The CIDOC CRM

and an

Integrated Approach to Semantic Interoperability

Martin Doerr

Information Systems Lab
Institute of Computer Science
Foundation for Research and Technology - Hellas

Heraklion
February 8, 2006

Global Information Integration
Research Challenges

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

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.

Global Information Integration
Research Challenges

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".

Problems:

- No support to solve a problem,
  - e.g., which ecosystem had the Easter Island in the 17th century?
- 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 objects were found together with 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?).

Global Information Integration
Grand Challenge

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

But How ?

Requisites for a (virtual) global network of knowledge:

1. **A sufficiently generic global model (core ontology with the revelant relationships) for metadata schema integration.**
2. **Automated methods to populate the network: metadata generation by knowledge extraction / data transformation / query mediation from/to existing sources**
3. **Massive, distributed, semiautomatic creation of co-reference relations across contexts in order to connect facts into knowledge networks (curation of co-reference relations as a generalization of data cleaning and authority data).**

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

Global Information Integration
About Knowledge Management

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 (subject of this talk).**
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.**

Global Information Integration
Example: the core ontology ISO21127

The CIDOC Conceptual Reference Model (ISO/FDIS 21127)

- is an extensible core ontology describing the underlying semantics of data schemata and structures from all museum disciplines and archives. Now being merged with generic library concepts from IFLA FRBR.
- 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.
- In particular a core model of scientific observation and related artefacts. Applications have been demonstrated in biodiversity and medicine.
- It can generate huge, meaningful networks of knowledge by a simple abstraction: history as meetings of people, things and information.
- Effective metadata structures and information integration schemes can be derived from it.

Global Information Integration

Example: Meetings and Metadata

Type:	**Text**
Title:	**Protocol of Proceedings of Crimea Conference**
Title.Subtitle:	**II. Declaration of Liberated Europe**
Date:	**February 11, 1945.**
Creator:	**The Premier of the Union of Soviet Socialist Republics**
	The Prime Minister of the United Kingdom
	The President of the United States of America
Publisher:	**State Department**
Subject:	**Postwar division of Europe and Japan**

Metadata

Documents

"The following declaration has been approved:
The Premier of the Union of Soviet Socialist Republics, the Prime Minister of the United Kingdom and the President of the United States of America have consulted with each other in the common interests of the people of their countries and those of liberated Europe. They jointly declare their mutual agreement to concert…
….and to ensure that Germany will never again be able to disturb the peace of the world…… "

Global Information Integration

Example: Meetings and Metadata

Type:	**Image**
Title:	**Allied Leaders at Yalta**
Date:	**1945**
Publisher:	**United Press International (UPI)**
Source:	**The Bettmann Archive**
Copyright:	**Corbis**
References:	**Churchill, Roosevelt, Stalin**

Metadata

About…

Photos, Persons

Global Information Integration

Places and Objects

TGN Id: 7012124
Names: Yalta (C,V), Jalta (C,V)
Types: inhabited place(C), city (C)
Position: Lat: 44 30 N,Long: 034 10 E
Hierarchy: Europe (continent) <– Ukrayina (nation) <– Krym (autonomous republic)
Note: …Site of conference between Allied powers in WW II in 1945; ….
Source: TGN, Thesaurus of Geographic Names

