europeana
think culture

The Europeana Data Model, current status

Carlo Meghini

Europeana v1.0 WP3 Meeting

Berlin, January 25-26, 2010

Outline

- Part I
  - Background
  - Requirements
  - Status
- Part II
  - The general picture
  - Classes
  - Properties
  - Examples
  - Future work

Outline

- Part I
  - Background
  - Requirements
  - Status
- Part II
  - The general picture
  - Classes
  - Properties
  - Examples
  - Future work

Background

- Why
  - to define what information is necessary in order to enable the functionality of Europeana
- What
  - Classes, arranged in a taxonomy
  - Properties, arranged in a taxonomy
  - Constraints: domain/range, cardinality of properties
- Who
  - The Europeana experts
- When
  - July 2010, Danube specs

Outline

- Part I
  - Background
  - Requirements
  - Status
- Part II
  - The general picture
  - Classes
  - Properties
  - Examples
  - Future work

Requirements

- Data integration
- Support rich functionality (*e.g.*, semantic search)
- Optimize the use of resources in time

Requirements: Data integration

- Standard approach in a sound software development process:
  - Requirement
    - Collection
    - Specification
  - Design
    - Analysis of the functionality
    - Algorithms
    - Required data
  - Implementation
  - Testing
  - Validation

Requirements: Data integration

- Europeana is a data integration system
  - A living organism, consisting of
    - Central Repository
    - Local Sources
  - In continuous expansion:
    - More data coming from the local sources
    - More sources being added
    - More users
    - More functionality

Consistency, data scalability

Extensibility in data model

Workload scalability

Extensibility in function

Requirements: Data integration

- A data integration system is built by taking into account the data models of the sources
  - At requirement collection time: collect the model of each source
  - At design time:
    - How to integrate the existing data in order to achieve the required functionality
    - May lead to: revision of requirements or addition of extra functionality
- In the present case, the sources are:
  - Large and important = lots of data, users, expectations
  - In different domains = significantly different data models
  - Very many = lots of significantly different data models
  - An open set = **who knows what data may come tomorrow**

Requirements: Data integration

- Two possible venues for data modeling:
  - Cross-domain element set
    - a common set of properties capturing features shared by all objects, e.g. the Dublin Core Element Set
  - An ontology
    - a **complete** conceptualization, emphasizing the fundamental notions around Cultural Heritage Objects that allows Europeana to accommodate the data coming from providers **regardless of the original models**
- Cross-domain venue: Rhine, set up the basic infrastructure
  - Europeana Semantic Elements
- What about Danube?

Requirements: Support rich functionality

- Europeana must outdo the competition in the Cultural Heritage domain, notably web search engines
  - richness: collect **all the data there is**
  - intelligence: connect data to Knowledge Organization Systems
  - coverage: multilingualism
  - For Danube, we need to go the ontology venue in order to support rich functionality
  - richness: a special ontological entity to represent aggregates
  - intelligence: classes to represent knowledge and properties to connect knowledge to objects
  - coverage: multilingualism is core in Europeana (more on this later)

Requirements: Optimize resources

- Minimize and protect the investment required for accumulating knowledge:
  - Re-use existing models
    - ontology is a controversial area of philosophy
    - recently, the controversy has reached computer science
    - very recently, the controversy has reached Europeana too
  - Build on standards
    - Institutions are making their data and their Knowledge Organization Systems available in the Web, using URIs, RDF/S, SKOS, Linked Data, and more
    - Need to buy into the Web Architecture and standards
    - Europeana wants to follow institutions rather to push them

Requirements: wrap up

- In sum, the EDM must:
  - be a simple ontology for capturing all relevant aspects of Cultural Heritage Objects
    - integrate the providers' data
    - support rich functionality
  - offer a structure for collecting data from contributors
  - re-use existing ontology and models
  - buy into the Web architecture and models
- Not obvious at first, a result

Outline

- Part I
  - Background
  - Requirements
  - Status
- Part II
  - The general picture
  - Classes
  - Properties
  - Examples
  - Future work

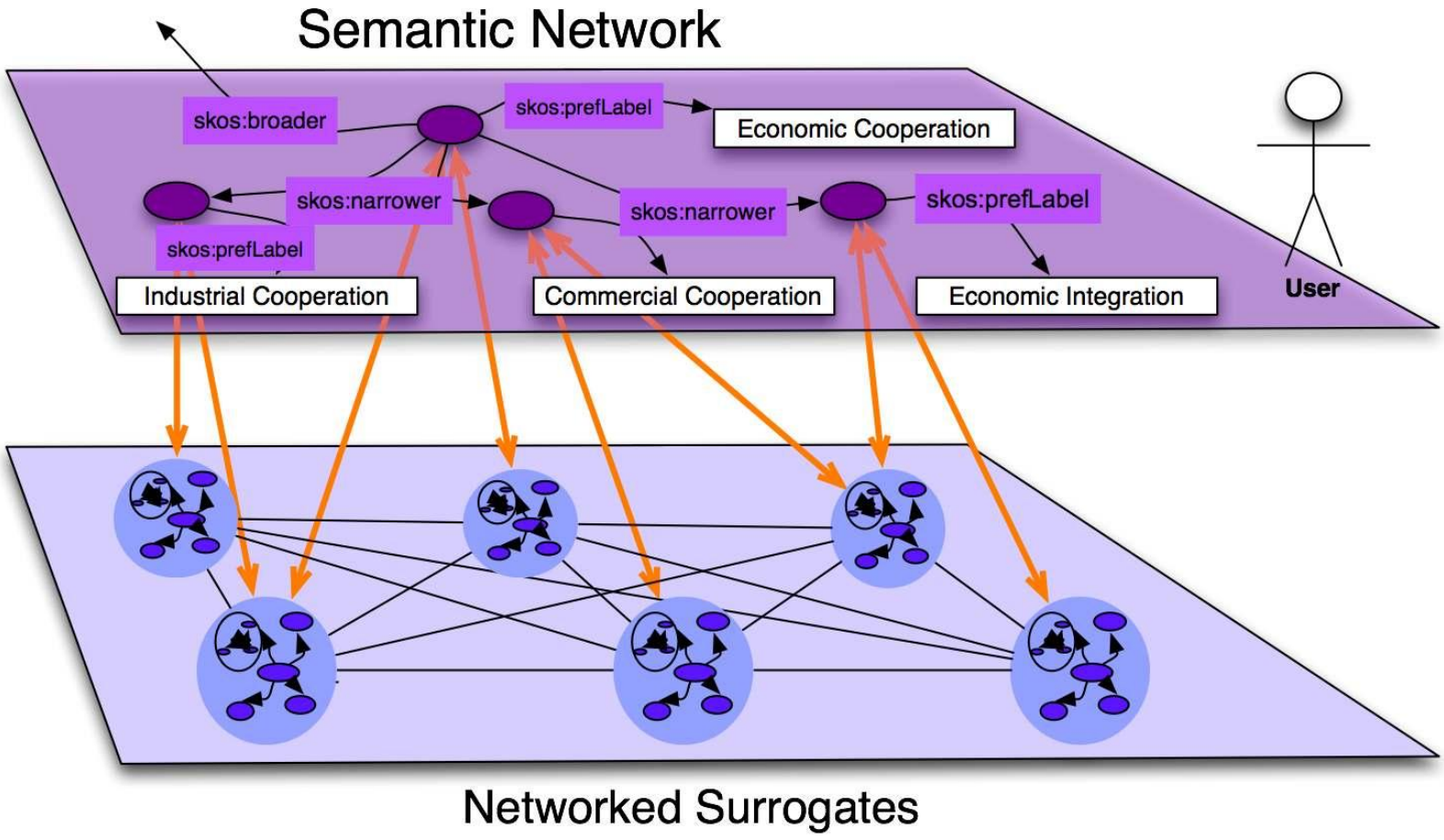
Status

- The present one is the fourth version of EDM
  - version 1, largely inspired to CIDOC CRM (June ‘09)
  - version 2, slight revision of version 1 (July ‘09)
  - version 3, evolution of version 2 towards the web (Sept. ‘09)
    - including contribution from Europeana:connect
  - version 4, revision of version 3, with more web in it (Dec. ’09)
- Outlook
  - Danube specs due July ’10
  - We need to evaluate the EDM
    - based on usage by domains (four meetings foreseen within March ‘10)
    - prototyping (basic functions and searches)

Outline

- Part I
  - Background
  - Requirements
  - Status
- Part II
  - The general picture
  - Classes
  - Properties
  - Examples
  - Future work

