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
Via Moruzzi 1, 56124 Pisa, Italy
name.surname@isti.cnr.it

[2]National Documentation Centre / National Hellenic Research Foundation
48 Vassileos Konstantinou Avenue, 116 35, Athens, Greece
nhoussos@ekt.gr

[3]Innovation Support Center, UKOLN, University of Bath
BA2 7AY, Bath, UK
b.joerg@ukoln.ac.uk

[4]EuroCRIS
Cor van Osnabruggelaan 61, 2251 RE Voorschoten, The Netherlands

Abstract. The OpenAIREplus project aims to further develop and operate the OpenAIRE e-infrastructure, in order to provide a central entry point to Open Access and non-Open Access publications and datasets funded by the European Commission and National agencies. The infrastructure provides the services to populate, curate, and enrich an Information Space by collecting metadata descriptions relative to organizations, data sources, projects, funding programmes, persons, publications, and datasets. Stakeholders in the research process and scientific communication, such as researchers, funding agencies, organizations involved in projects, project coordinators, can here find the information to improve their research and statistics to measure the impact of Open Access and funding schemes over research. In this paper, we introduce the functional requirements to be satisfied and describe the OpenAIREplus data model entities and relationships required to represent information capable of meeting them.

Keywords: open access, data model, infrastructure, CERIF, DataCite.

1 Introduction

A fundamental requirement in scholarly communication systems is the representation of metadata about scientific publications. A typical approach for such metadata is based on variants of Dublin Core, MARC or MODS. A common underlying principle of these solutions is that metadata is represented in essentially flat, monolithic records

adfa, p. 1, 2011.

with limited facilities (e.g. references to authority files) for capturing relationships to autonomous entities external to the publication like persons and publishers.

Emerging needs regarding capturing, publishing and preserving research output have shown limitations of these approaches. Two increasingly significant aspects can be identified that raise the bar for metadata representation techniques.

First, scientific output that needs to be captured is not limited to traditional publications that have the form of text documents but extends to data sets. Therefore, metadata and services that enable easy discovery and reuse of data sets are essential.

Secondly, contextual metadata about scientific output is very important for the provision of value-added services to end-users. Linking publications and data sets with specific projects, funding programmes, organisations and persons enables a range of services for monitoring and assessing research activity at various levels (e.g. organisation/research group, funding programme). Furthermore, it significantly improves common functions like discovery and browsing, since the additional contextual and provenance information can assist end users in evaluating the reuse potential of research work and ultimately in reusing it. Important requirements for contextual metadata are thus the ability to unambiguously represent the semantics of the relationships between entities (e.g. whether an organisation is a participant or a coordinator of a project) as well as the temporal aspect of relationships (e.g. date range a person has been the coordinator of the project). Another critical aspect is the constantly evolving nature of the research domain, where new types of results, tools, data sources and new semantic relationships between them ask for contextual metadata that is able to seamlessly reflect these real-world changes with minimal effort.

The OpenAIREplus project [1] needs to address both of these challenging requirements [6]. The project aims to further develop the OpenAIRE e-infrastructure, which in its current initial phase provides a central point of access to open access publications funded by the European Union Framework Programme 7 projects in a range of thematic areas. The publication metadata is harvested from institutional repositories across Europe and international thematic repositories. The next, and quite ambitious, steps evolution of the OpenAIRE infrastructure to (a) include metadata describing data sets and their semantic links to publications and to (b) incorporate research output produced all over Europe through any type of funding, thus not restricted to EU FP7, including linking of outputs and projects with funding programmes.

In order to address these important challenges (data sets and contextual metadata) the need was recognised for a substantial upgrade of the data model that had been specified and used within OpenAIRE [7]. The approach followed for the upgrade was to first take into account existing initiatives and standards that could be reused. One of them is the Common European Research Information Format (CERIF) [8][9], an EU Recommendation to Member States [1] continuously developed and maintained by euroCRIS (www.eurocris.org), a standard data model addressing representation of contextual research information, which has been adopted to cover this key aspect in the OpenAIREplus data model. CERIF inherently captures contextual metadata about data sets (e.g. semantic links of data sets to publications) and is hence used in that

