SEVENTH FRAMEWORK PROGRAMME

CAPACITIES

Research Infrastructures

INFRA-2009-1 Research Infrastructures

OPENAIRE

Grant Agreement 246686

"Open Access Infrastructure for Research in Europe"

OpenAIRE Specification and Release Plan
Deliverable Code: **D6.1**

Document Description

Project	
Title:	OpenAIRE, Open Access Infrastructure for Research in Europe
Start date:	1st December 2009
Call/Instrument:	INFRA-2007-1.2.1
Grant Agreement:	**246686**

Document	
Deliverable number:	D6.1
Deliverable title:	OpenAIRE Specification and Release Plan
Contractual Date of Delivery:	31st of January 2010
Actual Date of Delivery:	31st of March 2010
Editor(s):	Paolo Manghi
Author(s):	Paolo Manghi, Marko Mikulicic
Reviewer(s):	Natalia Manola
Participant(s):	
Workpackage:	WP6
Workpackage title:	OpenAIRE System operation and maintenance
Workpackage leader:	ICM
Workpackage participants:	CNR, CERN, NKUA, ICM, SURF, UNIBI
Distribution:	Public
Nature:	
Version/Revision:	V1.4
Draft/Final:	Final
Total number of pages: (including cover)	
File name:	
Key words:	

Disclaimer

This document contains description of the OPENAIRE project findings, work and products. Certain parts of it might be under partner Intellectual Property Right (IPR) rules so, prior to using its content please contact the consortium head for approval.

In case you believe that this document harms in any way IPR held by you as a person or as a representative of an entity, please do notify us immediately.

The authors of this document have taken any available measure in order for its content to be accurate, consistent and lawful. However, neither the project consortium as a whole nor the individual partners that implicitly or explicitly participated in the creation and article of this document hold any sort of responsibility that might occur as a result of using its content.

This article has been produced with the assistance of the European Union. The content of this article is the sole responsibility of the OPENAIRE consortium and can in no way be taken to reflect the views of the European Union.

The European Union is established in accordance with the Treaty on European Union (Maastricht). There are currently 27 Member States of the Union. It is based on the European Communities and the member states cooperation in the fields of Common Foreign and Security Policy and Justice and Home Affairs. The five main institutions of the European Union are the European Parliament, the Council of Ministers, the European Commission, the Court of Justice and the Court of Auditors. (http://europa.eu.int/)

OPENAIRE is a project funded by the European Union

Table of Contents

Table of Figures

Summary

This deliverable describes the design and development plan of the OpenAIRE System.

1 Introduction

In this deliverable we present the plan of design and development of the OpenAIRE System for the next 34 months. The aim of this document is to explain how the overall activity plan of the project (as illustrated in Figure 1) and the milestones relative to software and production system releases (as listed in Table 1) will be accomplished.

The technical activities will be supervised and led by CNR and carried out across the six technical partners CNR, CERN, NKUA, ICM, SURF and UNIBI.

M6.1	OpenAIRE Candidate v1.0	WP4, WP5, WP6	**CNR**	Beta release of the OpenAIRE portal	M5
M6.2	OpenAIRE Production v1.0 – Public release	WP4, WP5, WP6	**CNR**	Public release of the OpenAIRE portal	M6
M6.3	OpenAIRE Candidate v2.0	WP4, WP5, WP6	**CNR**	Candidate release	M10
M6.4	OpenAIRE Production v2.0 – Public release	WP4, WP5, WP6	**CNR**	Web release	M12
M6.5	OpenAIRE Candidate v3.0	WP4, WP5, WP6	**CNR**	Software release	M22
M6.6	OpenAIRE release v3.0	WP4, WP5, WP6	**CNR**	Web release	M24
M6.7	OpenAIRE Production final candidate release	WP4, WP5, WP6	**CNR**	Candidate release of all services	M34
M6.8	OpenAIRE Production final release	WP4, WP5, WP6	**CNR**	Web release of all services	M36

