Waking from a Dogmatic Slumber -

A Different View on Knowledge Management for DL's

NKOS Workshop

Martin Doerr

Center for Cultural Informatics
Institute of Computer Science
Foundation for Research and Technology - Hellas

Alicante, Spain
September 21, 2006

Knowledge Management for DLs

Traditional Use Cases

“There are no new research challenges in DL. There are only the ones from 30 years ago we still have not solved” (anonymous, ECDL2005)

Apologies: *I’ll be deliberately provocative and possibly incomplete. Don’t take me too serious.*

What are Digital Libraries (or more generally *Digital Memories*)?

Information systems preserving and providing access to source material, scientific and scholarly information, such as libraries of publications, experimental data collections, scholarly and scientific encyclopedic or thematic databases or knowledge bases.

Knowledge Management for DLs

Traditional Use Cases

The traditional library task:

- Collect and preserve documents and provide finding aids
- The job is solved, when *the* (one, best) document is handed out. “All you want is in this document”.

Implementing the finding aids:

- Assumption: User knows a topic, characterized by a noun, or knows associations of the topic uncorrelated to the problem to be solved (e.g. “organic farming” for “host-parasite studies”.)
- Semantic interoperability is limited to the aggregation task: Metadata are mainly homogeneous (DC, MARC etc.), challenge is the matching of terminology (KOS).

Knowledge Management for DLs
Problems

- No support to solve a problem,
  - e.g., what species is this object?

- No support to learn from the aggregated source, to retrieve by contexts,
  - e.g., Which professions had the relatives of van Gogh?
  - e.g., Which excavation drawings show the finding of this object?
  - e.g., Which resolution had Galileo's telescope when he observed... (in general how reliable was a scientific observation, can we correct the values found?).

- No support to integrate complementary information in multiple sources into new insight,
  - e.g., Which where the clients of van Gogh's paintings?

- No support for cross-disciplinary search.
  - e.g. Ecology, ethnology and biodiversity. Biology and archaeology.

Knowledge Management for DLs

Grand Challenge

DLs should become integral parts of work environments as sources to find integrated knowledge and produce new knowledge.

But How ?

Employing “global networks of knowledge”….

Is that a dream ?

“Isn’t Digital information and human knowledge is too diverse, fuzzy, case-dependent?”

“Is the Semantic Web much further than AI decades before?”

Knowledge Management for DLs

Grand Challenge

We regard suitable knowledge management as the key.

We distinguish:

1. Core ontologies for “schema semantics”, such as: “part-of”,”located at”,”used for”, “made from”. They are small and rich in relationships that structure information and relate content.

2. Ontologies that are used as “categorical data” for reference and agreement on sets of things, rather than as means of reasoning, such as: “basket ball shoe”, “whiskey tumbler”, “burma cat”, “terramycine”. They do not structure information. They aggregate, more than integrate.

3. Factual background knowledge for reference and agreement as objects of discourse, such as particular persons, places, material and immaterial objects, events, periods, names.

Knowledge Management for DLs
Preconceptions and Solutions

“Libraries should not depend on domain specific needs. Domains are too many and too diverse. DLs need a generic approach.”

- This seduces us to only employ intuitive top-down approaches for generic metadata schemata. As a result, when the fantasy is exhausted, research stops.
- We need deep knowledge engineering, generalizing in a bottom-up manner from real, specific cases to find the true generic structures across multiple domains. We need interdisciplinary work on real research scenarios.

“Ontologies are huge, messy, idiosyncratic and domain dependent. Mapping is the only generic thing we can do”

- We are transfixed with ontologies used as “categorical data” (term lists), for which this statement is mainly true. We oversee the different character of ontologies describing “schema semantics”. They may pertain to generic classes of discourse. We need interdisciplinary work.

Knowledge Management for DLs

Preconceptions and Solutions

“Queries are mainly about classes. The main challenge of information integration is the integration of classes (terms).”

- We believe this is not sufficiently supported by empirical studies. Query parameters pertain to universals and particulars and relationships. We need to systematically analyze original research questions.

