Università
Ca' Foscari
Venezia

# Computer Science Applications to Cultural Heritage

## Metadata

Filippo Bergamasco (filippo.bergamasco@unive.it)
http://www.dais.unive.it/~bergamasco
DAIS, Ca' Foscari University of Venice
Academic year 2018/2019

Università
Ca' Foscari
Venezia

# Metadata in cultural heritage

World Wide Web radically eased the way in which information can be distributed and presented

Since its creation, there has been an increasing interest to make available the deluge of information kept in museums, archives and libraries.

Initial focus was on its **presentation**: creation of **websites**, apps and graphical interfaces **to local databases** of the so called “memory organizations”

What about global search?

Università Ca' Foscari Venezia

# Metadata in cultural heritage

New challenges beyond data presentation:

- Search for resources/concepts
- Comparative studies
- Data transfer
- Data migration

**Problem?**

Heterogeneous sources of cultural contents

Università Ca' Foscari Venezia

# Metadata in cultural heritage

Data can be stored and filed efficiently in different kind of databases… but is not enough!

Databases are focused on the **logic** level of data management not on the **semantic**.

Describing the semantic of data, in a way in which **both computers and humans can understand**, allows:

- Data exchange between different domains
- Data integration between different archives
- Long term data preservation

Università
Ca' Foscari
Venezia

# The importance of context

*"I saw a man on a hill with a telescope"*

How many alternate meanings can you spot?

Università Ca' Foscari Venezia

# The importance of context

*"I saw a man on a hill with a telescope"*

1. There is a man on a hill, and I'm watching him with a telescope

Università Ca' Foscari Venezia

# The importance of context

*"I saw a man on a hill with a telescope"*

1. There is a man on a hill, and I'm watching him with a telescope

2. There is a man on a hill, who I'm seeing, and he has a telescope

Università Ca' Foscari Venezia

# The importance of context

*"I saw a man on a hill with a telescope"*

1. There is a man on a hill, and I'm watching him with a telescope

2. There is a man on a hill, who I'm seeing, and he has a telescope

3. There's a man, and he's on a hill that also has a telescope on it

Università Ca' Foscari Venezia

# The importance of context

*"I saw a man on a hill with a telescope"*

1. There is a man on a hill, and I'm watching him with a telescope
2. There is a man on a hill, who I'm seeing, and he has a telescope
3. There's a man, and he's on a hill that also has a telescope on it
4. I'm on a hill, I saw a man using a telescope

Università Ca' Foscari Venezia

# The importance of context

*"I saw a man on a hill with a telescope"*

1. There is a man on a hill, and I'm watching him with a telescope

2. There is a man on a hill, who I'm seeing, and he has a telescope

3. There's a man, and he's on a hill that also has a telescope on it

4. I'm on a hill, I saw a man using a telescope

5. I'm on a hill with a telescope, and I saw a man

# Ontologies

The meaning we associate to a sentence is based on a shared **ontology**:

"... a formal naming and definition of the types, properties, and interrelationships of the entities that really exist in a particular domain of discourse."

https://en.wikipedia.org/wiki/Ontology_(information_science)

Knowledge provided by an ontology allow us to reason and understand the concepts

What about computers?

# Ontologies

There is a strong need to have shared ontologies between humans and machines

How?

- By describing data (in a formal way) to the machine -> **metadata**
- By creating algorithms and tools to infer new information from previous knowledge -> **KDD**

Why?

To have better answers to our questions

# Metadata

Usually the term is described as "data about data". To simplify, metadata are used to specify the *who*, *what*, *when*, and *where* about the items you want to describe.

Metadata can be general or application-specific:

**General**: used for any kind of information (examples: describing books, photos, pictures,etc)

**Application specific**: used to specify important attributes in a specific context (examples: the sender of a letter, the style of a music track)

Università Ca' Foscari Venezia

# Metadata: formal definition

*"Metadata consists of statements we make about resources to help us find, identify, use, manage, evaluate and preserve them"*

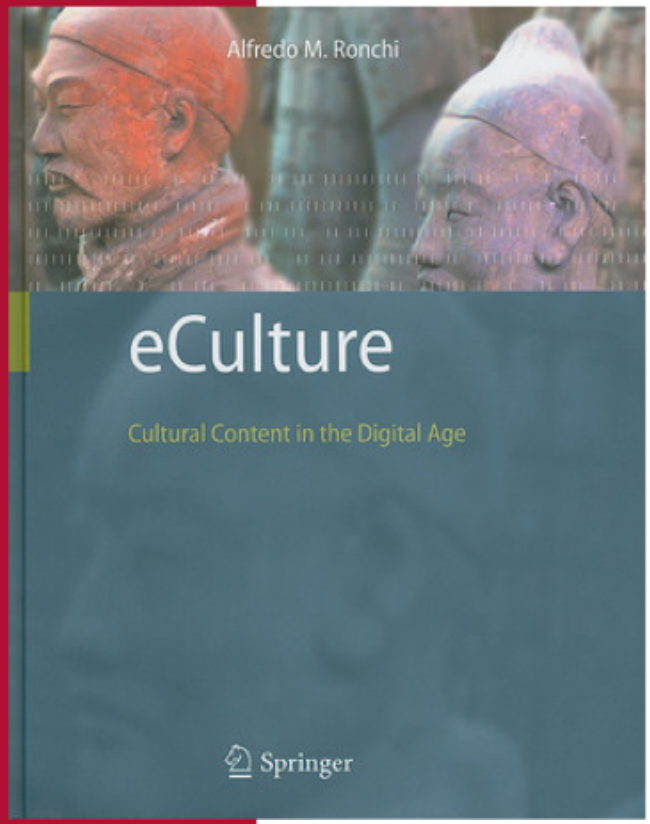
Marty Kurth, lecture on "Basic DC Semantics"

Types of metadata:

- **Descriptive:** give information about the content and the context
- **Structural:** give information about the data structure (types, multiplicity, relationships, etc)
- **Administrative:** give information about the rights managements, formats, etc.

