Markup,Text and Digital Humanities

Dino Buzzetti

University of Bologna

Associazione per l'Informatica Umanistica e la Cultura Digitale

a Rationale

Markup is a **technology**

So, think of the mutual **relationships** between

technology

text and **digital humanities**

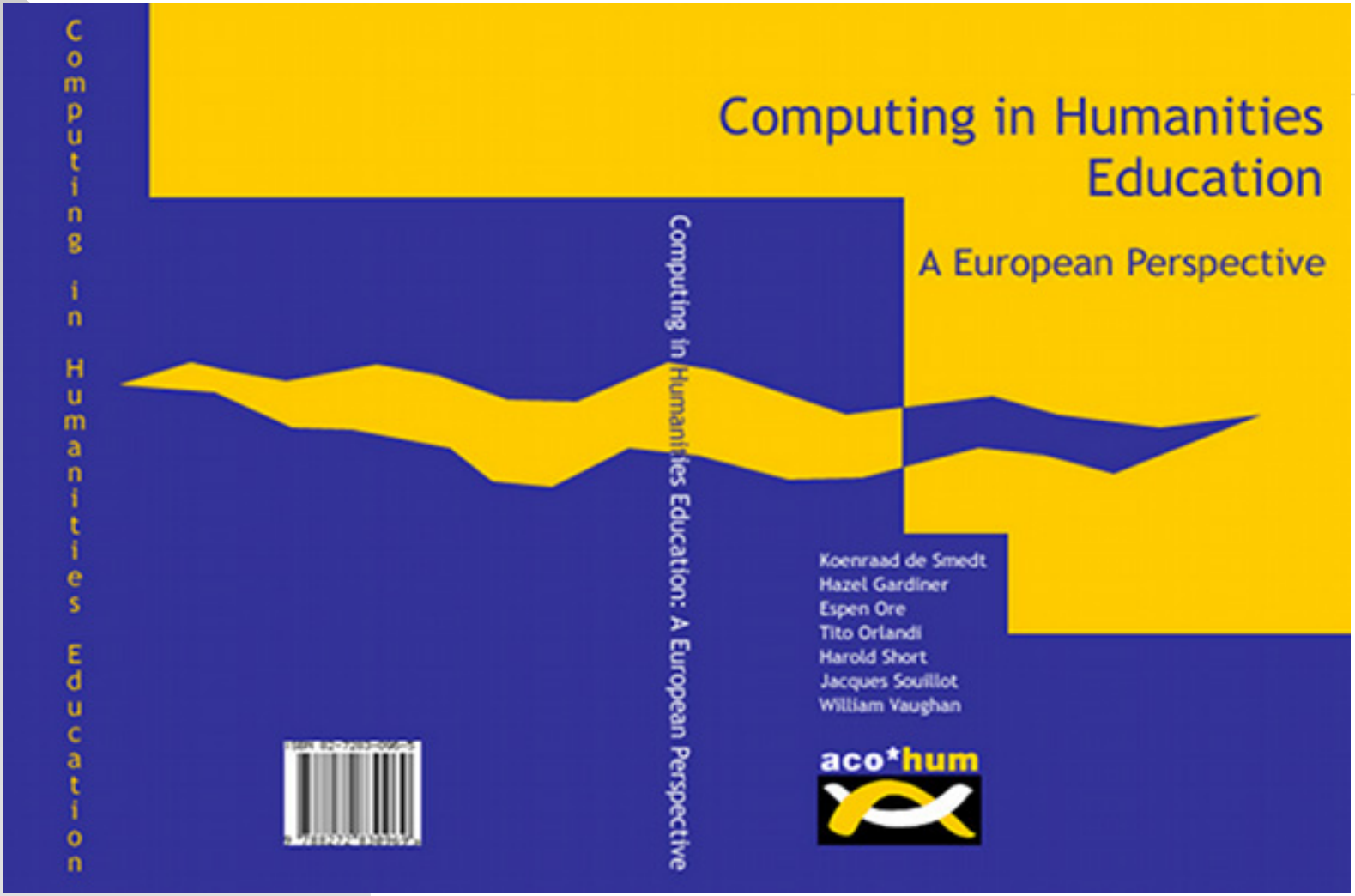
Why?

Digital Humanities: What ?

