The OpenAIRE Workflows for Data Management

Claudio Atzori(✉), Alessia Bardi, Paolo Manghi, and Andrea Mannocci

Istituto di Scienza e Tecnologie dell'Informazione, "A. Faedo" - CNR, Pisa, Italy
{claudio.atzori, alessia.bardi, paolo.manghi, andrea.mannocci}@isti.cnr.it

Abstract. The OpenAIRE initiative is the point of reference for Open Access in Europe and aims at the creation of an e-Infrastructure for the free flow, access, sharing, and re-use of research outcomes, services and processes for the advancement of research and the dissemination of scientific knowledge. OpenAIRE makes openly accessible a rich Information Space Graph (ISG) where products of the research life-cycle (e.g. publications, datasets, projects) are semantically linked to each other. Such an information space graph is constructed by a set of autonomic (orchestrated) workflows operating in a regimen of continuous data integration. This paper discusses the principal workflows operated by the OpenAIRE technical infrastructure in its different functional areas and provides the reader with the extent of the several challenges faced and the solutions realized.

Keywords: Aggregation · Workflows · e-Infrastructure · Metadata · Open science · Open access · De-duplication · Data mining · Information space

1 Introduction

The OpenAIRE initiative is the point of reference for Open Access in Europe [3, 4]. Its mission is to foster an Open Science e-Infrastructure that links people, ideas and resources for the free flow, access, sharing, and re-use of research outcomes, services and processes for the advancement of research and the dissemination of scientific knowledge. OpenAIRE operates an open, collaborative, service oriented infrastructure that supports (i) the realization of a pan-European network for the definition, promotion and implementation of shared interoperability guidelines and best practices for managing, sharing, re-using, and preserving research outcomes of different typologies; (ii) the promotion of Open Science policies and practices at all stages of the research life-cycle and across research communities belonging to different application domains and geographical areas; (iii) the provision of measurements of the impact of Open Science and the return of investment of national and international funding agencies; (iv) the development and operation of a technical infrastructure supporting services for the discovery of and access to research outcomes via a centralized entry point, where research outcomes are enriched with contextual information via links to objects relevant to the research life-cycle. This paper focuses on the workflows operated by the

C. Grana and L. Baraldi (Eds.): IRCDL 2017, CCIS 733, pp. 95–107, 2017.
DOI: 10.1007/978-3-319-68130-6_8

OpenAIRE technical infrastructure for the management of the OpenAIRE information space. The OpenAIRE technical infrastructure includes services dedicated to the aggregation of information about objects of the research life-cycle. In order to help repository managers to integrate their data with OpenAIRE, the guidelines (https://guidelines.openaire.eu) describe how to expose such information (publications, datasets, CRIS metadata) via the OAI-PMH protocol. Relationships between objects are collected from the data sources, but also automatically detected by inference algorithms [1] and added by users, who can insert links between publications, datasets and projects via the claiming procedure available from the OpenAIRE web portal. The Information Space is available for human and machine consumption via the OpenAIRE web portal and different kinds of APIs. Among the challenges emerging in this scenario, one is relative to the orchestration of the different workflows characterizing the OpenAIRE system. In fact, a key factor for its sustainability is represented by the system capability of being autonomic and extensible, i.e. the possibility to easily define and implement autonomous workflows. The OpenAIRE workflows orchestration is delegated to D-NET, a software toolkit for constructing and operating aggregative infrastructures in a cost-effective way as instances of service-oriented data infra-structures [6].

Outline. The following sections describe the OpenAIRE technical infrastructure by introducing the OpenAIRE data model (Sect. 2.1) and the general system architecture (Sect. 2.2). The remaining sections introduces the infrastructure workflows, intended as both automated and human activities aimed to (i) aggregate content (metadata and full-text) (Sect. 3), (ii) populate the OpenAIRE ISG (Sect. 4), (iii) de-duplicate it (Sect. 5), (iv) infer new valuable information from the full-text files (Sect. 6), (v) monitor and publish the ISG in order to make it available to both end users on the portal and third party services via the OpenAIRE API (Sect. 7).