Places, Objects

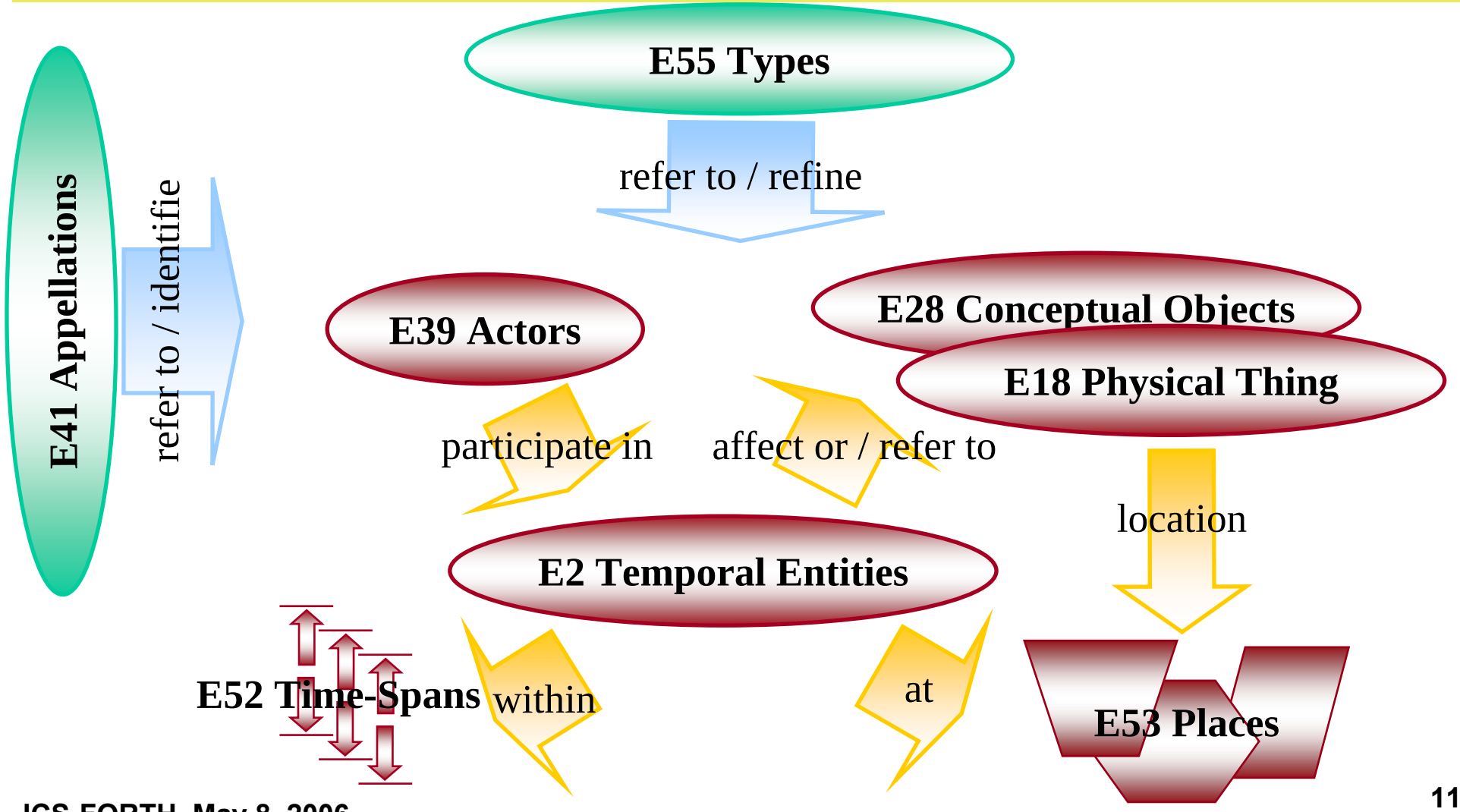
About…

Title: Yalta, Crimean Peninsula
Publisher: Kurgan-Lisnet
Source: Liaison Agency

Global Information Integration

Example: The ISO21127 Solution

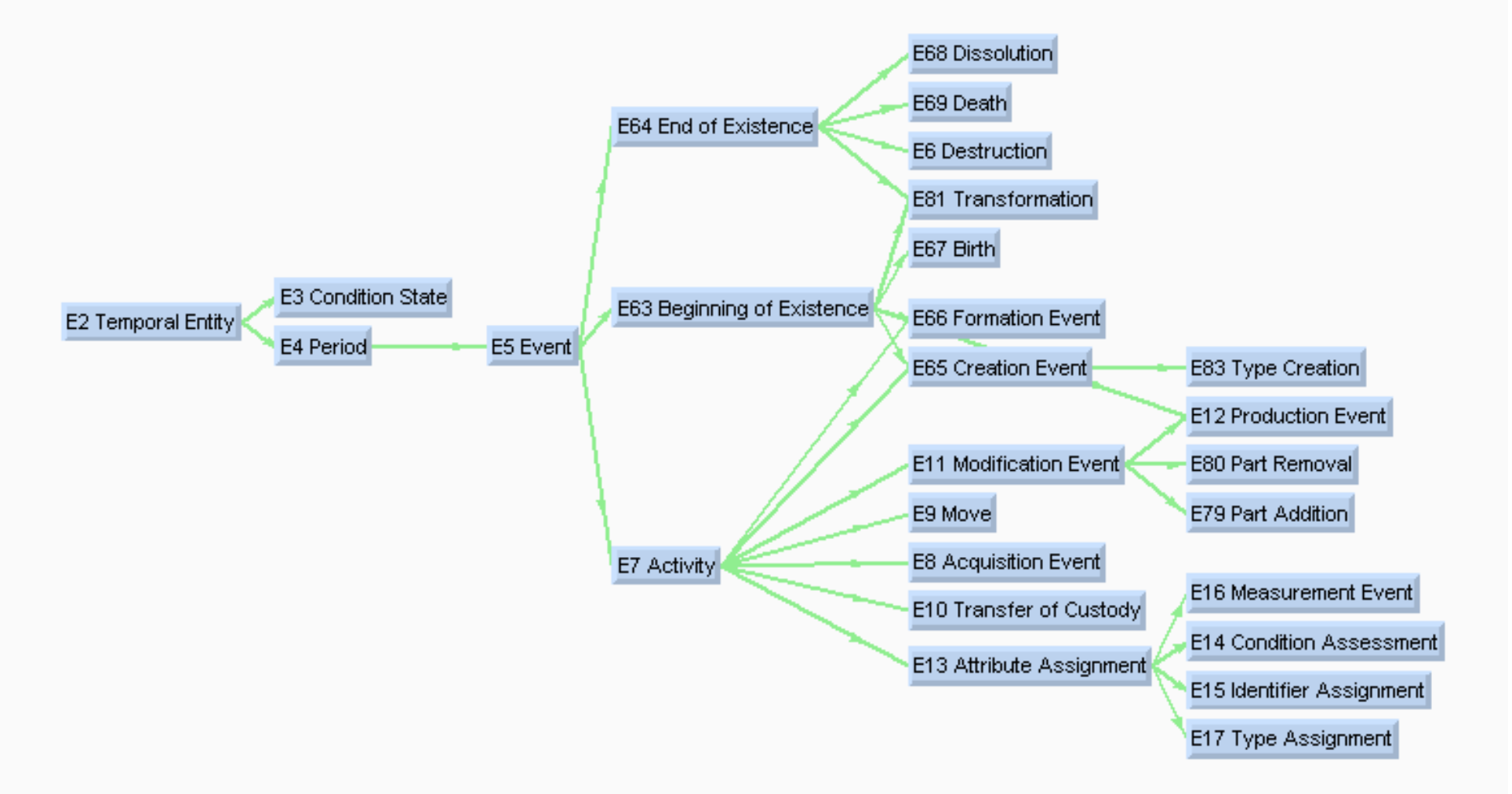
Knowledge Management for DLs
CIDOC CRM *Top-level Entities*



The CIDOC CRM
A Classification of its Relationships

- **Identification** of real world items by real world names.
- **Classification** of real world items.
- **Part-decomposition** and structural properties of Conceptual & Physical Objects, Periods, Actors, Places and Times.
- **Participation** of persistent items in temporal entities.
  - creates a notion of history: “world-lines” meeting in space-time.
- **Location** of periods/events in space-time and physical objects in space.
- **Influence** of objects on activities and products and vice-versa.
- **Reference** of information objects to any real-world item.

The CIDOC CRM
Example: The Temporal Entity Hierarchy



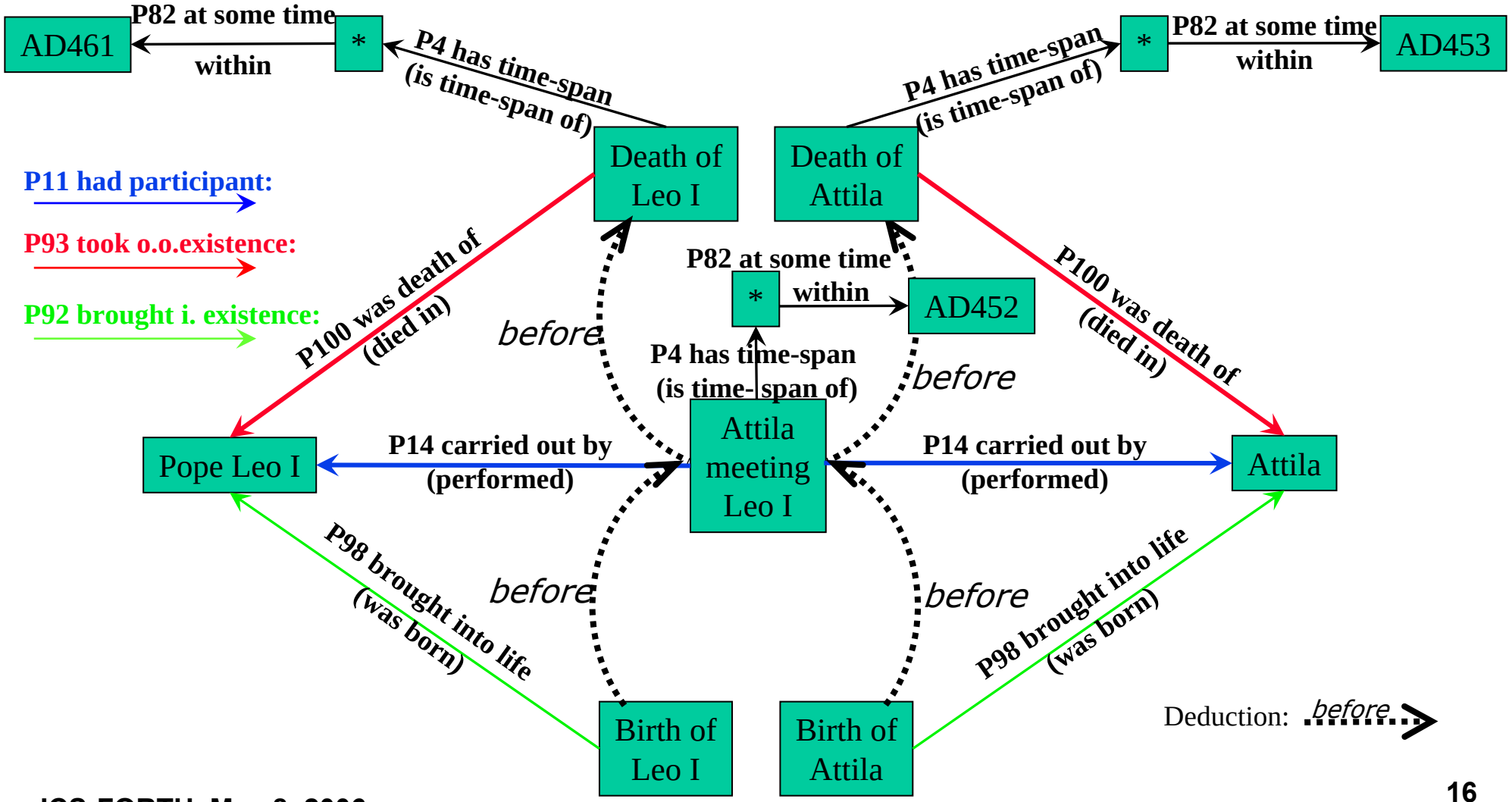
The CIDOC CRM
Temporal Entity- Main Properties

