



Content Enrichment

Jef Malliet

PCCE, Provincie Limburg



Content

1. Enrichment
2. Technology
3. Knowledge base
4. Thesauri
5. Ontology



Enrichment

- We have ...
 - Data from various sources
 - Source data made within specific contexts
- We want to ... Enrich
 - Relate the described objects to other objects across the borders of their original context
 - Reconstruct a new broader context
 - Enhance available data in the process
- We need to ...
 - Make data understandable outside their context
 - Refer to a larger model of the world
 - Make accurate connections

Enrichment

Semantic Web:

*I have a **dream** for the Web [in which computers] become capable of analyzing all the data on the Web – the content, links, and transactions between people and computers.*

*A ‘Semantic Web’, which should make this possible, **has yet to emerge**, but when it does, the day-to-day mechanisms of trade, bureaucracy and our daily lives will be handled by machines talking to machines.*

The ‘intelligent agents’ people have touted for ages will finally materialize.

[Tim Berners-Lee, 1999]



Enrichment

*The **Semantic Web** is about two things.*

- *It is about common formats for integration and combination of data drawn from diverse sources, where the original Web mainly concentrated on the interchange of documents.*
- *It is also about language for recording how the data relates to real world objects. That allows a person, or a machine, to start off in one database, and then move through an unending set of databases which are connected not by wires but by being about the same thing.*

[www.w3.org/2001/sw]

Technology

Resources

- Anything that can be described and identified

➤ RDF

- ‘Resource Description Framework’
- Triples: Subject – Predicate – Object
- Properties of resources

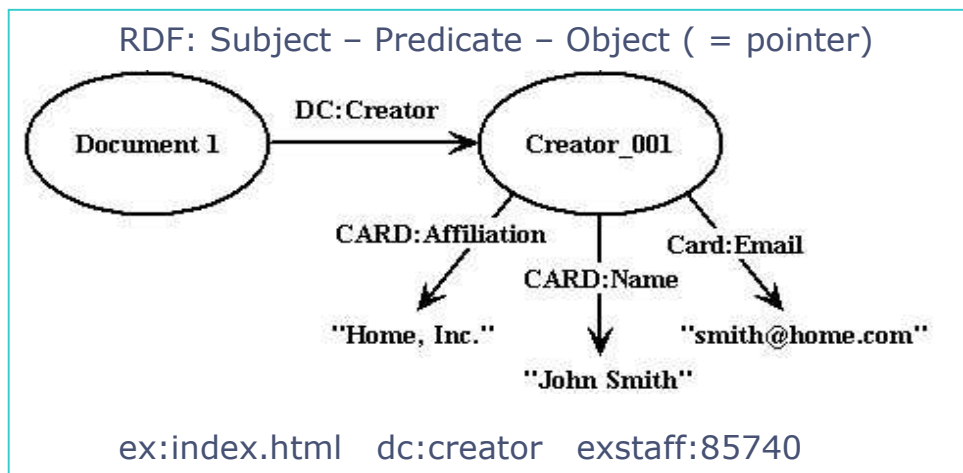
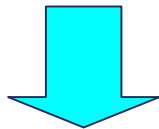
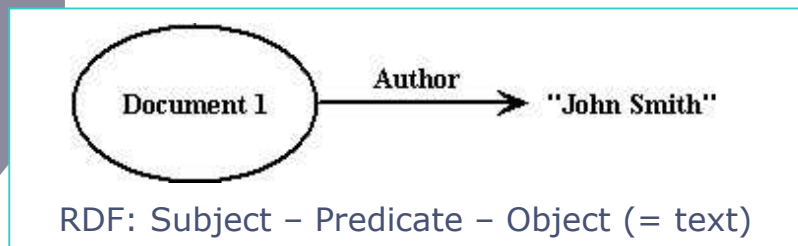


➤ URI

- ‘Uniform Resource Identifier’
- Combine with RDF
- Properties = relationships

Technology

RDF or RDF?



