The CIDOC CRM

- a Key to Digital Scholarly Knowledge

Martin Doerr

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

Washington, October 25, 2009

The CIDOC CRM:

Outline

Topic:

> *"How the CRM fits to the scholarly processes, IT tools and the new technological opportunities, as a key to manage scholarly knowledge in digital form."*

Outline:

- The scholarly process: collecting, connecting, interpreting
- Information integration
- What we have done
- What we can do now

The CIDOC CRM:

The Scholarly Process -Collection

Primary Knowledge Acquisition: The "Evidence layer"

- Memory institutions and researchers may collect and document

  butterflies, stamps, letters, proceedings. classical Greek vases, Ming porcelain, cars, ships, costumes, medical instruments, birds, lacquer ware, jades, witchcraft, curate excavation sites, nature parks, rock art……
- Organized in collections of extreme specialization: No global schema possible!

IT Situation:

- basically solved: dedicated IT systems since 1960, collection management S/W
- hardly any communication across the systems

Information Acquisition Needs:

- Questionnaire form: sequence and order, control of completeness and plausibility.
- Ergonomic units of documentation, specialized to a domain & its language
- Work-flow on series of analogous items, item-centric.
- Versioning: old opinions may count at any time.

The CIDOC CRM:
The Scholarly Process –Connection

Scholarly research starts with aggregating facts supporting a topic:

- Traditionally, creating personal paper card indices, dossiers, notes,
- physically visiting memory institutions, sites and people,
- After research is done, all non-published facts and correlations are lost!

IT Situation

- Search engines are a poor support! No notion of identity, no notion of facts!
  (such as: "I am here")

The greatest IT challenge: Create a global network of the latest stage of knowledge:

- an epistemic network of related facts, annotated by source/evidence
- comprehensive across contexts, categories and disciplines,
- break up document boundaries by matching co-referring identifiers of items
- use a global schema

Now Semantic Web technology is taking off!

The CIDOC CRM:
The Scholarly Process -Interpretation

The "Interpretation Layer":

Scholarly research continues with hypothesis building

- Connecting events to stories – find the "most probable past" (close gaps, resolve alternatives)
- From individual facts to the greater picture, such as shipwrecks to trade routes etc.
- present in order (storytelling!), illustrate, exhibit, publish (libraries collections!!)

IT situation:

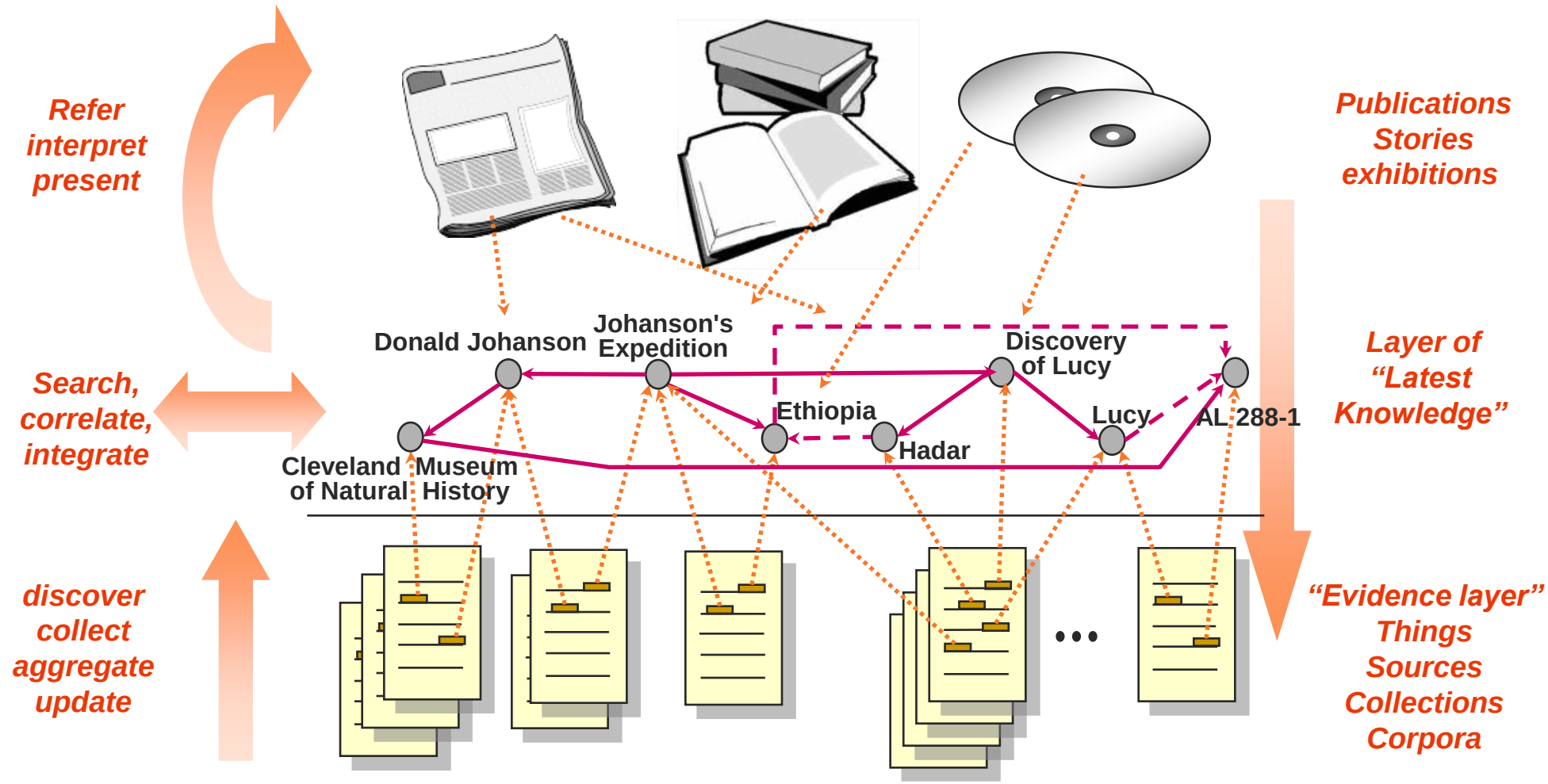
- Dedicated analysis systems, diverse presentation methods, still no integration!