- **E2 Temporal Entity**
  - Properties: P4 has time-span (is time-span of): E52 Time-Span
- **E4 Period**
  - Properties: P7 took place at (witnessed): E53 Place
    P9 consists of (forms part of): E4 Period
    P10 falls within (contains): E4 Period
- **E5 Event**
  - Properties: P11 had participant (participated in): E39 Actor
    P12 occurred in the presence of (was present at): E77 Persistent Item
- **E7 Activity**
  - Properties: P14 carried out by (performed): E39 Actor
    P20 had specific purpose (was purpose of): E7 Activity
    P21 had general purpose (was purpose of): E55 Type

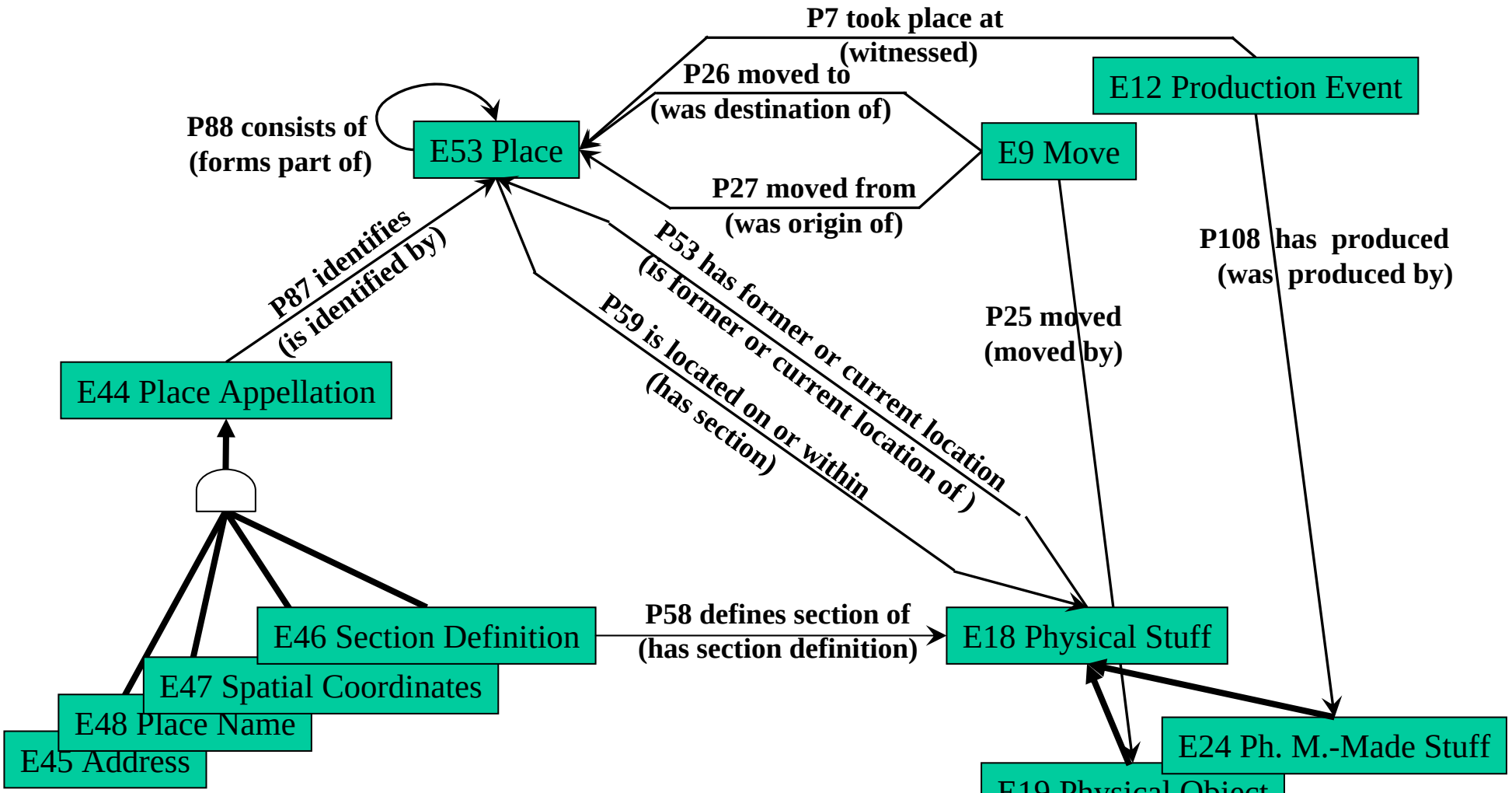
The CIDOC CRM
The Participation Properties

P12 occurred in the presence of (was present at)	E5 Event →	E77 Persistent Item
↘ P11 had participant (participated in)	E5 Event →	E39 Actor
↘ P14 carried out by (performed)	E7 Activity →	E39 Actor
↘ P22 transferred title to (acquired title through)	E8 Acquisition Event →	E39 Actor
↘ P23 transferred title from (surrendered title of)	E8 Acquisition Event →	E39 Actor
↘ P28 custody surrendered by (surrendered custody through)	E10 Transfer of Custody →	E39 Actor
↘ P29 custody received by (received custody through)	E10 Transfer of Custody →	E39 Actor
↘ P96 by mother (gave birth)	E67 Birth →	E21 Person
↘ P99 dissolved (was dissolved by)	E68 Dissolution →	E74 Group