Table 1 - Milestones

In order to better understand the technical steps described in this document, the deliverable will initially introduce an overview of the functional requirements of the OpenAIRE system, given in terms of expected end-user interaction scenarios with the system itself (detailed requirements and scenarios will be illustrated in D4.1 "Presentation and deposition Services Integration Plan" and D5.2 "Storage and Mediation Services Integration Plan"). Once the objectives are clarified, the document will describe the execution plan, organized in a number of "work threads".
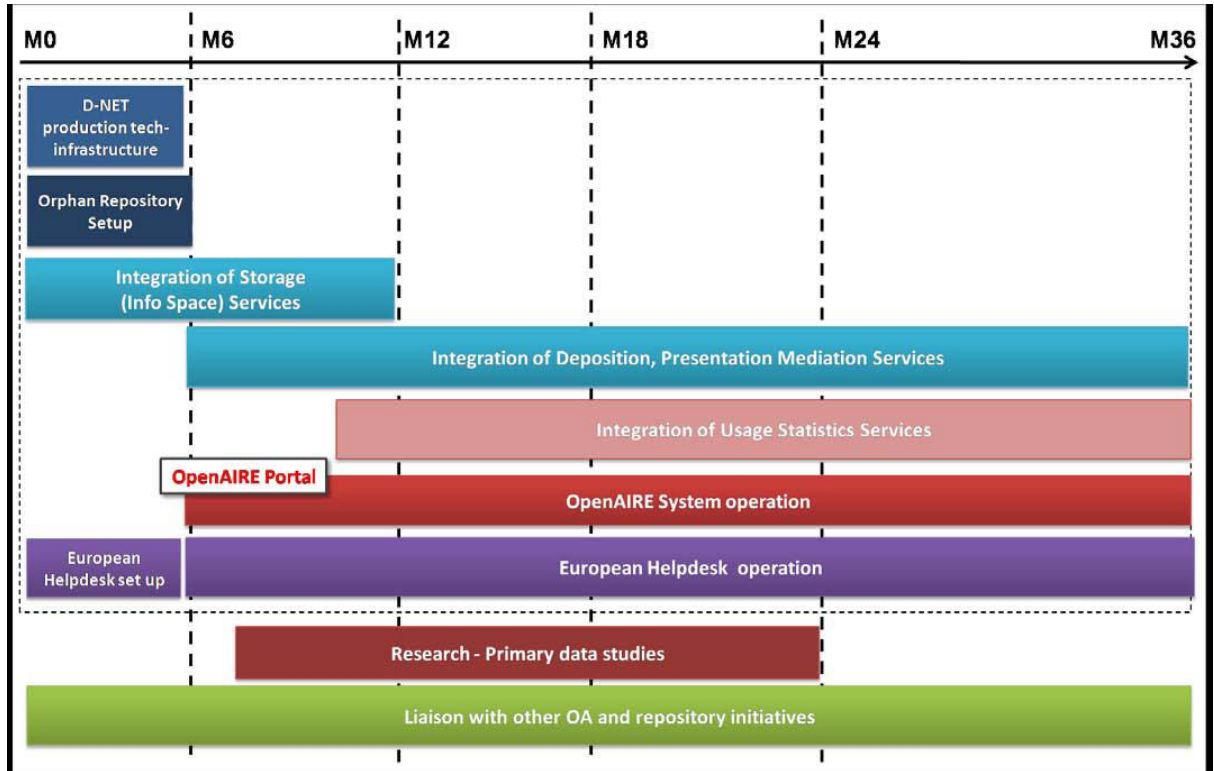
Figure 1 – OpenAIRE project: Overall Activity Plan