Università
Ca' Foscari
Venezia

# Descriptive metadata




## Bibliographic Information

**Book Title**
eCulture

**Book Subtitle**
Cultural Content in the Digital Age

**Authors**
Alfredo M. Ronchi

**Copyright**
2009

**Publisher**
Springer-Verlag Berlin Heidelberg

**Copyright Holder**
Springer-Verlag Berlin Heidelberg

**Distribution Rights**
Distribution rights In India: Aditya Books (P) Ltd., New Delhi, India

**eBook ISBN**
978-3-540-75276-9

**DOI**
10.1007/978-3-540-75276-9

**Hardcover ISBN**
978-3-540-75273-8

**Softcover ISBN**
978-3-642-09455-2

**Edition Number**
1

Università Ca' Foscari Venezia

# Descriptive metadata

**Title:** Mona Lisa

**Creator:** Leonardo Da Vinci

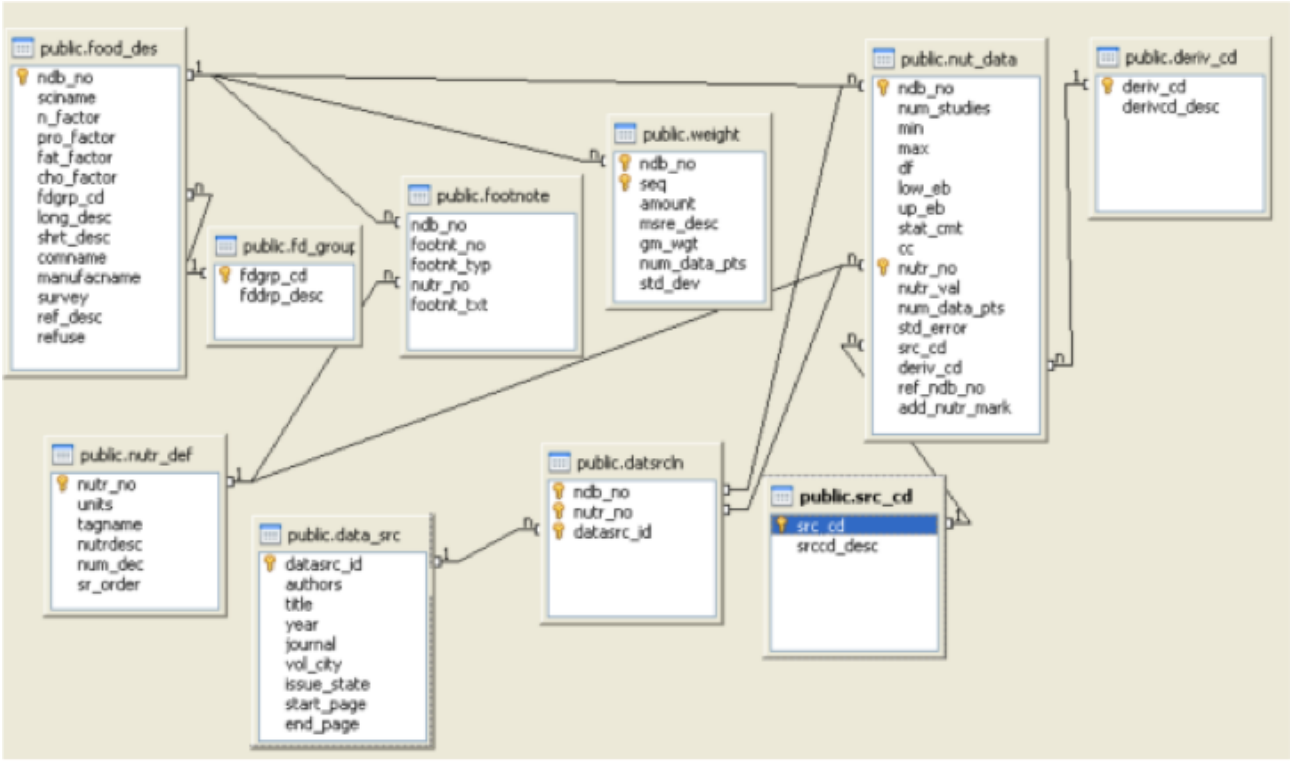
**Subject:** Portrait of Lisa Gherardini, wife of Francesco del Giocondo

**Keywords:** La Gioconda, Leonardo, Italy, Louvre, Renaissance

**Description:** The monumentality of the composition, the subtle modelling of forms and the atmospheric illusionism were novel qualities that have contributed to the continuing fascination and study of the work

Università
Ca' Foscari
Venezia

# Structural metadata




Relationships between tables in a relational database

Università Ca' Foscari Venezia

# Administrative metadata

Filename: report.docx

Checksum: 73f48840b60ab6da68b03acd322445ee

Copyright: 2012

File format: Microsoft Word 2011

Università Ca' Foscari Venezia

# Good metadata

To be useful, metadata should:

- Use controlled vocabularies and avoid ambiguous words
- Give appropriate description
- Provide persistent identifiers uniquely locate a resource
- Give **simple**, **complete** and **consistent** information
- **Use standards**:
  - Enable data interoperability
  - Provide a structured way to describe the data
  - Allows computer interpretation of the metadata

Università Ca' Foscari Venezia

# Metadata & CH

Cultural heritage information poses additional challenges in the definition of meaningful metadata because of **high diversity** and intrinsic **incompleteness** of information about the past

- At the **encoding level**, markup languages born for the WWW are commonly used as designed to be both human-readable and machine-readable
- At **semantics level**, standard terminology systems and ontologies have been created to support concepts of cultural heritage and museum documentation

# XML

One of the most common languages to encode document metadata in a format suitable to be read and modified by both humans and computers is the e**X**tensible **M**arkup **L**anguage.

Despite the name, it is not a language by itself but a specification to create custom markup languages

What is a markup language?

A system to put **annotations** on a document in a way that are distinguishable from the text but carry additional information (like the way in which the text has to be presented)

# XML

Since we are talking about markup languages, every XML document is by definition **text-based**.