The CIDOC CRM
Termini postquem / antequem

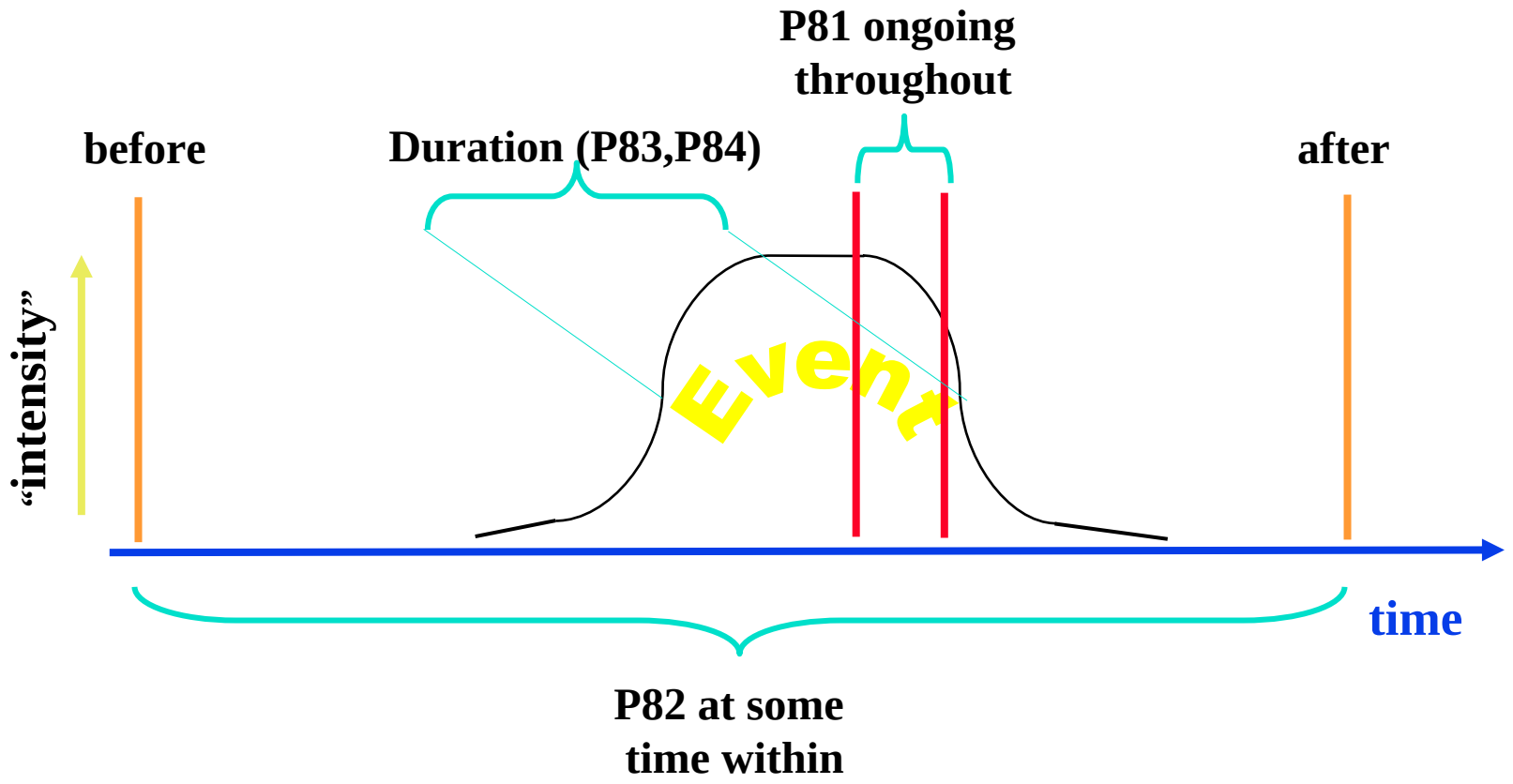


The CIDOC CRM
Place

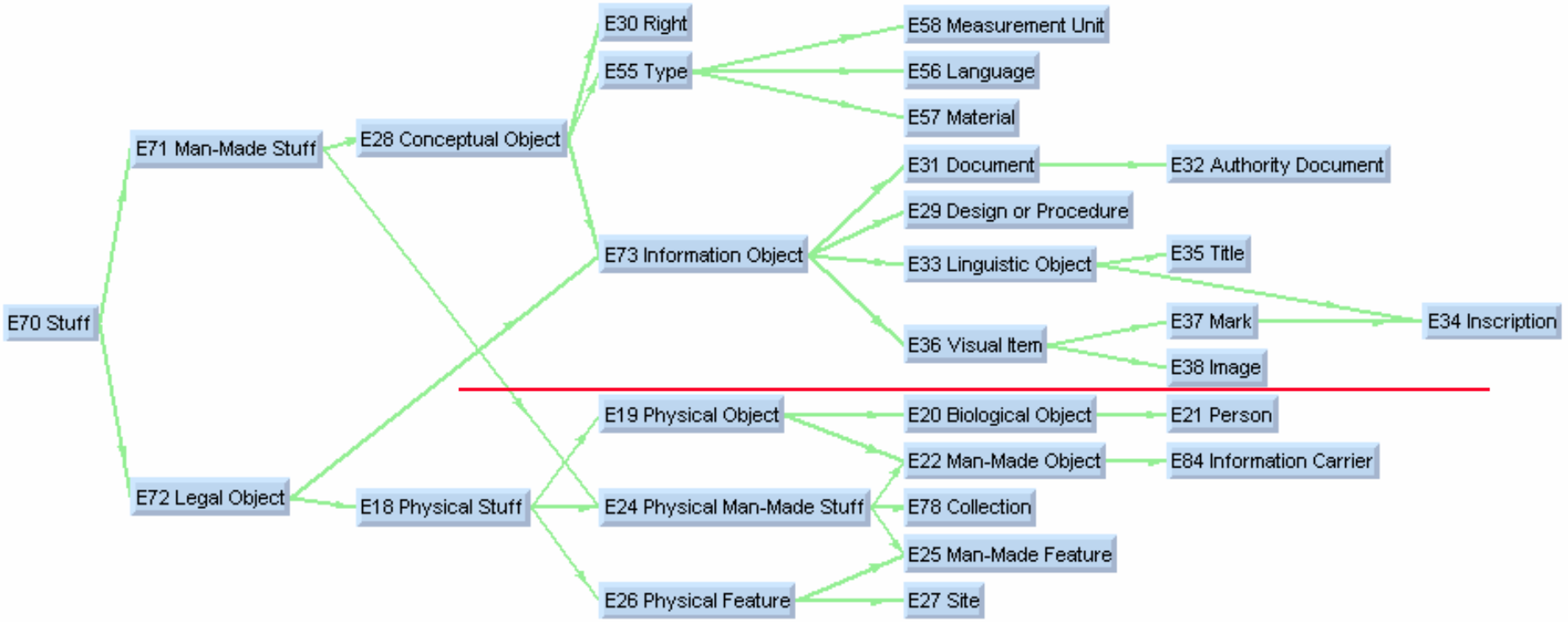


The CIDOC CRM

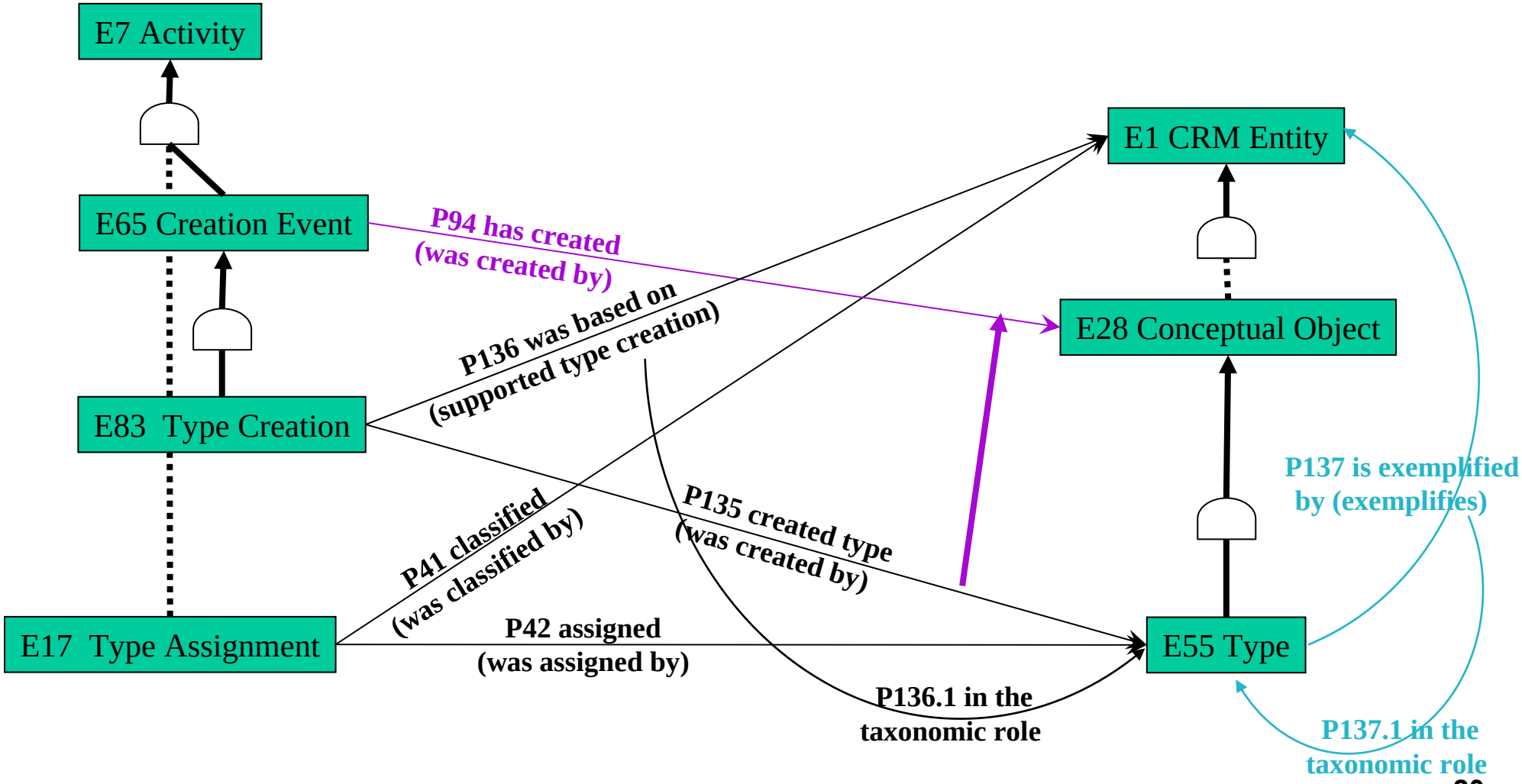
Time Uncertainty, Certainty and Duration



The CIDOC CRM Stuff



The CIDOC CRM
Taxonomic discourse



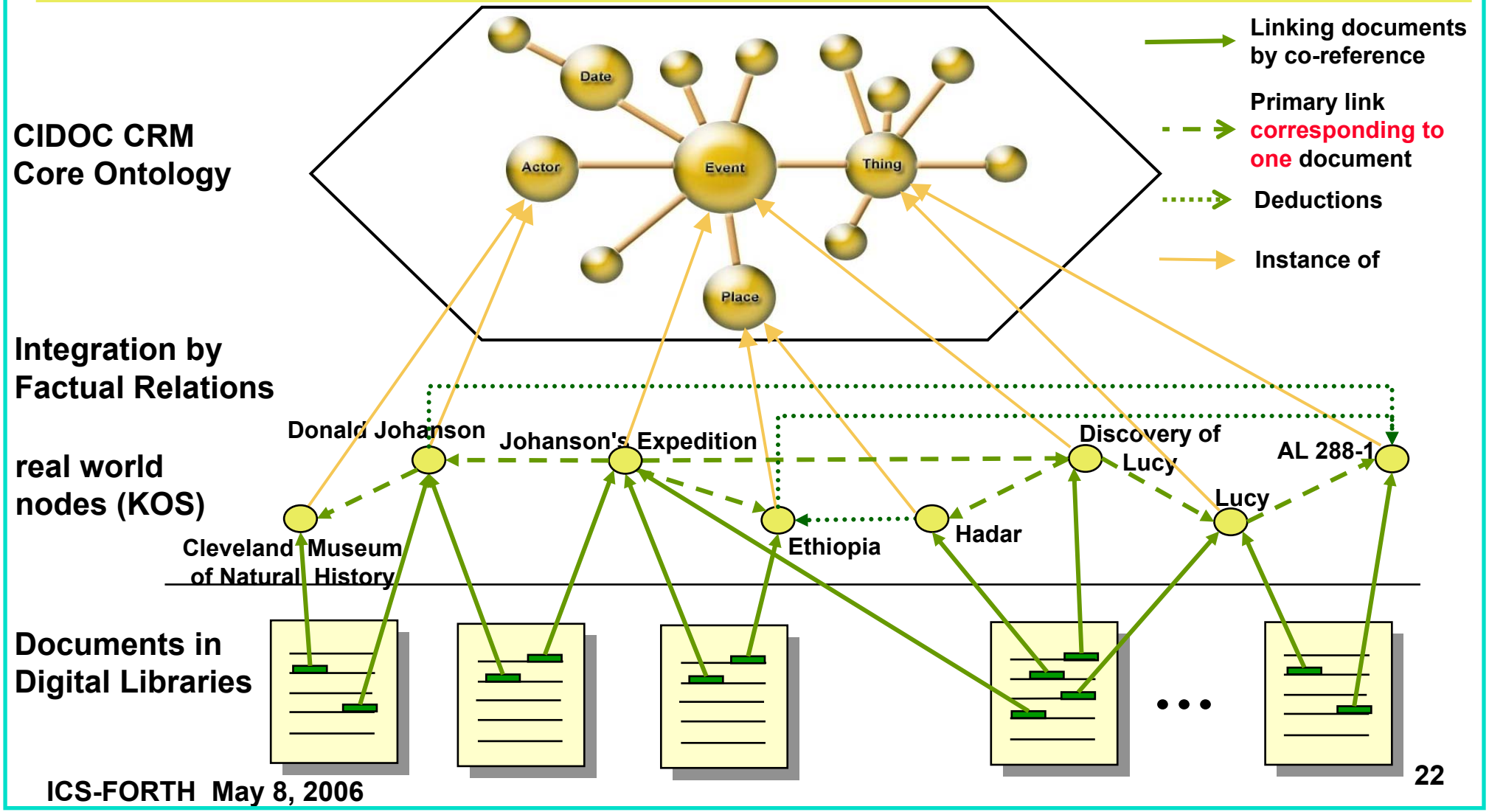
The CIDOC CRM Outcomes

- The CIDOC CRM is a formal ontology (defined in TELOS, RDFS, OWL)
  - An ontology of 80 classes and 132 properties for culture and more
  - With the capacity to explain dozens of (meta)data formats
  - Accepted by ISO TC46 in Sept. 2000, since 2006: ISO 21127
- Serving as:
  - intellectual guide to create schemata, formats, profiles
  - A language for analysis of existing sources for integration

