The Data Model of the OpenAIRE Scientific Communication e-Infrastructure

Paolo Manghi[1], Nikos Houssos[2,4], Marko Mikulicic[1], Brigitte Jörg[3,4]

[1]ISTI - Consiglio Nazionale delle Ricerche, Italy
[2]National Documentation Centre / National Hellenic Research Foundation, Greece
[3]Innovation Support Center, UKOLN, University of Bath, UK
[4] euroCRIS

6th Metadata and Semantics Research Conference (MTSR 2012), Cadiz, 28-30 November 2012

Agenda

- Introduction / background
- The OpenAIREplus information space and data modelling requirements
- Reuse of known data models
  - CERIF
  - DataCite
- The OpenAIREplus data model
- Modelling use cases within OpenAIREplus
- Summary – conclusions

Metadata in scholarly communication systems

- Metadata in scholarly communication systems: describes mainly publications typically using variants of Dublin Core, MARC or MODS.
- Flat metadata structures with limited facilities (e.g. links to authority files) to represent relationships among entities
- For example, linking publication and organisation (e.g. in roles publisher, author affiliation, commisioner)

Emerging metadata needs

- Two emerging requirements ask for more sophisticated metadata solutions:
  - Metadata needs to describe also data sets (heterogeneous, more complex than publications)
  - Contextual metadata is highly important – need for relationships of publications and data sets with projects, funding programmes, organisations, etc. to provide more sophisticated services to the end user
- OpenAIRE (2nd phase) needs to address both these challenges!
- Approach: Reuse existing data modelling approaches and standards
  - CERIF
  - DataCite

The OpenAIRE e-infrastructure

- 1st phase (in operation since 2010): a central point of access to OA publications funded by the EU FP7 projects in a range of thematic areas
- 2nd phase (OpenAIREplus project)
  - include metadata describing data sets and their semantic links to publications
  - incorporate research output produced all over Europe through any type of funding (not restricted to EU FP) including linking of outputs and projects with funding programmes
- Substantial upgrade of the data model required to address these challenges

The OpenAIREplus information space

- Includes entities such as publications, datasets, projects, licenses, persons, data sources, organizations, funding programmes.
- Captures semantics relationships among entities
- Data collection from various data source types:
  - Publication repositories
  - Data repositories
  - CRIS systems
  - Entity registries (e.g. ORCID, CORDA, OpenDOAR)

OpenAIREplus services

- Services to end-users
  - Researchers
  - Data source managers
  - Project coordinators
  - Funding agencies
- Services to applications
  - APIs
  - Data retrieval in standard data formats (CERIF XML, DataCite)