Advantages:

- Easily interoperable between different computer systems
- Can be used without a dedicated software
- Can be easily shared, or embedded in other text-based documents (mail, messages, etc.)

Additionally, XML is vendor neutral standard and is designed to organize text data in a **structured way**.

# XML

A markup language must specify:

1. What markup is allowed (ie what are the annotations that can be used in a document)
2. What markup is required
3. How the markup (annotations) can be distinguished from the text (ie. the rest of the data)
4. What is the meaning of the markup

XML specifies 1,2 and 3. Point 4 is specified by another language, usually associated with XML, called DTD.

# XML

XML defines the set of rules (specifications) by which it is possible to create a custom markup language, depending on the application.

XML is a set of syntactic rules and standards to model the structure of documents and data.

XML is not focused on the presentation of the data. The focus is "what data is" and not "how data looks"

# XML: prolog

Every XML document should start with a prolog that gives information about:

- the XML version
- the text encoding
- details if all the entity definitions can be found inside the document itself

```
<?xml version="1.0" encoding="UTF-8"
standalone="yes"?>
```

# XML: tags

XML is composed by **tags**, used to specify a name for a given piece of information

A tag is made by two angular brackets enclosing the name of the tag

Every tag must start with a letter or the underscore, and can only contain [ A-Z, a-z, 1-9, -, _, .]

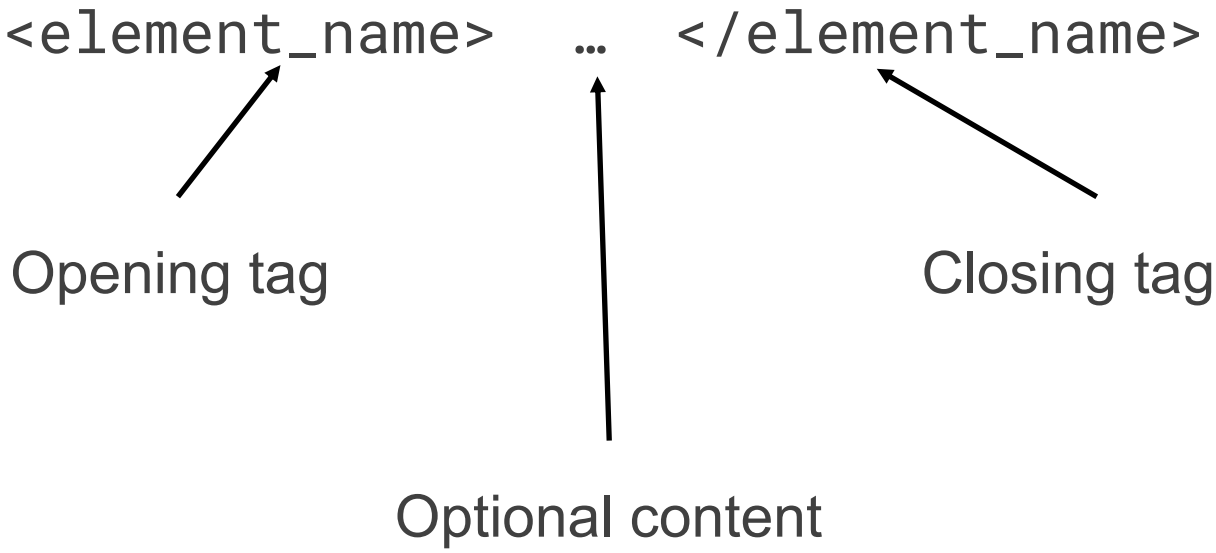
Examples:

```
<author>  <Name>  <_artefact2>
```

# XML: elements

The basic building block to structure data in XML format are the **elements**.

An element is represented using tags:



# XML: elements

An element can contain simple text, other elements or a mix of elements and text

XML elements have relationships! Since one element can contain some others, they are related hierarchically as parents and children

Example:

```
<student>
   <name>Mario Rossi</name>
   <age>28</age>
</student>
```

# XML: attributes

An attribute can be associated to any XML element to provide additional information which usually is not part of the data.

Example:

```
<student>
   <name>Mario Rossi</name>
   <age>28</age>
   <phone type="mobile">3481232912</phone>
</student>
```

# XML: comments

Comments are part of XML file that are ignored by the computer but are useful to explain the meaning of data

Example:

```
<student>
      <!-- Student properties -->
   <name>Mario Rossi</name>
   <age>28</age>
   <phone type="mobile">3481232912</phone>
</student>
```

# XML: entities

Some characters (like <>”) are not allowed as the text content of an XML file. To overcome this, special objects called entities can be defined and used.

Built-in entities:

| | |
|---|---|
| < | < |
| > | > |
| & | & |
| ‘ | ' |
| “ | " |

```
<speech>
Mario said &quot;hello&quot;
</speech>
```

# XML well-formed

An XML file is well formed if it satisfies the following rules:

1. It must contain only one high-level (root) element. All other elements must be child of the root element
2. Every element must have an opening tag with a corresponding closing tag. An empty element can use the abbreviated form `<elementname />`
3. All the attributes must be enclosed into quotes "..."
4. Elements must be nested properly. The ordering of the closing tags must be the inverse of the opening

   `<a><b><c> </c></b></a>`

   Note: XML is case-sensitive

# XML bad-formed examples

```
<title>
   <text>test
   </title>
</text>
```

Incorrect nesting

```
<news title=test>
  This is a news
</News>
```

Attribute not enclosed in “”

Not the same tag since XML is case sensitive

# XML: valid

In addition to being well-formed, an XML file may be valid or not.

It is possible to specify the grammatical rules of a specific class of XML files, ie. all the **elements** that can be used together with their expected **structure**.

If an XML file satisfies all the rules of a specific grammar than it is **valid**.

The two common ways to define such grammatical rules are called DTD and XML Schema

# DTD vs. XML Schema

- Use a simple syntax different than XML
- Can be embedded locally in an XML file
- Can be used to define new entities

- XML-based syntax
- Can define and use data types
- Can limit the occurrence of certain elements
- Can define the order of the elements
- Can use enumerations as attributes

Università Ca' Foscari Venezia

# Metadata standards for CH

We have seen that, at the encoding level, markup languages like XML can be used to define metadata that are both readable by computers and humans.

