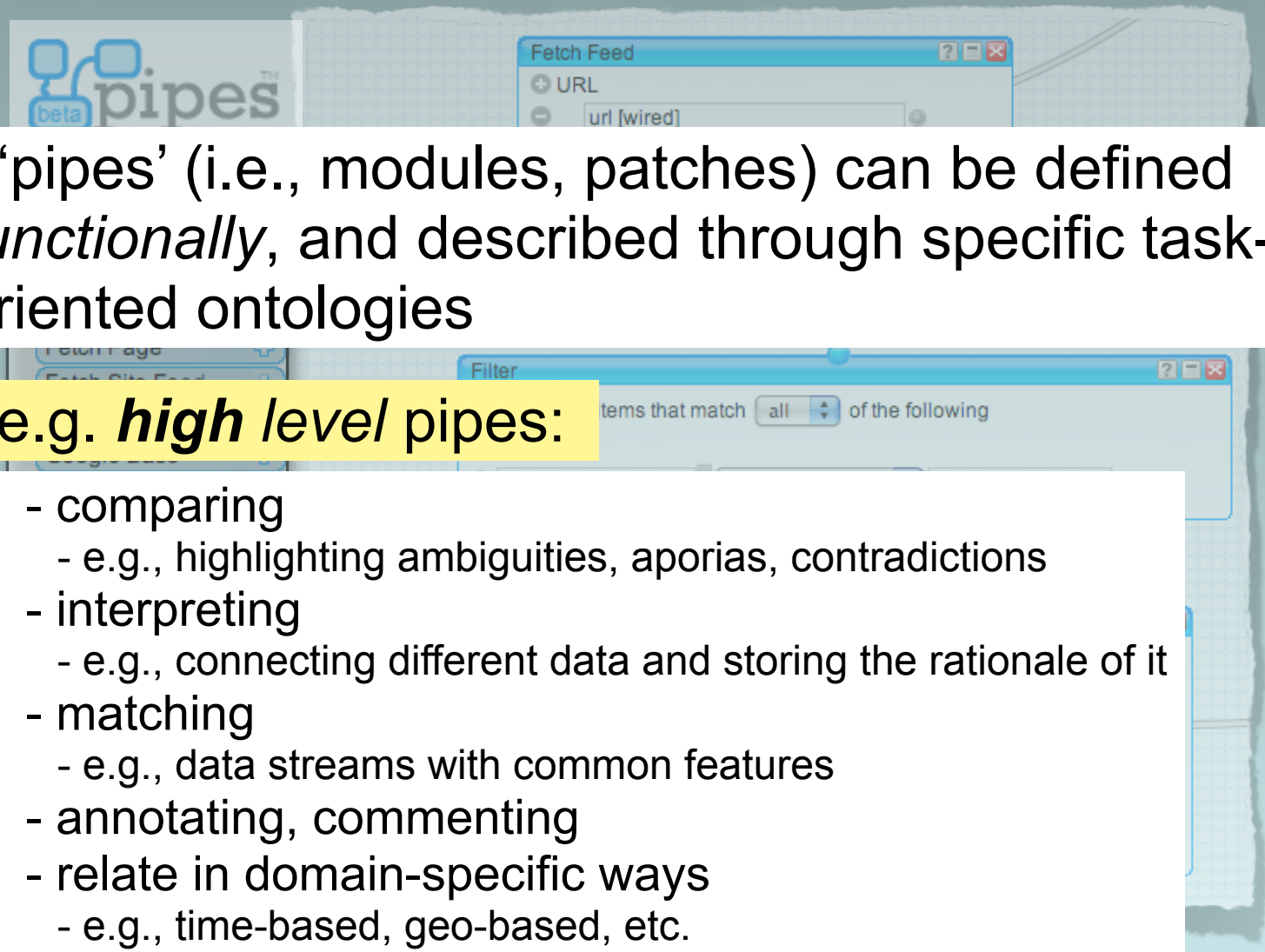
Let's find the right pipes (for the hum.scholar)



- 'pipes' (i.e., modules, patches) can be defined *functionally*, and described through specific task-oriented ontologies

- e.g. ***low*** *level* pipes:
  - tokenize, segment
  - name and rename parts
  - modify the notation of the original content
  - sort, rearrange according to different criteria
  - identify and extract patterns of data

Let's find the right pipes (for the hum.scholar)



- 'pipes' (i.e., modules, patches) can be defined *functionally*, and described through specific task-oriented ontologies

- e.g. ***high*** *level* pipes:
  - comparing
    - e.g., highlighting ambiguities, aporias, contradictions
  - interpreting
    - e.g., connecting different data and storing the rationale of it
  - matching
    - e.g., data streams with common features
  - annotating, commenting
  - relate in domain-specific ways
    - e.g., time-based, geo-based, etc.

Conclusions

- main points:
  - the web of data is quickly emerging
  - ontologies allow reuse and sharing
  - solid ontologies need to be carefully crafted
  - communities of practice to improve the modeling of common entities
  - importance of modeling also the *interpretative* connections

Some references

Auer, S. et al. Dbpedia: A Nucleus for a Web of Open Data. 6th International Semantic Web Conference (ISWC 2007) (2007).

Mizoguchi, R. Tutorial on Ontological Engineering - Part 1, 2, 3: Advanced Course of Ontological Engineering. *New Generation Computing* **22**, 198-220 (2004).

Guarino, N. & Welty, C. Evaluating Ontological Decisions With Ontoclean. *Commun. ACM* **45**, 61-65 (2002).

Gruber, T. It Is What It Does: The Pragmatics of Ontology. Invited presentation to the meeting of the CIDOC Conceptual Reference Model committee (2003).

Jones, A. (ed) Summit on Digital Tools for the Humanities: Report on Summit Accomplishments. (2006). Retrieved 20 Feb. 2009,

http://staff.cch.kcl.ac.uk/~mpasin/

thanks for the attention!