Reusing known data models

- CERIF
  - Common European Research Information Format
- DataCite

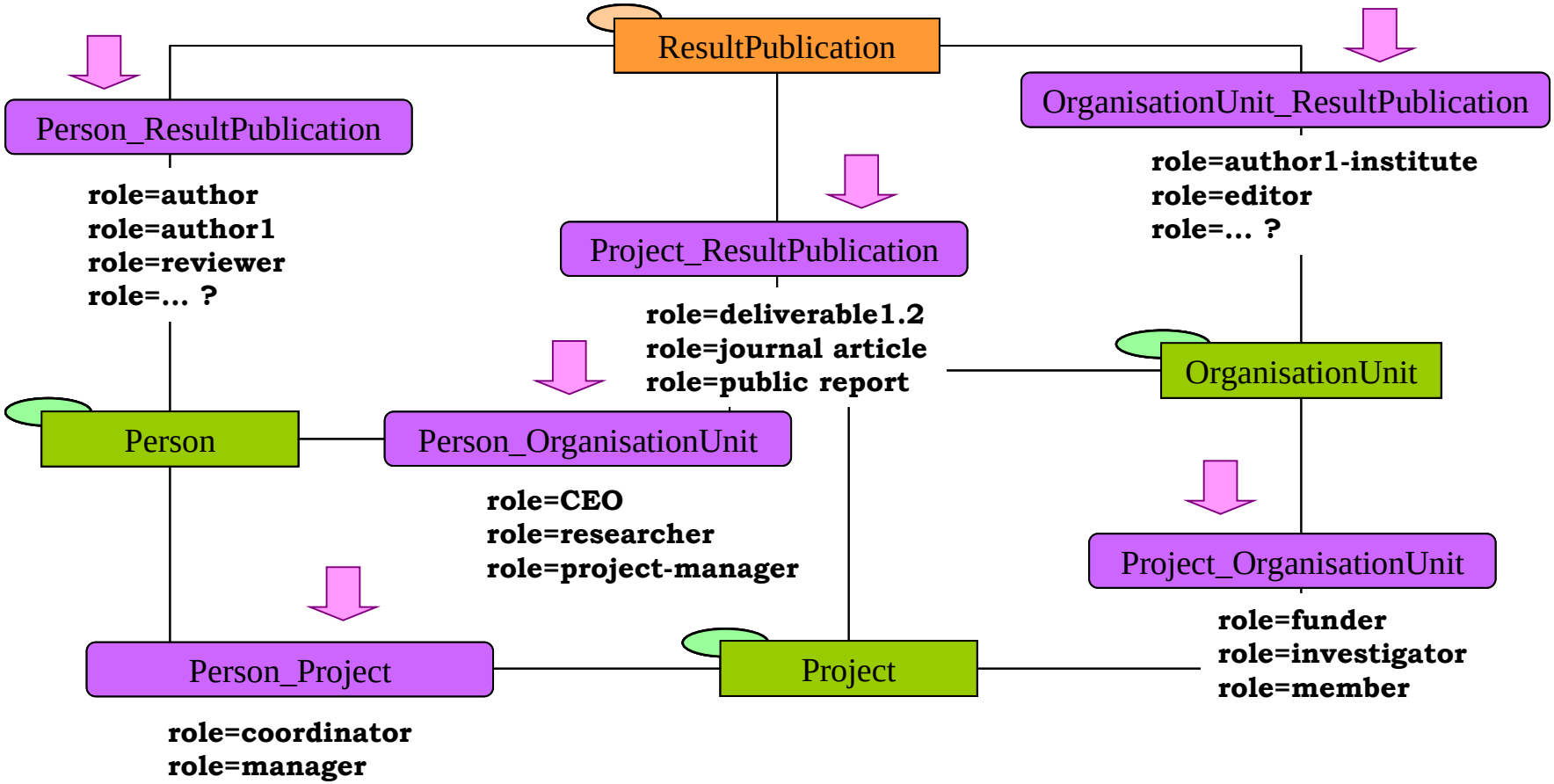
CERIF

- Conceptual model, mostly used in the research domain
  - Maintained by euroCRIS - used by 100s of institutional and national systems in Europe and beyond
  - Other application domains exist in the industry and the public sector
- Clear separation of syntax and semantics
- Explicit definition of the semantic of relationships among entities
- Temporal aspects of relationships captured
- Many data properties represented as semantic relationships – not rigid data fields
- Inherent support for multi-linguality (field values in different languages)

CERIF Semantic Layer

- Links/relationships among entities and classifications of entities
- Roles, timestamping (date range), fractions
- Definition of terms, vocabularies and relationships among them
- Examples:
  - OrgUnit X ***merged with*** OrgUnit Y in April 2011
  - Person X was ***Project Manager*** of Project Y from January 2010 to April 2011
- Facilitates role-typed, timestamped links to entities in other systems (e.g. identifier systems, registries, authority files)

CERIF Semantic Layer example



CERIF in OpenAIRE

- CERIF has been adopted in OpenAIRE (second phase) to represent contextual metadata about publications and datasets
- CERIF Semantic Layer used to represent relationships with defined semantics
- Ability to dynamically inject into the system vocabularies and terms without altering the data model structure
- Ability to represent arbitrary funding structures and their connections with publications and data sets