- Object = text
 - Inventory / Catalogue / Database
 - Careful:
 - What with homonyms?
 - What with different versions of a name?
 - "John Smith" is not the author, but the name of the author
 - Object has no meaning as such
 - Inaccuracy

- Object = pointer
 - Relation between 2 resources
 - O-O, Object-oriented
 - Ontology
 - Truly semantic
- Predicate = pointer
 - Type of relationship is part of a common system

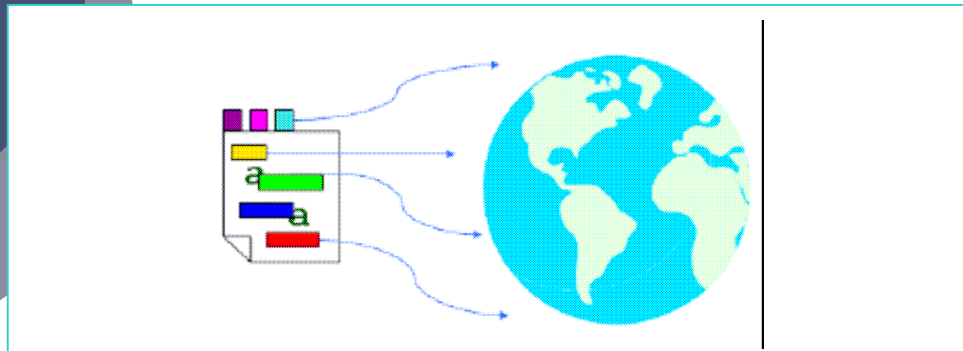


Technology

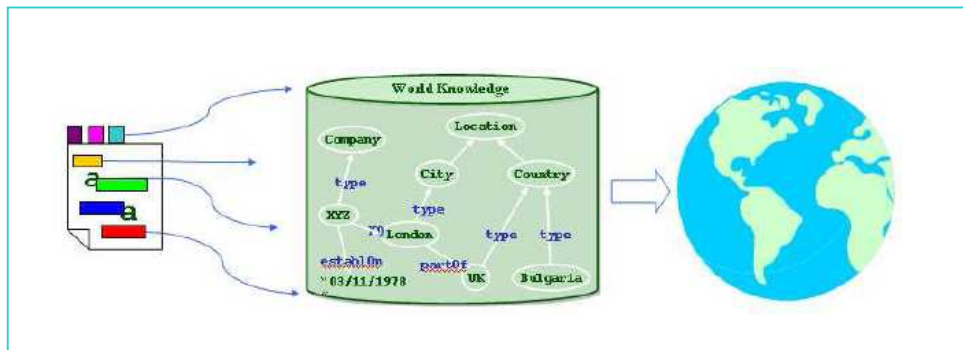
$\text{Data}(A+B) > \text{Data}(A) + \text{Data}(B)$

- New - Derived information, e.g.:
 - Source A: Object 1 was made in Place x
 - Source B: Place x is within Place y
 - Derived: Object 1 was made in Place y
- Accuracy required
 - Statements derived from unreliable premises become even less reliable
 - Resources = a real objects
 - Not just data, but objects (O-O, Object-oriented)
 - Ontology, for accurate datamodelling
 - Truly semantic

Knowledge base



- Relate to real-world objects?
- Only through a model of the real world: a 'Knowledge base'
- Precise models can be made for specific domains
- More general models are required to link several specific models



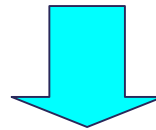


Knowledge base

- Required:
 - Identifiers, for pointers
 - Order – classification – taxonomy >> **thesauri**
 - Hierarchies: from more general to more specific
 - Accurate definition of concepts for disambiguation
 - Many synonyms, as commonly used, for searching by external users
 - Agreement on possible relations >> **ontology**

Thesauri

- ISO 2788 & ISO 5964
 - Taxonomy of **TERMS**
 - BT / NT / RT
 - UF / USE: point to correct term
 - ScopeNote: defines the scope of the term



- SKOS
 - Taxonomy of **CONCEPTS**
 - skos:broader / skos:narrower / skos:related
 - skos:prefLabel / skos:altLabel: possible names for the concept
 - skos:definition: defines the meaning of the concept
 - skos:scopeNote: additional information about limitations of use of the concept (e.g. in indexing environment)



Thesauri: AAT

Getty Art & Architecture Thesaurus

- Work started in 1970s, first published in 1990
- Scope: 'Terminology needed to catalog and retrieve information about the visual arts and architecture'
- Covers the broad domain of cultural heritage
- Compliant with ISO standards
- Often used as example when thesaurus technology is discussed

- Strictly hierarchical database (polyhierarchy)
- Structured in facets
- Focus is concepts, identified by ID number
- Framework is not context-specific
- Translations in Dutch, Spanish, Chinese, French, (German)...
- Detailed Editorial Guidelines



Thesauri

Maintenance

- Thesaurus is in constant evolution: expansion, correction
 - Evolution in heritage field
 - Advance in knowledge
 - Evolution in language
- Experts are in the heritage institutions: collection managers and scientific staff
 - Their contribution must be respected and valued
 - Quick response required