2 Functional Requirements/Scenarios

Technically speaking, the OpenAIRE System will be realized to meet the functional requirements of the three end-user scenarios described in this section. The scenarios describe different ways of interacting with a so-called OpenAIRE Information Space (OIS), i.e, the FP7 pilot and ERC related publications, devised to provide storage and access facilities to project metadata, article metadata and connections between the two. The patterns of interaction regard "depositing" information and "accessing" information, both from the perspective of end-users and applications. The underlying assumption is that authors (or others on their behalf) of articles funded by FP7 projects in one of the seven areas of the EC-OA-pilots, will have to deposit into the OIS metadata relative to such articles, completed with information regarding the projects, the license and the repositories of reference, i.e., the repository where the file of the article can be found (if the article is not "orphan" of a repository). This process should be eased by the OpenAIRE Portal interfaces and, whenever possible, reduced to an automated OAI-PMH harvesting from repositories which directly provide OpenAIRE-enriched article metadata from OpenAIRE compliant repositores, with no need for end-users to access the system.

Scenario 1: Depositing into OIS

The deposition architecture is depicted in Figure 2Figure 4. Project metadata will be stored into the OIS by accessing and fetching data from SESAM[1], the European Commission online reporting tool for research and technological projects. As such, SESAM will act as the authoritative source for project information in the OIS.

Article metadata, instead, will be deposited in the OIS in three ways:

- *End-User deposit:* end-users ingest, through the OpenAIRE Portal (www.openaire.eu) metadata regarding articles. This operation can be performed in two ways:

    o Articles whose author has a repository of reference in his/her institution: supported by tools, the user browses through public article information spaces (such as DRIVER Infrastructure and Google Scholar) searching for his/her publications. When these are found, the user double-checks the data, cleans it with the automated support of the user interface, adds the project information funding the article (selects from the authoritative list of projects), adds repository information (selects from a list of repositories where the publication file is present, if any) and eventually deposits the metadata. For articles not available through browsing, the user proceeds manually to the insertion of the metadata records.

    o Article with an "orphan" author, i.e. an author who has no repository in his/her institution: users will deposit bibliographic metadata, project metadata, license metadata and upload the file of the article into the OpenAIRE Orphan Repository.

    Most importantly, for the case of end-user deposit, the system should carefully guide users through their possibilities of depositing publications, and support them and their institutions towards adopting and supporting automatic deposition (described below).

---

[1] http://webgate.ec.europa.eu/sesam/index.do

- *Automatic deposit*: the OpenAIRE System registers repositories which are compliant to the "OpenAIRE Guidlines for repositories". In short, the guidelines require the repositories to provide through OAI-PMH interfaces, article metadata accompanied by project and license information, through an OpenAIRE OAI-Set. Note that both project and license information are encoded through special vocabularies, to be shared by OpenAIRE with the repository community. The OpenAIRE System will incrementally harvest from the respositories to ingest article records into the OIS.

- *Refinement deposit*: another form of deposition is that of "refinement". The OpenAIRE System should allow anonymous users at suggesting corrections in the OIS in order for administrators to clean, refine and enrich the information therein. So for example, end users searching for articles will be able to suggest corrections to the name of the authors or of the institutions relative to articles in their query results; or, while depositing article information, they will be able to add new institutions to the system, as well as repositories related with the articles, and so on.
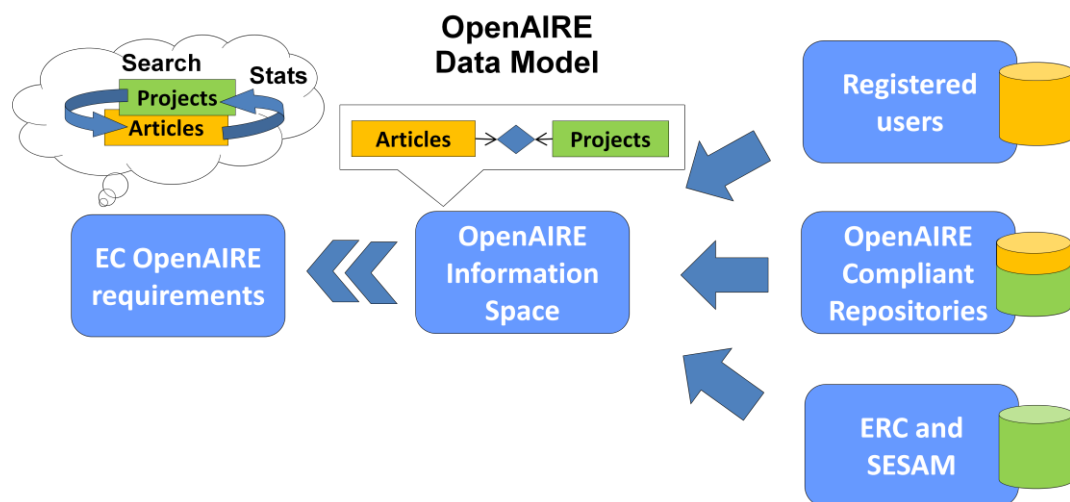