2 OpenAIRE Technical Infrastructure

In this section, we introduce the OpenAIRE technical infrastructure by describing the OpenAIRE data model, and the general architecture of the system.

2.1 The OpenAIRE Data Model

The OpenAIRE technological infrastructure provides aggregation services capable of collecting content from data sources available on the web in order to populate the so-called OpenAIRE Information Space, a graph-like information space (ISG - Information Space Graph) describing the relationships between scientific articles, their authors, the research datasets related with them, their funders, the relative grants and associated beneficiaries. By searching, browsing, and post processing the graph, funders can find the information they require to evaluate research impact (i.e. return on investment, RoI) at the level of grants and funding schemes, organized by disciplines and access rights, while scientists can find the Open Access versions of scientific trends of interest. The ISG is then made available for programmatic access via several APIs (Search HTTP APIs, OAI-PMH, and Linked Open Data), for search, browse and statistics consultation via the OpenAIRE portal, and soon for data sources with the

Literature Broker Service [8]. The graph data model is inspired by the standards for research data description and research management (e.g. organizations, projects, facilities) description provided by DataCite and CERIF, respectively. Its main entities are Results (datasets and publications), Persons, Organizations, Funders, Funding Streams, Projects, and Data Sources:

Results are intended as the outcome of research activities and may be related to Pro-jects. OpenAIRE supports two kinds of research outcome: *Datasets* (e.g. experimental data) and *Publications* (other research products, such as Patents and Software will be introduced). As a result of merging equivalent objects collected from separate data sources, a Result object may have several physical manifestations, called instances; instances indicate URL(s) of the payload file, access rights (i.e. open, embargo, restricted, closed), and a relationship to the data source that hosts the file (i.e. provenance).

Persons are individuals that have one (or more) role(s) in the research domain, such as authors of a Result or coordinator of a Project.

Organizations include companies, research centers or institutions involved as project partners or that are responsible for operating data sources.

Funders (e.g. European Commission, Wellcome Trust, FCT Portugal, Australian Research Council) are Organizations responsible for a list of Funding Streams (e.g. FP7 and H2020 for the EC), which are strands of investments. Funding Streams identify the strands of funding managed by a Funder and can be nested to form a tree of sub-funding streams (e.g. FP7-IDEAS, FP7-HEALTH).

Projects are research projects funded by a Funding Stream managed by a Funder. Investigations and studies conducted in the context of a Project may lead to one or more Results.

Data Sources, e.g. publication repositories, dataset repositories, journals, publishers, are the sources on the web from which OpenAIRE collects the objects populating the OpenAIRE graph. Each object is associated to the data source from which it was collected. More specifically, in order to give visibility to the contributing data sources, OpenAIRE keeps provenance information about each piece of aggregated information. Since de-duplication merges objects collected from different sources and inference enriches such objects, provenance information is kept at the granularity of the object itself, its properties, and its relationships. Object level provenance describes the origin of the object consisting of the data sources from which its different manifestations were collected. Property and relationship level provenance tells the origin of a specific property or relationship when inference algorithms derive these (e.g. algorithm name).

2.2 General Architecture

The OpenAIRE system depicted in Fig. 1 illustrates the system architecture from a high-level perspective, highlighting the data flows occurring within the subsystem, conceived as decoupled components. The aggregator is intended as the set of services responsible for the collection, validation, semantic and structural transformation of the metadata records, and the collection of the full-texts relative to the Open Access publications. The data provision pipeline consists of (i) a mapping layer used to

populate the ISG, and (ii) the Action Manager Service, the implementation of a framework responsible for the management of the enrichments introduced to the ISG. They can be new nodes of the graph, property of existing nodes, or relationships among nodes. Such *Actions* and are organized in *Action Sets*, a logical container for all the *Actions* produced by a given process. The system associates dedicated Action Sets to the processes contributing at the ISG enrichment, such as deduplication, and the different mining algorithms described in the followings. All the components described in Fig. 1 are defined as decoupled subsystem, and in order to realize the data management workflows OpenAIRE relies on, the orchestration mechanism is provided by the D-NET software toolkit.