[1] CORDIS Archive: http://cordis.europa.eu/cerif/

respect also in OpenAIREplus. Furthermore, the OpenAIRE representation of data sets information is compliant with the metadata schema of the international DataCite[2] initiative. Notably, OpenAIREplus focuses on domain-independent contextual metadata for datasets, not handling vertical, domain-specific dataset representation. In that respect, the OpenAIREplus data modelling scope differs from other initiatives, such as the ENGAGE Public Sector Information datasets infrastructure, which employs a multi-level metadata architecture that handles to some extent both detailed, discipline-specific metadata, besides domain-independent contextual CERIF metadata [10]. Furthermore, the CERIF for Datasets (C4D) project aims at the representation of a detailed, discipline-specific datasets metadata standard using CERIF [11].

This article presents the results of the OpenAIREplus data modelling effort with particular emphasis on aspects related to representing semantically rich contextual metadata integrating information from many different research contexts (e.g. countries). It is structured as follows: Section 2 is an overview of the OpenAIREplus information space. Section 3 provides background information on CERIF and DataCite. Section 4 presents in detail the OpenAIREplus data model and elaborates on key aspects. Section 5 provides concrete examples of complex information representation requirements seamlessly addressed by the model through CERIF. The paper concludes with a summary of its main contributions and ideas for future directions.

2 The OpenAIREplus information space and data modelling requirements

As mentioned in the introduction, the OpenAIRE infrastructure is conceived to support and promote modern workflows of scientific communication. In such context research datasets become as important as textual publications and Open Access policies play a major role, to be observed and measured. The main objective of the infrastructure is therefore to deliver to all actors involved in the research process and scientific communication chain an Information Space aggregating metadata descriptions and pointers to the scientific research output, together with information relative to license and research funding. To this aim, it will offer services for the registration of data sources containing metadata descriptions of research output (e.g., publications and datasets) and their contextual information (e.g., projects, funding schemes), for the collection and aggregation of such metadata, and for inferring meaningful relationships between them. Further services will provide interested actors with portals (end-users) and APIs (third-party applications) to access to the resulting aggregated Information Space. To support the effective operation of such services, the data model of the Information Space including the entities, entity properties and entity relationships must be capable of capturing the functional requirements of such actors.

End-users. In particular, end-users may belong to the following categories:

[2] DataCite: http://www.datacite.org/

- The generic user (researcher): interested in finding publications of his/her own or other’s publications and datasets and investigate on how these are interrelated with further publication and datasets, projects, other researchers, etc.;
- The data source manager: interested in observing statistics measuring how the metadata content of his/her data source is balanced and possibly connected to others; also interested to have its content visible and linked to the original data source, so as to increase its visibility;
- The project coordinator: interested in observing the research output of his/her project and its comparison with other projects in the same subject of investigation;
- The funding agency officer: interested in measuring the impact of research funding in terms of project outputs and, in some cases, in terms of Open Access vs non-Open Access production; also interested in contacting coordinators of projects which meet specific criteria, e.g., for dissemination purposes.

In order to build services meeting such demands, the information space data model should include entities such as publications, datasets, projects, licenses, persons (e.g., authors of publications and datasets and project coordinators), data sources (e.g., source of origin of the entities), and organizations (e.g., responsible of data sources, project participants), together with relationships between them.

The process of modelling such entities has “boundaries” partly imposed by the best practices and standards adopted by the data sources from which this information can be collected from. Indeed, quality information can only be found in very specific kinds of data sources, serving the needs of well-established communities, with standards data models and services. In particular, infrastructure services will collect information from four main categories of data sources: publication repositories, data repositories, CRISs, and so-called “entity registries” – entity registries are intended as sources of authoritative lists of relevant entities such as persons, e.g., ORCID[3] for researchers, projects, e.g., European Commission CORDA database[4], and data sources, e.g., OpenDOAR[5] for repositories. The architectural assumption is that the infrastructure will collect, via several standard protocols, metadata records (e.g., XML files) from such sources assuming their compatibility with the *OpenAIREplus guidelines for data source providers*[6]. The guidelines establish, for each data source typology, the metadata format to be expected from the data sources. The format consists of an XML schema and a set of vocabularies to be used for given crucial properties (paths in the XML schema).[7] The XML schemas correspond to standard formats in the given data source application domain: Dublin Core for publication repositories,[8]

[3] ORCID, http://about.orcid.org/

[4] European Commission: COmmon Research DAta Warehouse (CORDA), https://webgate.ec.europa.eu/e-corda/resources/pdf/Confidentiality_rules_FP_data.pdf

[5] The Directory of Open Access Repositories – *Open*DOAR, http://www.opendoar.org/

[6] OpenAIREplus Guidelines: http://www.openaire.eu/en/component/attachments/download/79

[7] In general, both structure and semantics of the incoming records will be “massaged” by transformation and cleaning services, in order to ingest quality and uniform metadata records.