At the semantic level, we need ontologies and tools to explain the meaning of the data (ie. to define metadata) in an effective and interoperable way

**> We need metadata standards**

One of the most common standards in cultural heritage is the **The Dublin Core Metadata Element Set** proposed by the **Dublin Core Metadata Initiative** (**DCMI**)

Università Ca' Foscari Venezia

# DCMI

The The Dublin Core Metadata Initiative is an open organization engaged in the development of interoperable online metadata standards.

- Their work started in 1995 with an invitational workshop in Dublin, Ohio.
- Workshop brought together librarians, digital library researchers, content providers, and text markup experts in order to **improve discovery standards for information resources**.

DCMI started by developing a set of descriptors originally intended for bibliographic material but later extended to generic resources

Università Ca' Foscari Venezia

# DCMI Objectives

DCMI worked with the following objectives in mind:

1. Simplicity of creation and maintenance
    a. Simple descriptive records should be created either by non specialists
    b. Resources should be annotated easily and inexpensively
2. Universally understood semantics
    a. A common set of elements should be defined whose semantics is universally understood and supported
3. International scope
    a. Versions in many different languages should be available
4. Extensibility
    a. Descriptors should be designed to be further extended

# Dublin Core

The Dublin Core standard includes two levels: Simple and Qualified.

The Simple Dublin Core standard comprises The Dublin Core Metadata Element Set (DCMES), which emerged as a **vocabulary of fifteen properties** for use in resource description.

The Qualified standard extends DCMES with:

- Additional elements
- Element refinements
- Value encoding schemes

# DCMES

The 15 properties of the simple Dublin Core:

**Title**, **Creator**, **Subject**, **Description**, **Publisher**, **Contributor**, **Date**, **Type**, **Format**, **Identifier**, **Source**, **Language**, **Relation**, **Coverage**, and **Rights**

- Each element is optional and can be repeated
- The ordering of multiple occurrences of the same element may have a significance intended by the provider, but preservation of the ordering is not guaranteed in every system.

Università Ca' Foscari Venezia

# DCMES

*Element Name:* **Title**
*Label:* Title
*Definition:* A name given to the resource.
*Comment:* Typically, Title will be a name by which the resource is formally known.

*Element Name:* **Creator**
*Label:* Creator
*Definition:* An entity primarily responsible for making the content of the resource.
*Comment:* Examples of Creator include a person, an organisation, or a service. Typically, the name of a Creator should be used to indicate the entity.

Università Ca' Foscari Venezia

# DCMES

*Element Name:* **Subject**
*Label:* Subject and keywords
*Definition:* A topic of the content of the resource.
*Comment:* Subject is expressed as keywords, key phrases, or classification codes that describe a topic of the resource. Best practice is to select a value from a controlled vocabulary or formal classification scheme.

*Element Name:* **Description**
*Label:* Description
*Definition:* An account of the content of the resource.
*Comment:* Examples of Description include, but are not limited to, an abstract, table of contents, reference to a graphical representation of content, or free-text account of the content.

Università Ca' Foscari Venezia

# DCMES

*Element Name:* **Publisher**
*Label:* Publisher
*Definition:* An entity responsible for making the resource available.

*Comment:* Examples of Publisher include a person, an organisation, or a service. Typically, the name of a Publisher should be used to indicate the entity.

*Element Name:* **Contributor**
*Label:* Contributor
*Definition:* An entity responsible for making contributions to the content of the resource.
*Comment:* Examples of Contributor include a person, an organization, or a service. Typically, the name of a Contributor should be used to indicate the entity.

# DCMES

*Element Name:* **Date**
*Label:* Date
*Definition:* A date of an event
*Comment:* Date will be associated with the creation or availability of the resource. Recommended best practice for encoding the date includes (among others) dates of the form YYYYMMDD.

*Element Name:* **Type**
*Label:* Resource Type
*Definition:* The nature or genre of the content of the resource.
*Comment:* Type includes terms describing general categories, functions, genres, or aggregation levels for content.

Università Ca' Foscari Venezia

# DCMES

*Element Name:* **Format**
*Label:* Format
*Definition:* The physical or digital manifestation of the resource.
*Comment:* Typically, Format will include the media type or the dimensions of the resource. Format may be used to identify the software, hardware, or other equipment needed to display or operate the resource.

*Element Name:* **Identifier**
*Label:* Resource identifier
*Definition:* An unambiguous reference to the resource within a given context.
*Comment:* Recommended best practice is to identify the resource by means of a string or number conforming to a formal identification system (URI,DOI,ISBN,etc)

Università Ca' Foscari Venezia

# DCMES

*Element Name:* **Source**

*Label:* Source

*Definition:* A reference to a resource from which the present resource is derived.

*Comment:* The present resource may be derived from the Source resource in whole or in part.

*Element Name:* **Language**

*Label:* Language

*Definition:* A language of the intellectual content of the resource.

*Comment:* Recommended best practice is to use RFC 3066, which, defines primary language tags with optional subtags. Examples include “en” or “eng” for English, “akk” for Akkadian, and “en-GB” for English used in the United Kingdom.

Università Ca' Foscari Venezia

# DCMES

*Element Name:* **Relation**
*Label:* Relation
*Definition:* A reference to a related resource.
*Comment:* Recommended best practice is to identify the referenced resource by means of a string or number conforming to a formal identification system.

*Element Name:* **Coverage**
*Label:* Coverage
*Definition:* The extent or scope of the content of the resource.
*Comment:* Coverage will include spatial location (a place name or geographic coordinates), temporal period (a period label, date, or date range), or jurisdiction (such as a named administrative entity). Recommended best practice is to select a value from a controlled vocabulary

Università Ca' Foscari Venezia

# DCMES

*Element Name:* **Rights**

*Label:* Rights management

*Definition:* Information about rights held in and over the resource.