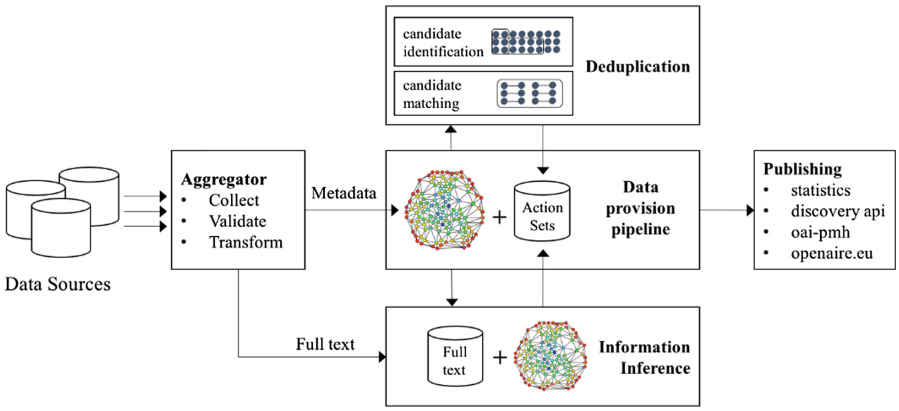

Fig. 1. High level architecture.

3 Content Aggregation Workflow

OpenAIRE aggregates metadata and full-texts according to a well-defined content acquisition policy: metadata records and full-texts of open access publications, metadata records of publications funded by EC projects or national funding schemes, metadata records about datasets that are outcomes of a funded research project or related to a publication already in the OpenAIRE ISG [12]. To ensure a minimum level of quality of the aggregation, OpenAIRE requires data sources to comply with the OpenAIRE guidelines. The OpenAIRE aggregation services support the OpenAIRE data managers in implementing the content acquisition policy and in supervising the content aggregation activity. This consists of (i) registration of a new data source, (ii) validation of its content with respect to the OpenAIRE guidelines, (iii) configuration of existing data sources in terms of access parameters and workflow scheduling, (iv) configuration of the rules for transforming input objects according to the OpenAIRE data model, (v) monitor and tracking the history of workflow executions.

3.1 Metadata Aggregation

Objects and relationships in the OpenAIRE ISG are extracted from information pack-ages, i.e. metadata records, collected from data sources of the following kinds:

Institutional or thematic repositories (aggregated 543). Information systems where scientists upload the bibliographic metadata and PDFs of their articles, because of either obligations from their organization or community practices (e.g. ArXiv, Europe PMC).

Open Access Publishers and journals (aggregated 6676). Information system of open access publishers or relative journals, which offer bibliographic metadata and PDFs of their published articles.

CRIS (aggregation starting by end of 2017). Information systems adopted by research and academic organizations to keep track of their research administration records and relative results; examples of CRIS content are articles or datasets funded by projects, their principal investigators, facilities acquired thanks to funding, etc.

Data archives (aggregated 59). Information systems where scientists deposit descriptive metadata and files about their research data (also known as scientific data, datasets, etc.); data archives are in some cases supported by research and academic organizations and in some cases supported by research communities and/or associations of publishers.

Aggregator services (aggregated 16). Information systems that, like OpenAIRE, collect descriptive metadata about publications or datasets from multiple sources in order to enable cross-data source discovery of given research products; aggregators tend to be driven by research community needs or to target the larger audience of researchers across several disciplines; examples are DataCite for all research data with DOIs as persistent identifiers, BASE for scientific publications, DOAJ for OA journals publications.

Entity Registries (aggregated 13). Information systems created with the intent of maintaining authoritative registries of given entities in the scholarly communication, such as OpenDOAR for the institutional repositories or re3data for the data repositories.