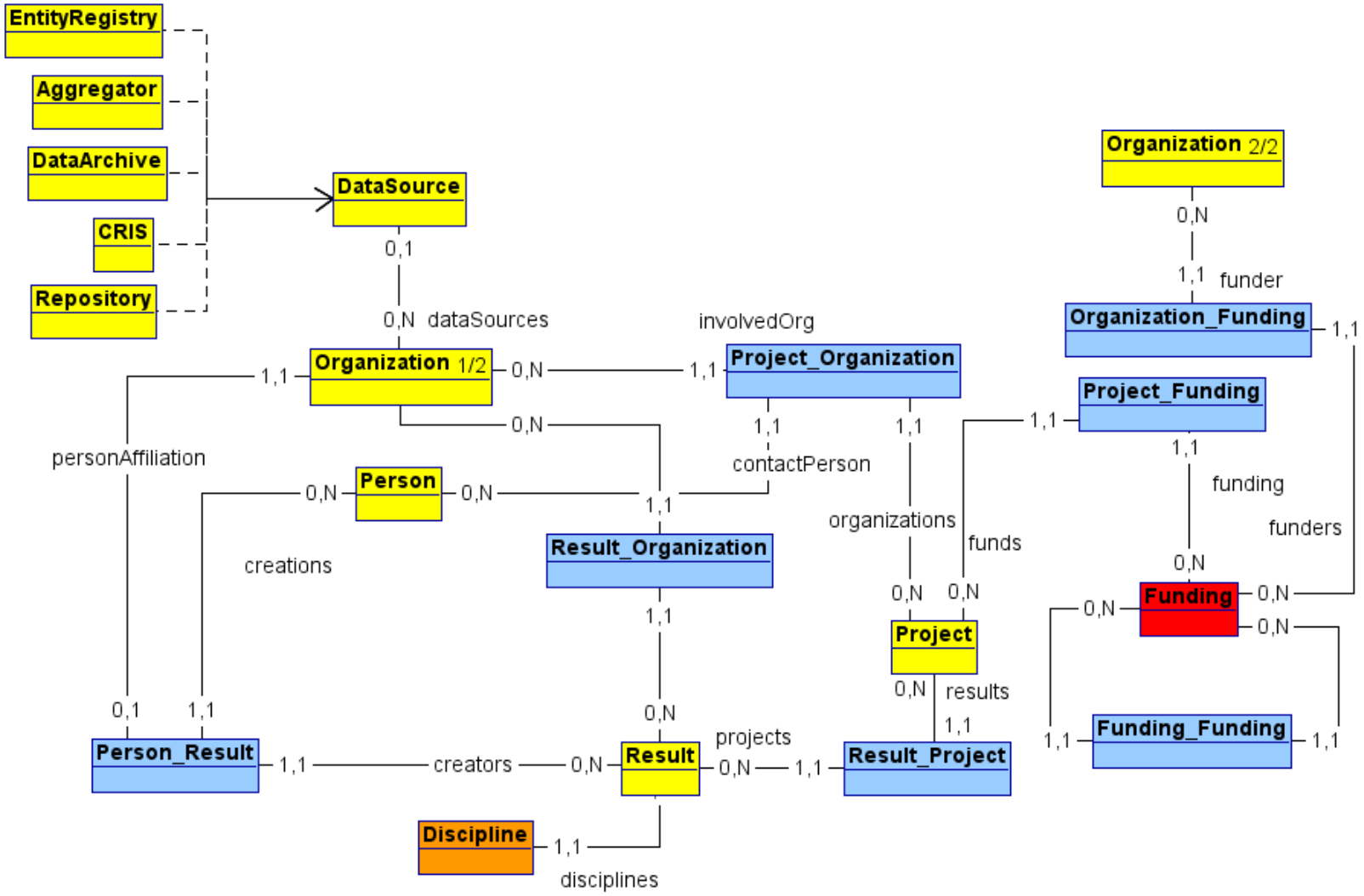
DataCite

- International consortium
- Aims at making data citable in a harmonized, interoperable and persistent way
- DOIs must be assigned to datasets
- Standard DataCite metadata format
  - Mandatory properties: title, authors, publishing year, distributor, persistent identifier
  - Optional properties, including links to other datasets and publications
- Used by many data repositories (e.g. PANGAEA, DANS)