*Comment:* Rights will contain a rights management statement for the resource, or reference a service providing such information. Rights information often encompasses intellectual property rights (IPR), copyright, and various property rights.If the Rights element is absent, no assumptions may be made about any rights held in or over the resource.

Università Ca' Foscari Venezia

# Qualified Dublin Core

- Adds the following elements:

  Audience, Provenance, RightsHolder, InstructionalMethod, AccrualMethod, AccrualPeriodicity, AccrualPolicy

- Allows element refinements:

  Qualifiers to narrow the meaning of an element (for example to distinguish the creation date to the modification date)

- Value encoding schemes

  Schemes that aid the interpretation of an element value

Università Ca' Foscari Venezia

# Dublin Core Principles

**One-to-One principle:**

Metadata describe one single manifestation of a resource. For example, the jpeg image of the Mona Lisa is different than the original painting. The relationship between the metadata for the original and the reproduction is also part of the metadata

**Dumb-down principle:**

Qualification is only used to refine a property, not to extend the semantic scope

**Appropriate values principle:** just use the metadata that are useful for resource discovery

Università Ca' Foscari Venezia

# Dublin Core Implementation

Metadata defined with Dublin Core can be either:

- Embedded with the resource itself (for example metadata used in webpages)
- Stored in any kind of database and linked to the original resource

Dublin core is usually implemented using XML or XML/RDF

http://nsteffel.github.io/dublin_core_generator/generator_nq.html

Università
Ca' Foscari
Venezia

# From DC to ontologies

Dublin Core is a remarkable attempt to formalize the metadata structure to be applied to different kind of contents.

Its popularity derives mostly from its simplicity and breadth of the attempted scope

It is a good starting point but not properly a formal ontology...

# Back to ontologies..

An ontology provides a **vocabulary** to represent knowledge

The vocabulary allow us to specify:

- **Entities** of our knowledge domain
- How the entities can be grouped together (**Classes**)
- The **relationships** that relate entity together

Usually expressed using triplets:

(Venice, is_a, place), (coin, made_of, silver)

Subject-verb-object

Università Ca' Foscari Venezia

# CIDOC-CRM

**C**onceptual **R**eference **M**odel (CRM) made by CIDOC Documentation Standards Group in the International Committee for Documentation of the International Council of Museums

International ISO standard aims to be a an ontology for cultural heritage and museum documentation

Recent Version:

http://www.cidoc-crm.org/Version/version-6.2.2

# CIDOC-CRM

Composed of several classes and properties:

- A **class** is a category of items that share one or more common traits, serving as criteria to identify the items belonging to the class
- A **property** serves to define a relationship of a specific kind between two classes. A property plays a role analogous to a grammatical verb, in that it must be defined with reference to both its domain and range, which are analogous to the subject and object in grammar

# CIDOC-CRM

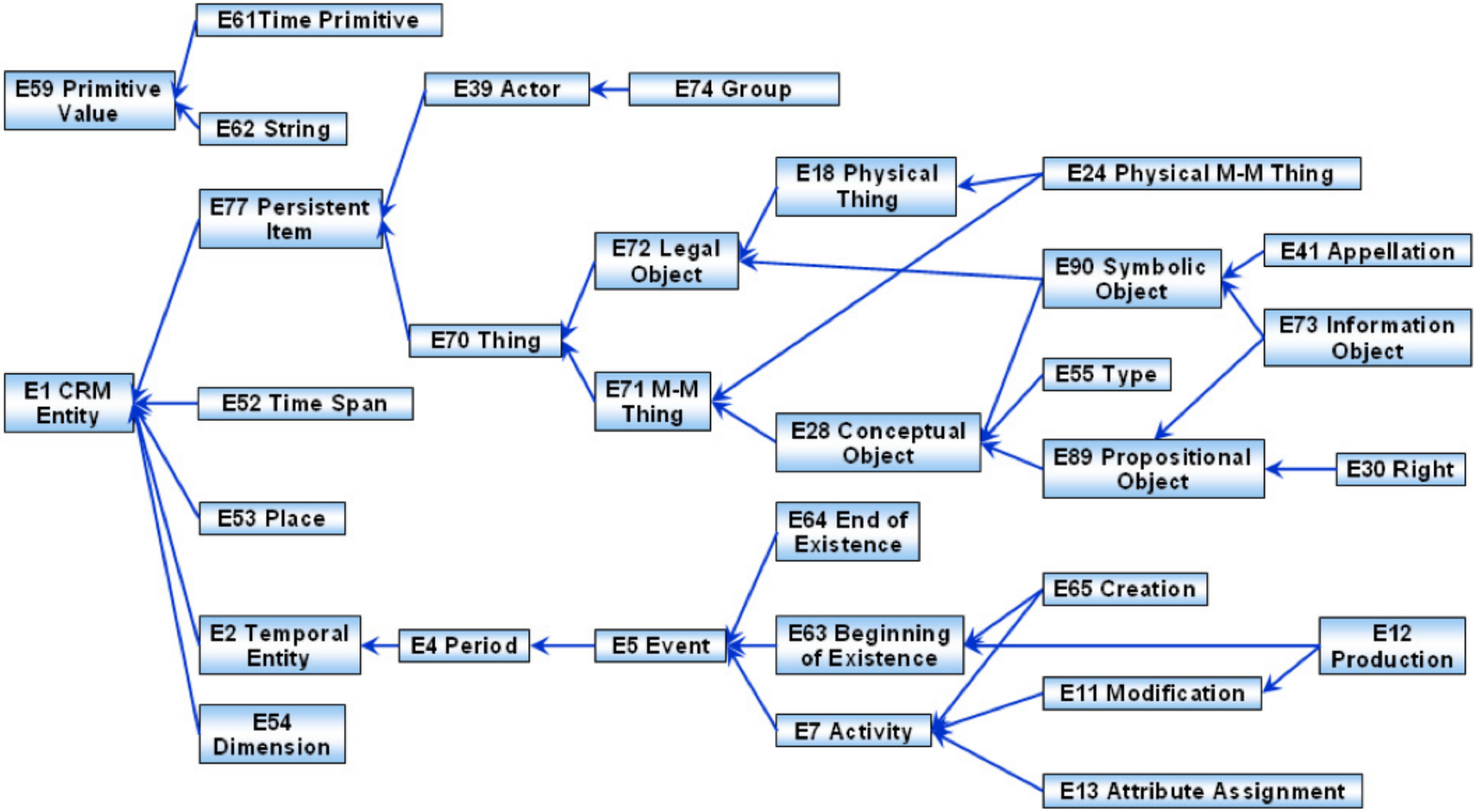
The "core" classes include:

- Space-time concepts: era/period, places, time-spans, relationships to persistent things
- Events: beginning/ending of existence, participants, creation/modification, etc
- Material things: place, information content carried by the material, relationships to persistent items
- Immaterial things: conceptional things

Classes are organized in a hierarchical manner in which each sub-class refines the concepts and properties of the super-class.