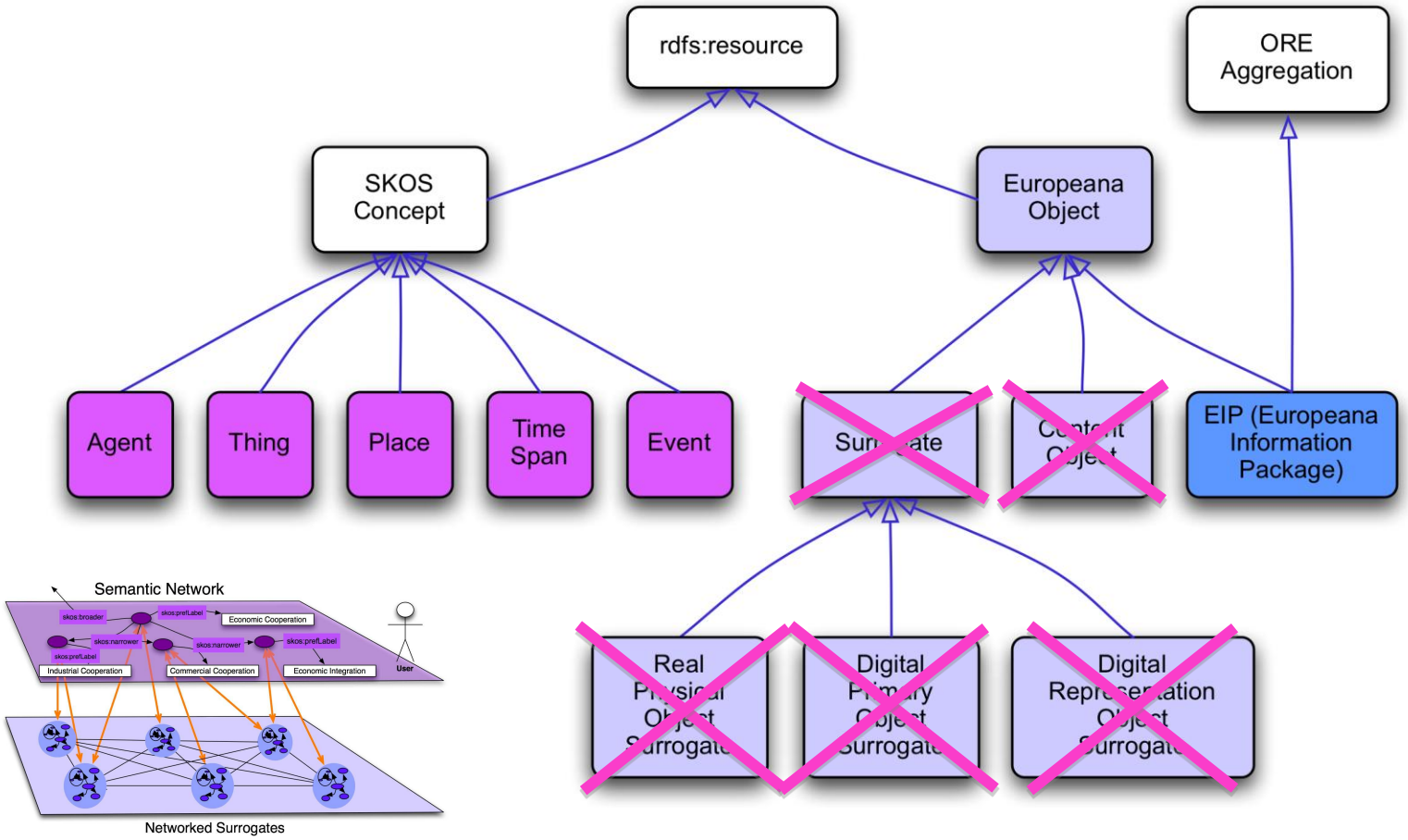
The general picture



Outline

- Part I
  - Background
  - Requirements
  - Status
- Part II
  - The general picture
  - Classes
  - Properties
  - Examples
  - Future work

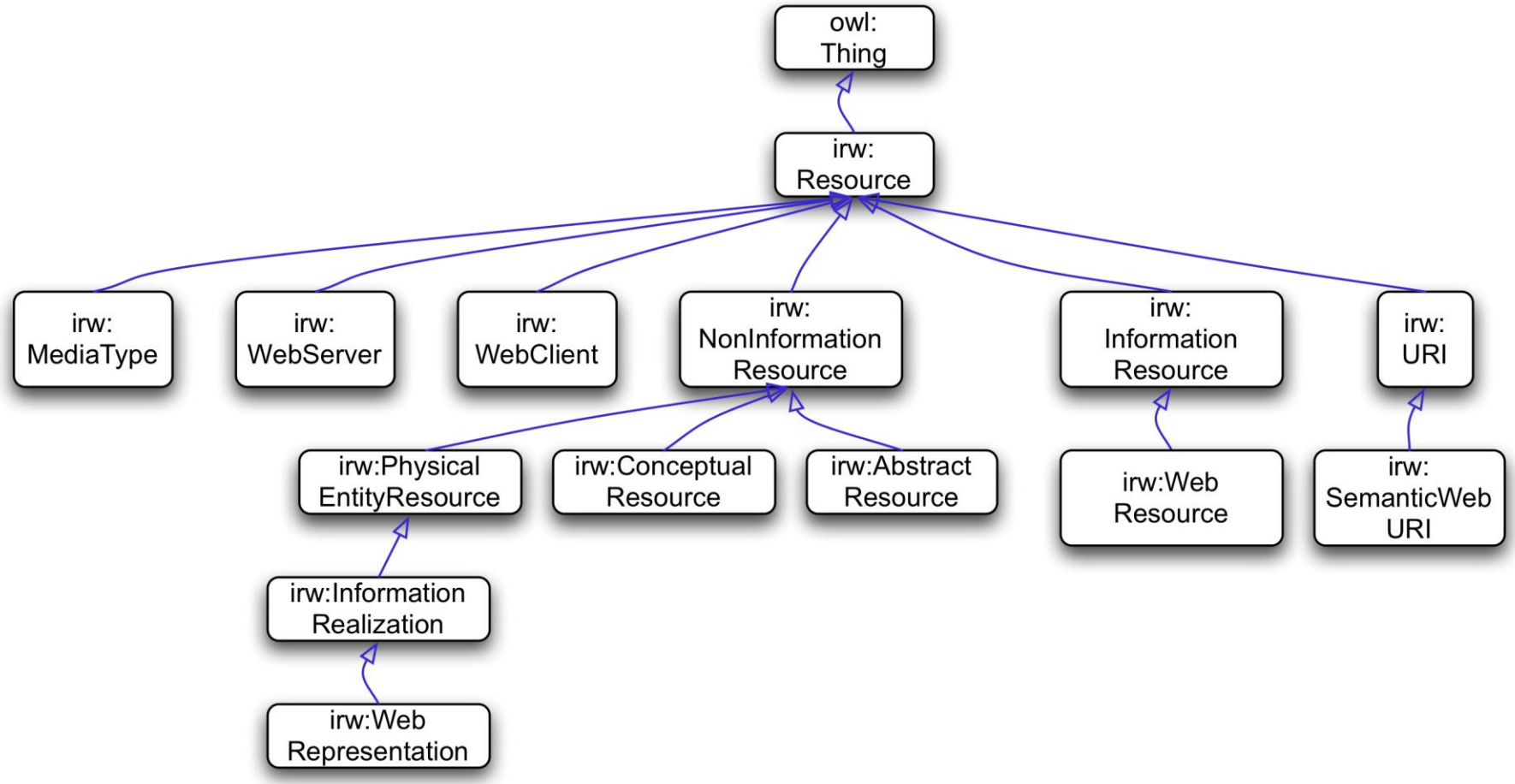
The class taxonomy in version 3



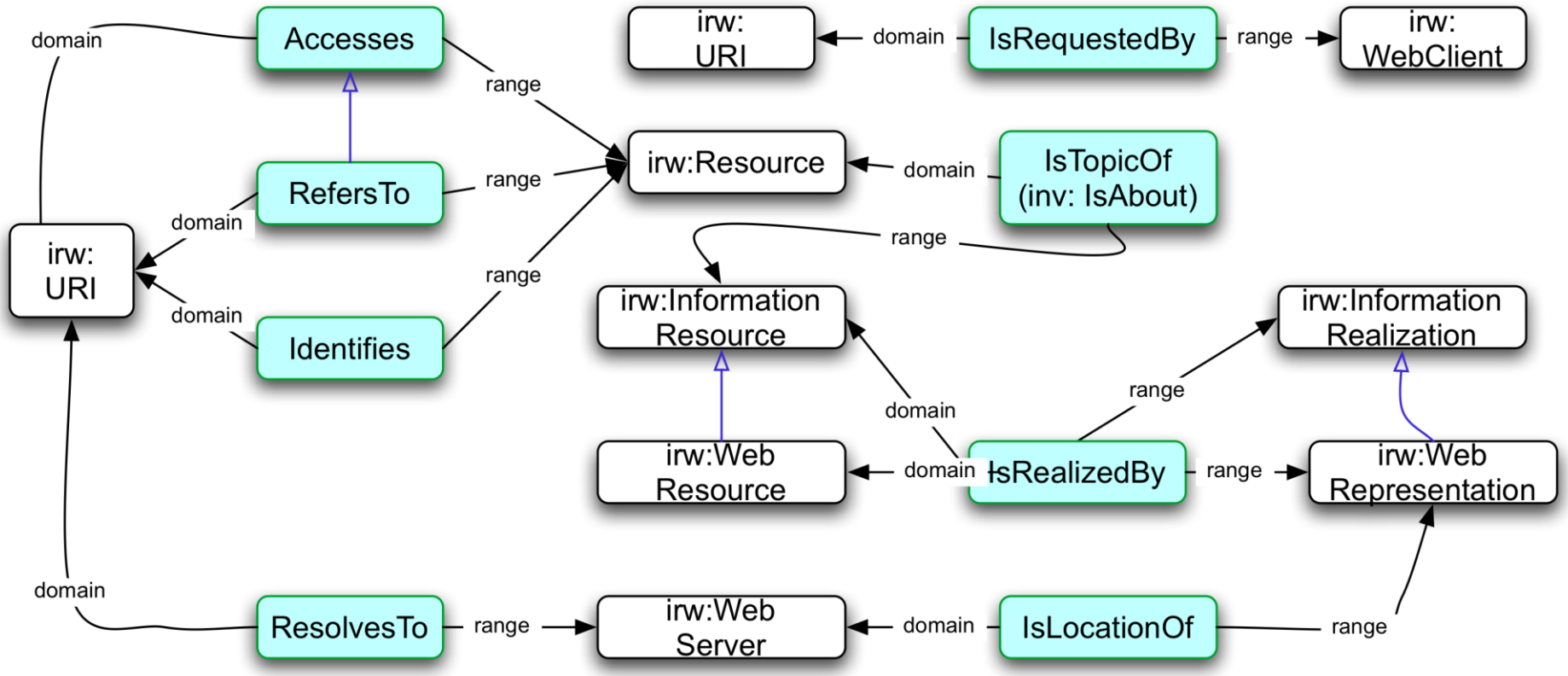
A more web-oriented classification

- To use the web ontology as a basis for Europeana
  - Information Resources
    - Web Resources
  - Non Information Resources
    - Information Realizations
      - Web page (HTML document)
- IRW ontology
  - An ontology that emphasizes a conceptualization of the Web entities
  - Derived from DOLCE Ultra Light (DUL)
  - Being developed