    "Identify elements with common meaning"
  - Transportation format for data integration / migration / preservation
  - A global schema for heterogeneous access (query mediation, data warehousing)

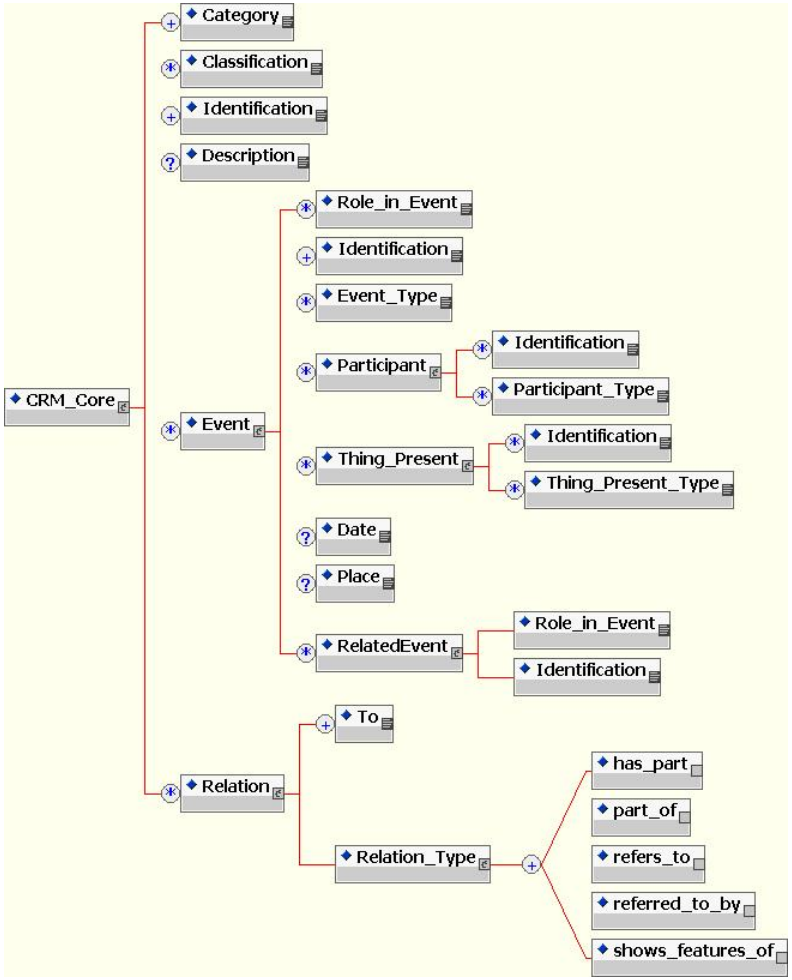
Global Information Integration

A Symbolic Architecture



CRM Core Metadata Schema

- *An ontology **is not a** database schema (data structure)*
- *An ontology can be used as virtual global model (with specific adaptations)*
- *Multiple schemata can be used under a global ontology. The challenge is to use one core ontology and many schemata.*
- *CRM Core is a very simple CRM – compliant metadata schema*
- *There can be very complex CRM – compliant data structures*



Example using CRM Core

Three records combine into one network

E84 Information Carrier
The "Monument to Balzac"(S1296)

P62 depicts

E21 Person
Honoré de Balzac

P62 depicts

E84 Information Carrier
The "Monument to Balzac" (plaster)

P108B was produced by

P16B was used for

E12 Production
Bronze casting"Monument to Balzac" in 1925

P134 continued

P108B was produced by

P2 has type

P2 has type

E55 Type
plaster

E55 Type
bronze

P4 has time-span

P120B occurs after

E52 Time-Span
1925

P14 carried out by

E12 Production
Rodin making "Monument to Balzac" in 1898

P7 took place at

E52 Time-Span
1917

P4 has time-span

E40 Legal Body
Rudier (Vve Alexis) et Fils

P4 has time-span

E52 Time-Span
1898

E69 Death
Rodin's death

P2 has type

E55 Type
companies

E55 Type
sculptors

P2 has type

P14 carried out by

E53 Place
France (nation)