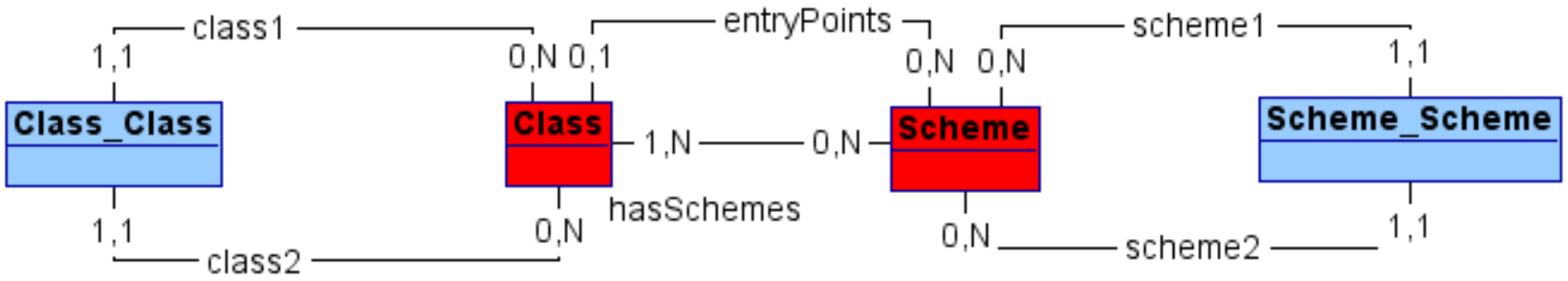
DataCite in OpenAIRE

- DataCite will be the standard metadata used by data repository data sources to contribute content to OpenAIRE
- DataCite data elements have been embedded to the OpenAIRE data model
- OpenAIRE will be able to export dataset metadata as DataCite metadata records
- OpenAIRE plans to exchange with DataCite the following types of data
  - Dataset metadata
  - Dataset-dataset and dataset-publication relationships