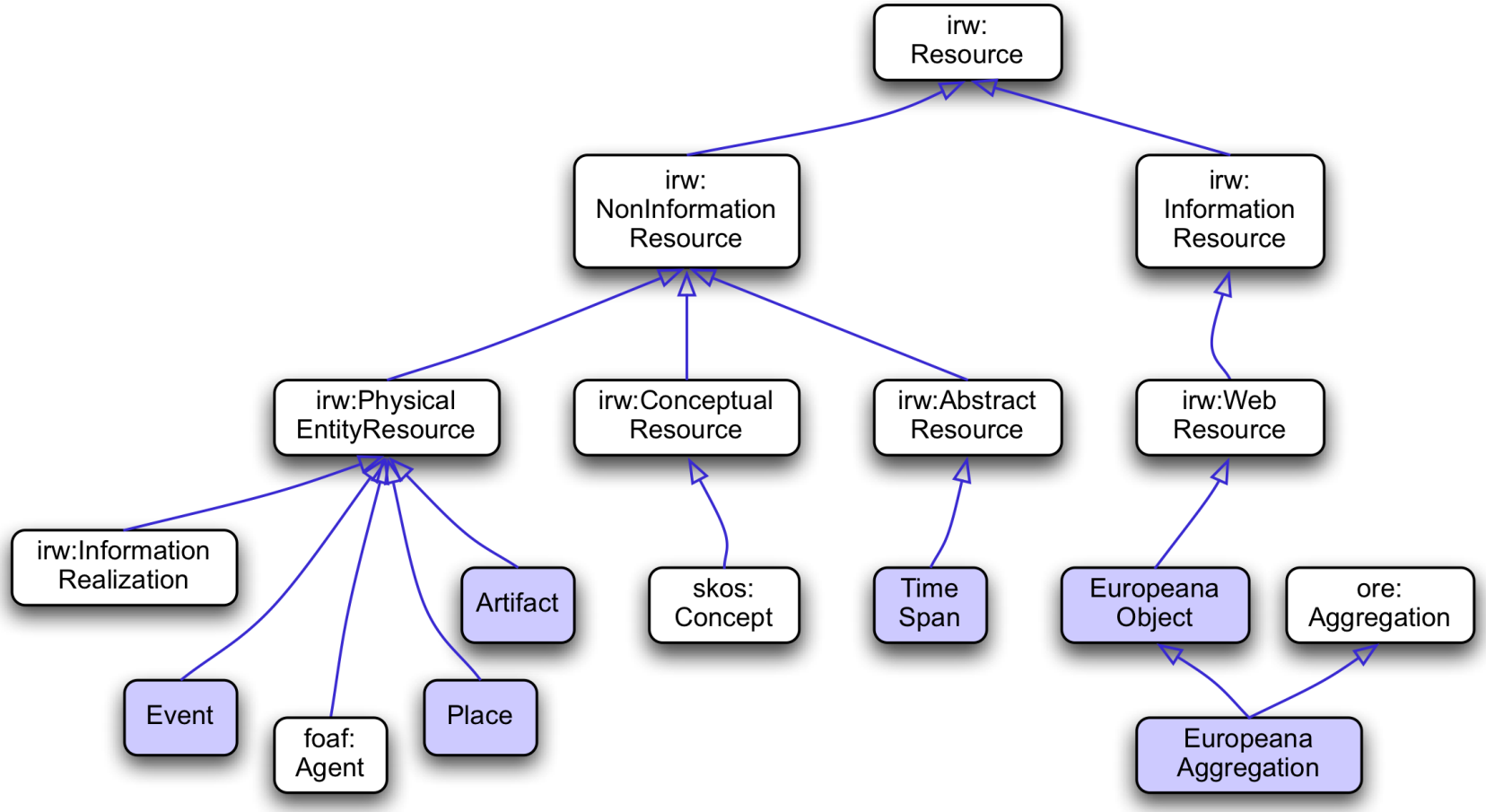
IRW classes


owl:
Thing
irw:
Resource
irw:
MediaType
irw:
WebServer
irw:
WebClient
irw:
NonInformation
Resource
irw:
Information
Resource
irw:
URI
irw:Physical
EntityResource
irw:Conceptual
Resource
irw:Abstract
Resource
irw:Web
Resource
irw:
SemanticWeb
URI
irw:Information
Realization
irw:Web
Representation

IRW Properties



The class taxonomy in version 4



Europeana Aggregation

- The set of resources related to a single Cultural Heritage Object that collectively represent that object in Europeana.
  - all descriptions about the object that Europeana collects from (possibly different) content providers
    - including thumbnails and other abstractions
  - the description of the object that Europeana builds
- Every Cultural Heritage Object known to Europeana is represented by an instance of EuropeanaAggregation
- Every instance of EuropeanaAggregation represents a Cultural Heritage Object.

Examples

- The painting Mona Lisa is a Cultural Heritage Object represented in Europeana (only) by the EuropeanaAggregation instance ens:MonaLisa
- The title “Le Temps” is a Cultural Heritage Object represented in Europeana (only) by the EuropeanaAggregation instance ens:LeTemps
- The 56th issue of “Le Temps” is a Cultural Heritage Object represented in Europeana (only) by the EuropeanaAggregation instance ens:LeTemps-n56

Europeana Object

- Any digital object on which Europeana has rights
  - Aggregations
  - Europeana content
    - Annotations (this class is the range of ens:hasAnnotation)
    - Deliverable of one the Europeana projects
  - Any content provider's object on which Europeana has acquired some right
    - A thumbnail of the painting Mona Lisa owned by the Louvre and offered to Europeana as an illustration of the painting, along with some rights (e.g., display)
    - A digitization of a photograph of the first page of issue number 56 of the title "Le Temps"
    - The text of the first page of issue number 56 of the title "Le Temps" s

Outline

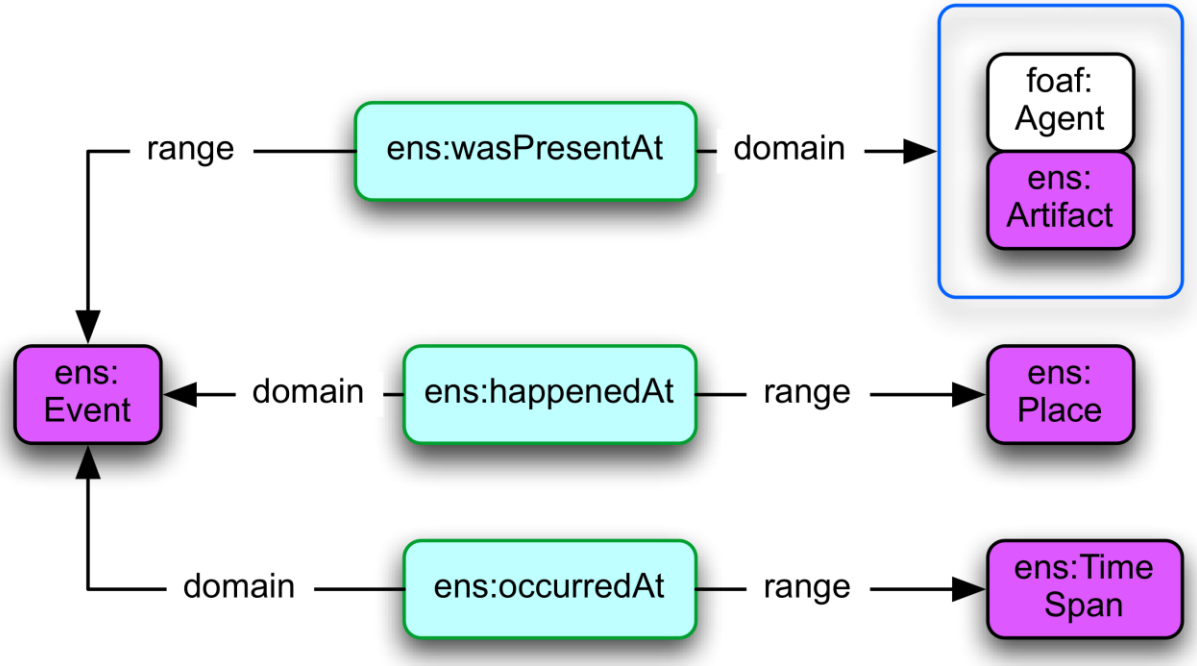
- Part I
  - Background
  - Requirements
  - Status
- Part II
  - The general picture
  - Classes
  - Properties
  - Examples
  - Future work

Properties

- A few properties in version 3 have been replaced in version 4 by properties imported from other schemas
  - there will probably be others
- Many properties in version 3 have changed domain and range in version 4
  - due to the changes in the classes of the model

Was Present At, Happened At, Occurred At

- Same as in version 3

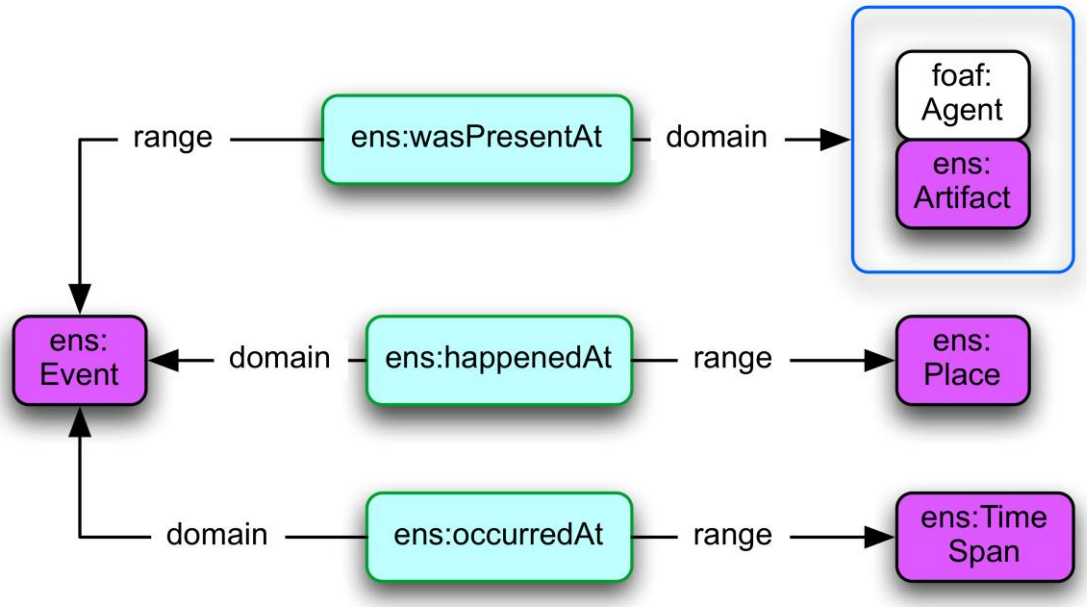


Rationale

- wasPresentAt (L. da Vinci wasPresentAt the creation of Mona Lisa)
  - discovery by persons, together with other properties (who query)
  - browsing artifacts
  - browsing events
- happenedAt (the creation of Mona Lisa happenedAt Florence)
  - discovery by places (where query)
  - browsing events
- occurredAt (the creation of Mona Lisa occurredAt 1503-1506)
  - discovery by time (when query)
  - time-line browsing
  - browsing events