“Manual work is not scalable or affordable. Only fully automated methods have a chance”

- This seduces us to discard the quality of manual, intellectual decisions. Yet billions of people produce content manually. Wikipedia demonstrates, that the above is not true.
- We need to design the interactive processes and the awarding of users to massively involve Virtual Communities / Organisations in cataloguing, data cleaning and ontology development. We need semiautomatic, highly distributed algorithms. We need interdisciplinary work.

Knowledge Management for DLs

Do we talk about the same thing?

“We need more reasoning!”

- This is true. But what sort of reasoning? And before any reasoning can be done, data must be connected, in a “global network of knowledge”. We must first clarify:

Do we talk about the same thing?

Requisites for a global network of knowledge:

1. A sufficiently generic global model (core ontology with the revelant relationships).
2. Methods to populate the network: knowledge extraction / data transformation.
3. Massive, distributed, semiautomatic detection of co-reference relations (data cleaning) across contexts and to
4. Curate referential integrity of co-reference in order to create, maintain and improve the consistency of global networks of knowledge as a continuous process (not making yet another database).

- And only then we can do advanced reasoning and intelligent query processing ...

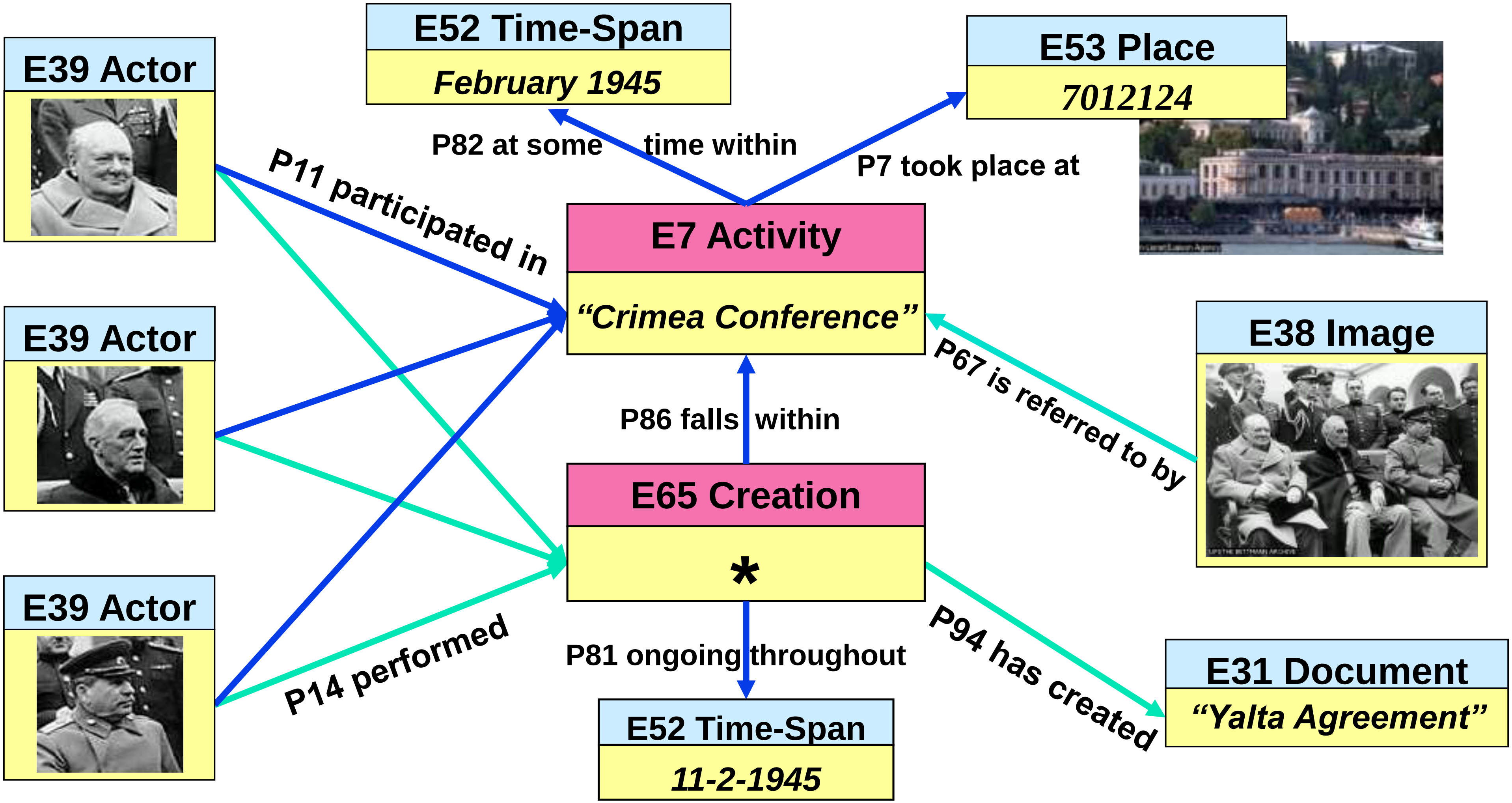
Knowledge Management for DLs
A nearly global model: ISO21127