[8] Dublin Core will be qualified to include the representation of optional relationships with license schemes, projects, publications (e.g., citations) and datasets.

DataCite for data repositories, CERIF XML for CRIS systems, and arbitrary structured representations for entity registries. On the one hand, such formats suggest the entities and the relationships curated by domain experts and available to the information space, that is to the data model. On the other hand, the infrastructure includes services to infer new relationships by mining the information space and therefore enrich it with content not explicitly available from any data sources. In this sense, since new inference algorithms can be added to the infrastructure at any time, hence new relationships between entities can be inferred, the data model should consider the possibility to dynamically include new semantic relationships between entities, without breaking the consistency of services operating over the information space.

Applications. Third-party applications require access to the information space through standard protocols and standard exchange formats. While the first requirement has to do with the implementation of the export services according to given API specifications, the second can impact the data model definition. In fact, the more the data model is aligned with the data models of given standard export formats, the easier and straightforward it is to map information space content onto such formats. Avoiding cumbersome structural and semantics rewriting avoids maintenance issues relative to the mappings, minimizes ambiguity and loss of information due to complex mapping rules and delivers to applications data which neatly matches the one accessible through the portal. In OpenAIREplus, this requirement will be addressed by reflecting in the information space data model the entities, the properties and the relationships identified by the standard export formats adopted by the infrastructure data sources (listed above). For example, the DataCite data model finds a straightforward mapping onto the OpenAIREplus data model. A dataset metadata record is mapped onto a set of OpenAIREplus entities and relationships: the dataset entity represented by the record with relationships to persons (e.g., dataset authors) and possibly other datasets and publications. This property allows directly exporting the subparts of the OpenAIREplus information space corresponding to dataset descriptions as DataCite records, hence to channel out incoming record formats as record export format.

3 Re-using known Data Models

3.1 Research Information – CERIF

CERIF is a conceptual model of the research domain, typically applied in Common Research Information Systems (CRIS). It captures research results (publications, patents, products – the latter covering datasets, software and other types of output) as well as entities constituting the research context, like persons, organizations, projects, funding programmes, facilities, services. Every entity instance in CERIF is associated with a URI; the latest CERIF release allows for multiple federated identifiers.

A key feature of CERIF is the ability to represent semantic relationships (e.g. person-publication, organization-project, project-funding programme), including recursive links, e.g. connecting two project instances. Relationships in CERIF are called link entities and contain temporal information specifying the date range within which a specific semantic relationship applies, for example person A was coordinator of project X between 01-Feb-2012 to 29-Jun-2012. The semantics of each relationship instance (e.g. the role of a person in a project) and the associated vocabularies (i.e. potential values for roles) are not static components of the CERIF entity structure, but can be dynamically injected into a CERIF database. This is accomplished using the *CERIF Semantic Layer*, which enables the specification and maintenance of controlled vocabularies, called classification schemes, and their terms, called classes, as well as their association with entities. CERIF is able to represent any vocabulary structure (e.g. thesaurus) and the mapping among terms in different vocabularies. The semantic layer is also used to directly represent classifications of CERIF entities, example.g. terms from a subject classification vocabulary can be assigned to a publication, organisations can be typed. While a CERIF-based system is extensible to include any vocabulary, a set of common vocabularies is published as a separate component of the CERIF standard. The design and structure of the semantic layer facilitates the generation of Linked Open data from CERIF databases [12], which is being standardized by the Linked Open Data Task Group of euroCRIS.