Interpretation needs:

- Standardized access to the integrated "epistemic network"
- Externalization of argument: Web 2.0 systems/annotators on the network.
- Creative presentation methods – diverse and innovative – feed back into the network!

The CIDOC CRM:

The Scholarly Process



The CIDOC CRM:

But how to find the Global Schema?

Data integration systems in the mid 90's has shown:

- impossible to integrate RDBMS under RDBMS
- equally impossible are the "application profiles"!

First problem to solve: integrate more detailed and more general databases

- RDF/OWL can generalize over tables **AND** fields (but the ooDBMS technology fails!) (also could older knowledge representation languages)

Simple idea:

- Break data up into independent propositions (using URIs!), "boxes and arrows"
- describe facts by broader/narrower categories and broader/narrower relationships!
- The **CIDOC CRM (ISO21127)** is a schema based on that!

The CIDOC CRM:

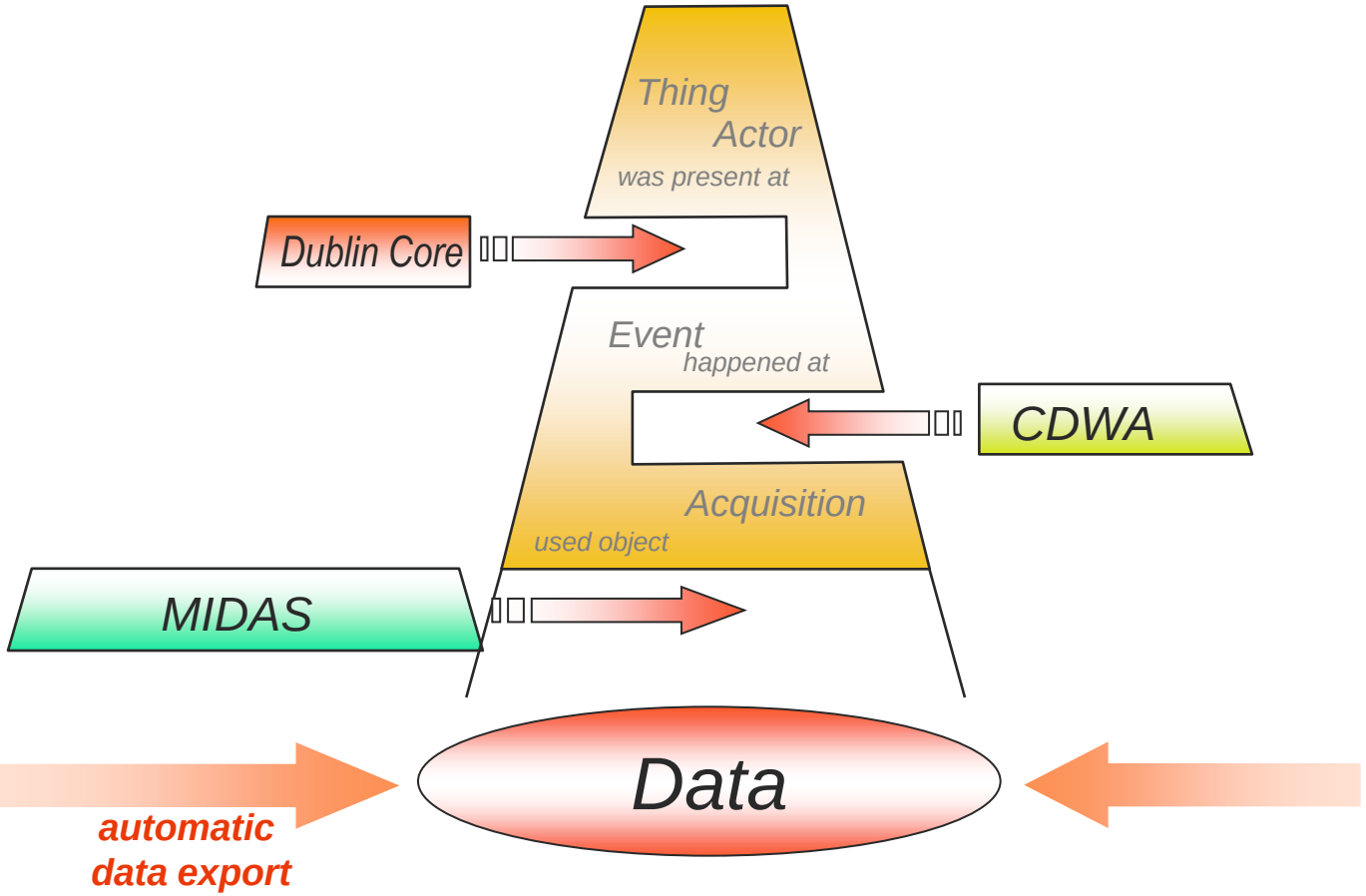
Challenge: Integrating Poor and Rich…

Access all data from any level:

Few concepts, high recall

Special concepts, high precision

CIDOC
Conceptual Reference Model (CRM)



The CIDOC CRM:

But how to find the Global Schema?

Foresight:

- We started 1994-1996 with CIDOC/ICOM Documentation Standards Working Group
- Addressed challenge: integrating poor and rich schemata to built the connection layer
- Convinced Group to use new knowledge representation technology
- => became later compatible with the RDF/RDFS/OWL!
- Proactive, in order to be ahead of industrial solutions.

The CIDOC CRM:

But how to find the Global Schema?