The CIDOC Conceptual Reference Model (ISO/FDIS 21127)

- is a core ontology describing the underlying semantics of data schemata and structures from all museum disciplines and archives. Now being merged with IFLA FRBR concepts.
- It is result of long-term interdisciplinary work and agreement.
- In essence, it is a generic model of recording of “what has happened” in human scale, i.e. a class of discourse.
- It can generate huge, meaningful networks of knowledge by a simple abstraction: history as meetings of people, things and information.
- It bears surprise: more effective metadata structures, and linking schemes can be created from it.

Knowledge Management for DLs

Example: The ISO21127 Solution



E39 Actor

E39 Actor

E39 Actor

P11 participated in

P14 performed

E52 Time-Span

February 1945

P82 at some time within

E53 Place

7012124

P7 took place at

E7 Activity

"Crimea Conference"

P86 falls within

E65 Creation

*

P81 ongoing throughout

E52 Time-Span

11-2-1945

E38 Image

P67 is referred to by

E31 Document

"Yalta Agreement"

P94 has created

Knowledge Management for DLs

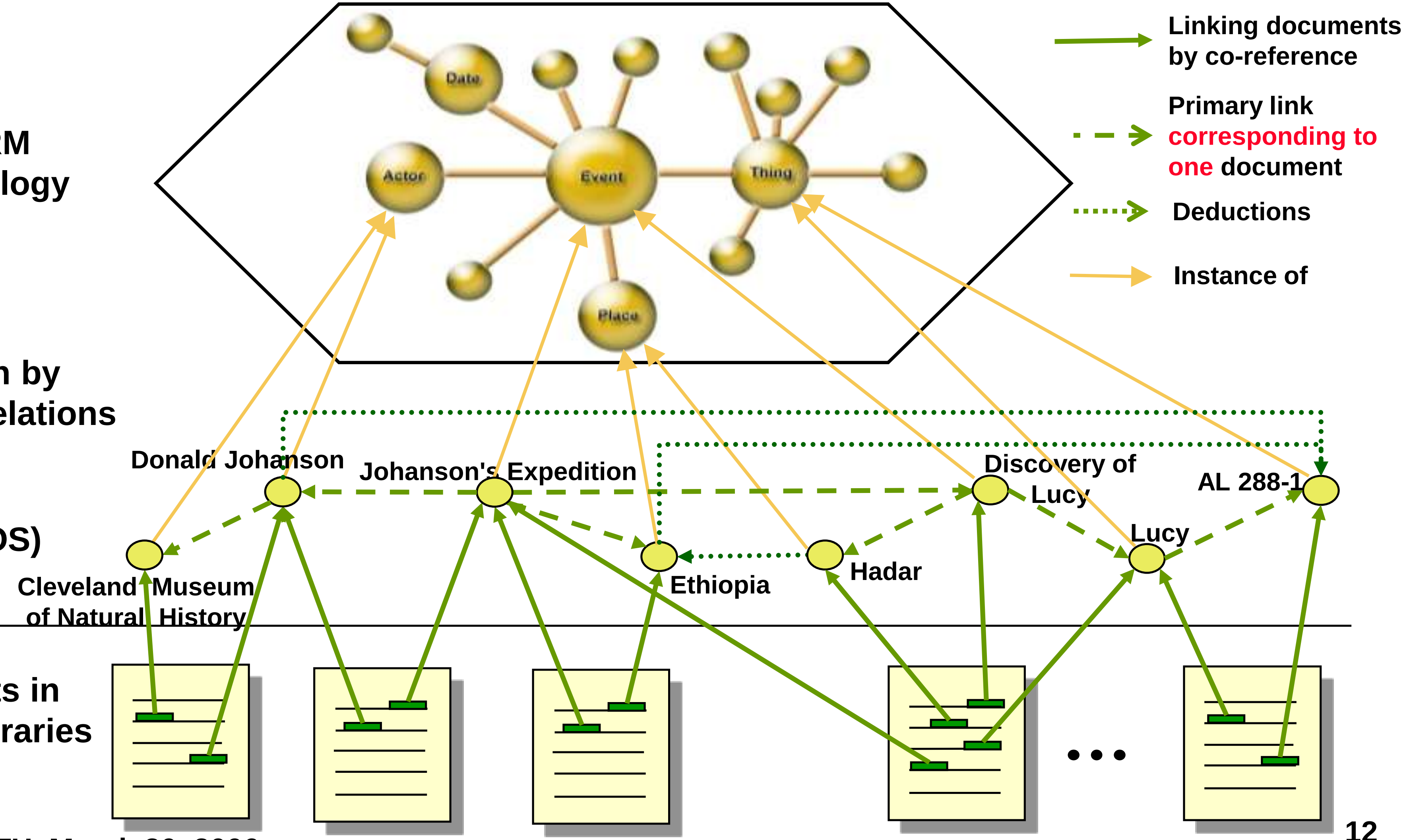
Hypertext is wrong: Documents contain links!

CIDOC CRM
Core Ontology

Integration by
Factual Relations

real world
nodes (KOS)