Thesauri

Maintenance = Dynamic process

- Heritage sector needs common ground
- Heritage sector wants to contribute
- Coordination required
- Authority required
- Web 2.0 can help but is not enough



Ontology

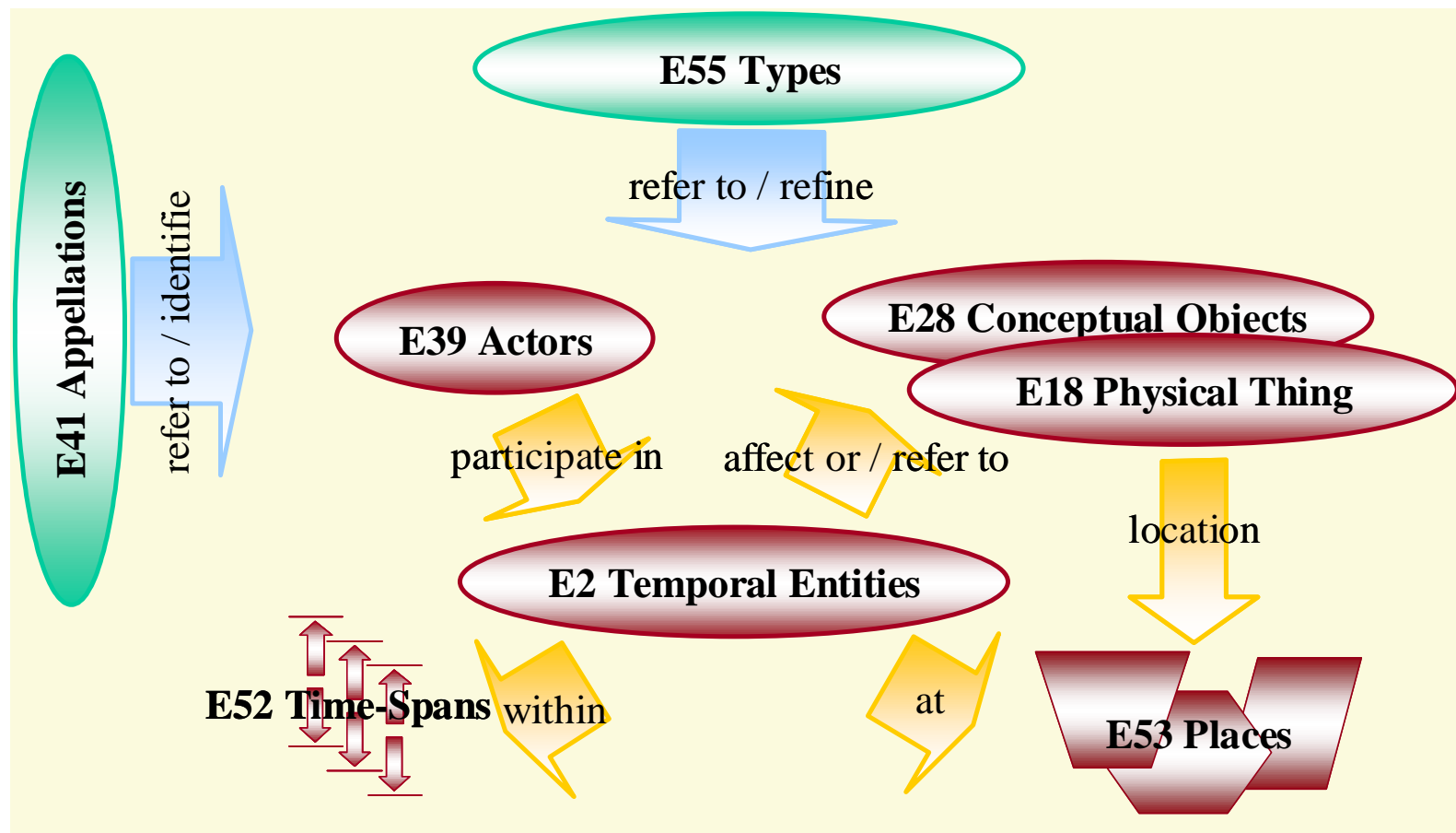
- **Level of classes**
 - **Classes of entities**
 - Object-oriented approach
 - Hierarchy of classes: super- & subclasses
 - Subclasses inherit properties from superclasses
 - **Properties**
 - Possible relationships between entities of certain classes
 - Hierarchy of properties
- **Level of instances**
 - Real entities belong to a class
 - Can have relationships with other entities as described in the properties catalogue