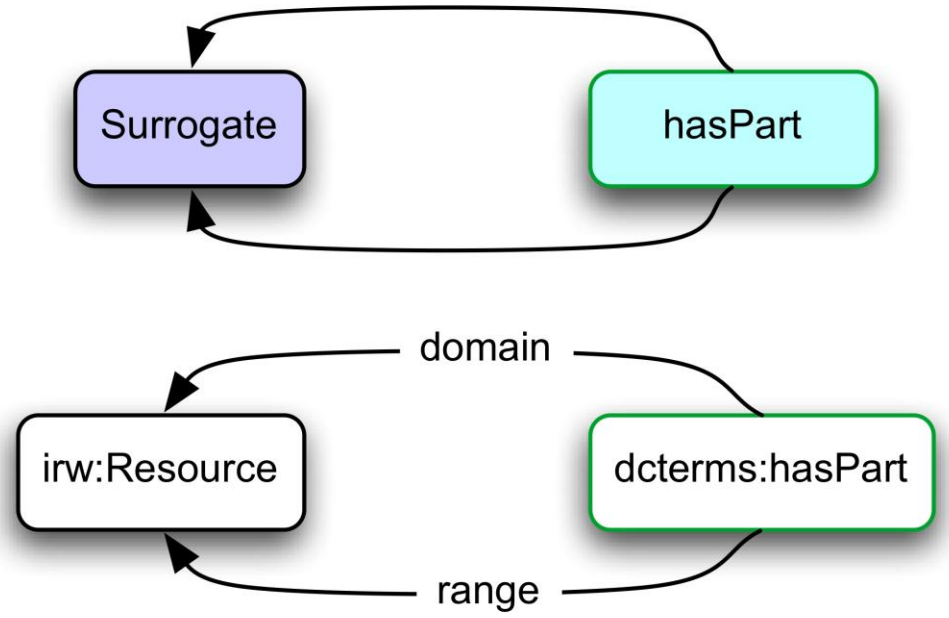
Rationale

- The combined usage of these properties is very important!
- e.g., from persons to events, from events to the involved:
  - artifacts
  - people
  - time
  - places



Has Part

- Structure of objects

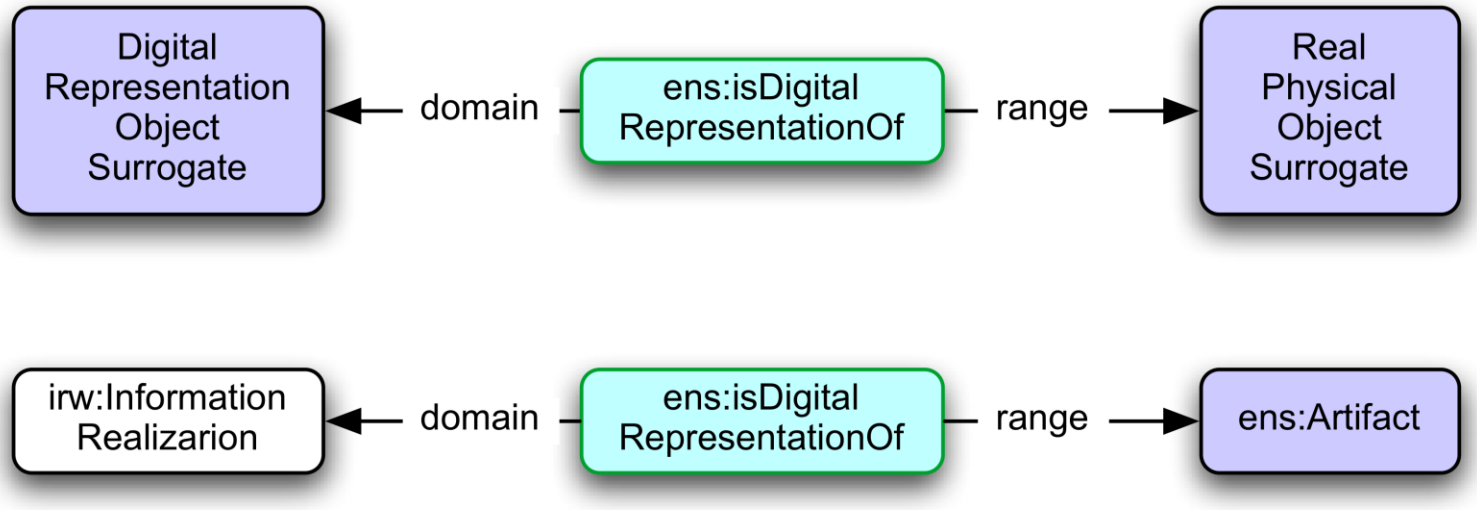


Rationale

- hasPart (the title “Le Temps” hasPart issue 56 of “Le Temps”)
  - Identification of parts
  - Integration of structural properties used within the descriptions contributed by content providers.
    - any such property should be declared to be a (direct or indirect) sub-property of dcterm:hasPart

Is Digital Representation Of

- Is digital representation of



RPO = ens:Artifact, DRO = irw:InformationRealization, because digitization produces a physical object (in digital form)

Rationale

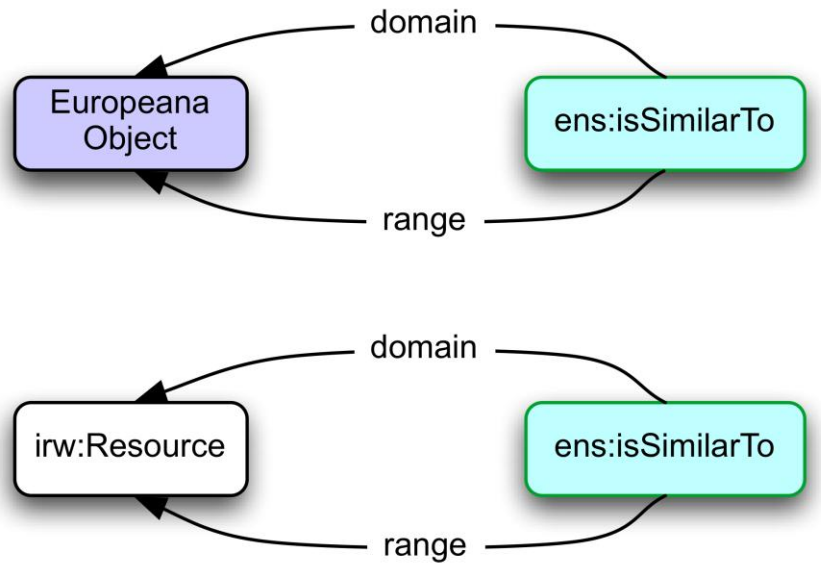
- This property allows to properly represent the information inside Europeana.
- For instance, if Europeana collects information about both Mona Lisa and a high resolution digital image of Mona Lisa, it needs:
  - to distinguish between the two objects
  - to properly relate them

Same As

- The property ens:europeanaSameAs has been removed
  - redundant
  - can use owl:sameAS

Properties for Versioning

- Is Similar To



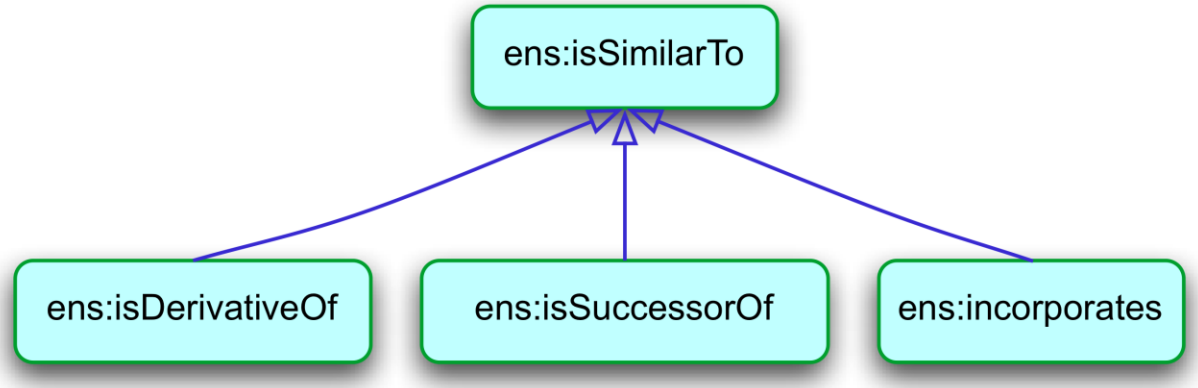
- previously, similarity related (one another) surrogates or packages.
- now similarity relates resources

Rationale

- Discovery of objects that are similar to a given one
- Integration of all properties used in content providers' descriptions that capture the notion of relatedness
  - dc:relation

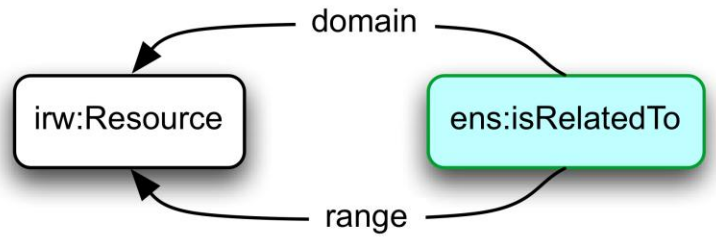
Versioning sub-properties

- The Italian translation of Moby Dick isDerivativeOf the original work.
- The issue 57 of “Le Temps” isSuccessorOf issue 56
- The movie “A Clockwork Orange” incorporates Rossini’s symphony from “La Gazza Ladra” in its original soundtrack. “E.A.Poe, The Raven (poem)” is incorporated in “Emerson Lake & Palmers Tales of Mystery (music)” which is incorporated in “Concert Recording 1973 (vinyl)”.