Figure 2 – Deposition into the OIS

Scenario 2: Accessing the OIS

End-users want to retrieve information on articles and project data. Key functionalities, to be delivered through the OpenAIRE Portal (www.openaire.eu), are:

- *Search/browse*: over the OIS through specific or pre-tailored views, e.g. breakdown and display by research, projects, people, funds, articles, subjects, etc.,

- *Format conversion*: convert articles (i.e. their full-text payload), which have been deposited or selected through searches into OpenAIRE, on-demand into other types of MIME formats;

- *Notifications*: receive notifications on events related to their expertise and type of work, such as:

  o impact (derived from overall usage statistics) on their articles or projects;

  o new projects or articles appearing within their scientific research area;

  o articles released from specific projects or more general research areas or on usage statistics (calculated by thresholds) on specific research areas/projects.

- *Statistics:* monitor and assess impact through usage statistics: users want to measure the impact of articles and projects through usage statistics (obtained from the orphan repository and all other compliant, participating repositories) and their combination of various quantitative web indicators involving external link references and public search engines data.

Scenario 3: Third-party services accessing the Information Space

The OIS may be re-used by third-party services, e.g. aggregators, distributed search engines, e-Science text-mining applications. Such services must be able to retrieve content from OpenAIRE through standard access protocols, such as OAI-PMH, OAI-ORE and SRW. Moreover, OpenAIRE may be able to export content to external services, e.g. Google-Scholar, e-infrastructures like Europeana and DRIVER.

3 Release Plan

In order to achieve the goals depicted by the scenarios in the previous section, the work is splitted in three main "chunks", identified by software milestones and system public releases, shown in Table 1. In this document we detail how the technical partners will proceed in achieving such results in the next 20 months (until M24 when all main technological components are in place). In particular, the work is organized into nine possibly parallel threads of work, each one aimed at particular tasks/components of the system. Note that, although such threads are intended to provide firm guidelines for the partners involved, the plan is open to modifications in the years to come in order to adapt to unexpected requirements and issues that may arise on the way.