P100B died in

E21 Person
Rodin Auguste

P98B was born

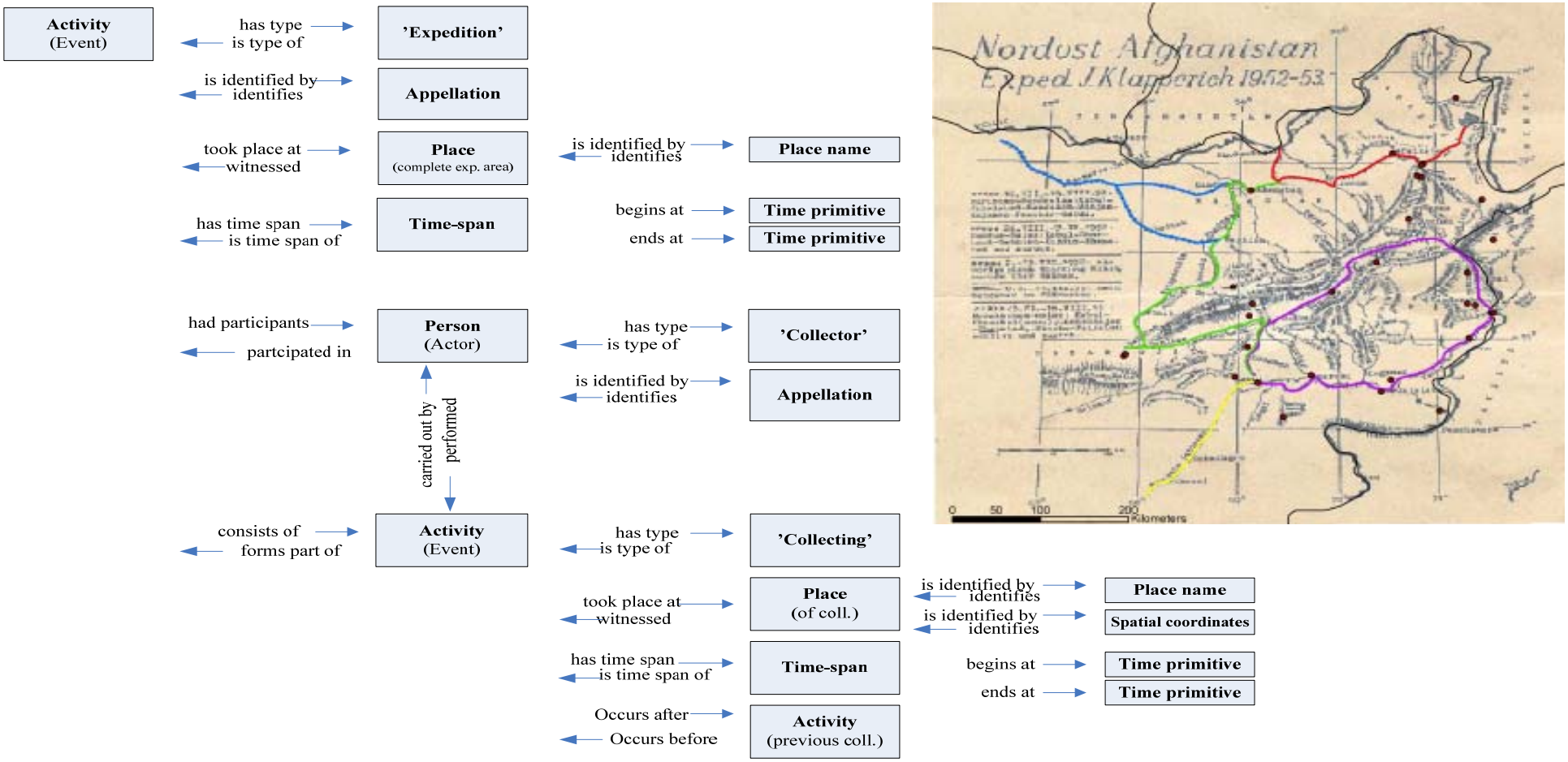
E67 Birth
Rodin's birth

P4 has time -span

E52 Time-Span
1840

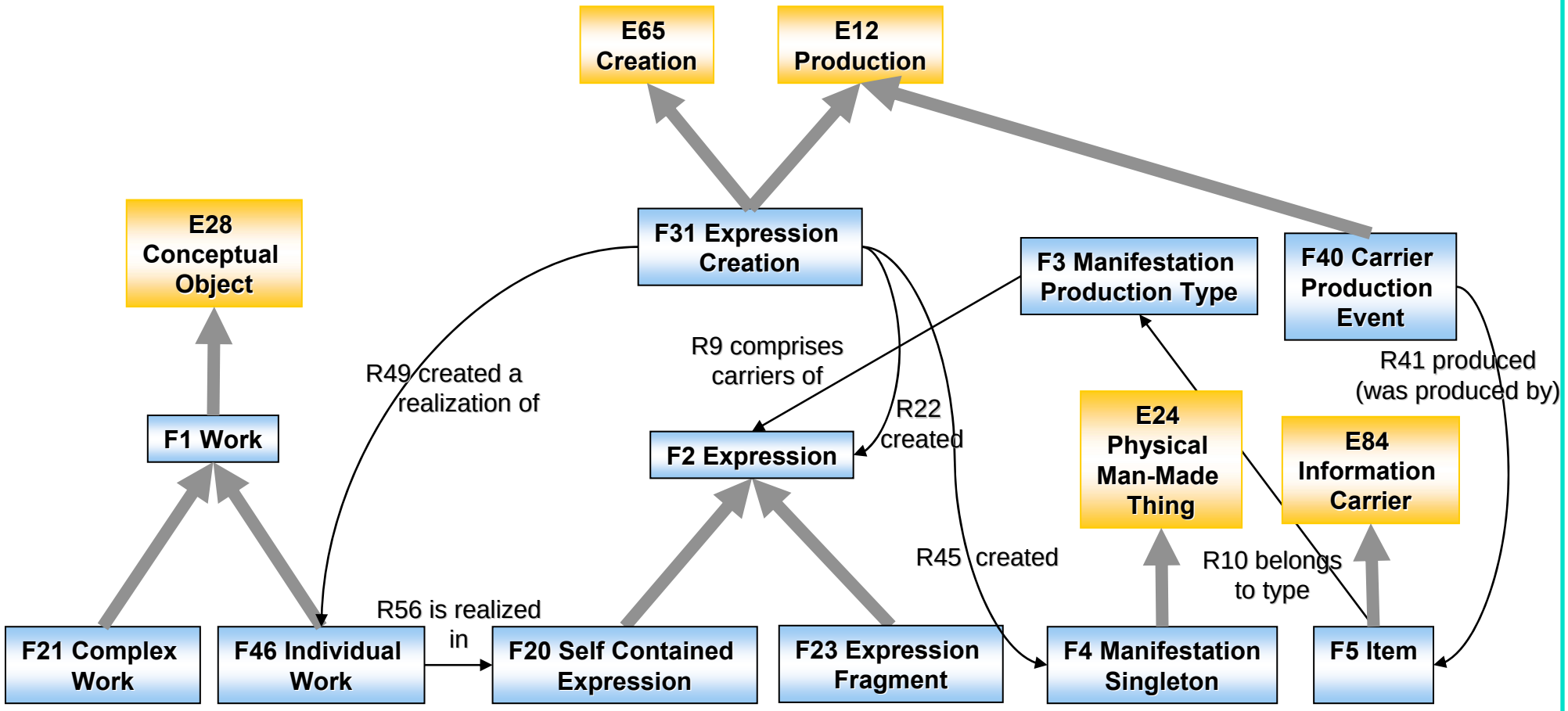
Collector‘s itineraries & expedition routes

(by courtesy of K.H. Lampe, ZMFK Bonn)