The distinctive characteristics and modeling philosophy of CERIF can be briefly summarized as follows:

- The model is highly extensible and flexible, since a significant part of the information is not hard coded but specified through the semantic layer. The latter allows customization to a particular environment (e.g. research system of a country) or even the co-existence of specifics of different contexts (e.g. different European countries) in the same system, without the loss of CERIF compatibility, via the evolution of semantic definitions. This is particularly significant in the rapidly evolving research domain, where the emergence of new types of output, tools, research methods, funding schemes are commonplace. It facilitates also maintenance, since data schema evolution can be achieved to a large extent without changing the underlying database structure, since for example updating vocabularies does not require modifying table definitions in a relational database back-end.
- Entity properties can be modelled as relationships between entities with declared semantics specified within the semantic layer instead of data fields with the CERIF specification. This avoids the need for the proliferation of rigidly defined data fields. For instance, the creator property of a dataset can be a relationship of product with persons, while the creation date can be captured as temporal information in the “creator” relationship. In combination with the extensible semantic layer, this approach facilitates the generality of CERIF.
- Multi-linguality (field values in different languages) is inherently supported.

3.2 Research dataset modelling – DataCite

The DataCite initiative forms an international consortium addressing the challenges of making data citable in a harmonized, interoperable and persistent way. In particular DataCite supports data centers by providing persistent identifiers for datasets, workflows and standards for data publication and journal publishers by enabling research articles to be linked to the underlying data. As such, DataCite targets a wide audience and focuses on the minimal infrastructural aspects to enable cross-discipline best practices for data citation. DataCite members must assign Digital Object Identifiers[9] (DOIs) [2] to their data sets and export metadata descriptions conforming to the DataCite metadata format (data model) specification [3]. DataCite objects mandatorily include the properties: title, authors, publishing year, distributor, and persistent identifier (it is a subset of the Dataverse mandatory fields [4], without the property UNF). Such properties may be structured, e.g., creators can be more than one, have separate name separate from surname property, and may have a unique persistent identifier. Moreover, the data model includes a rich set of optional properties. For example, it includes properties to classify the data based on subject, format, typology, its access rights, language, and how it is interlinked with other datasets and publications. Many data repositories are today part of DataCite and follow its directives; for example, PANGAEA[10] (geo-referenced data from earth system research) and DANS[11] (data for social science research), which are already liaising with the OpenAIRE infrastructure, and many others. In OpenAIRE, DataCite has been adopted as the standard metadata to be used by data repository data sources to be able to contribute content to the infrastructure and its data model has been embedded into the information space data model, not to lose relevant information, and also to be able to export dataset information as DataCite metadata records. Furthermore, OpenAIRE is liaising with DataCite to exchange dataset metadata and dataset-dataset and dataset-publication relationships.

4 The OpenAIREplus data model

In order to match the aforementioned requirements the OpenAIREplus data model includes five main entities, visible as yellow boxes in see Figure 2: result (encompassing publications and datasets), person, organization, project, and data source. Furthermore, the funding entity represents funding programmes. In order to support the evolution in time of relationship-inference algorithms, the model adopts the CERIF semantic layer approach to specify semantic-agnostic relationships between publications-datasets, publications-publications, datasets-datasets, person-results, organizations-results, projects-funding, organizations-funding, funding-funding and organizations-projects. Their intended semantics will be injected at run-time, when required, thanks to a Class entity of a Scheme entity (see Figure 1).

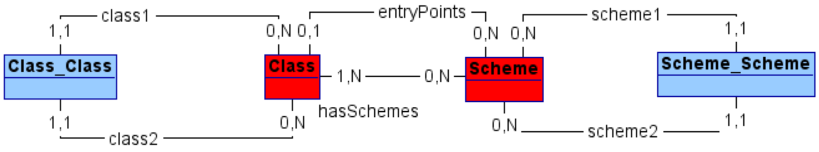

Figure 1 – E-R model: semantic layer entities

[9] *Digital Object Identifier System*, http://www.doi.org

[10] *PANGAEA*, http://www.pangaea.de

[11] *DANS*, http://www.dans.knaw.nl/

Similarly, whenever entities need to be classified based on a property value (e.g., nationality of a person), property and values are modeled by an association to a *Class* (e.g., *nationalityClass*) and one to the relative *Scheme* (e.g., *nationalityScheme*). The benefit of the approach is that applications can be written in such a way they cope with the dynamic addition, removal, or deletion of *Classes* and *Schemes*.