As of December 2016, OpenAIRE aggregates about 21 million of information pack-ages describing publications and datasets. OpenAIRE features three workflows for metadata aggregation: (i) for the aggregation from data sources whose content is known to comply with the OpenAIRE content acquisition policy, (ii) for the aggregation of content that is not known to be eligible according to the policy, (iii) for the aggregation of information packages from entity registries.

Workflow for OpenAIRE compliant data sources. This workflow is for data sources that comply with the OpenAIRE guidelines and thus it is executed for the majority of data sources. The workflow consists of three phases: collection, validation, and transformation. The collection phase collects information packages in form of XML metadata records from an OAI-PMH endpoint of the data source (as the OpenAIRE guidelines mandate) and stores them in a metadata store. The validation phase is an optional phase that can be enabled to validate the collected metadata records according to the OpenAIRE guidelines. Finally, the transformation phase transforms

the collected records according to the OpenAIRE data model and stores them in another metadata store, ready to be read for populating the OpenAIRE ISG.

Workflow for data sources with unknown compliance. This workflow applies to data sources that are registered into OpenAIRE but are not known to be OpenAIRE compliant. This is the typical case for aggregators of data repositories (e.g. Datacite). According to the content acquisition policies, OpenAIRE can include a dataset into the ISG only if it has a link to an object (project or publication) already in the ISG. Therefore, OpenAIRE collects all metadata records and transforms them according to the OpenAIRE data model, but the records are marked so that the ISG population workflow will not use them for the creation of the ISG. In fact, the inference workflow (see Sect. 6) will use those objects and will add to the ISG only those that have been detected as eligible according to the content acquisition policy.

Workflow for entity registries. This workflow applies to data sources offering authoritative lists of entities. The workflow consists of two phases: collection and transformation. The collection phase collects information packages in the form of files in some machine-readable format (e.g. XML, JSON, CSV) via one of the supported exchange protocols (OAI-PMH, SFTP, FTP(S), HTTP, REST). The transformation phase transforms the packages according to the OpenAIRE data model and stores them into a metadata store ready to be read for populating the OpenAIRE ISG.

3.2 Full-Text Aggregation

The full-text aggregation workflow has a twofold goal: (i) collect and store the files described by publication metadata records, and (ii) extract their full-texts so that they can be used by full-text mining algorithms (Sect. 6). When collecting a file, it is crucial to preserve the association between the file and the corresponding metadata record. This association plays a crucial role in the inference workflow as it determines the possibility to correctly associate the inference results produced by mining a given full-text, to the corresponding object in the OpenAIRE ISG. While in case of metadata records describing publications the aggregation system can rely on well-established formats and exchange protocols such as Dublin Core [7] and OAI-PMH [8] respectively, in case of full-text files the aggregation system often needs to crawl the landing page referred in a metadata record to discover the link to the actual file. The full-text collection system is therefore designed to be extensible with new plugins, capable to manage specific html page structures or to be configured to recognize specific URL patterns.

The large majority of full-texts collected by the system are PDF files, a for-mat well suited for printing and human reading, but less tractable by machines. For this reason, the full-text collection workflow includes a final phase designed to automatically extract structured metadata from such PDF files using CERMINE [2]. The extracted full-texts are then stored in dedicated caches that are accessible by the OpenAIRE Information Inference System. As of December 2016, OpenAIRE collects about 4.5 million of full-text files responding to different formats: PDF, JATS, HTML.

4 Information Space Graph Population

An information package collected from a data source is a file in some machine-readable format (e.g. XML, JSON, CSV), which contains a data source-assigned identifier (mandatory) and information (e.g. properties) relative to one or more primary object. Beyond the primary object, an information package may contain information (but not necessarily the identifier) relative to other entities, called derived objects, which must be directly or indirectly associated with the primary object. Such association represents a link between the objects, which collectively form the OpenAIRE Information Space Graph. For example, a Dublin Core bibliographic metadata record describing a scientific article will yield one OpenAIRE result object (of Publication typology) and a set of OpenAIRE person objects (one per author) with relationships between them. In OpenAIRE we opted for representing the ISG with an adjacency list, as we believe this choice can cope well with a large class of scenarios. The storage system identified to persist the ISG is Apache HBase [17]. By supporting horizontal scalability and featuring full support for the Hadoop MapReduce framework, its columnar storage system is well suited to persist and process the adjacency list exploiting the parallelism offered by the MapReduce framework.