The OpenAIREplus data model

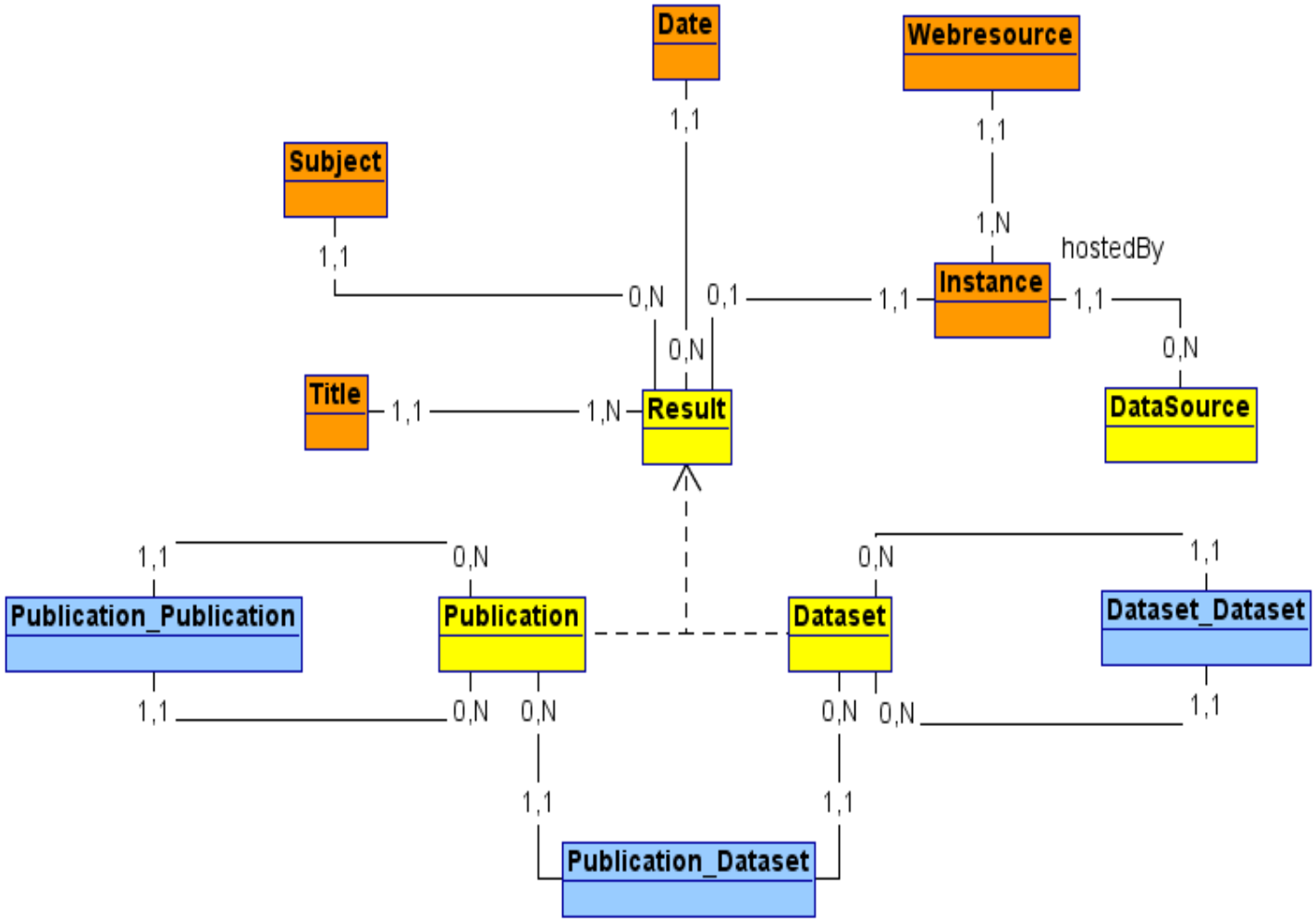


Semantic layer entities

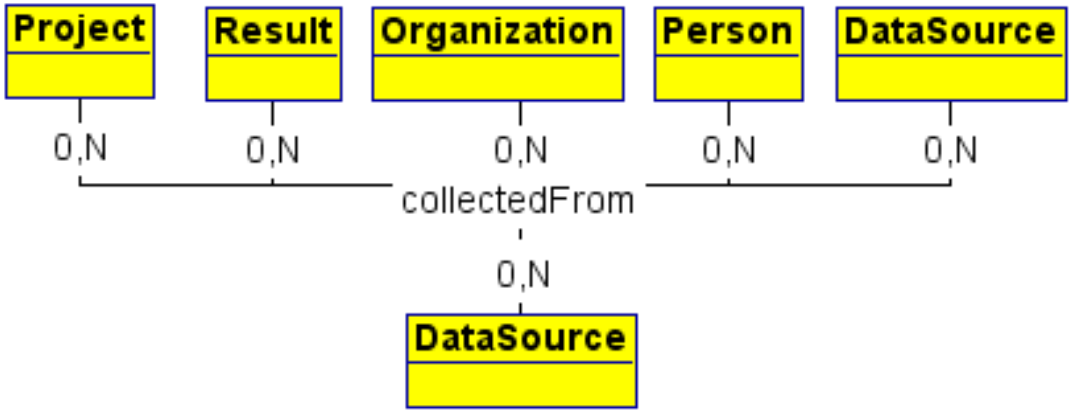


- CERIF Semantic Layer applied for capturing the semantics of relationships and classifications of entities
- Ability to represent vocabularies (Scheme) and terms (Class)