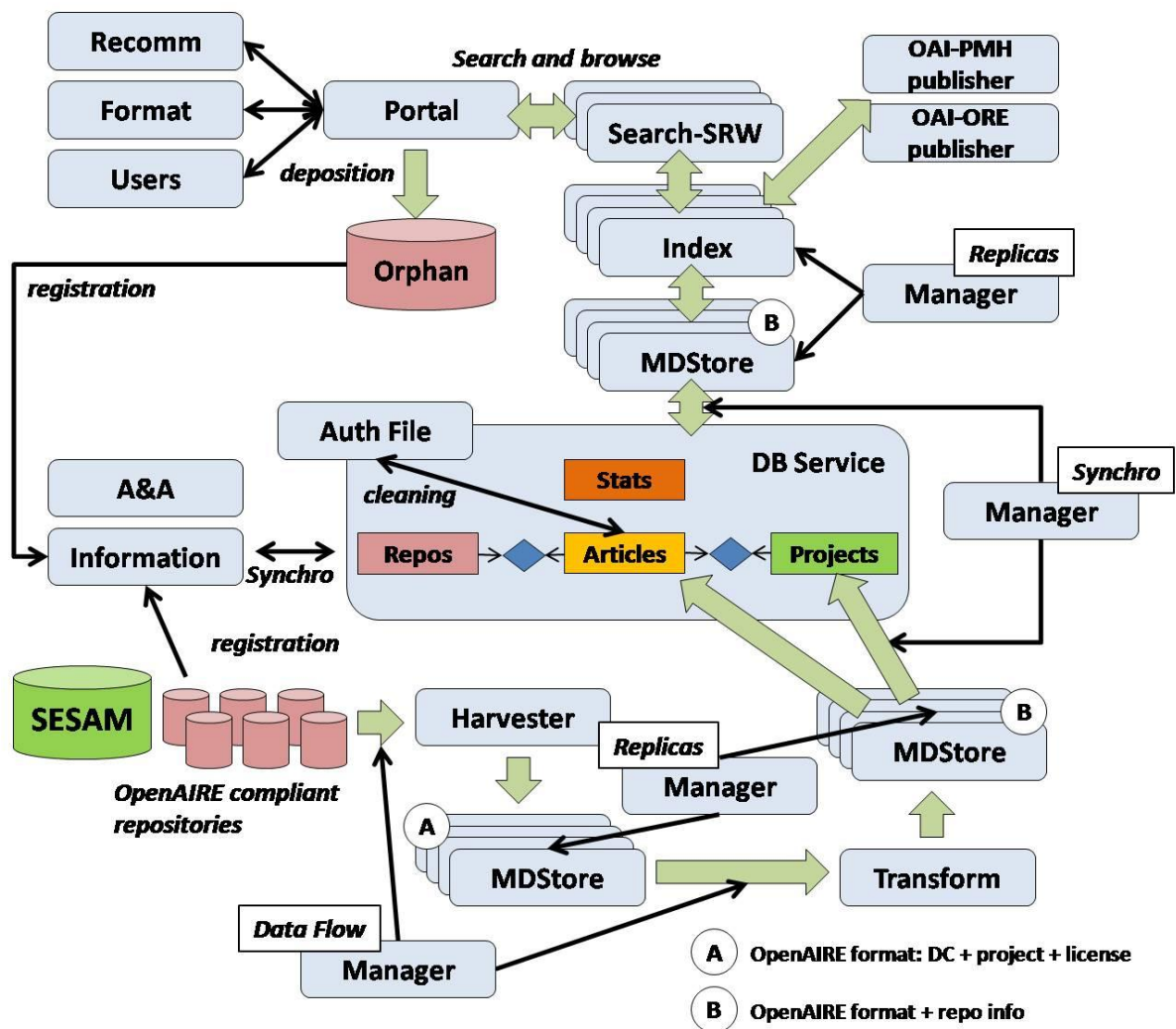The architecture of the OpenAIRE System, to be referred to in the following sections, is illustrated in Figure 1.

Figure 3 – OpenAIRE System architecture

3.1 OpenAIRE Production System v1.0 – Month 6

The first production release of the OpenAIRE portal is planned at Month 6 and will include:

- The OpenAIRE Portal with Deposition User Interfaces for Scenario 1 supporting deposition of article whose authors do not have a repository of reference and deposit in the Orphan Repository;

- Access services supporting Scenario 2, providing end-users with basic functionalities for searching/ retrieving articles;

3.1.1 Thread #1 – OIS: orphan repository, OpenAIRE Portal (M0-M6, CNR, CERN, NKUA)

This thread will be completed in the first 6 months and will target Scenario 1, for end-user deposit into orphan repository. The activities, which are also preparatory to the OIS implementation, consists of:

- *Data model specification*: CNR, with the cooperation of NKUA, and interacting with the IT department responsible (ref. Claudio Montanari) for SESAM (EC Participant Portal), has to deliver at Month 4 the definition of the OIS data model.