The FRBR - CRM Harmonization

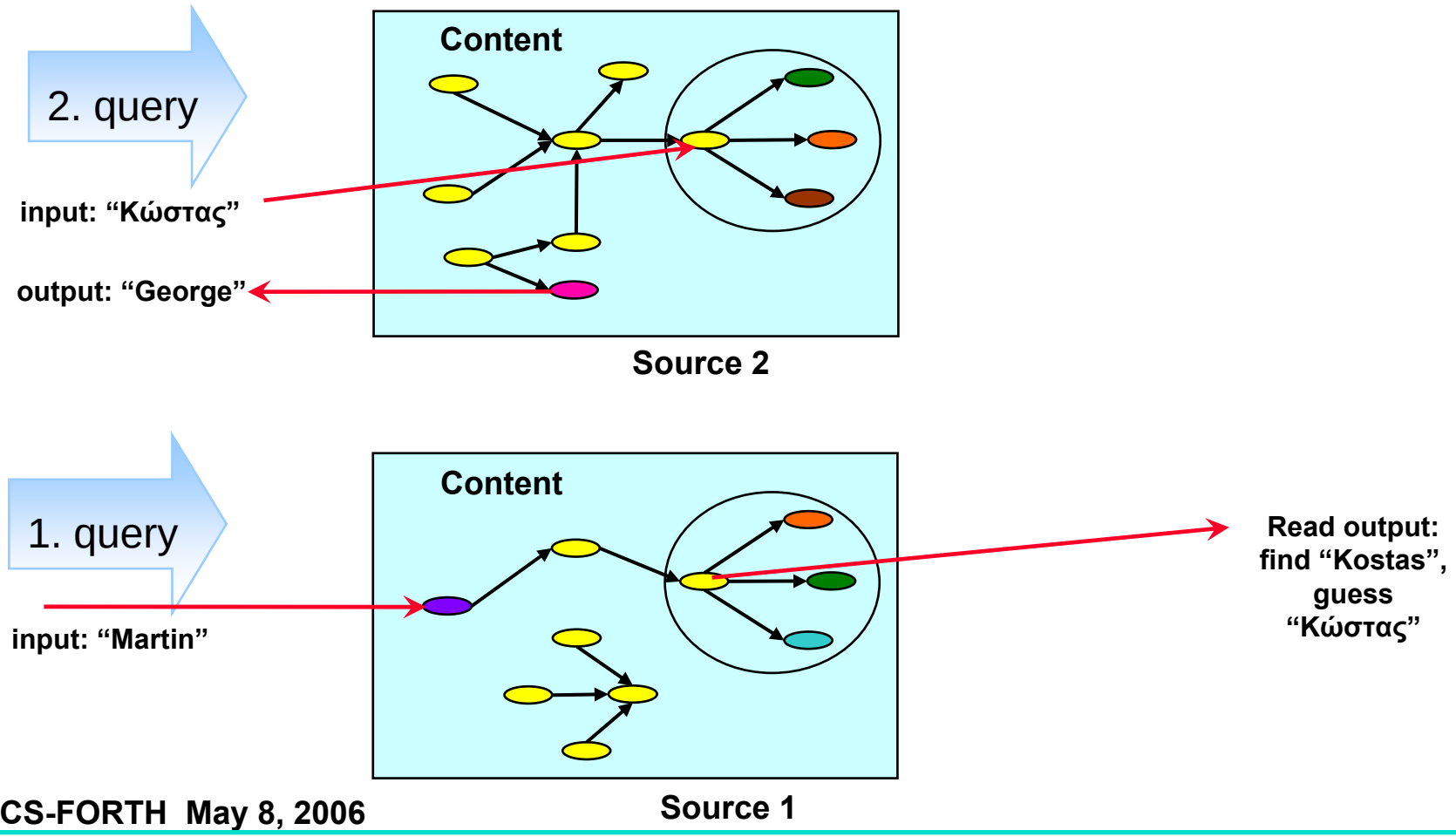
The "Externalization", material / immaterial



Global Information Integration

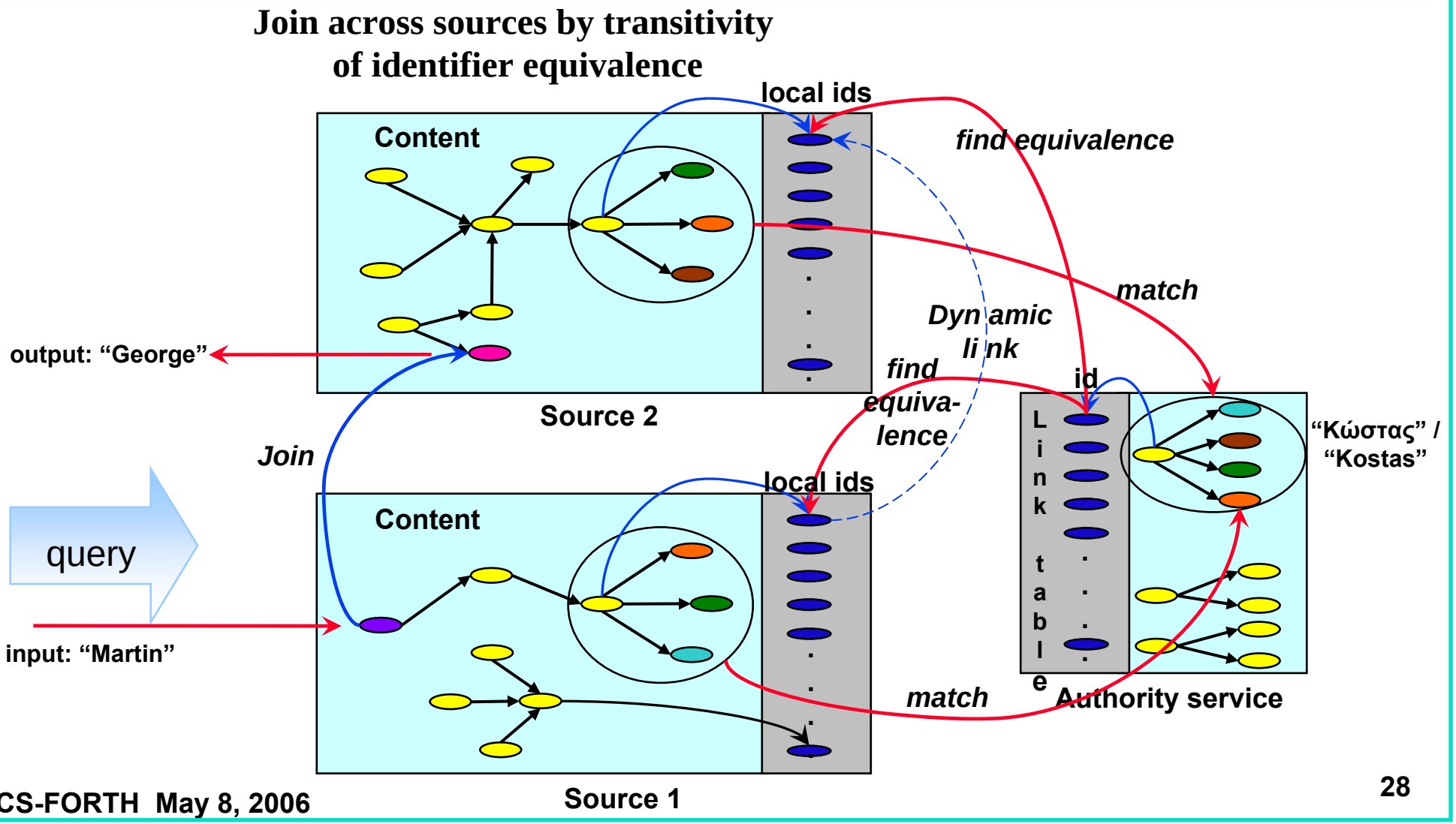
Only if facts are connected…

Query "Friends of a Friend"



Global Information Integration

….*we can do automated reasoning!*



Global Information Integration
Conclusions

The CIDOC CRM is more generally applicable to e-science

- Humanities collect factual knowledge. The CRM is a model of factual relationships at first.
- Sciences aim at categorical knowledge. But we oversee the record of experimental data, which justifies this knowledge and is far larger than the resulting categorical knowledge.
- Descriptive sciences produce both categorical and factual knowledge.

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 and metadata in Archives, Libraries, Museums and scientific observation.
- The CRM can be applied to mediate between multiple data structures of varying complexity.

Thesis:

- The need for multiple core ontologies in DLs should be empirically supported by identifying the necessity of incommensurable concepts in practice (and not by student examples).
- We overestimate the relevance of domain categories (and natural language), and completely oversee the relevance of the historical, factual relationships in our scientific and scholarly reasoning.
- Try the CRM or any extensions of it, as a starting model, and see how far it can be used, and what else (extensions, mapping, transformation etc.) is needed.