Result entity. The *Result* entity, depicted in Figure 3, generalizes over the concept of research output and currently includes the sub-entities *datasets* and *publications*. Datasets are intended to describe any digital object that may result from a research process or be meaningful for its completion and at the same time could be useful for others to re-use or better understand research results. Examples are scientific experimental or secondary data, sensors data, proteins, but also software products. Examples of publication types are conference and journal papers, PhD theses, technical reports, project deliverables, but also emerging "enhanced publications" [5]. Other kinds of results may be added in the future to the model, as further sub-entities (e.g., patents).

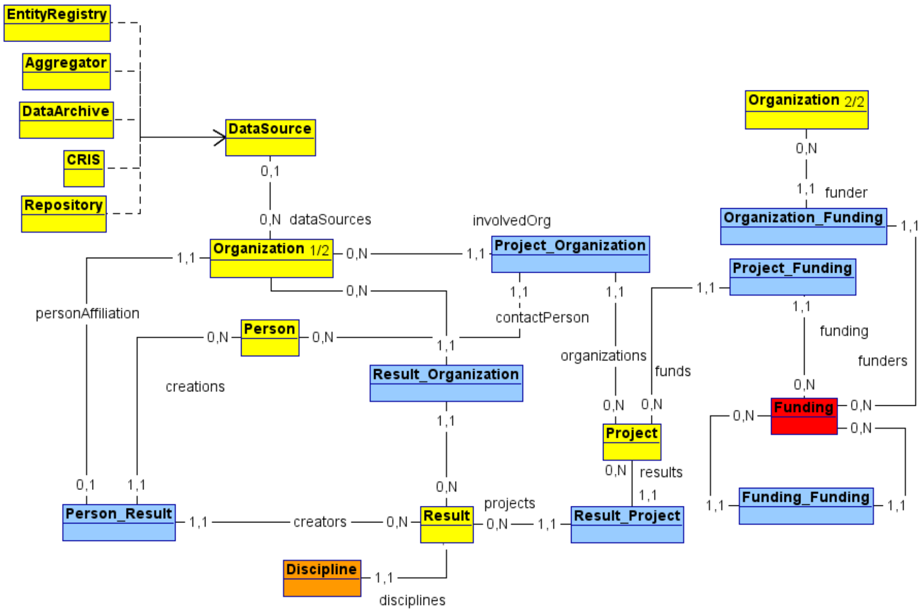

Figure 2 – E-R model: main, linked, static and structural entities

Results are also associated to a set of so-called *structural entities*, which logically describe structured properties of an entity as a hierarchy of objects "private" to a result object, i.e., not shared by other results.[12] In particular, the same result is associated to one or more *instances*. For example, the same publication may be kept in two

[12] An alternative conceptual representation could have been possible using the notion of structured property of an entity.

different repositories. Hence, an instance of a result is associated (relationship *hostedBy)* to one or more *web resources* relative to the sub-parts of the result and of the *data source* object from which such resources are made available.