Result entities



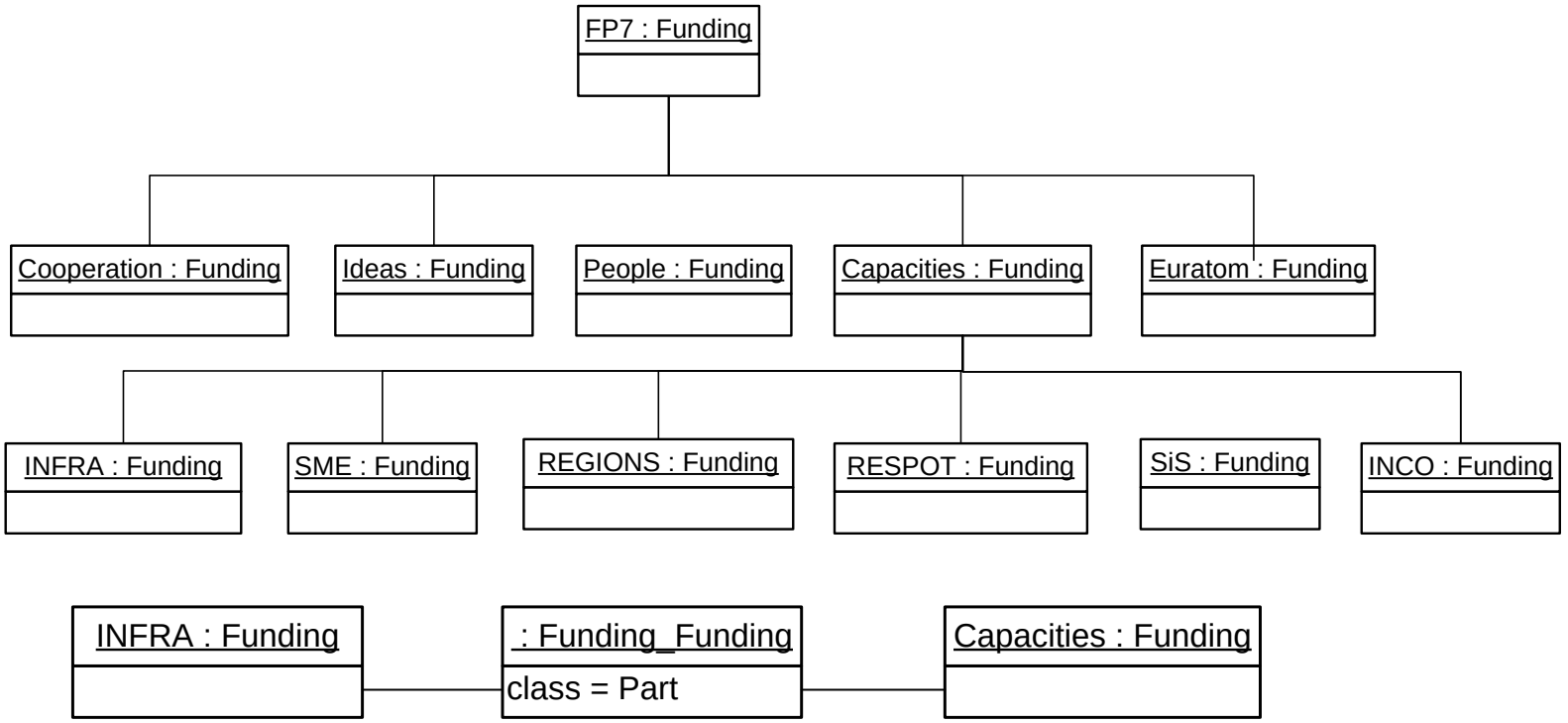
Data sources – provenance relationships



- A link to the originating data source is maintained for all entity instances collected and inserted into OpenAIRE

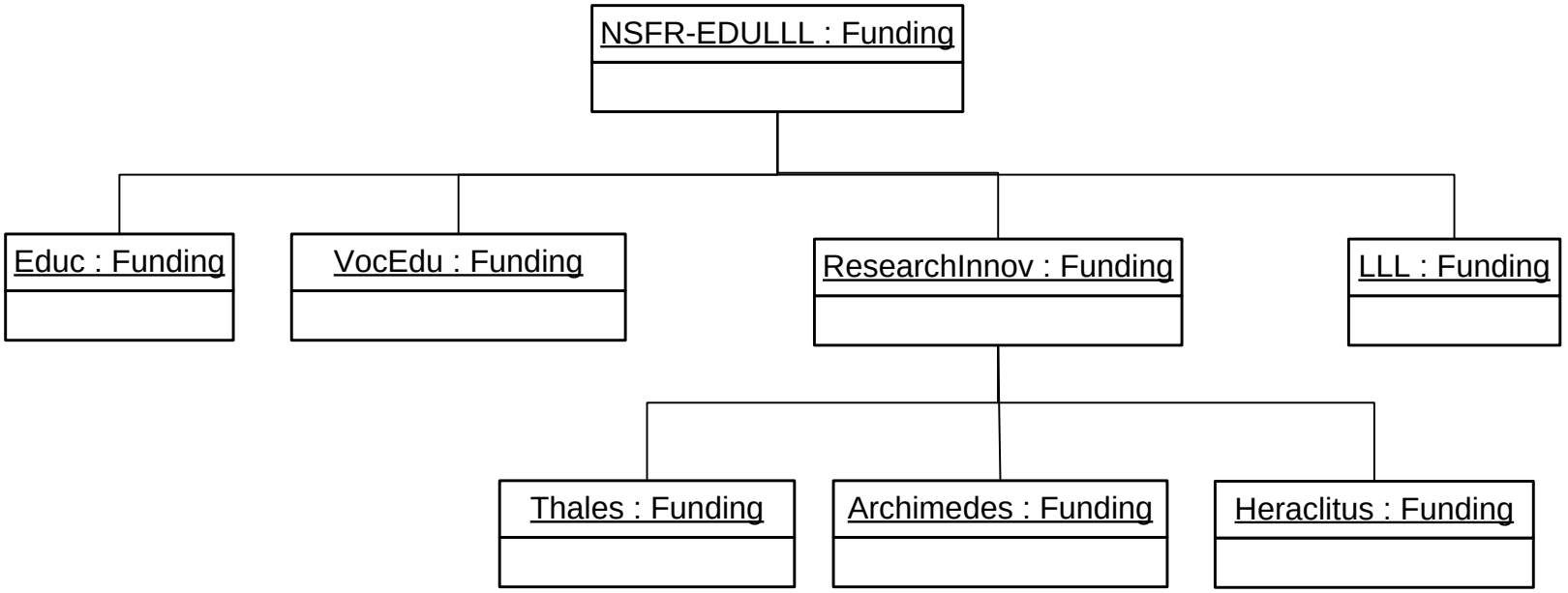
Modelling use cases in OpenAIREplus

- Representing multiple different funding structures (e.g. fron EU and national programmes) simultaneously – EU FP7