Contextualization properties

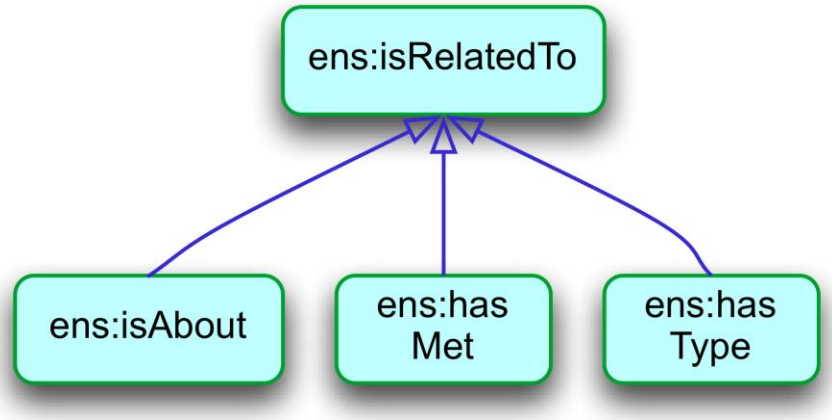
- Is Related To (Moby Dick is related to XIX century literature. Mona Lisa is related to Renaissance Art)

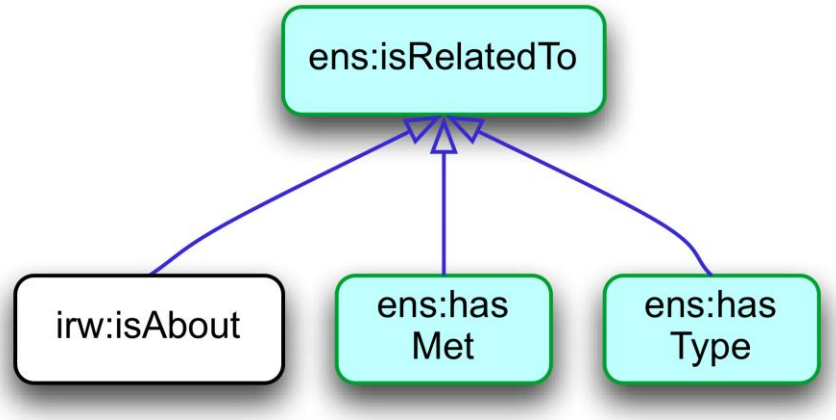


- domain: Cultural Heritage Objects known to Europeana
  - the ORE proxies that are ore:proxyIn a Europeana Aggregation" (easier in version 3)

Rationale

- Basic semantic discovery
- Sub-properties:





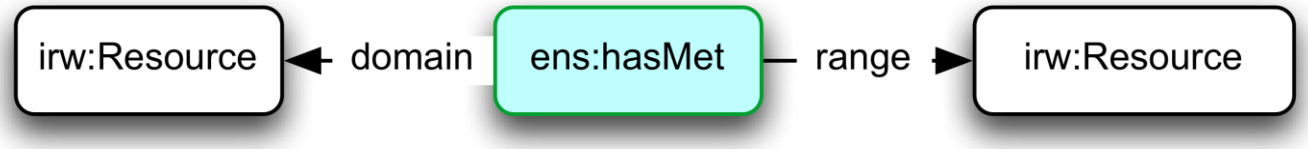
Is About

- “irw:isAbout expresses the relationship of an information resource to the resources the information is ‘about.’”

- In Europeana, irw:isAbout captures reference in the more general sense, i.e., anything a resource refers to by its shape, form or features in a figural or encoded form.
- in particular, the association between a Europeana Aggregation and the CHO it refers to
- sub-properties: frbr:hasAsSubject, dc:subject, dc:coverage

Has Met

- Relates a resource with the objects or phenomena that have happened to or have happened together with the resource under consideration.



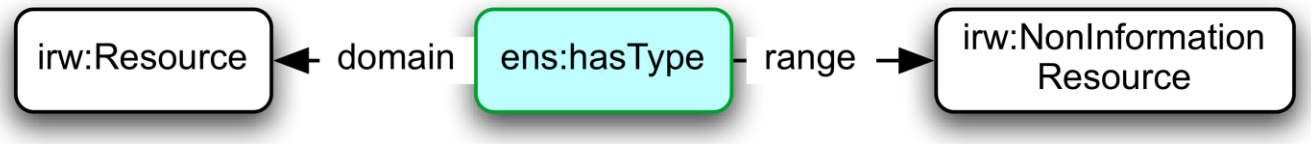
- Allows for “who, when, where, what” queries, without specifying if the “who” matches the “when”, such as a (fictitious) object made by Praxiteles and found in 1865.
  - sub-properties: dc:creator, dc:publisher, dc:contributor, dc:date.

Rationale

- "has met: Egypt AND Crete AND 2000-1000 BC" would return all Egytian artefacts found in Crete
  - the scarabs that are extraordinarily important for dating Minoan archaeological layers.
- "has met: Tapio Wirkkala" would return objects collected by the designer, his photosand his designs.
- "has met: Gauguin" would return South Sea objects collected by Gauguin and his works of art, his letters etc.
- "has met: Samoa AND 1800-1900AD" would return objects made in Samoa and objects sold to Samoa in Colonial times.

Has Type

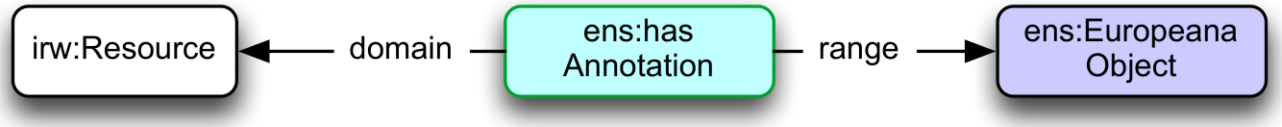
- Relates a resource with the concepts it belongs to in a suitable type system such as MIME.
  - The type of Mona Lisa is (AAT) Painting. The type of a digital image of Mona Lisa may be JPEG.



- Support “what” queries
  - Sub-properties: dc:type, dc:format, dc:language

Has Annotation

- Relates a resource with a Europeana Object that annotates it



- Users of the Europeana portal have the possibility of annotating any resource.
- The annotation is an object owned by Europeana
- We are considering a richer annotation model proposed by N. Ferro

Landing Page

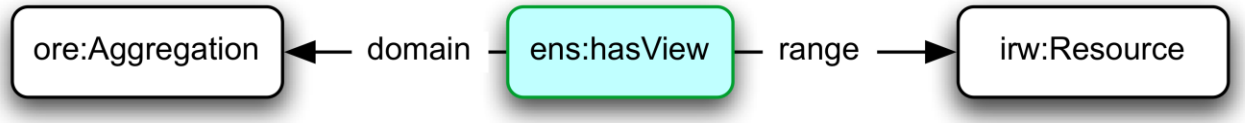
- Relates a Europeana aggregation to the page of the institution where the corresponding object is accessible

Has View

- Relates an ORE aggregation about a CHO with a digital object providing a view of the CHO
  - a thumbnail
  - a textual abstract
  - a table of contents.

Has View

- The ORE aggregation may be
  - a Europeana Aggregation, in which case the view is an object owned by Europeana (i.e., an instance of ens:EuropeanaObject)
  - an aggregation contributed by a content provider, in which case the view is an instance of irw:InformationRealization.
- In order to capture both these cases, the domain of ens:hasView is ore:Aggregation and its range is irw:Resource

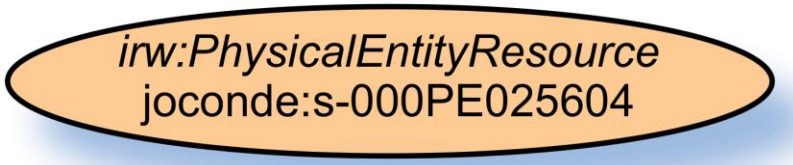


Outline

- Part I
  - Background
  - Requirements
  - Status
- Part II
  - The general picture
  - Classes
  - Properties
  - Examples
  - Future work

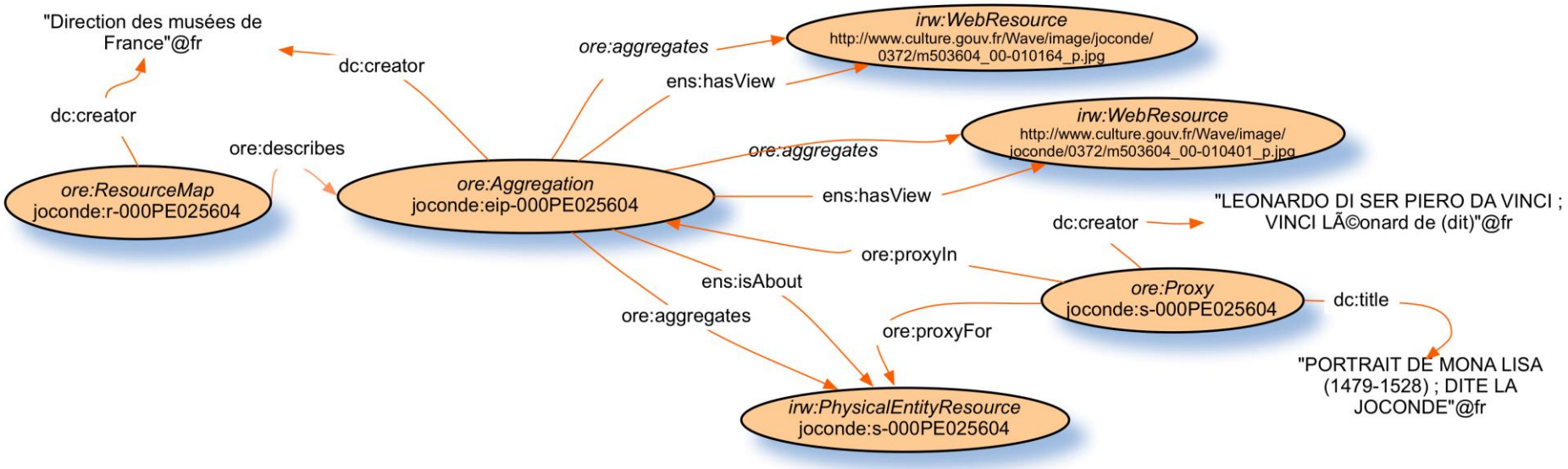
Modeling Mona Lisa

- The Cultural Digital Object representing Mona Lisa in the institution that holds the painting (Direction des musées de France)