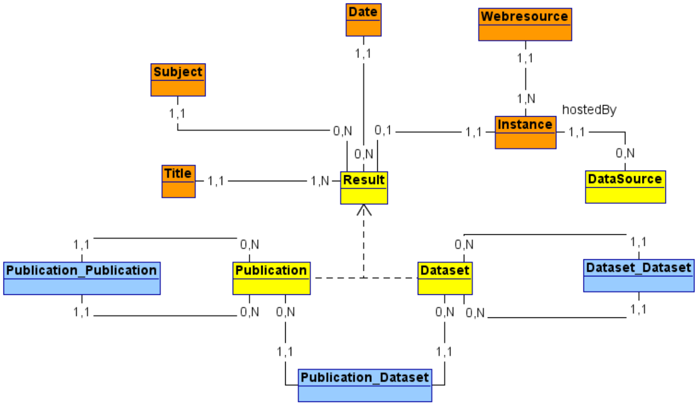

Figure 3 - E-R model: Result entities

Data source entity. The data source entity includes objects describing the data sources registered to the infrastructure and contributing with their content to populate the information space. OpenAIREplus objects of any entity are created out of data collected from various data sources of different kinds, such as publication repositories, dataset archives, CRIS systems, entity registries, and aggregators modelled as sub-entities of the data source entity. Data sources export to the OpenAIRE infrastructure information packages (e.g., XML records, HTTP responses, RDF data) which may contain information on one or more of such entities and possibly relationships between them. It is important, once each piece of information is extracted from such packages and inserted into the information space as an entity, for such pieces to be linked to the originating data source. This is to give visibility to the data source, but also to enable the reconstruction of the very same piece of information if problems arise. The model includes a relationship *collectedFrom*, which models such dependency (see Figure 4).[13] Initially, information relative to repository data sources will be collected by the OpenDOAR directory, which will act as main entity registry for (literature) repositories in Europe, but other data sources may join OpenAIREplus in the future. Analogous centralized directories for dataset archives and CRIS systems might become available; meanwhile, their administrators need to provide data about them to OpenAIREplus upon registering their data sources.

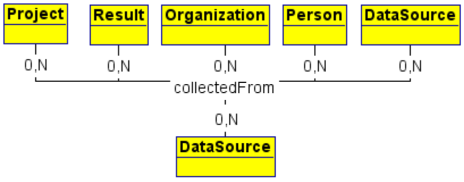

Figure 4 – E-R model: provenance relationships

Person entity. The person entity includes all person objects describing authors or persons covering roles in project management and organizations. As such, person objects are mainly created out of data source information packages (e.g., Dublin Core records from repositories) which do not provide rich properties and unique identifiers. OpenAIRE is liaising with ORCID, which will register as an entity registry to feed and fetch clean authoritative information. There are also plans for OpenAIRE to exchange publication-author links and license information with Mendeley.[14]

Project entity. The project entity includes objects describing funding resources (co-)granted by funding agencies, such as the European Commission or National governments. Of crucial interest to OpenAIREplus is also the identification of the funding programmes (called *funding* in Figure 2) which co-funded the projects that have led to a given result. Initially, EC FP7 programme projects data will be fetched from the authoritative EC CORDA database, together with the organizations or persons which are participants of such projects. Data relative to National funding schemes and relative projects will be instead fetched from CRIS systems, together with other entities which may be typically kept within a CRIS system (e.g., publications, datasets, etc.).

Organization entity. The organization entity includes objects describing companies, research centers or institutions involved as project partners or as responsible of operating data sources. Information about organizations will be initially collected from the information packages collected from the entity registry of CORDA and various CRIS systems. For the future, OpenAIREplus is liaising with UK RepositoryNET+[15] to open and exchange of entity registries including organizations and authors.

5 Modelling use cases within OpenAIREplus

The present section provides a characteristic example of the capabilities of the OpenAIREplus data model, in particular accommodating data about national funding schemes across Europe and diverse funding programme structures from different countries, and provides links to other entities (e.g. projects, organizations). To address this, funding programmes and funding programme components are represented as instances of the Funding entity, while the recursive Funding_Funding link entity enables the representation of arbitrarily complex funding programme structures, for example hierarchies of any depth or even graphs, using an appropriate vocabulary for the classification of each instance of this relationship. The most common class term is currently “Part” upon the Funding_Funding link entity; in the CERIF vocabulary this denotes that a funding programme is a sub-programme of another one.

As an example, Figure 5 depicts – in a highly simplified form – the representation in OpenAIREplus of a part of the European Commission Framework Programme Seven (FP7), which comprises five sub-programmes, each containing many subdivisions. For instance, the Capacities sub-programme has 6 subdivisions. It is presented as a UML Object Diagram, where each box is an instance of the Funding entity. Due to the economy of presentation the Funding_Funding entities in the figure

[14] Mendeley, http://www.mendeley.com/

[15] UK RepositoryNET+, http://www.repositorynet.ac.uk

appear only as lines connecting these instances. Figure 6 shows explicitly the link entity between two Funding instances with the class term specifying the relationship semantics (classification scheme values and timestamps are omitted for simplicity).