As of December 2016, the OpenAIRE ISG counts about 21 million publications, 600,000 projects, 30,000 datasets, 80,000 organizations, 16 million persons, 18,000 data sources and more than 90 million of relationships.

5 Information Space De-duplication

The OpenAIRE ISG possibly contains, by construction, different objects representing the same publication. In fact, metadata about one publication can be collected from different data sources. For the disambiguation of publications in the graph, OpenAIRE features a de-duplication system based on the GDup software [16] implementing a workflow in three phases: (i) candidate identification: considering the number of publication objects participating to the graph (about 21 million), matching all pairs of publications to identify the duplicates is by no means feasible: heuristics are needed to compare only publications that are likely to be duplicates; (ii) candidate matching: once the candidates are identified, their properties are compared and a similarity mark assigned; (iii) graph disambiguation: groups of duplicates are identified and, for each group, one unique publication is created to represent all members. In the following paragraphs, the three phases are explained in details with respect to the challenges posed by the de-duplication of publications. However, the system is configured to treat the same problem also for the organization entities, aggregated from different sources, and suffering from the same duplication issue.

5.1 Candidate Identification

Matching all possible pairs of 21 million publications is by no means tractable. To address this issue, candidate identification is the phase entitled of providing the heuristics and technological support necessary to avoid such “brute force” solution.

Candidate identification is solved using clustering techniques based on functions that associates to each publication one or more key values, out of its properties, to be used for clustering. The idea is that publications whose keys fall in the same cluster are more likely to be similar than across different clusters. This action narrows down the number of pairwise matches to perform within the clusters of publications, thereby reducing the complexity of the problem. Ideally, the definition of a good clustering function for de-duplication should avoid false negatives (i.e. making sure that obvious duplicates fall in the same cluster), avoid false positives (i.e. making sure that clearly different publications do not fall in the same cluster), and make the number of matches to be tractable for the technology at hand. The definition of a good clustering function for de-duplication of publications starts from the properties available in the publications metadata. From the analysis of the publication properties, the only always present and informative enough is the title. Clustering publications starting from their title may be done according to different strategies, which avoid or tolerate minor differences in the values, typically caused by typos or the partial or full presence of words. Some examples are: removing stop words, blank spaces, etc.; lower-casing all words; using combination of prefixes or postfixes of title words; using n-grams of relevant words; using hashing functions. Using any of these strategies has implications that depend on the features of titles in the ISG. For example, the heavy presence of short titles (consisting at most of a few short words) may find in the hashing function a better solution than using prefixes of words. On the other hand, the adoption of high performance technologies may allow for a greedier approach, which allows for more matches to be performed hence avoid false negatives.

The OpenAIRE ISG is very heterogeneous as both data sources and disciplines behind publications are of different kinds. As a consequence, the preferred approach is the one that combines the first letters of words (like an acronym) into a clustering key and the last letters of words into another clustering key. The approach is quite typo-safe and proves to exclude the majority of false negatives, on the other hand it includes false positives, which shall be excluded with the subsequent detailed similarity match.

5.2 Candidate Matching

The method described above is well known in record linkage literature as Blocking. It is well suited to address the de-duplication problem in large datasets [14, 15], and to further narrow the number of pairwise comparisons can be followed by the so-called Sliding Window method. The sliding window is based on the idea that publications in the cluster are ordered according to a defined function, to maximize the probability that similar publications are as close as possible. Publications in the cluster are then pairwise matched only if they are part of the same sliding window of length K. When all the publications have been matched, the sliding window is moved to the next element of the ordering and a new set of pairs is matched. Sliding windows introduce false negatives, since they exclude from the match publications in the same cluster, but control performance (especially in terms of memory and execution time) by setting an upper bound to the number of matches in each cluster. In order to define a solid candidate matching function, we need to identify which properties are most influential in the matching process, i.e. those that best contribute to establish publication