A possible **understanding**

A possible **definition**

ACO*HUM (1996-1999)



Formal Methods

Chapter 2
European studies on formal methods in the humanities

Tito Orlandi, Università di Roma La Sapienza

Joseph Norment Bell, University of Bergen - **Lou Burnard**, Oxford University - **Dino Buzzetti**, University of Bologna - **Koenraad de Smedt**, University of Bergen **- Ingo Kropac**, University of Graz - **Jacques Souillot**, CRIM-INALCO Paris - **Manfred Thaller**, University of Bergen

http://www.hd.uib.no/AcoHum/book/fm-chapter-final.html

2.3 Defining humanities computing methodology

[...] we will attempt to **define** the core of all **applied computer sciences** in terms of the traditional combination of *data structures* and *algorithms*, applied to the requirements of a discipline:

The methods needed to **represent the information** within a specific domain of knowledge in such a way that this information can be processed by computational systems result in the *data structures* required by a specific discipline.

The methods needed to formulate the research questions and specific procedures of a given domain of knowledge in such a way as to benefit from the application of **computational processing** result in the *algorithms* applicable to a given discipline.

[Manfred Thaller]

I. Text & Markup

What kind of **relationship** ?

a **problematic** one

What is text ? (1)

A **technological** answer:

information **coded as** characters or **sequences of characters**

not

literary material as originally written by an author

A. C. Day, *Text Processing*, Cambridge, Cambridge University Press, 1984, p. 1.

What is text ? (2)

A **literary critic's** answer:

"The text is **not a physical reality** at all but a concept-limit [*Grenzbegriff*]." "The nature of the text is not material [...] the text is only' and 'always an **image**."

C. Segre, Introduction to the analysis of the literary text, Bloomington, Ind., Indiana University Press, 1988, pp. 301, 315 .

Adequacy

Does **markup technology** succeed in bringing the **digital representation** of the **text** – this particular **image** of the text – in line with its **literary apprehension**?

Why markup ?

The pure and simple **character sequence** is **not** adequate **enough** to represent all of the information contained in the "**literary material** as originally written by an author"

- **graphic code**
- **paratext**

the answer: **Markup languages**

Origin of Markup

A **typographic** origin: proof correction

Document production systems

specific markup *vs* **generic** markup

procedural markup *vs* **declarative** markup

a **standard** for **declarative** markup: **SGML** (**XML**)

A closer characterization

Markup is […] simply the **denotation of** specific **positions** in a text [string of characters] with some assigned tokens [**tags**]

(**D. R. Raymond et al.** (1992), 'Markup Reconsidered,' p. 4.)

XML **markup**

- is **embedded** : its position in the data is information bearing
- assigns **structure** to the data
- the assigned structure is a **hierarchical** tree structure

The OHCO thesis

S. J. DeRose, D. G. Durand, E. Mylonas, and A. H. Renear (1990), "**What Is Text, Really?**," p. 6:

an **ordered hierarchy** of content objects, or '**OHCO**'

Difficulties :

(a) **overlapping** hierarchies
(b) no **semantics**

(a) Overlapping

Difficulty in dealing with **textual variants**

A proposed **solution**: **MVD** format

MVD

How to represent textual variants

A Data Structure for Representing Multi-version Texts Online

Desmond Schmidt [a], Robert Colomb [b]

[a] School of ITEE, University of Queensland, Brisbane, Australia

[b] Centre for Advanced Software Engineering, University of Technology, Malaysia

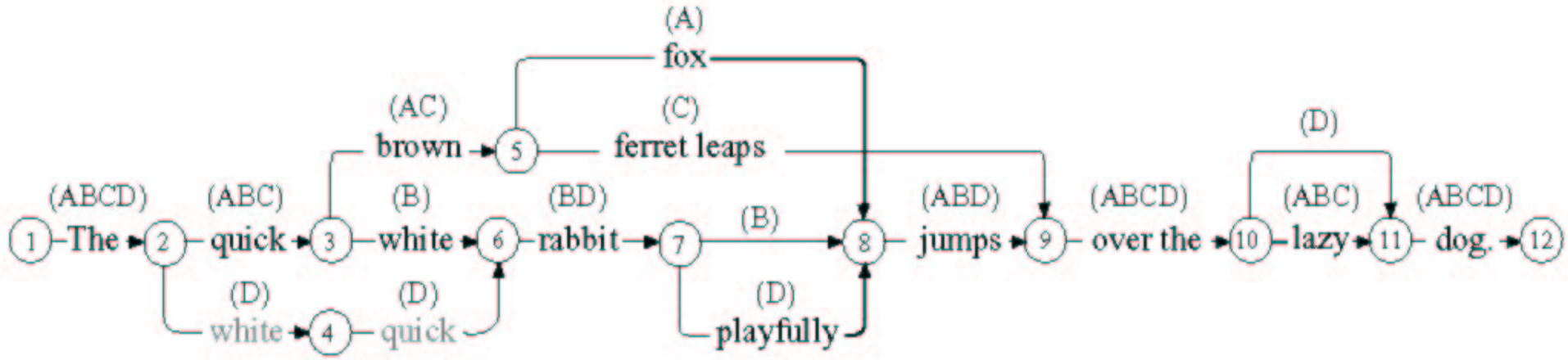
International Journal of Human-Computer Studies (2009)