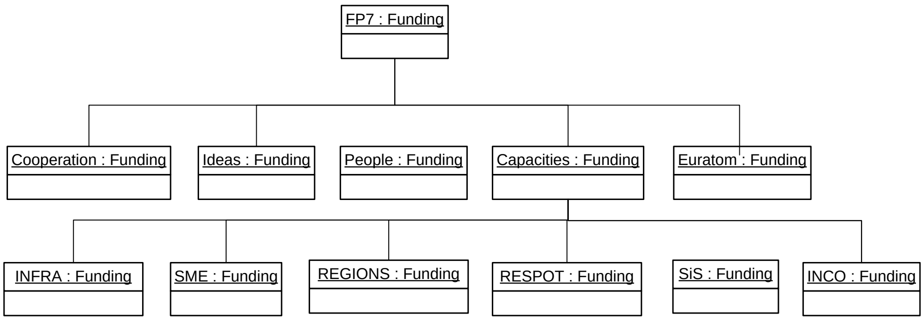

Figure 5. OpenAIREplus data model: fragment of the FP7 funding programme structure.

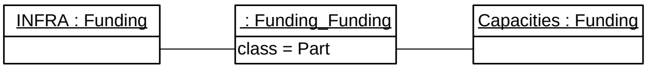

Figure 6. Funding_Funding relationship with declared semantics (highly simplified).

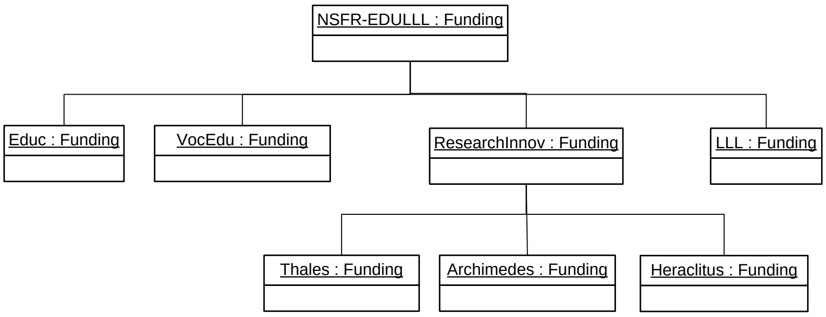

Figure 7. OpenAIREplus data model: fragment of the Greek EDULLL funding programme structure.

Figure 7 presents the entity instances used to represent in OpenAIREplus a Greek national funding programme, supporting education, lifelong learning and research. Lines between instances of Funding correspond also in this case to Funding_Funding “Part” linkages. The two funding structures in Figures 5 and 7 smoothly co-exist in the OpenAIREplus data model and are linked with other entities without requiring any changes in the logical and physical model of the underlying relational database.

As an example of linking other entities to Funding, a project is connected with funding programmes through the Project_Funding link entity. Such a link is shown in Figure 8, where the OpenAIREplus project instance is related to the Research Infrastructures (INFRA) FP7 programme sub-division via two instances of the Project_Funding link entity: one stating that the INFRA programme is the FundingProgramme of OpenAIREplus (i.e. OpenAIREplus is funded by INFRA) and that OpenAIREplus, in terms of funding instrument, is a Combination of Collaborative Project and Coordination and Support Action (CPCSA). The class term may take values from a specific classification scheme that contains terms for all possible instruments. In a similar example a project X, of type CollaborativeResearch, is connected to a sub-division of the EDULLL programme called Thales.

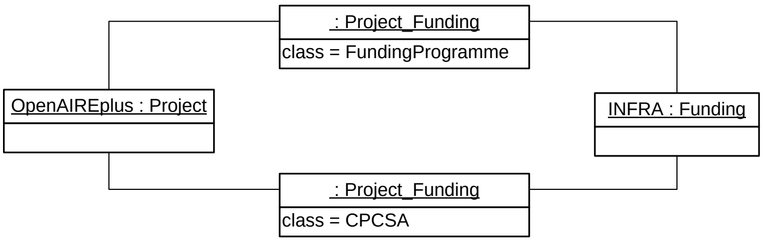

Figure 8. Example relationships of Project to the FP7 Research Infrastructures programme.