equivalence introducing lower computational cost, while allowing clear cut decisions, and are often present in the publications. As in the previous phase of candidate identification, the title is again a good choice: it is present in (almost) all objects and consists of a relatively short text, which can be fast and reliably processed by known string matching functions. In general, if the titles of two publications are not similar "enough" (according to a given threshold) then no other property-to-property match may revise this decision. Conversely, sufficient similarity in titles (or even equivalence) alone is not enough as one of the following cases may occur: (i) very short titles, composed of few, commonly used words may lead to obvious equivalence; e.g. the title "Report on the project XYZ" may be recurrent, the only difference being the name "XYZ" of the project; (ii) recurrent titles; e.g. the title "Introduction" of some chapters is very common and introduces ambiguity in the decision, and (iii) presence of numbers in titles of different published works; e.g. the title "A Cat's perspective of the Mouse" is likely referring to a publication different from "A Cat's perspective of the Mouse v2", but not different from "A Cat perspective of the Mouse"; As a consequence, the decision process must be supported by further matches that may strengthen the final conclusion, possibly based on one or more of the following publication properties: author names, date of acceptance, abstract, language, subjects, PID. Those are all features that could contribute to the matching on different levels, however their contribution mostly depends on data quality. In OpenAIRE case, PIDs are significantly contributing to the matching process. Unluckily they are present only in a subpart of our publication objects (between 30–40%), but on the bright side they contribute allowing to take strong decisions on the equivalence of two publications: if two publications provide the same DOI they are indeed duplicates. Therefore, a similarity function, based on the availability of certain properties can take straightforward decisions on equivalence or difference between publications, while in other cases can only come up with a rank of confidence that depends on the availability and weights of the properties above.

5.3 Graph Disambiguation

Duplicate identification terminates providing a set of pairs of duplicate publications. In order to disambiguate the ISG, duplicates should be hidden and replaced by a "representative object" that links to the duplicates it represents (and vice versa). The representative publication becomes the hub of all incoming and outgoing relationships relative to the publications it hides. As a result, the graph is disambiguated but still keeps track of its original topology, hence allowing data managers to measure the duplicates percentage for a given data source. The graph disambiguation phase consists of two steps: duplicates grouping and duplicates merging. Grouping duplicates requires the identification of the connected components formed by the equivalence relationships identified by duplicate identification. Merging the groups of duplicates requires instead the creation of a representative publication for each connected component (or group of duplicates) and the propagation towards this new object of all incoming and outgoing relationships of the object it merges. Both actions have serious performance implications, which depend on the topology of the graph (e.g. fragmentation and density of graph, edge distance of publications in the graph, number of the duplicates). For

example, the number of duplicated publications depends on the replication of the publication across different data sources, e.g. institutional repository of the author, thematic repository, and a number of aggregators, but it is in general not very high (e.g. co-authors, each depositing in their respective institutional repositories which are in turn harvested by OpenAIRE).

6 Information Inference Workflow

The OpenAIRE information inference system (IIS) is based on an instantiation of the Information Inference Framework (IIF) [1]. The IIS is responsible for enriching the ISG with new information produced by various types of data mining algorithms. The inference workflow has been designed to work on a snapshot of the entire ISG, in this sense the IIS is a stateless subsystem, whose outcome can be regenerated from scratch on each run. Each IIS inference algorithm is associated to an Action Set, i.e. a logical container that stores and versions the inference results. Such results consist of new objects (publications and datasets), new properties that enrich existing objects (such as document classification properties and citation lists), and semantically typed relationships among objects in the ISG. The IIS results versioning supported by Action Sets allows, in case of regressions in one or more mining algorithms, to reuse previously generated results without requiring to (i) rollback an algorithm to one of its previous versions, (ii) reintegrate it in the IIS, (iii) re-execute it to obtain consistent results. The inference workflow has been divided in two distinct phases: pre-processing and primary. The pre-processing phase supports the implementation of the OpenAIRE acquisition policy [12], according to which (i) a non-Open Access publication can be included into the ISG if it is funded by a project already in OpenAIRE; (ii) a dataset can be included into the ISG only if it is related to a publication already in OpenAIRE. To this aim the pre-processing phase includes the algorithms tailored to infer links between publications and projects, and between publications and datasets. The IIS main phase instead, operates over the ISG and executes all the full-text mining algorithms described in the Table 1 below, which summarizes the inference results produced in November 2016.