- *Orphan repository platform*: Based on the OIS data model, CERN will configure an instance of the Invenio Repository capable of hosting data responding to the model. The instance should be adapted to access content (e.g., vocabularies) from the D-NET Information Service.

- *OpenAIRE Portal for deposition*: NKUA and CERN will proceed with the operation of the OpenAIRE Portal and its integration with the Invenio Instance.The OpenAIRE orphan repository will have an autonomous UI in the beginning (month 6), which will be gradually integrated into the OpenAIRE portal until month 12 (a combination of Joomla and D-NET).

- *End-user registration and access*: end-users should properly register to the OpenAIRE Portal before they can modify the OIS (initially only depositing into Orphan Repository). The D-NET Authentication & Authorization Service (A&A Service) is capable of providing such functionalities and should thus be integrated with the OpenAIRE Portal and Invenio. Registration specifics and rights will be discussed in detail in deliverable 4.1.

3.2 OpenAIRE Production System v2.0 – Month 12

In parallel with thread #1 depicted above, other threads will be carried on, from month 0 to month 12. These will complete Scenario 1, by providing end-user deposit for authors whose articles have a repository of reference and automated deposit, add further functionalities for Scenario 2, and introduce access to third-party applications for Scenario 3.

3.2.1 Thread #2 – OIS implementation (M4-M8, CNR)

The result of this thread will result in one of the cornerstone components of the OpenAIRE system, as it will support the implementation of all scenarios depicted above. In particular, based on the data model resulting at Month 4, CNR will modify D-NET to host project information, article information and their relationships in a traditional database framework. To this aim a new D-NET service, called the *DB Service* will be designed and developed to host a relational database (Postgres) capable of serving SQL clients (posing DDL, DML and DQL requests) and capable of synchronizing with the D-NET Information Service (IS). The reason for storing OIS entities in a relational database and in the IS is twofold: on the one hand we want to provide fast and efficient responses to queries used in document-project

related statistics; on the other hand, the same OIS entities will be also treated as D-NET system resources in the Information Service, so that D-NET services are able to interact with them (e.g., subscribe to their actions) in the infrastructure eco-system. An example is a repository, whose information will be in the OIS to be used in statistics calculations, and perhaps for other quick information extraction reasons, but should also be in the Information Service, in order to be accessed by Harvesting Services in the need of fetching records from them.

The DB Service will contain ALL the information in the OIS, to be used as knowledge base for statistical caculation and as authoritative source for third-party applications interested in OpenAIRE content.

3.2.2 Thread #3 – OIS: automated deposition (M6-M12, CNR, UNIBI)

This thread targets Scenario 1, automated deposit. To this aim, the D-NET MDStore, transformation and harvesting services will be configured and deployed by CNR and UNIBI for:

- Repository registration: only OpenAIRE compliant repositories will be registered
- Record harvesting: a special OpenAIRE format will be harvested, consisting of Dublin Core bibliographic fields enriched with fields for the project identifiers and others describing the license of the publication (e.g., name of licence, months of embargo)
- Record transformation: initially, the transformation will consist of adding to OpenAIRE records fields describing the repository of origin;
- Record storage for future reuse: MDStore services will be used to host OpenAIRE records harvested from repositories and OpenAIRE transformed records
- Manager Services will be configured to:
  - o Feed OpenAIRE transformed records into the DB service
  - o Keep replicas of MDStore service content (robustness).

Note that *the first OpenAIRE compliant repository will be the OpenAIRE Orphan Repository.*

Another important aspect of automated deposition is concerned with accessing the SESAM system in order to keep project information up-to-date in the OIS (DB Service). To this aim, CNR will interact with SESAM IT department to establish and develop an API through which the OpenAIRE System will be able to access their databases (today, no API for external access are provided). The idea is to propose the adoption of an OAI-PMH interface to SESAM. Through this they can provide project and other information in the form of metadata records.