Multi-Version Documents

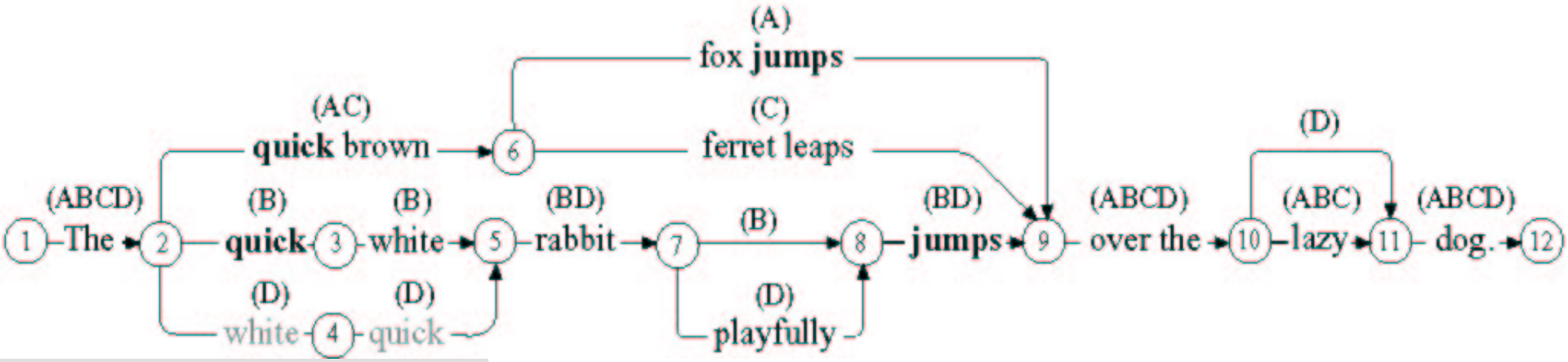
THIS PROJECT IS ABOUT CREATING A WIKI TO HANDLE DOCUMENTS CONSISTING OF MULTIPLE SIMULTANEOUS VERSIONS (MVDS) OR WHICH CONTAIN OVERLAPPING MARKUP.

http://multiversiondocs.blogspot.com/

MVD graph



Variant graph



Well-formed variant graph

(b) Semantics

database community *vs* **document community**

- the **database** community chose to standardize the **semantics** of data
- the **document** community chose to standardize the **representation** of data

The scholarly community

Attempts to **define semantics** in the scholarly community, most notably the Text Encoding Initiative, similarly met with **resistance**. Thus, the route proposed by **SGML** was a reasonable one: promote the notion of **application and machine independence**, and provide a base on which semantics could eventually be developed, but avoid actually specifying a semantics

D. Raymond et al. From Data Representation to Data Model: Meta-semantic issues in the evolution of SGML, in «Computer Standards & Interfaces», 18 (1996), 25-36, pp. 26-28.

Semantic markup ?

TEI markup :

from **markup** to **ontology**
from **ontology** to **markup**
but

- **not** machine-**processable**
- **no interoperability**

A possible **solution** : **stand-off markup**
semantic web technologies

II. Digital Humanities & Markup

What kind of **relationship** ?

an **influential** one

John Unsworth

J. Unsworth (2004), Forms of Attention: Digital Humanities Beyond Representation :

forms of attention (Frank Kermode) **change** over time

we are, I think, on the verge of what seems to me the **third major phase** in **humanities computing**, which has moved from **tools** in the 50s, 60s, and 70s, to **primary sources** in the 80s and 90s, and now seems to be moving back to **tools**

I think we are arriving at a moment when the form of the **attention** that we pay to primary source materials is shifting from **digitizing** to **analyzing**, from artifacts to aggregates, and from **representation** to **abstraction**