Table 1. Summary of IIS results, Nov 2016.

Phase	Description	Count
Pre-processing	Dataset references found in publications	88.592
Primary	Dataset references found in publications	78.086
	Publications enriched by protein data bank references	43.586
	Protein data bank references	196.462
Pre-processing	Project references found in publications	88.978
Primary	Project references found in publications	351.302
Primary	Citation references	15.319.346
	Publications enriched by citation references	2.632.059

(*continued*)

Table 1. (*continued*)

Phase	Description	Count
Primary	Software references	21.481
	Publications enriched by software references	15.592
Primary	Similarity references	164.602.477
Primary	Publications enriched by document classes ArXiv, Mesh, ACM	2.405.869

7 Information Publishing

The result of the workflows described in previous sections (content aggregation, ISG population, de-duplication and inference) are materialized by the data publishing workflow into four ISG projections: (i) a full-text index to support search and browse queries from the OpenAIRE Web portal and to expose subsets of the ISG on the OpenAIRE search API, (ii) a E-R database and a dedicated key-value cache for statistics, (iii) a NoSQL document storage in order to support OAI-PMH bulk export of subsets of the ISG in XML format, and finally (iv) a triple store in order to expose the ISG as LOD via a SPARQL endpoint. Every time the data publishing workflow executes, four new ISG projections are generated and persisted in a "pre-public status" before being accessible from the general public. The switch from pre-public to public, meaning that the currently accessible ISG projections and statistics will be dismissed and the new versions will take their place, is still manual for safety reasons. Whenever new pre-public ISG projections (pre-public ISG) are created, it is important to verify some constrains in order to evaluate whether the switch to public can be performed or some regressions in the overall data quality needs to be addressed first. Some constraints to be ensured regard the control of quality metrics extracted from the different projections of the ISG and may involve one or more projections (e.g. threshold checks, alignment of different ISG projections); other conditions regard instead the trend throughout time of such quality metrics (e.g. whether a certain trend is monotonic increasing/decreasing or not) and may involve one or more trends extracted from different projections.

The number of quality metrics that has to be extracted to ensure the quality of the ISG is large (about a hundred metrics) and cannot be covered here for the sake of brevity. However, it is important to notice that, since the processes for the generation of the four projections run in parallel, the aforementioned quality metrics will be evaluated in different time, as soon as it is possible; hence to enable a correct comparison among them a synchronization routine takes place in order to align them to the same "epoch".

The data publishing workflow of the OpenAIRE's production environment has been monitored since 2015 by a monitoring system implemented utilizing MoniQ [15], a data flow quality monitoring system resulting from an enhancement of the solution proposed in [10]. Despite the switch to public is still triggered manually, the collection and inspection of the quality metrics from ISG projections is performed automatically via MoniQ, hence dramatically decreasing the operational cost of the control phase.

8 Conclusions

The mission of OpenAIRE is to foster an Open Science e-infrastructure supporting the advancement of research by means of interlinking and disseminating scientific knowledge. Thanks to its growing network composed of different scholarly communication stakeholders (e.g. institutional repositories, data repositories, OA journals, libraries, and funders) and to its mature technological infrastructure, OpenAIRE makes openly accessible a rich Information Space Graph (ISG) where objects of the research life-cycle (e.g. publications, datasets, projects) are semantically linked to each other. The management of the OpenAIRE ISG is a complex operation realized by means of different types of workflows orchestrated by the D-NET framework: the content aggregation, the population of the ISG and its de-duplication, the mining of inferred knowledge from publications full-texts, the publication of the ISG and the monitoring of its quality metrics. The workflow automation represents an important advantage for the data managers work, who can focus on supervising the workflow executions, monitoring the data quality, and limiting their intervention only when necessary.