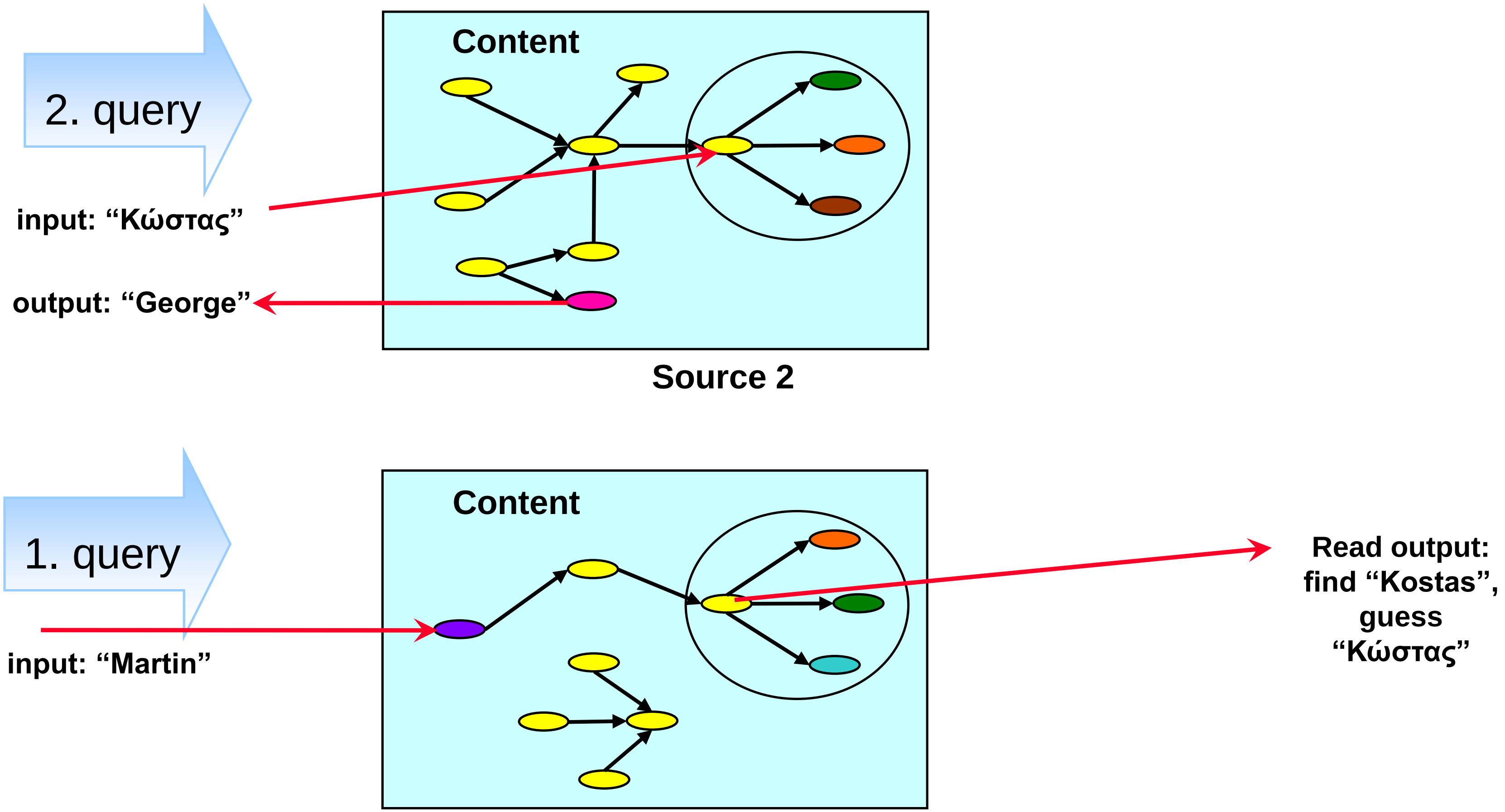
Documents in
Digital Libraries



Knowledge Management for DLs

Identifier Equivalence

Query "Friends of a Friend"



2. query

Content

input: "Κώστας"

output: "George"

Source 2

1. query

Content

input: "Martin"