Ontology: CIDOC-CRM

- ISO 21127
 - Conceptual Reference Model
 - ‘A formal ontology intended to facilitate the integration, mediation and interchange of heterogeneous cultural heritage information’
 - Formal ontology defining classes and properties
 - Supra-institutional, abstracted from any specific local context
 - ‘Provide an optimal analysis of the intellectual structure of cultural documentation in logical terms’

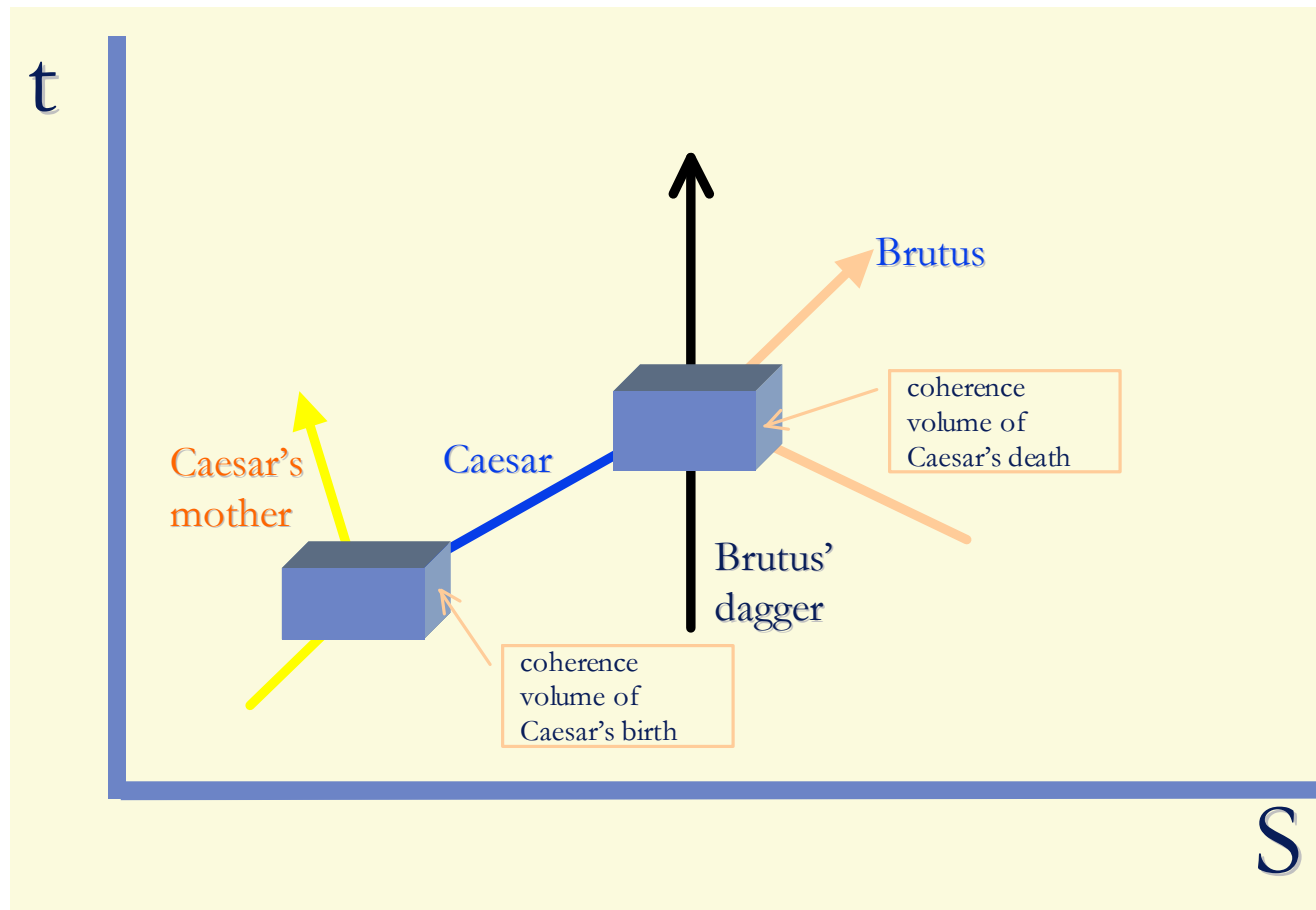
Ontology: CIDOC-CRM

Who? What? Where? When?