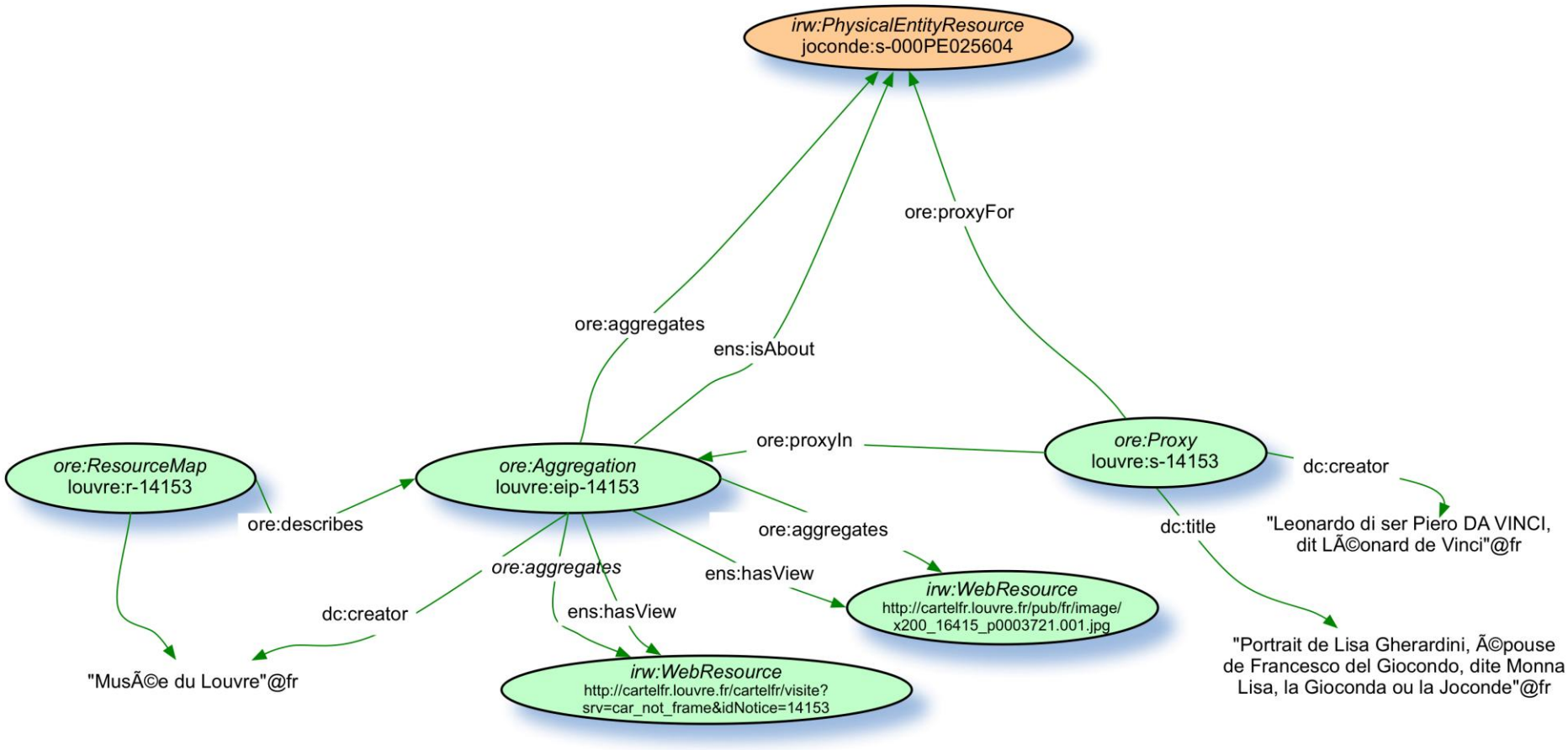


- identified by a URI assigned by Direction des musées de France
- classified according to the IRW ontology

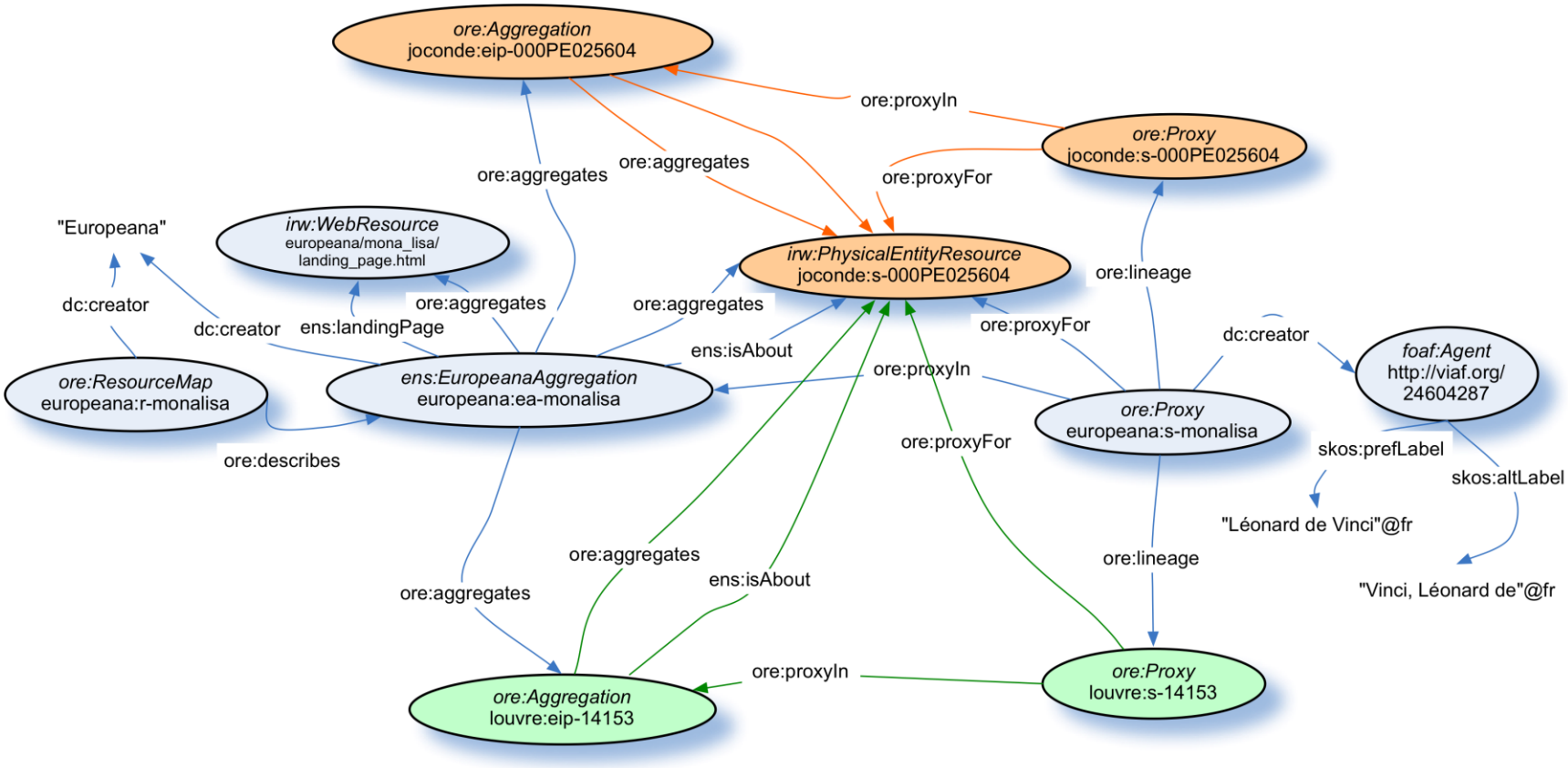
The object provided by Direction des musées de France



The object provided by Louvre



The object in Europeana



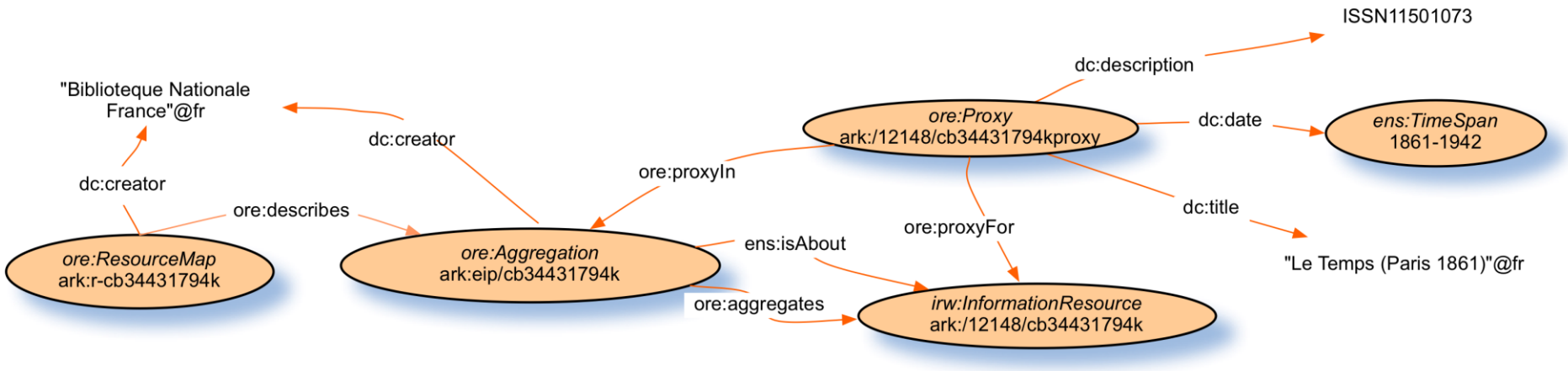
Second example: Le Temps

- The selected object is about a newspaper **title**, *Le Temps*, with a specific **issue**.
- For both the title and the issue we have metadata records from **BNF**.
- For the title only, we have an alternative record from **Wikipedia**, referring to the French, German and English Wikipedia pages about this newspaper.
- For both the title and the issue we also have **Europeana**-specific records. These are used
  - to record Europeana-specific metadata not in the BNF nor Wikipedia data (like the Europeana date of ingestion, the URI of the Europeana landing page, etc), and
  - to state that both BNF and Wikipedia are actually talking about the same newspaper title.