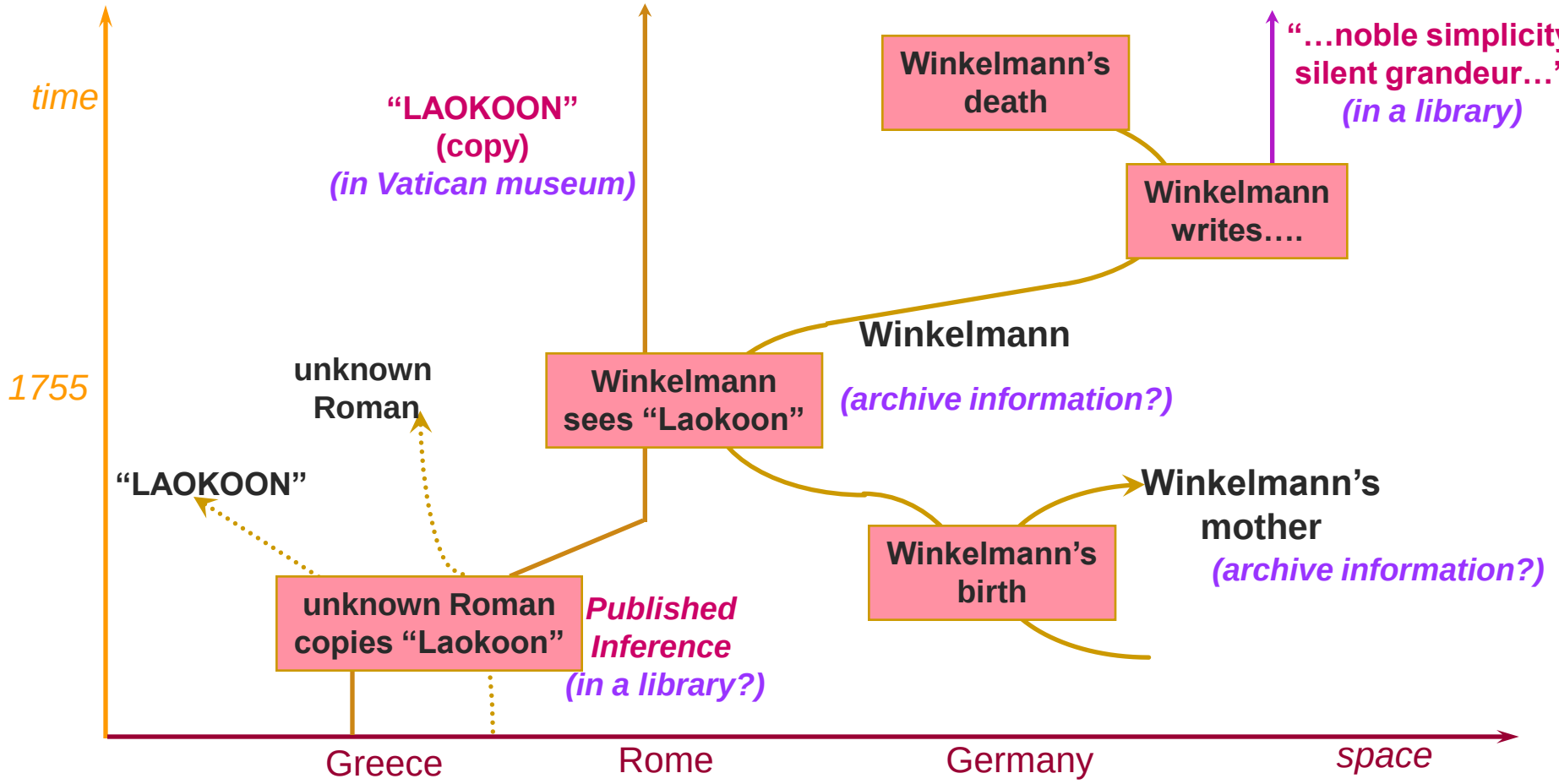
Second problem to solve: What are the shared concepts?

The simple idea of the CIDOC CRM:

- Use existing schemata in daily use as evidence for underlying concepts and as discourse restriction!!
- Describe only facts, that are relevant across documents.
- Deal with terminology as data!
- Explaining data needs much less words than a questionnaire.
- The data in the collection layer are already checked for consistency, plausibility and completeness.

The CIDOC CRM:

Networks of Historical of Knowledge



The CIDOC CRM:

The CIDOC CRM development

Break Through: We found the generic layer of concepts 1996-2000:

- Material facts: history seen as meetings of people, things and information.

- Material facts are the most objective information we have.
- Material facts are constraining the discourse about possible pasts.
- Material facts are carriers, evidence and constraints of the more vague and subjective social / intellectual processes and their interpretations.

- We do not model psychological/ spiritual processes and social / ethical interpretations!

The CIDOC CRM:

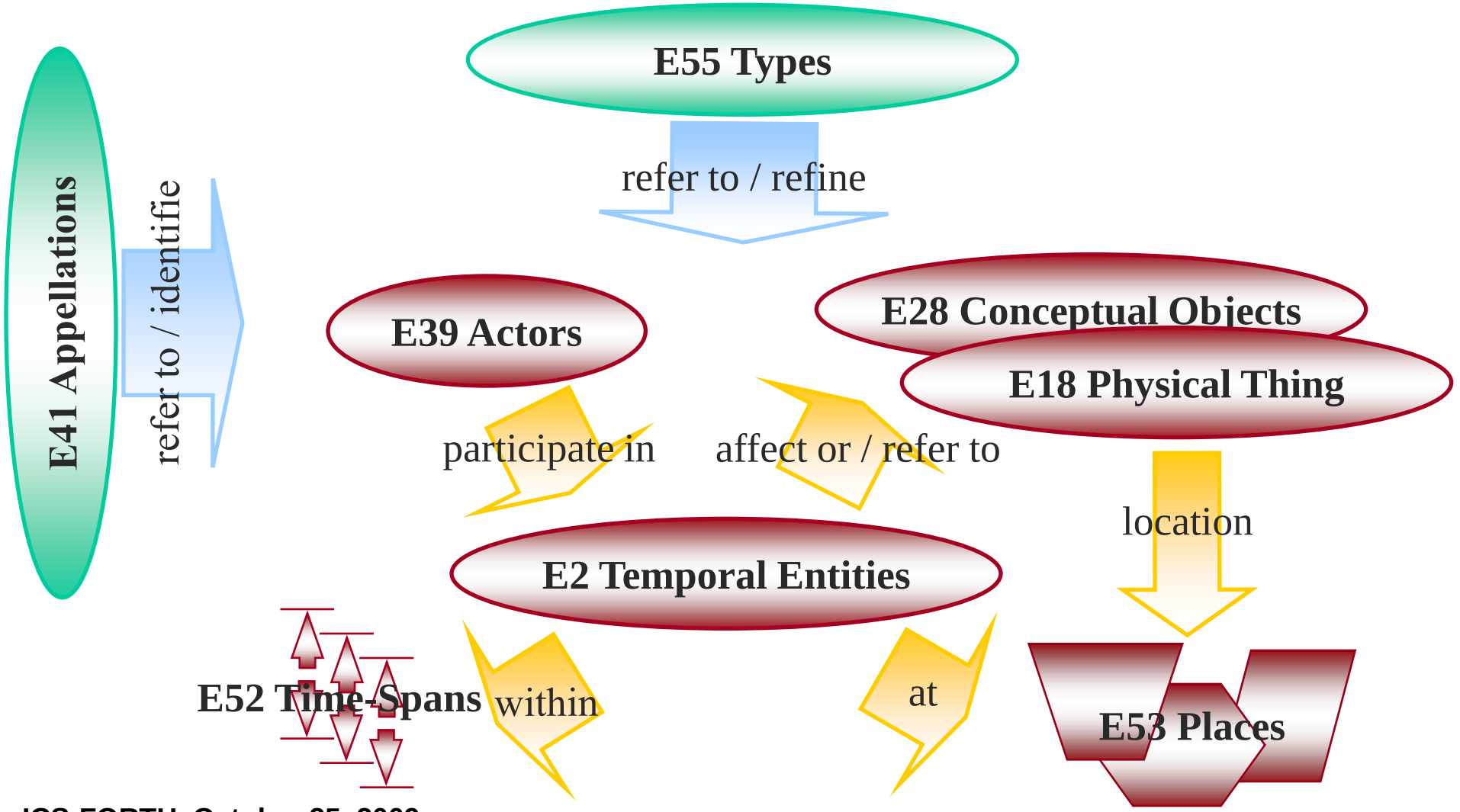
The CIDOC CRM development

The CIDOC CRM development

- A collaboration with the International Council of Museums
- An ontology of 86 classes and 137 properties for culture and more
- With the capacity to explain hundreds of (meta)data formats
- Accepted by ISO TC46 in September 2000
- International standard since 2006, ISO 21127:2006

The CIDOC CRM:

Top-level Entities relevant for Integration



The CIDOC CRM:

The CIDOC CRM development

Patience: We have been believing in the future!

- Work started 1996
- The model exists since 2003 as ISO submission.
- Take-up needs time: Standard in 2006, compatible systems are since then emerging
- Only recently we have free enabling technology: "Web Technology"
  (Even though demonstrated in 1990's!)
  - Native RDF technology ("quad stores"),
  - Linked Data Model,
  - URI conventions,
  - XML tools

The CIDOC CRM:

The CIDOC CRM development

Validation of genericity:

We could specialize the CRM to

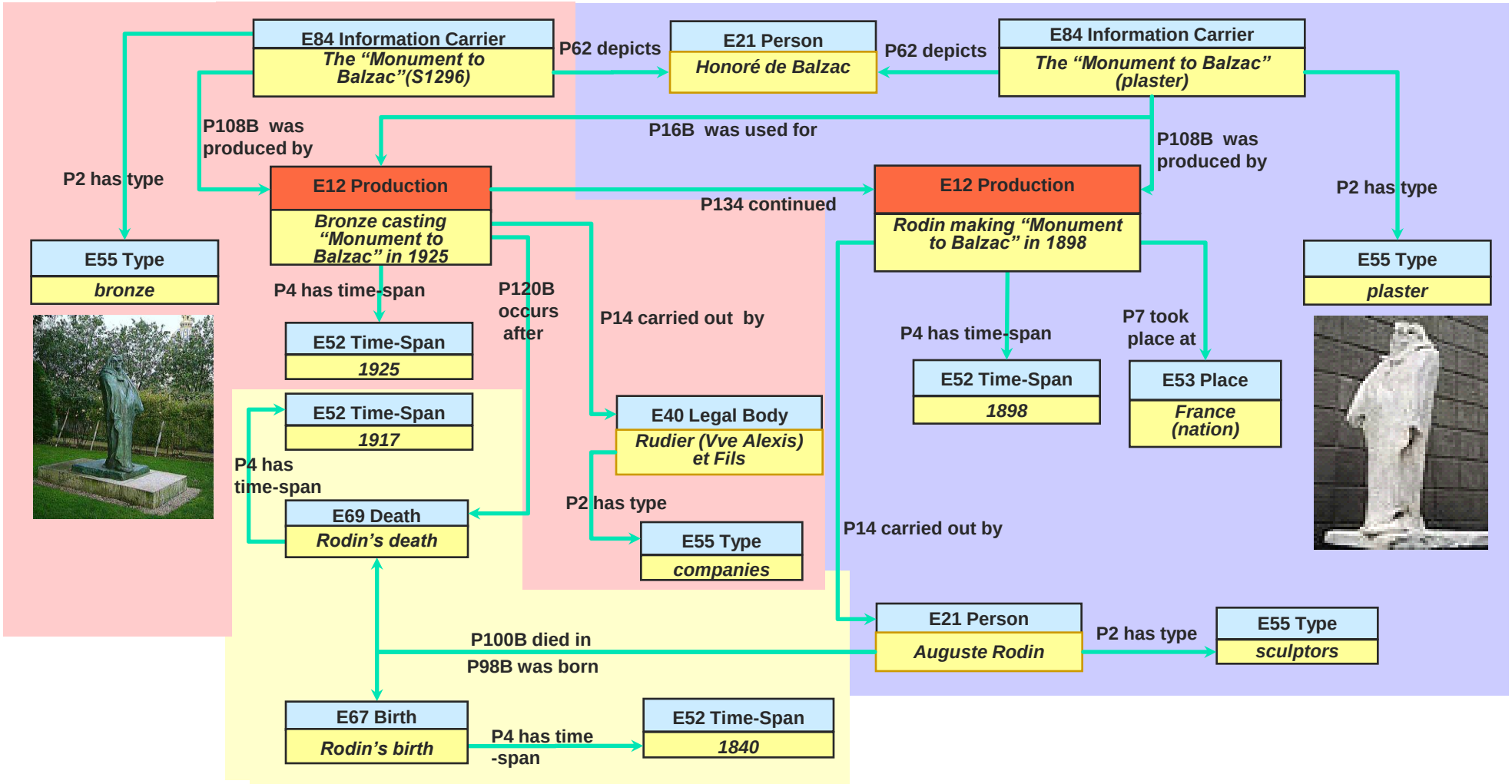
- Digital Provenance, e-science
- integration of empirical medical records (clinical reports and genomics data)
- library conceptualization: "FRBRoo"
- Digital Rights description
- ….

In validated semantics we are ahead of many others,

we are ready for the technology!

The CIDOC CRM:

Merging metadata contents with CRM



The CIDOC CRM:

What can we do now?

Make better collection management systems

- use less fields to say more, clearer semantics,
- more powerful queries (genericity and context)
- easy extension

Let others see your data

- Design compatible metadata and data export formats
- Implement data export from your system
- Open systems for querying by CRM concepts
- Make upper levels of your vocabularies compatible with CRM
  (use SKOS, but not for people!)

Access and use other data

- Exchange data in CRM compatible form
- Link, import, clean and merge data

Start trying out tools -> push technology demand
Support building/using CRM compatible "content aggregation" systems