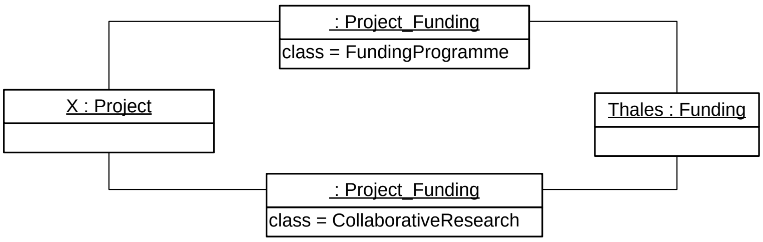

Figure 9. Example relationships of Project to the Greek EDULLL funding programme.

6 Summary and future work

In this paper we presented the OpenAIREplus data model, at the core of the OpenAIRE infrastructure. The intention was to focus on the informational aspects of the data model, hence on the requirements that led to the definition of its constituent parts. To this aim highlighted its modern scientific communication flavour in combination with the funding issues surrounding the research process. More, we stressed its flexibility and ability to cope with evolving requirements, in terms of entities to be modelled and relationships between them. Part of the data model, but out of the scope of this paper, is the additional scaffolding of entities, properties and relationships required to cope with data curation issues. The continuous synchronization of the information space with the data sources, removes, deletes, and updates information about one or more of the entities, as well as relationships between them. For example, some archives provide dataset metadata that also includes links to publications relevant for the dataset or vice versa. Such cross-entity and cross-sources data integration brings in data inference issues, which have mainly to do with information absence, duplication, and versioning (intended as replicas of the same entity). These issues will push into the data model a number of entities, properties, and relationships whose aim is to deliver to data curators the tools to maintain a clean, uniform, and consistent information space.

Acknowledgements. The work presented in this paper has been partly supported by the OpenAIREplus Project (Ref No: 283595) of the EU-funded FP7-INFRASTRUCTURES Programme. The authors wish to acknowledge the contribution of Keith Jeffery, Anne Asserson, Claudio Atzori, and Jochen Schirrwagen to the ideas and work presented in the article.

7 References

1. OpenAIREplus Project (2012), www.openaire.eu.
2. Natasha Simons, Implementing DOIs for Research Data, D-Lib Magazine,, Volume 18, Number 5/6, May/June 2012. doi:10.1045/may2012-simons
3. Joan Starr and Angela Gastlis, CitedBy: A Metadata Scheme for DataCite, D-Lib Magazine, January/February 2011, Volume 17, Number 1/2, doi:10.1045/january2011-starr
4. M. Altman and G. King "A Proposed Standard for the Scholarly Citation of Quantitative Data". D-Lib Magazine March/April 2007.
5. S. Woutersen-Windhouwer, R. Brandsma, P. Verhaar, A. Hogenaar, M. Hoogerwerf, P. Doorenbosch, E. Durr, J. Ludwig, B. Schmidt, B. Sierman, "Enhanced Publications", edited by M. Vernooy-Gerritsen, SURF Foundation, Amsterdam University Press, 2009
6. Paolo Manghi, Natalia Manola, Wolfram Horstmann, and Dale Peters. An Infrastructure for Managing EC Funded Research Output, The OpenAIRE Project. International Journal on Grey Literature (TGJ), 6(1), Spring 2010
7. Paolo Manghi. OpenAIRE Data Model Specification. Deliverable D5.1. Funded in call INFRA-2007-1.2.1, Grant Agreement Number 246686, May 2010.
8. Keith Jeffery and Anne Asserson. CERIF-CRIS for the European e-Infrastructure. Data Science Journal, 9:CRIS1–CRIS6, 2010.
9. Brigitte Jörg. CERIF: The Common European Research Information Format Model. Data Science Journal, 9:CRIS24–CRIS31, 2010.
10. N. Houssos, B Jörg, B Matthews. A multi-level metadata approach for a Public Sector Information data infrastructure. Proc. 11th International Conference on Current Research Information Systems (CRIS2012), Prague, Czech Republic, 06-09 Jun 2012.
11. K. Ginty, S. Kerridge, P. Fairley, R. Henderson, P. Cramer, A. Bokma, S. Garfield, CERIF for Datasets (C4D) - An Overview. Proc. 11th International Conference on Current Research Information Systems (CRIS2012), Prague, Czech Republic, 06-09 Jun 2012.
12. B. Jörg, et al. Connecting Closed World Research Information Systems through the Linked Open Data Web. International Journal of Software Engineering and Knowledge Engineering (IJSEKE), Volume 22, June, 2012.