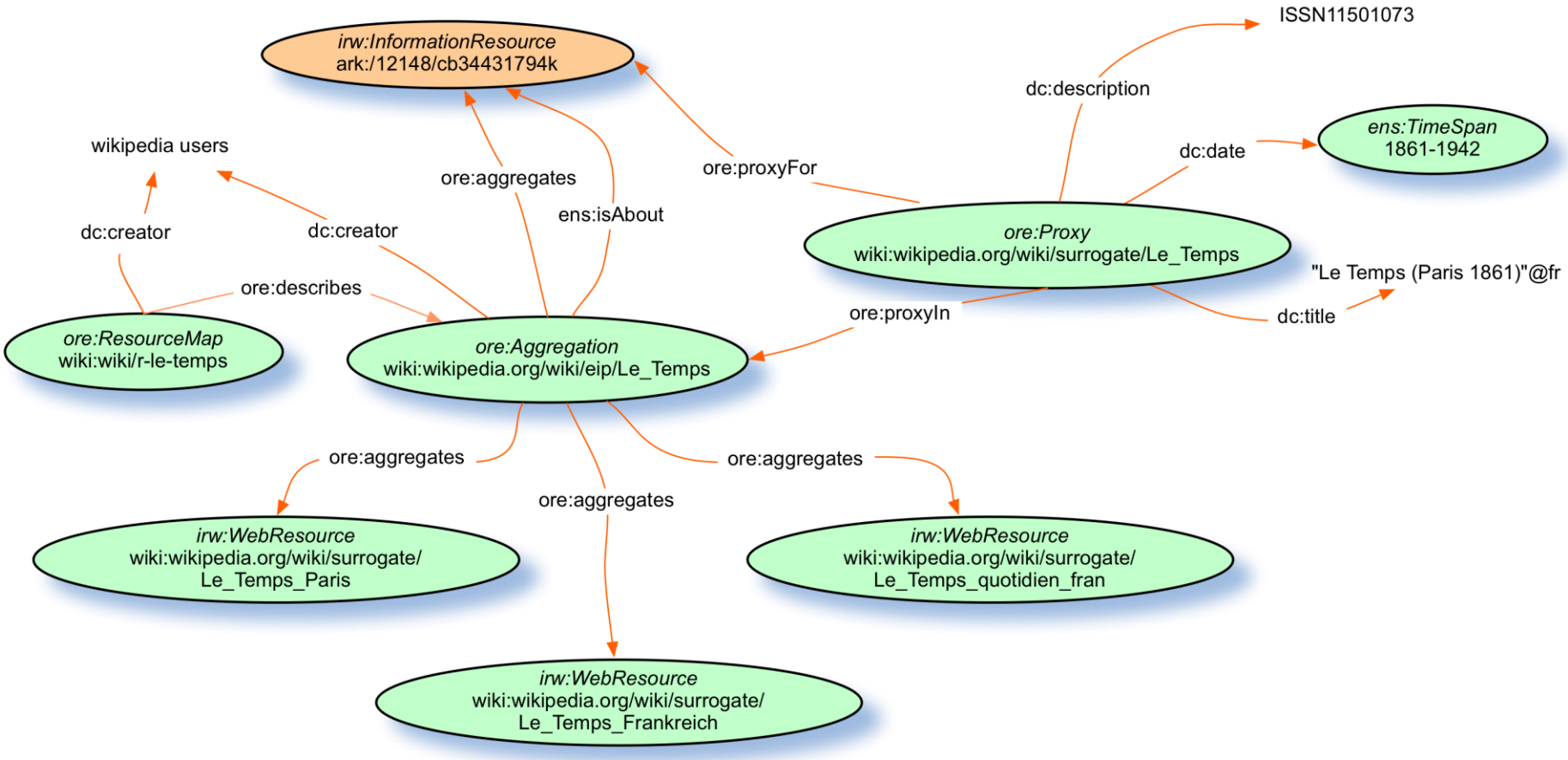
europeana
think culture

Le Temps in BNF

- The title in the institution that owns it (BNF)