Finally, both repository information and project information will be entities of the data model to be stored in the DB Service, as well as information to construct relative D-NET resources in the Information Service. In particular:

- projects will form a *vocabulary resource* of projects to be used by the OpenAIRE portal (all the replicas of the portal) to support article metadata deposition and to be disseminated to repository managers willing to make their repositories OpenAIRE compliant;

- repositories will be independent resources, to be accessed by harvesting and transformation services and by the OpenAIRE portal (all the replicas of the portal) to support article metadata deposition

3.2.3 Thread #4 – OIS: end-user access, part 1 (M6-M12, CNR, NKUA, UNIBI)

This thread targets Scenario 2, search/browse. To this aim the OpenAIRE portal will be integrated with D-NET services for search and browsing functionality, which in turn rely on D-NET Index services whose content will be taken from the DB Service. The idea is to maintain a layer of Index Services containing the portion of the OIS required by the Portal. To this aim, the indices will represent a customized materialization of the OIS, calculated and extracted by querying the DB Service. By using D-NET Index Services the OpenAIRE system will be able to (i) provide full-text search and IR ranked-results and (ii) scale up to arbitrary number of accesses by keeping automated index replicas. Specifically, CNR, NKUA and UNIBI will:

- Establish the structure of the data to be queried and accessed through the portal;
- Identify the queries to be run on the DB Service to extract such information;
- Configure and deploy D-NET index services and MDStore services capable of hosting the data whose data structure has been identified above;
- Manager Service for synchronizing DB service and Index Service: Build D-NET workflows capable of running the queries over the DB Service, feed the MDStores and indices with the results; the workflows should be scheduled and executed automatically by the D-NET Manager Services;
- Manager Service for replicas: build D-NET workflows capable of keeping K replicas of MDStore Services and Index Services;
- Build the OpenAIRE Portal pages supporting search and browse through the Search Services and operating over the materialization of the OIS into Index Services.

3.2.4 Thread #5 – OIS: end-user deposition (M6-M12, NKUA)

This thread targets Scenario 1, end-user deposit. As such, NKUA will realize OpenAIRE Portal pages to welcome the end-user at depositing their article metadata. The pages will have to help the end user at understanding "who he/she is" with respect to OpenAIRE:

- An author orphan of a repository of reference:
  - support the insertion into the Orphan Repository;
  - use the helpdesk to notify all potential institutions that their researchers are depositing in OpenAIRE orphan and might be in need of a repository of reference;
- An author who has a repository of reference that is not OpenAIRE compliant:
  - support the insertion into the OIS (DB Service);
  - remind to insert into own repository;
  - use the OpenAIRE helpdesk to notify to potential repositories of reference, that metadata of an article they might be interested in has been deposited into OpenAIRE, and guide them to be OpenAIRE compliantl
- An author who has a repository of reference that is OpenAIRE compliant

o Send user to own repository (the publication will be harvested in due course);

While depositing metadata, users can be helped by the following tools:

- Search and browse the DRIVER information space to select the records and speed up the insertion of values
- Access to menus of controlled vocabularies in the Information Service, such as Projects, Licenses, Organizations
- Semi-Automatic corrections of the values, e.g., based on vocabularies

3.3 OpenAIRE Production System v3.0 – Month 24

In the second year of the project, once data deposition and access to the OIS have been completed and offered in production, the activities will concentrate on threads that will complete the scenarios. Furthermore, investigations on how to deposit publications from the OpenAIRE portal onto remote repositories will be carried out.

3.3.1 Thread #6 – OIS: access to third-party applications (M13-M16, CNR)

This thread targets Scenario 3 and consists in providing the D-NET services for exporting data in the OIS according to OAI-PMH, OAI-ORE and SRW. The main issues are related with the configuration of the services and with the development of public APIs equipped with controlled and filtered access.