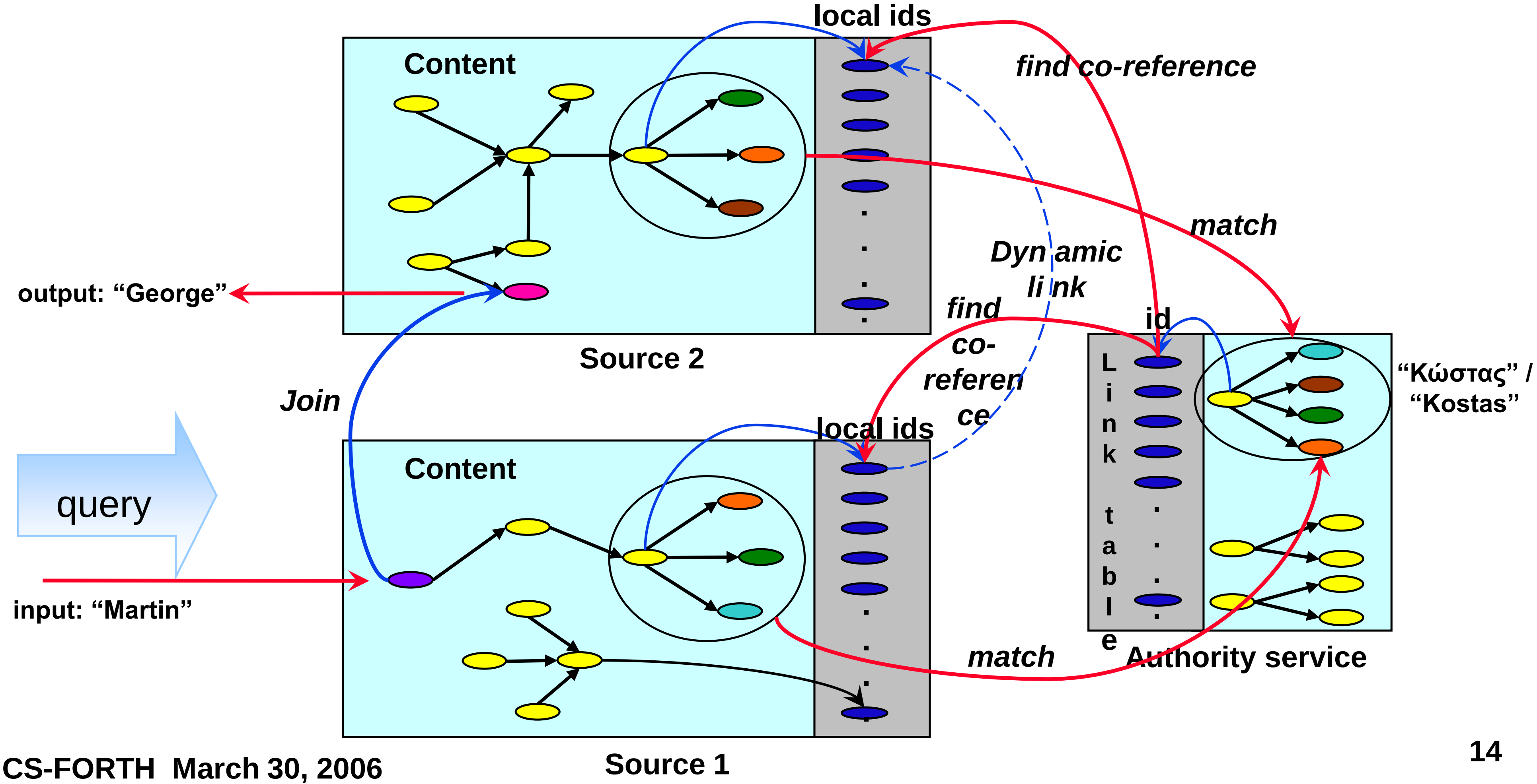
Read output:
find "Kostas",
guess
"Κώστας"