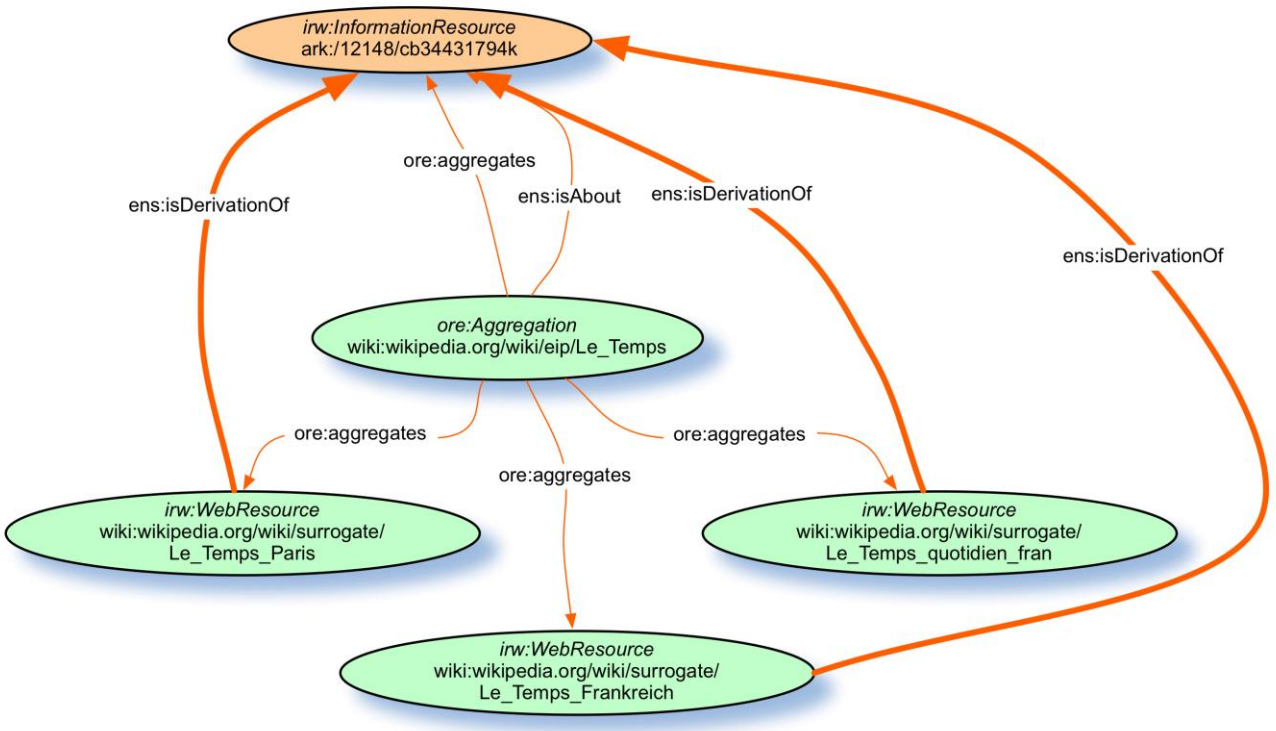


The title in Wikipedia

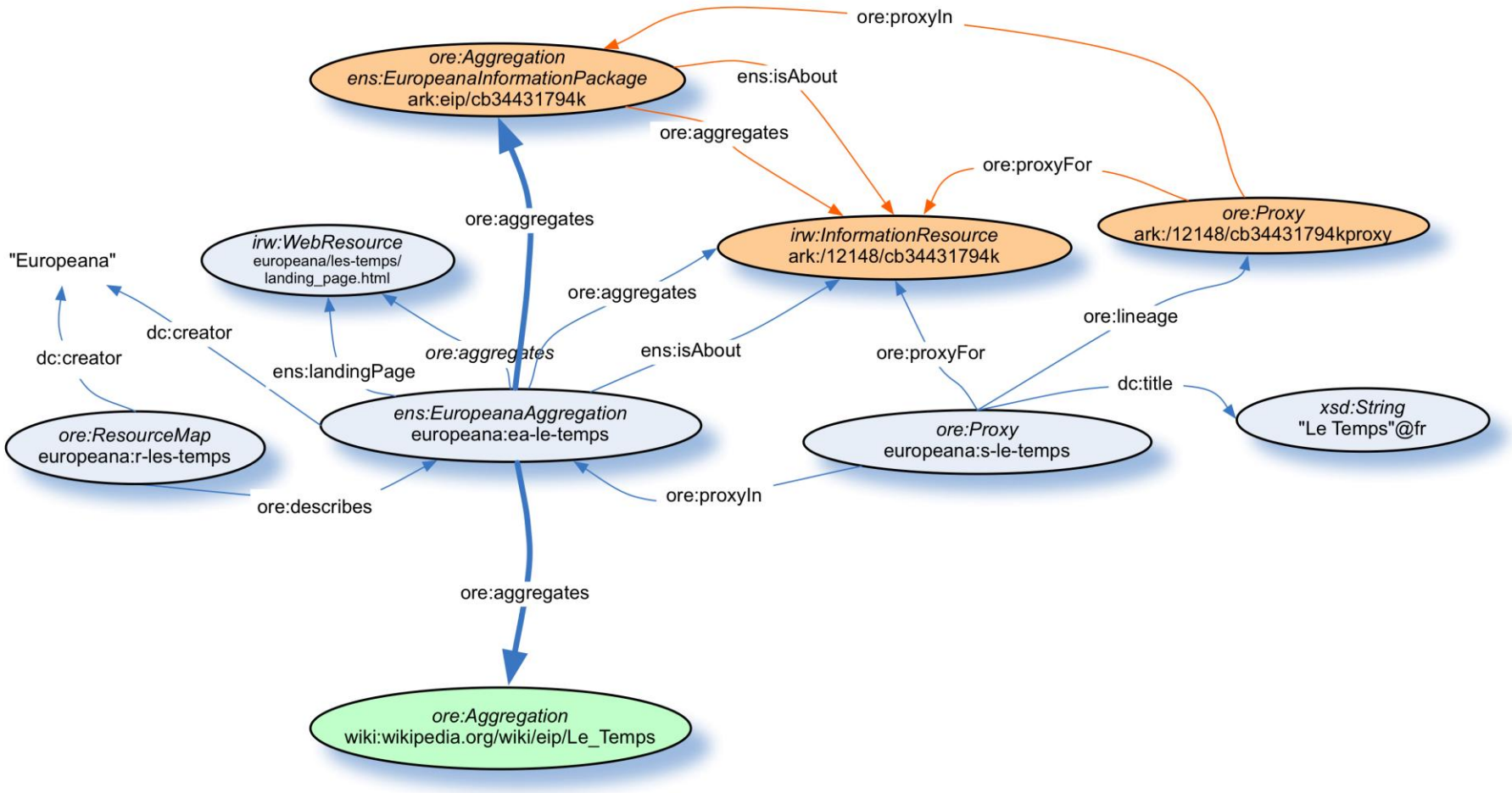


Derivation information

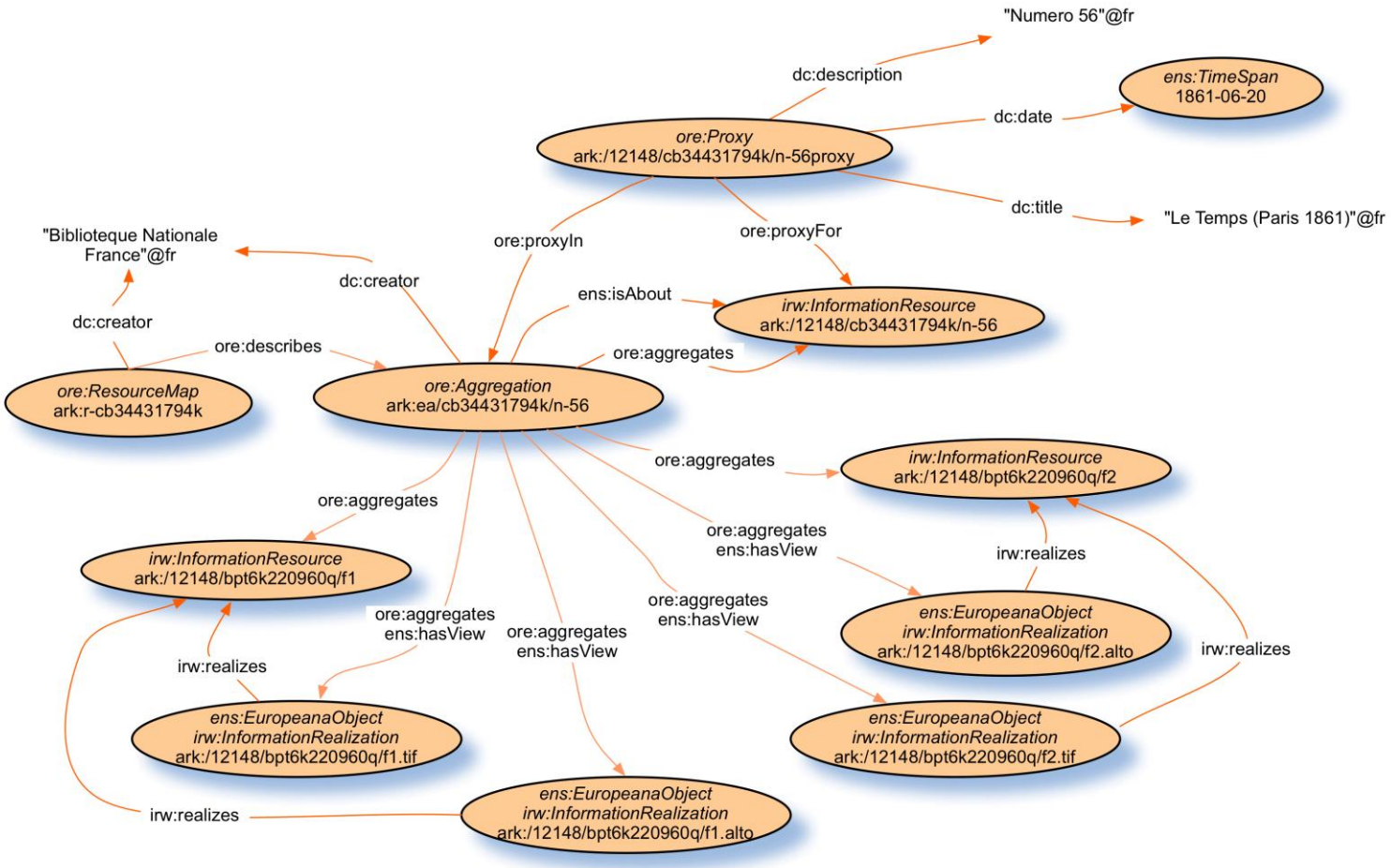
- to represent the fact that the title in wikipedia includes derivations of the original title, e.g. in different languages, we can use the ens:isDerivationOf property



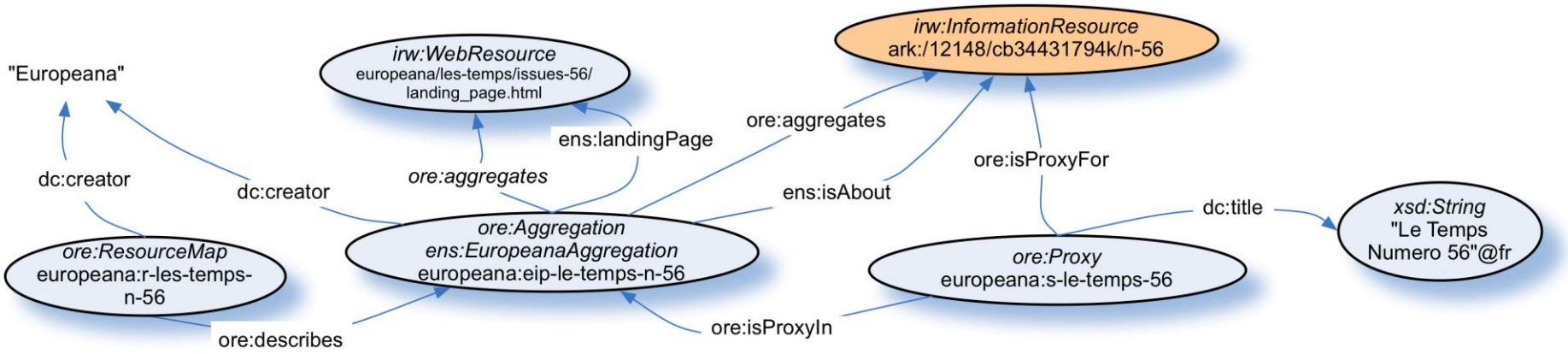
The title in Europeana



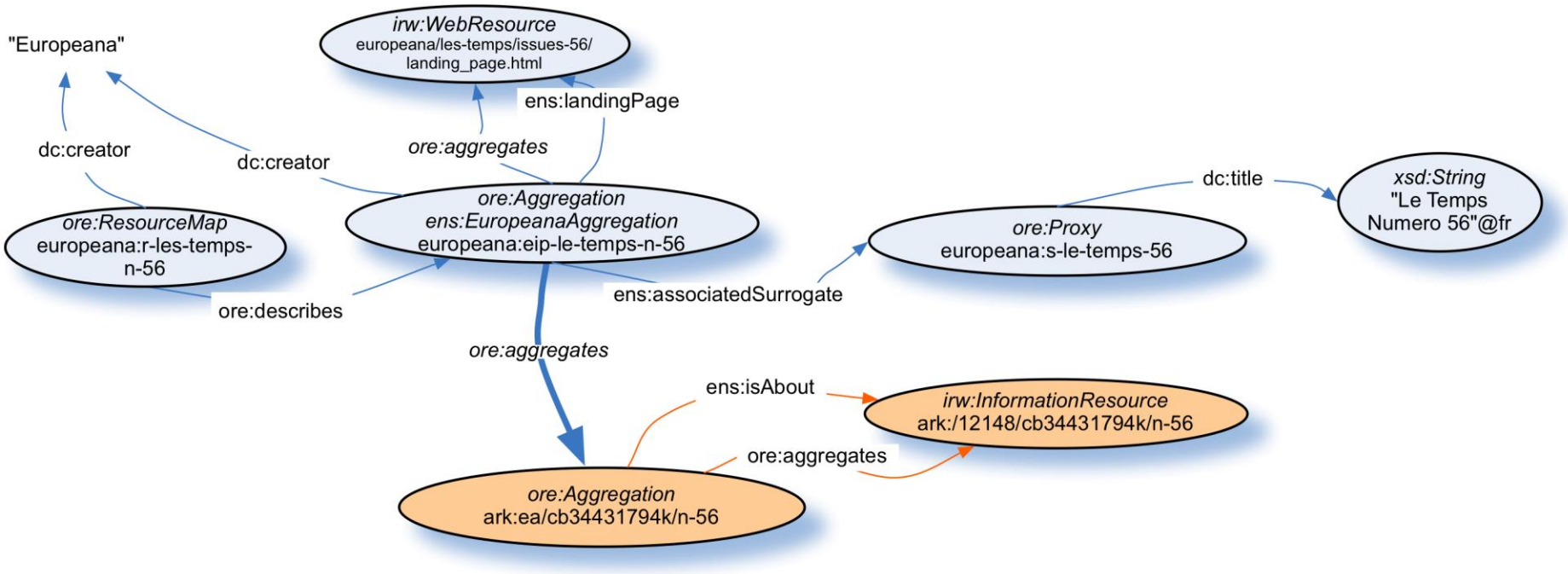
The issue in BNF



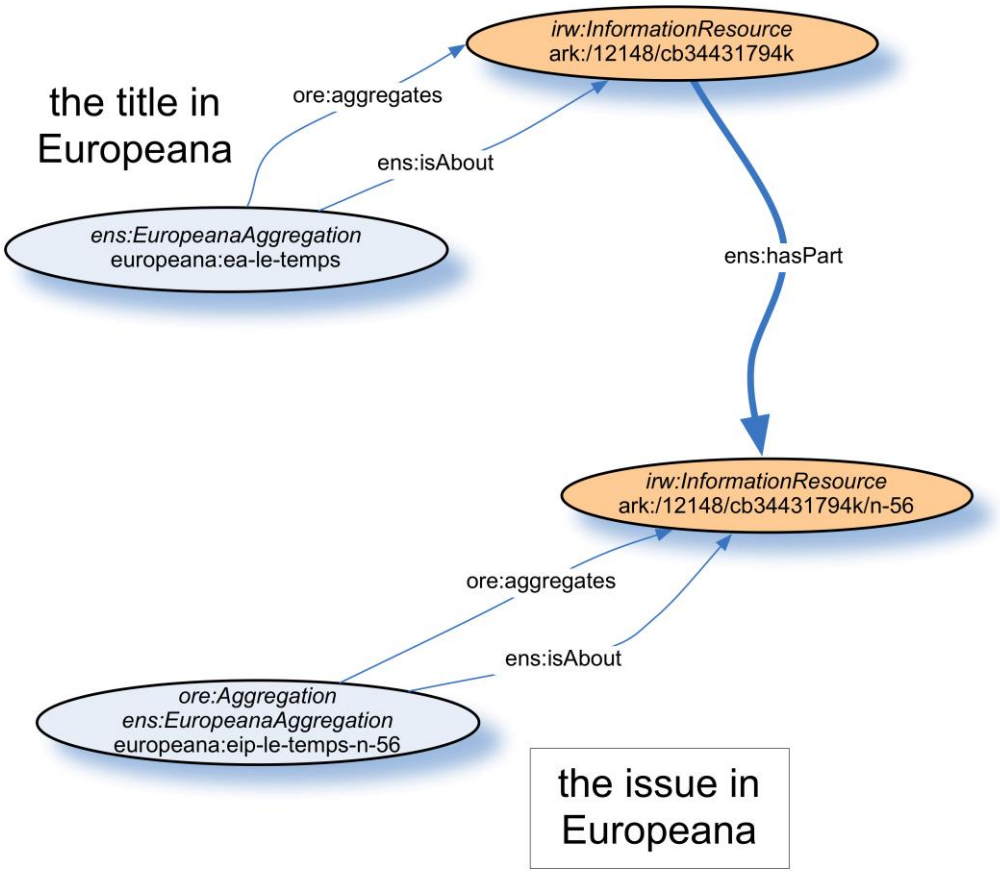
The issue in Europeana



The issue in Europeana



The issue is part of the title



Outline

- Part I
  - Background
  - Requirements
  - Status
- Part II
  - The general picture
  - Classes
  - Properties
  - Examples
  - Future work

Future work

- Discuss to stabilize
- Evaluation:
  - Mapping real data to EDM
  - Functional check
  - Implementation
- Revise
- Iterate