Digital Humanities phases

tools – **processing**

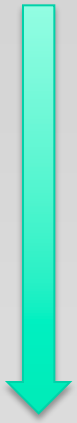
primary sources – **representation**

tools – **processing**

Technology phases

different phases

from



mainframes

PCs and stand-alone workstations

the World Wide Web

to

Web and markup

HTML and XML :

- data **representation** languages
- **not** data **processing** languages

XSLT :

- can be thought of as a complete **programming language**
- it transforms **tree structures** into other **tree structures**
- it processes the structure of the document, or of the **expression** of the text not of its information **content**

Beyond representation

We've spent a generation furiously building **digital libraries**, and I'm sure that we'll now be building **tools** to use in those libraries [...] I'm sure that the text won't go away while we do our tool-building—but I'm also certain that our tools will put us into new relationships with our texts.

John Unsworth

Forms of Attention: Digital Humanities Beyond Representation (2006)

Semantic Web

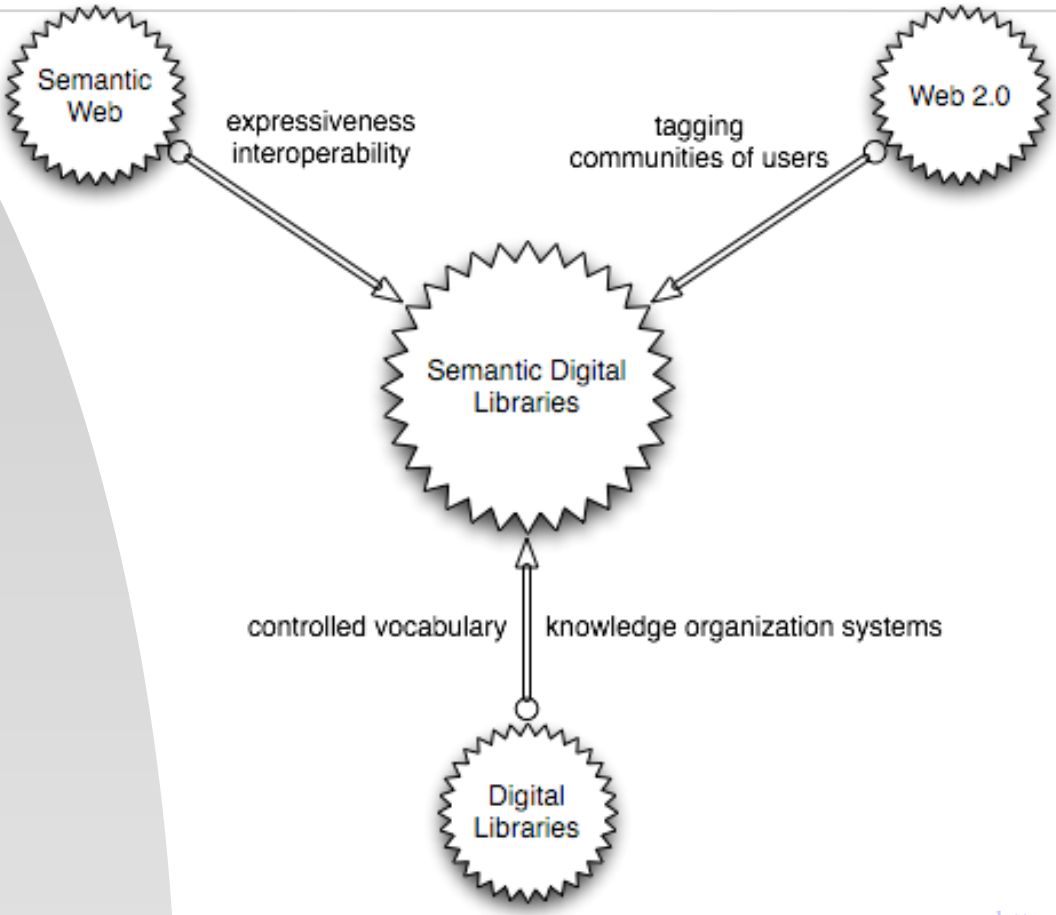
from

visualization

to

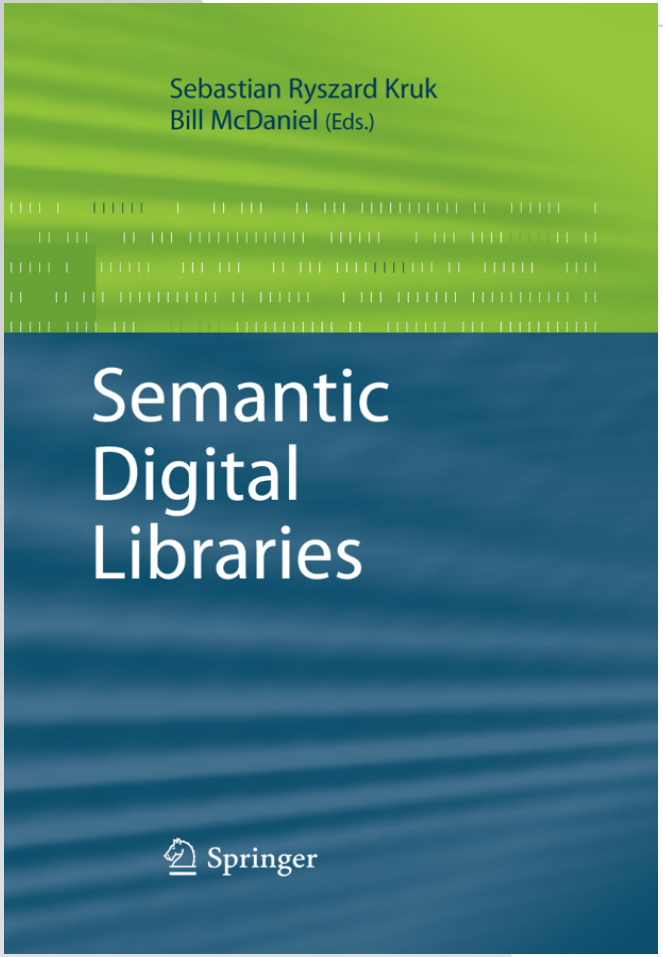
content processing

Semantic Digital Libraries



http://semdl.info/

Semantic Digital Libraries

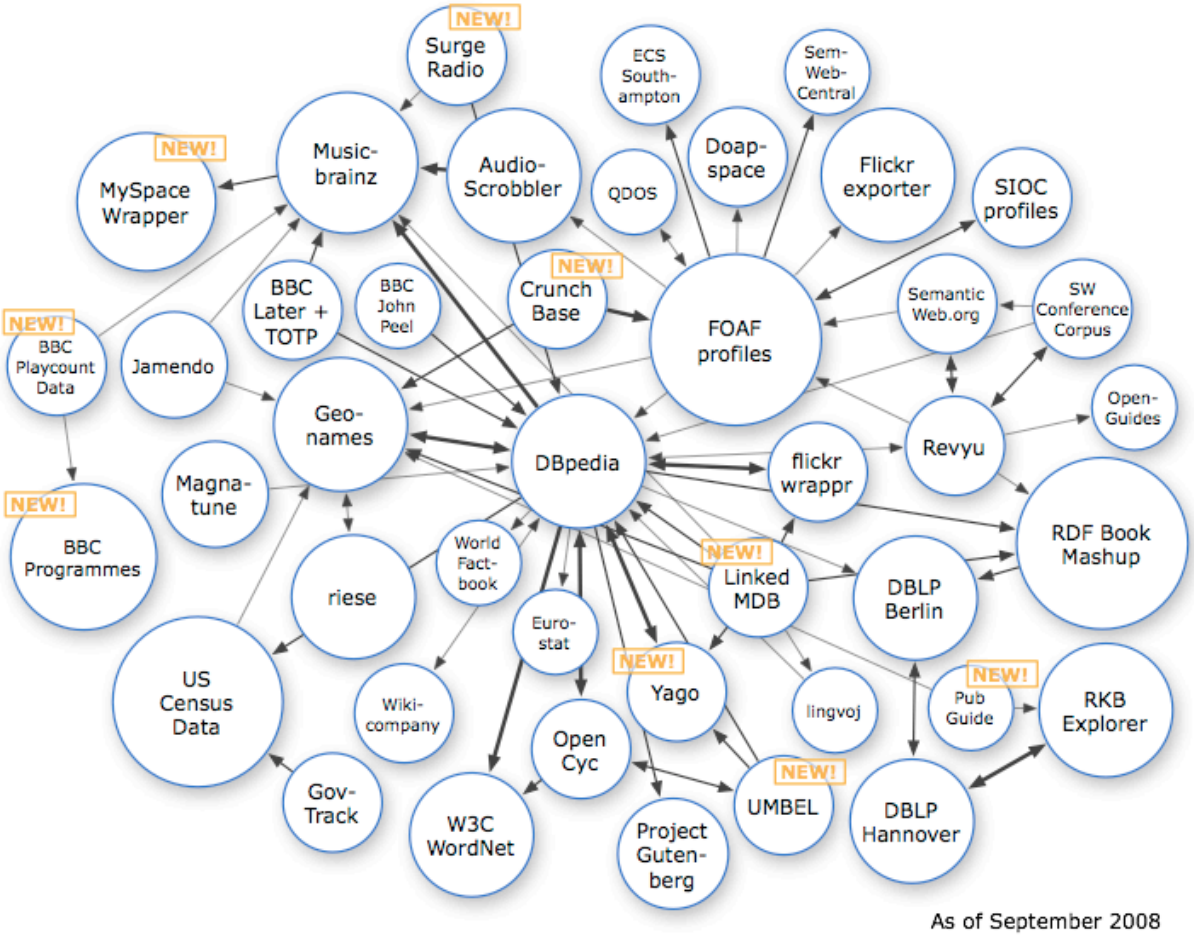


S. R. **Kruk** and B. **McDaniel** (eds),
Semantic Digital Libraries,
Berlin, Springer, 2009

Web:

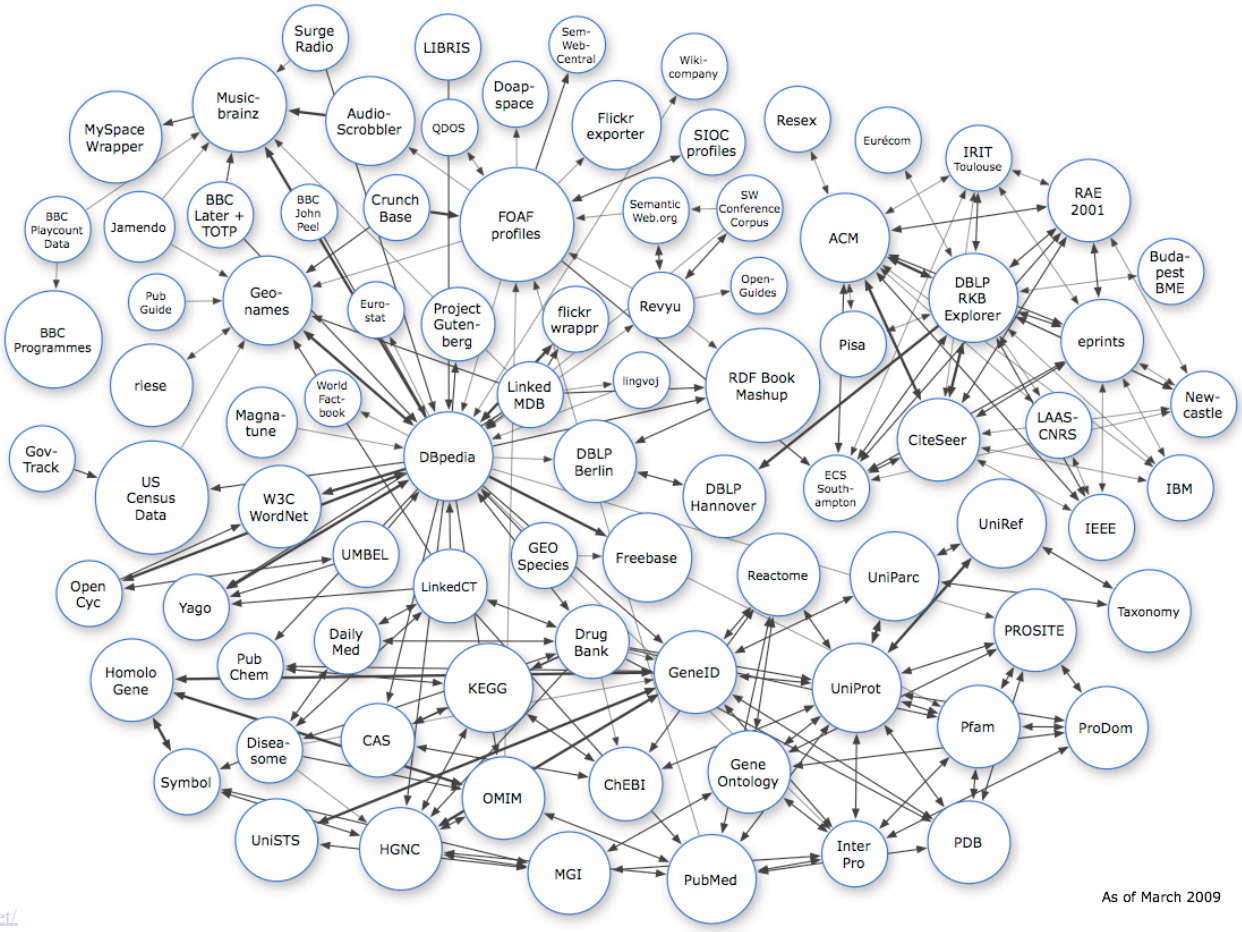
Semantic Digital Libraries
http://semdl.info/