Source 1

Knowledge Management for DLs

Co-reference via Authority

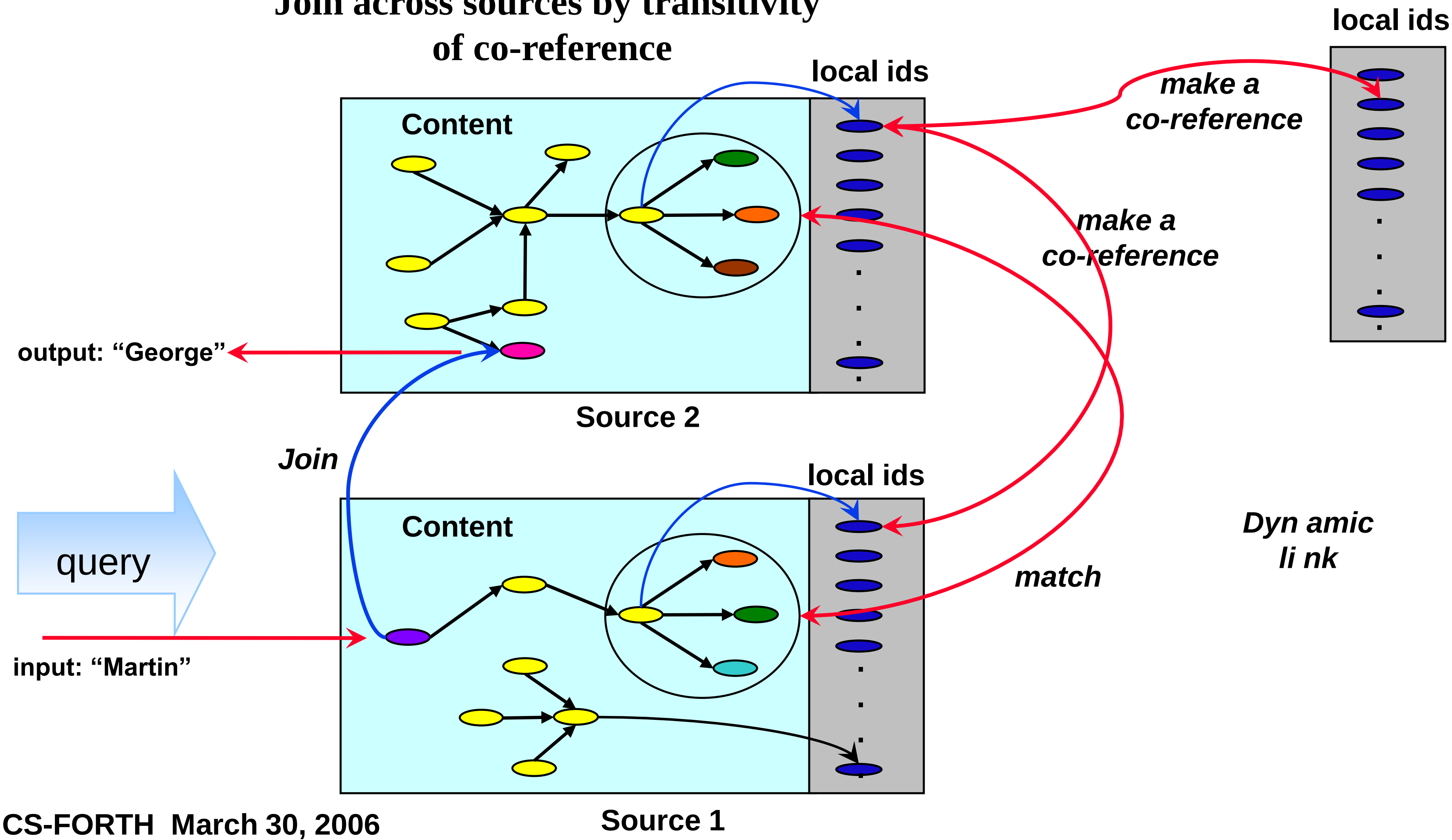
Join across sources by transitivity of co-reference



Knowledge Management for DLs

Curating Co-reference without Authority

Join across sources by transitivity of co-reference



Knowledge Management for DLs
Conclusions

It is feasible to create effective, sustainable, large-scale networks of knowledge:

- The CRM and its extensions seems to have the power to integrate historical knowledge in Archives, Libraries and Museums. Even e-Science applications have been tested.
- The CRM is a model of factual relationships at first. Humanities collect factual knowledge.
- Sciences collect categorical knowledge. But we oversee the record of experimental data, which justifies this knowledge and is by far larger than the resulting categorical knowledge.
- Descriptive sciences already produce both categorical and factual knowledge.

Thesis:

- Once there is a global model, we must invest in managing and preserving co-reference. Else no large-scale networks of knowledge will ever emerge.
- Co-reference clusters can be distributed and are scalable.

Knowledge Management for DLs
Conclusions

If we rethink old positions, we will find surprising new answers to

“..an information model for digital libraries that intentionally moves 'beyond search and access’, without ignoring these functions, and facilitates the creation of collaborative and contextual knowledge environments.”

(C.Lagoze, D-Lib Magazine 2005)

But:

- We need a massive investment in understanding and generalizing the intellectual processes and original research questions in interdisciplinary work.
- We have to do research in dynamic collaborative knowledge organization forms, formal processes and algorithms that converge to higher stages of knowledge integration via co-reference management.
- The large networks of integrated knowledge to come will need continuous maintenance with new, specific social organisation forms and GRID-like resource access, and they may look very different from our current systems…

(This is again a 30 years old challenge, are we closer now?)