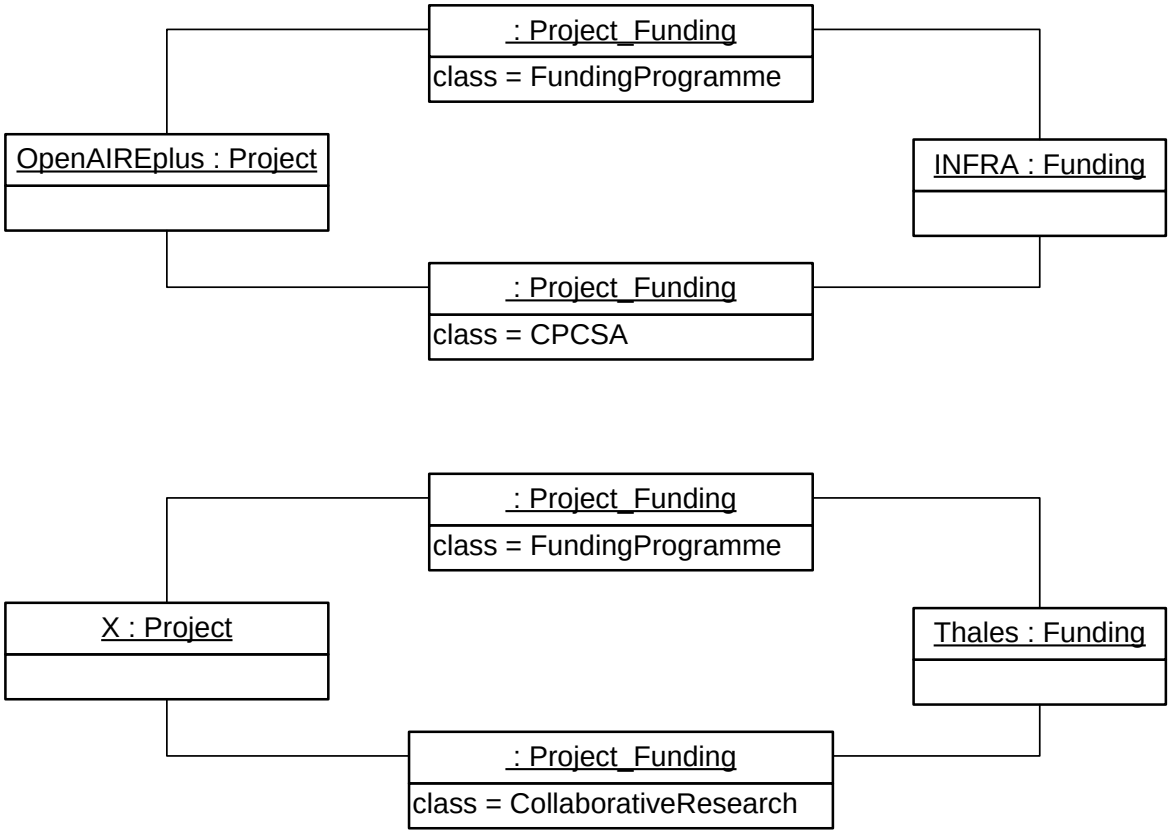
Modelling use cases in OpenAIREplus

- Representing multiple different funding structures (e.g. fron EU and national programmes) simultaneously – a Greek national funding programme



Modelling use cases in OpenAIREplus

- Linking projects to funding



Summary – future work

- The OpenAIREplus data model has been described – the core of the OpenAIRE e-infrastructure.
- Designed to address modelling of the increasingly complex scientific communications environment
- Constantly evolving research environment requires flexibility and adaptability
- Advanced functions and services in development – facilitating data curation, coping with data interference

Thank you for your attention!

- More info:

paolo.manghi AT isti.cnr.it

nhoussos AT ekt.gr

marko.mikulicic AT isti.cnr.it

brigitte.joerg@gmail.com