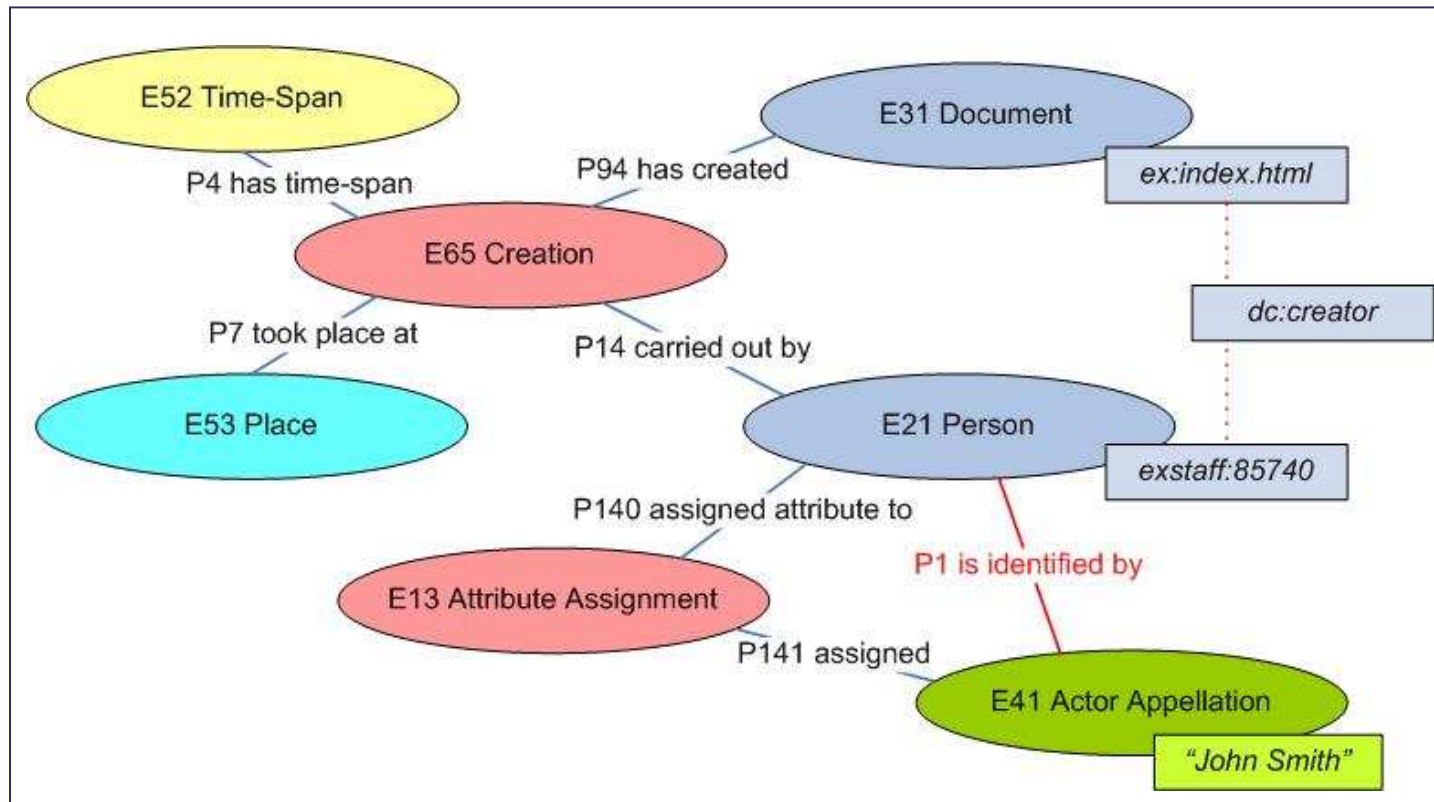
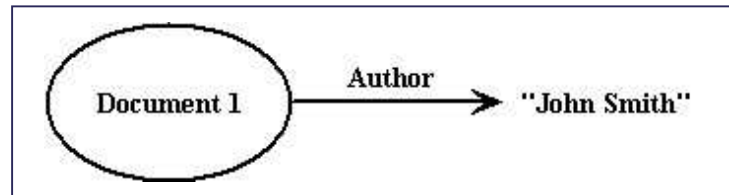
Ontology: CIDOC-CRM

Description = 'Biography' = sequence of events'



Ontology: CIDOC-CRM

Example:





Ontology

- Application:
 - For specialists
 - Not directly applicable to data input or navigation
 - Different from the way people are used to see data
 - To compare datastructures
 - To create common datastructure
 - To check consistency
 - To cross context borders



Conclusion

- In the cultural heritage sector we have excellent semantic tools
- They require
 - Ongoing maintenance of Knowledge base
 - Application requires better understanding by collection managers
 - A lot of work from humans
- Not a conclusion, but a beginning !!!



Thank you

jmalliet@limburg.be