Università Ca' Foscari Venezia

# Hierarchy of core classes



# CIDOC-CRM: Events

In CIDOC-CRM, the modeling is mostly based on **events**

Basic idea of CIDOC-CRM:

**Model a change and not the state.**

Example:

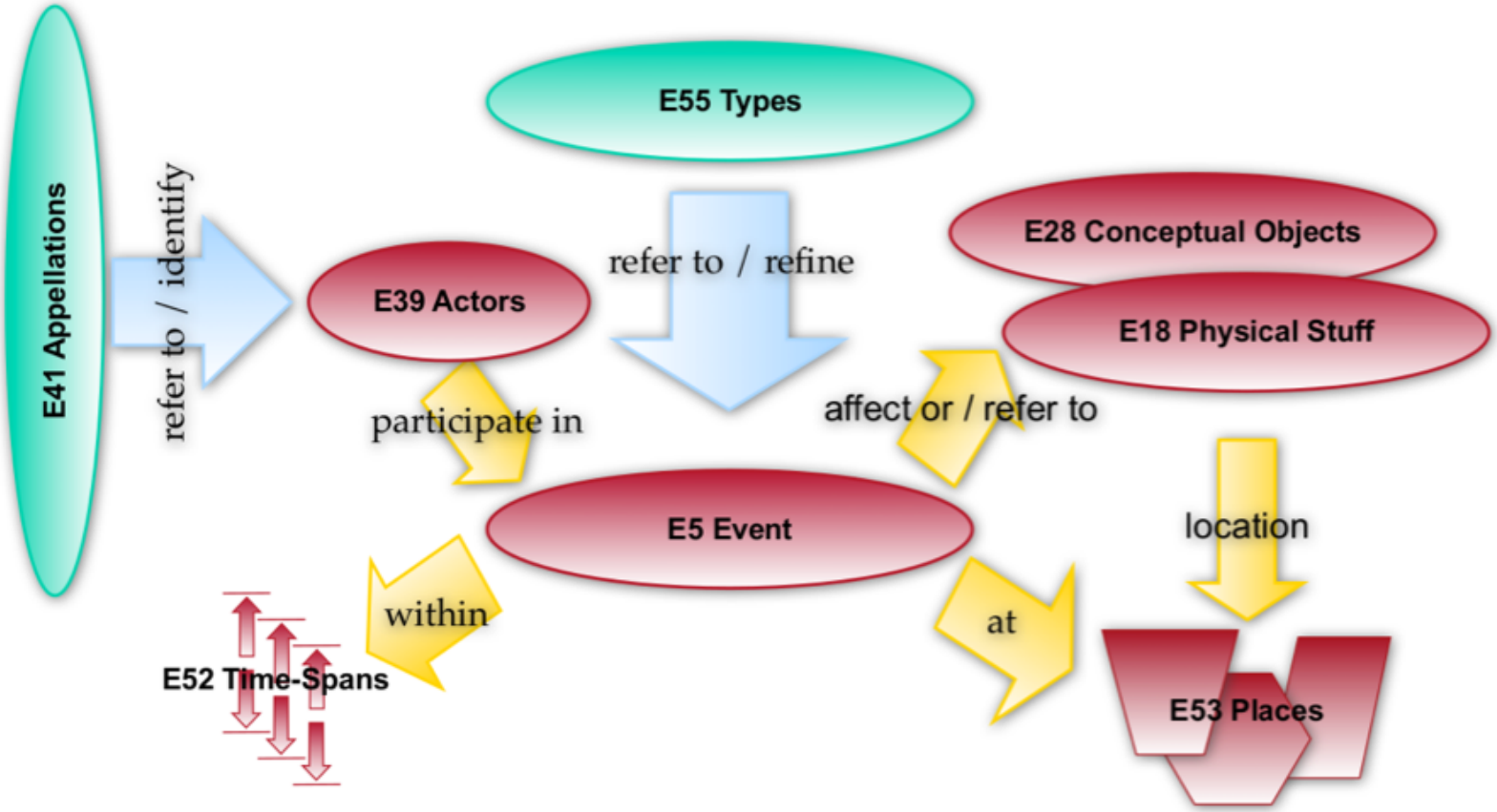
Instead of coding the birthdate associated to an author, code a new event called "birth" associated to that author

That's why in the CIDOC-CRM we have a lot of temporal entities (E67 Birth, E63 Beginning of Existence, E68 Dissolution, E69 Death, ...)

Università Ca' Foscari Venezia

# CIDOC-CRM: Events

Relations between "Actors" and "Objects" are defined only through temporal entities and events



# CIDOC-CRM: Events

Why instantiating events is a good idea? We can make **deductions** about time intervals.