LOD (Linking Open Data) cloud 2008



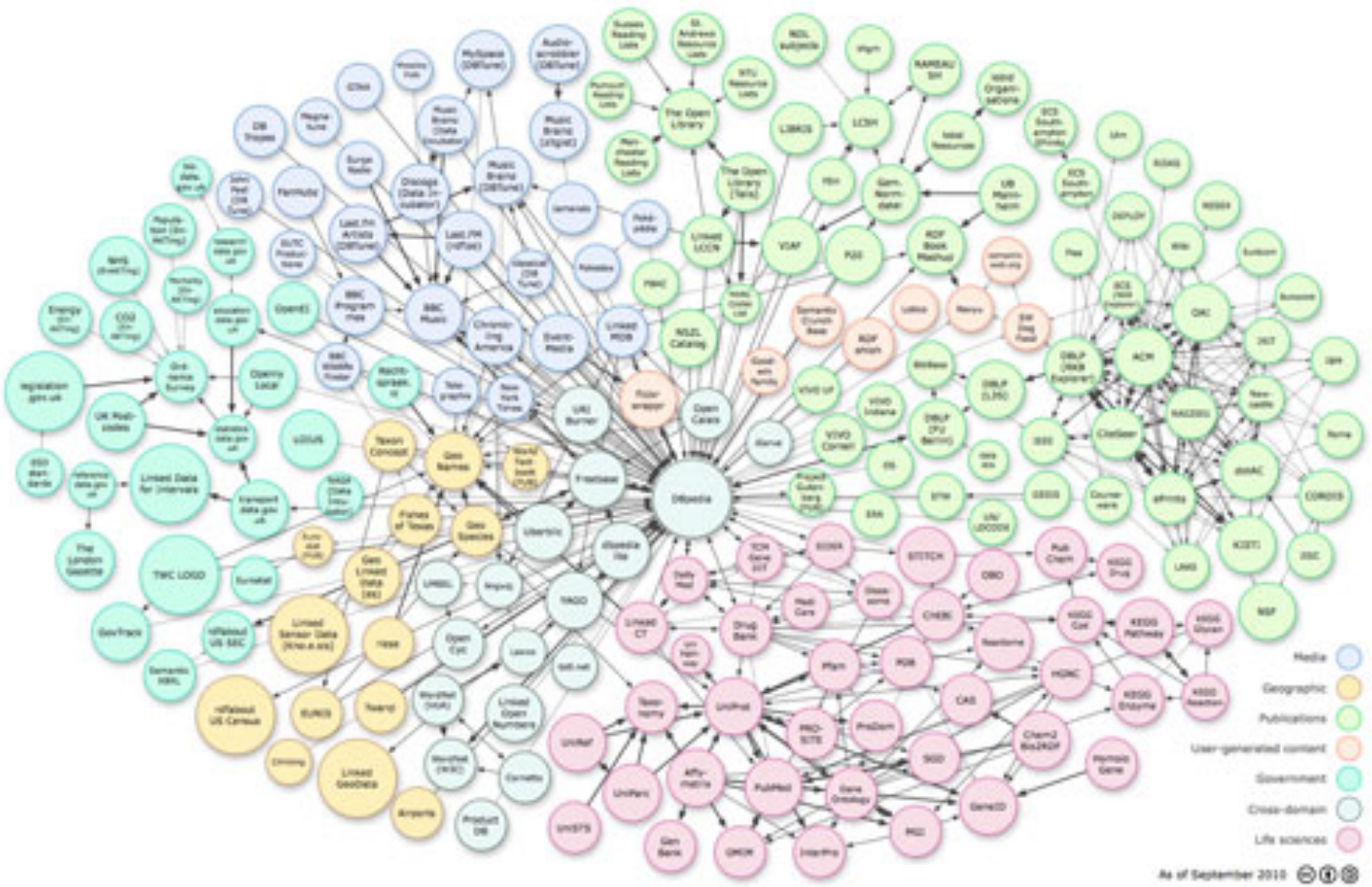
source: http://lod-cloud.net/

LOD (Linking Open Data) cloud 2009



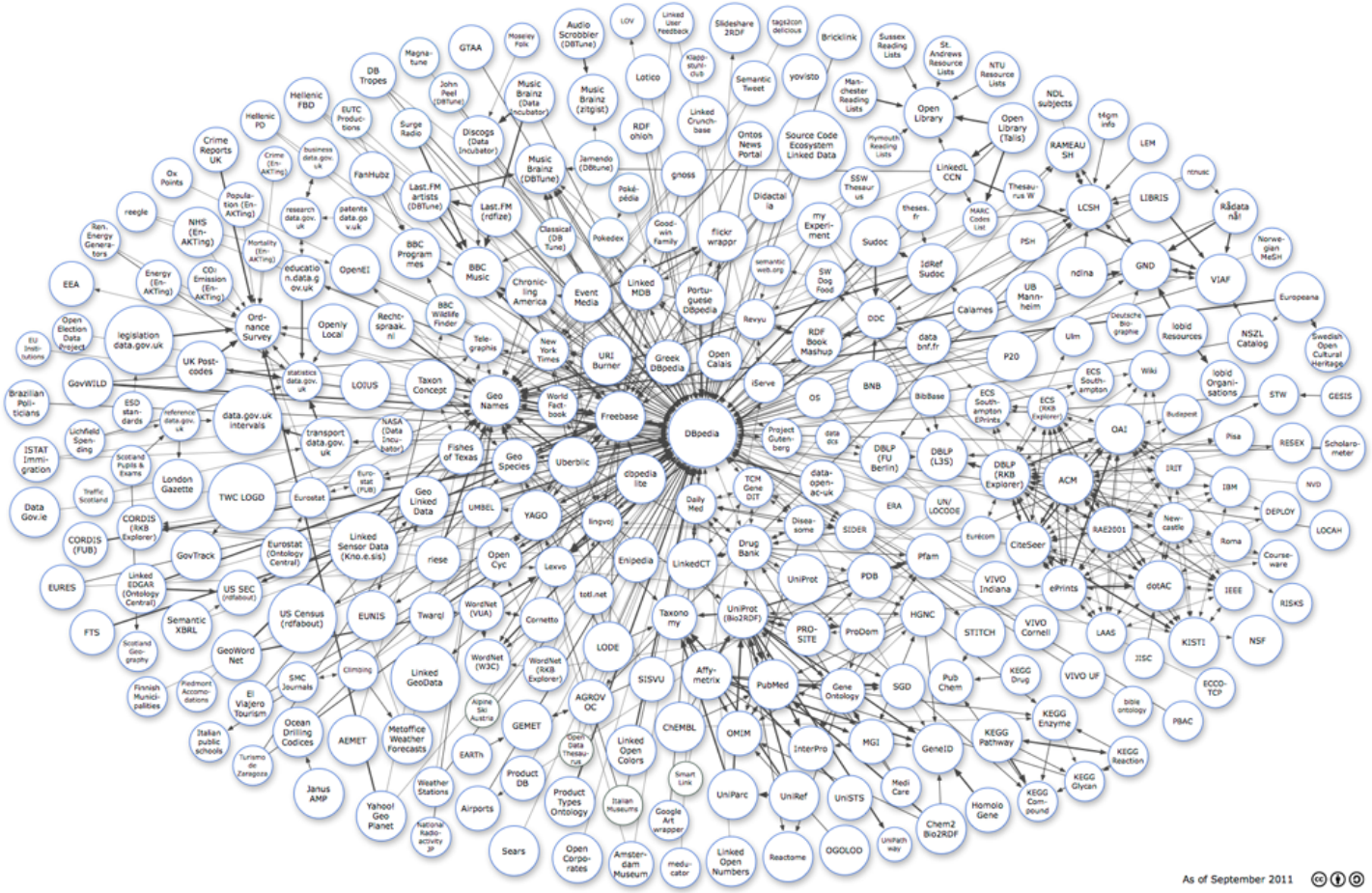
source: http://lod-cloud.net/

LOD (Linking Open Data) cloud 2010



source: http://lod-cloud.net/

LOD (Linking Open Data) cloud 2011



source: http://lod-cloud.net/

Gracias !

Title

ooooo ooooo ooooo ooooo ooooo

ooooo ooooo ooooo ooooo ooooo