Acknowledgments. Research partially supported by the EC H2020 project Open-AIRE2020 (Grant agreement: 643410, Call: H2020-EINFRA-2014-1).

References

1. Kobos, M., Bolikowski, Ł., Horst, M., Manghi, P., Manola, N., Schirrwagen, J.: Information inference in scholarly communication infrastructures: the OpenAIREplus project experience. Procedia Comput. Sci. **38**, 92–99 (2014). doi:10.1016/j.procs.2014.10.016
2. Tkaczyk, D., Szostek, P., Fedoryszak, M., Dendek, P.J., Bolikowski, Ł.: CERMINE: automatic extraction of structured metadata from scientific literature. Int. J. Doc. Anal. Recognit. (IJDAR) **18**(4), 317–335 (2015)
3. Manghi, P., Bolikowski, L., Manola, N., Schirrwagen, J., Smith, T.: OpenAIREplus: the European scholarly communication data infrastructure. D-Lib Magaz. **18**(9), 1 (2012)
4. Manghi, P., Manola, N., Horstmann, W., Peters, D.: An infrastructure for managing EC funded research output – the OpenAIRE project. Int. J. Grey Lit. **6**, 31–40 (2010)
5. Manghi, P., Artini, M., Atzori, C., Bardi, A., Mannocci, A., La Bruzzo, S., Candela, L., Castelli, D., Pagano, P.: The D-NET software toolkit: a framework for the realization, maintenance, and operation of aggregative infrastructures. Program **48**(4), 322–354 (2014)
6. Weibel, S., Kunze, J., Lagoze, C., Wolf, M.: Dublin core metadata for resource discovery (No. RFC 2413) (1998)
7. Sompel, H.V.D., Nelson, M.L., Lagoze, C., Warner, S.: Resource harvesting within the OAI-PMH framework. D-Lib Magaz. **10**(12), 1082–9873 (2004)
8. Artini, M., Atzori, C., Bardi, A., La Bruzzo, S., Manghi, P., Mannocci, A.: The OpenAIRE literature broker service for institutional repositories. D-Lib Magaz. **21**(11), 3 (2015)
9. Príncipe, P., Schirrwagen, J.: OpenAIRE guidelines for data source managers: aiming for metadata harmonization. In: CERN Workshop on Innovations in Scholarly Communication (OAI9) (2015)
10. Mannocci, A., Manghi, P.: DataQ: a data flow quality monitoring system for aggregative data infrastructures. In: Fuhr, N., Kovács, L., Risse, T., Nejdl, W. (eds.) TPDL 2016. LNCS, vol. 9819, pp. 357–369. Springer, Cham (2016). doi:10.1007/978-3-319-43997-6_28

11. Kolb, L., Thor, A., Rahm, E.: Parallel sorted neighborhood blocking with mapreduce. arXiv preprint (2010). arXiv:1010.3053
12. McNeill, N., Kardes, H., Borthwick, A.: Dynamic record blocking: efficient linking of massive databases in mapreduce. In: Proceedings of the 10th International Workshop on Quality in Databases (QDB) (2012)
13. Jaro, M.A.: Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. J. Am. Statist. Assoc. **84**(406), 414–420 (1989)
14. Fellegi, I.P., Sunter, A.B.: A theory for record linkage. J. Am. Statist. Assoc. **64**(328), 1183–1210 (1969)
15. Mannocci, A.: Data Flow Quality Monitoring in Data Infrastructures (2017)
16. Atzori, C.: gDup: an integrated and scalable graph deduplication system (2016)
17. George, L.: HBase: The Definitive Guide: Random Access to Your Planet-Size Data. O'Reilly Media, Inc., Sebastopol (2011)