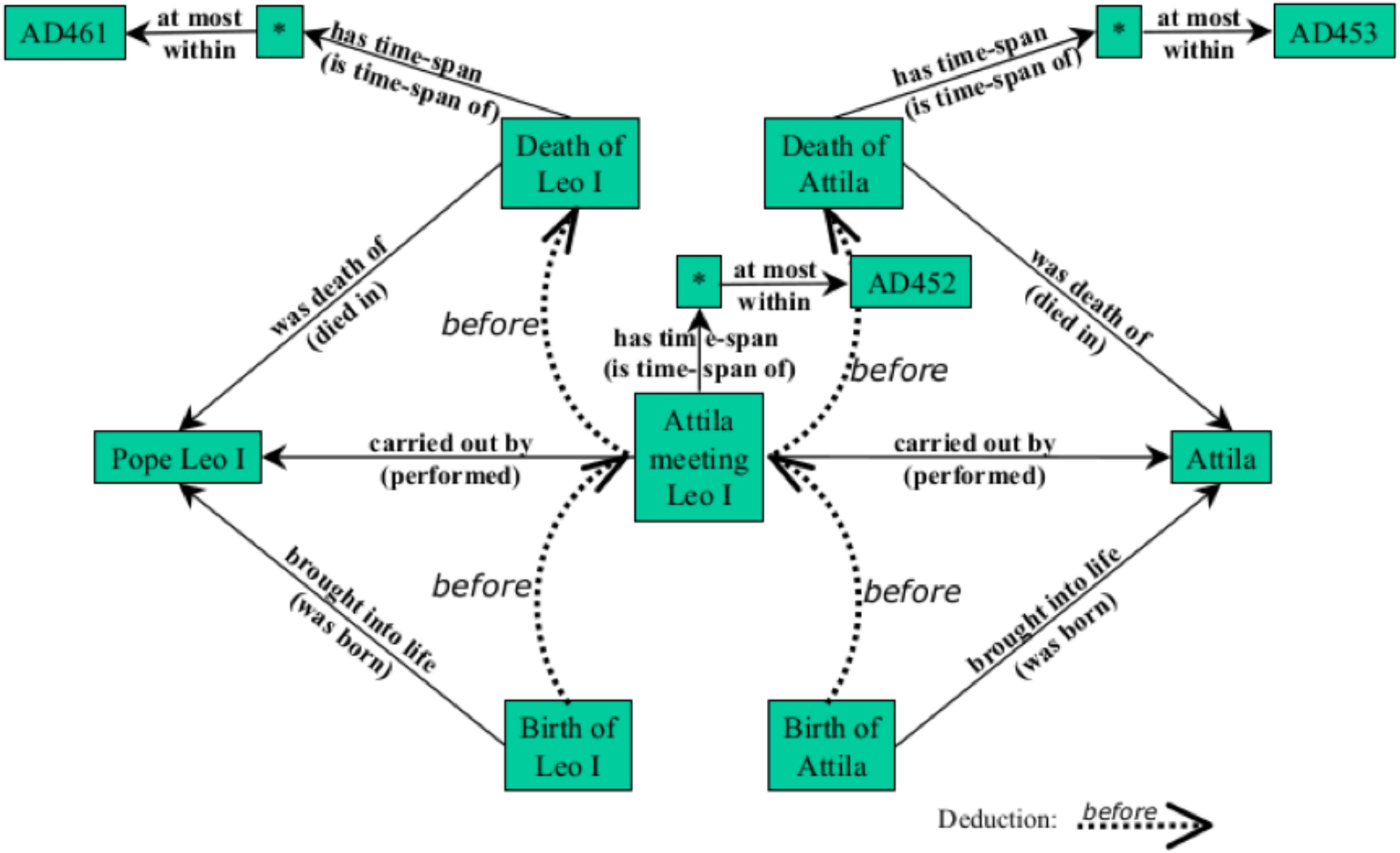
- The participation of several (non temporal) entities in an event E1 allows to conclude that they have been in the same time-interval and space even without knowing the particular time or space.

What else?

- They must have existed at that time
- They must have not been somewhere else at that time
- Creation of each entity must be happened before E1 and the destruction must happen after E1

Università Ca' Foscari Venezia

# CIDOC-CRM: Events



Università Ca' Foscari Venezia

# CIDOC-CRM: Important Properties

The property **P11** "Had participant" denote involvement of different actors to an event

The property **P12** "Occurred in presence of" just enumerates the object being present at the event

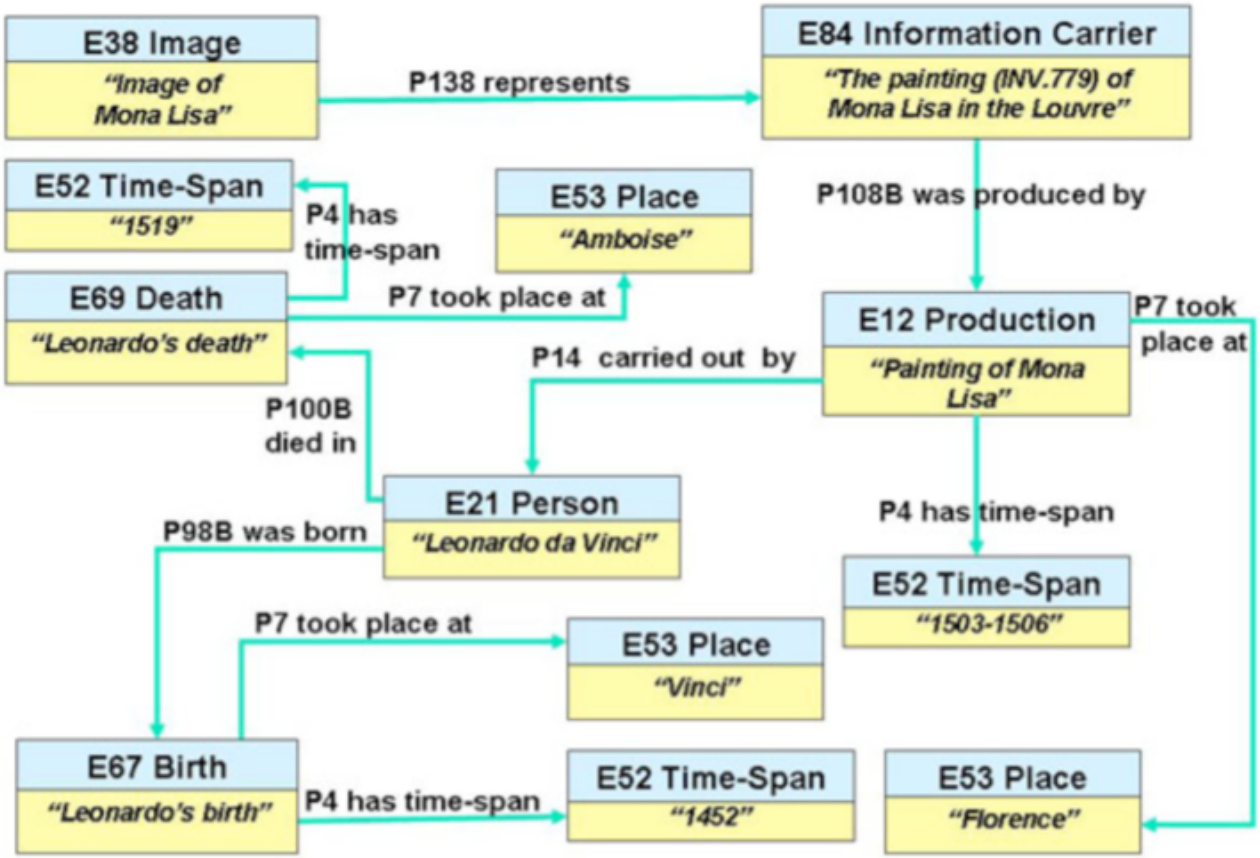
To limit the existence of entities:

The property **P92** "Brought into existence" allows a "start" to be attached to any Persistent Item being documented

The property **P93** "Took out of existence" allows an "end" to be attached to any Persistent Item being documented

Università Ca' Foscari Venezia

# CIDOC-CRM: example of digital content modeling



Università Ca' Foscari Venezia

# Adding semantic meaning to relational databases

Relational databases are a good way to store data in a structured way

The structure is formally defined in a database schema.

Example:

| Id | Artifact | Site |
|---|---|---|
| 0 | Coin | A1 |
| 1 | Vase | A2 |
| 2 | Coin | A2 |

| Id | Artifact | Room | Floor |
|---|---|---|---|
| 0 | Coin | 3 | 1 |
| 1 | Vase | 2 | 1 |
| 2 | Coin | 4 | 0 |

CIDOC-CRM allow us to create a conceptual system to describe the semantic our data

Università Ca' Foscari Venezia

# Adding semantic meaning to relational databases

CIDOC-CRM allow us to create a conceptual system to describe the semantic our data

Archeo DB

| Id | Artifact | Site |
|---|---|---|
| 0 | Coin | A1 |
| 1 | Vase | A2 |
| 2 | Coin | A2 |

Museum DB

| Id | Artifact | Room | Floor |
|---|---|---|---|
| 0 | Coin | 3 | 1 |
| 1 | Vase | 2 | 1 |
| 2 | Coin | 4 | 0 |

E22 Object

E53 Place

E22 Object

E53 Place

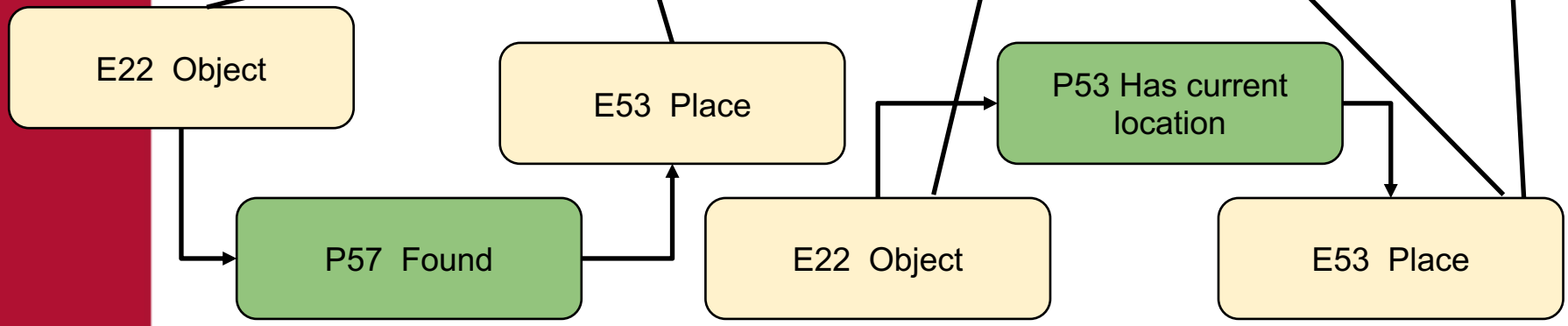
Università Ca' Foscari Venezia

# Adding semantic meaning to relational databases

… and map the implicit knowledge

| Id | Artifact | Site |
|---|---|---|
| 0 | Coin | A1 |
| 1 | Vase | A2 |
| 2 | Coin | A2 |

| Id | Artifact | Room | Floor |
|---|---|---|---|
| 0 | Coin | 3 | 1 |
| 1 | Vase | 2 | 1 |
| 2 | Coin | 4 | 0 |



Università
Ca' Foscari
Venezia

# CIDOC-CRM: encoding

At a coding level, CIDOC-CRM metadata are usually written in XML/RDF

RDF (Resource Description Framework) is an XML markup language used to generally describe resources.

Based on the idea of making subject-verb-object statements about resources

Università Ca' Foscari Venezia

# CIDOC-CRM in XML/RDF

Denario 103 is made of bronze and dates back to the second century AD.

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:coins="http://coins-project.eu/coins#"
xmlns:cidoc="http://cidoc.ics.forth.gr/rdfs/cidoc_v4.2.rdfs#"
>
<cidoc:E22.Man-Made_Object rdf:about="http://coins-project.eu/coins#Denario_103">
    <cidoc:P5.consists_of rdf:resource="http://coins-project.eu/coins#Bronzo"/>
    <cidoc:P4.has_time_span rdf:resource="II_d.C."/>
    </cidoc:E22.Man-Made_Object>
</rdf>
```