3.3.1 Thread #7 – OIS: statistics (M13-M24, NKUA, UNIBI and CERN)

This thread targets Scenario 2, usage statistics. NKUA will develop Usage Statistics services matching the requirements gathered from the EC Pilot and combined with the experimentation of usage statistics harvested from institutional repositories. On the side of article and project statistics, the service will have to interact with the DB Service. It is still to be decided if the statistical data will be materialized into Index Services or calculated on demand from the DB Service.

3.3.2 Thread #8 – OIS: refining the data (M13-M24, CNR, CERN, NKUA)

This thread targets Scenario 1, refinement deposition (see Figure 1). Since originally it was not part of the DoW, its accomplishment could be partial. NKUA and CNR will design and develop tools giving support to end-users willing to "refine" the data while depositing into and searching and browsing from the OIS. End-users will be able to give their advice by "tagging" any record and any "field" with an action and some parameters to go with it:

- Remove record or field;
- Update field;
- Add record or field;
- Merge two records;
- Split one record in two records.

Administrators of OpenAIRE will have to revise, hence *validate, suspend, reject* or *roll back* all end-users actions. To this aim:

- The DB Service will be extended with tables for supporting the (history of) tagging and validation model;
- The OpenAIRE portal will have to be upgraded to offer user interfaces allowing this interaction;
- D-NET authority file services will be employed to identify possible duplicates across different entities, e.g., authors. CERN will also contribute in this direction, by encapsulting duplication matching algorithms into new D-NET services specially devised for OpenAIRE.
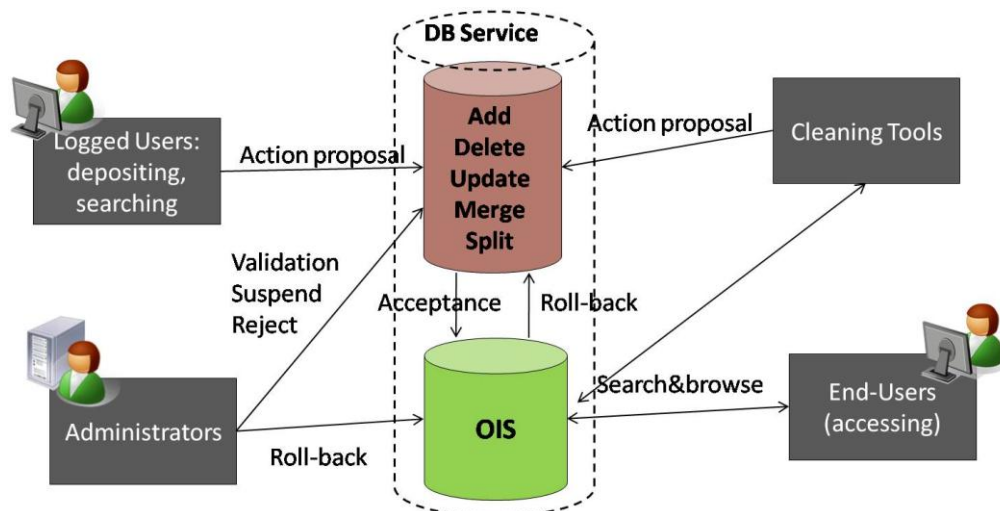

Figure 4 – Architecture of refinement deposition scenario

3.3.1 **Thread #9 – OIS: end-user access, part 2 (M18-M24, ICM, NKUA)**

This thread targets and completes Scenario 2, in the requirements of format conversion and end user notifications. To this aim NKUA and ICM will respectively configure and deploy D-NET reccomendation services and develop format conversion services. Both will have to be integrated to be made available to end users with the OpenAIRE Portal.

3.3.1 **Thread #10 – OIS: depositing publication metadata and file onto remote repositoriesa (M13-M24, CNR, UNIBI)**

This thread investigates the possibilities of metadata and file deposition from the OpenAIRE portal onto remote repositories adopting known platforms and willing to cooperate. The activities will take on from the results of the SWORD App (Simple Web-service Offering Repository Deposit) project.[1]

---

[1] http://www.swordapp.org/