DIPARTIMENTO SVILUPPO DI APPLICAZIONI DI GESTIONE DATI
UNITÀ OPERATIVA PER LA REALIZZAZIONE DI SERVIZI A SUPPORTO DELLE BIBLIOTECHE
E DELLA LORO UTENZA
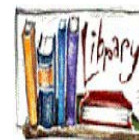
CILEA Open Archives Platform - Project plan

Project Title	**CILEA Open Archives Platform**
Start Date	**May 2003**
End Date	**October 2004**
Lead Institution	**CILEA**
Project Director	**Ing. Fabio Valenziano, CILEA**
Project Manager & Contact	**Dr. Susanna Mornati, CILEA** Via R. Sanzio 4, I-20090 SEGRATE MI tel. +39 0226995322, fax +39 022135520 email mornati@cilea.it
Document History	**2003.10.24, Version 1.3** 2003.10.08, Version 1.2 2003.09.29, Version 1.1 2003.09.19, Version 1.0
Author	**Dr. Susanna Mornati**

---

[1] http://www.rdn.ac.uk/projects/eprints-uk/
[2] De Robbio, Antonella, *Auto-archiviazione per la ricerca: problemi aperti e sviluppi futuri*, in *Proceedings Comunicazione scientifica ed editoria elettronica: la parola agli Autori: L'Utente-Autore nel circuito della comunicazione scientifica: editoria elettronica e valutazione della ricerca*, Milano, 20 Maggio 2003, http://eprints.rclis.org/archive/00000180/

CILEA Open Archives Platform - Project plan

0. Contents

1. Introduction[3] and context

A deep crisis has affected the traditional scholarly publishing models in the last decades. Free circulation of papers that contribute to the scientific and technological progress is blocked or at least slowed down by a number of factors, ranging from the limits of traditional publishing practices to publishers' rules for copyright assignement, from the dramatic rise of periodical prices to the shrinking of library budgets. In the age of the Internet and of electronic publishing, this paradoxically results in a progressive reduction of access to scientific research outputs. These events encouraged the birth of the Budapest Open Access Initiative[4], asserting the need to open access to scholarly publishing, and particularly of the Open Archives Initiative[5], suggesting that open access can be achieved via open archives, digital repositories where authors can self-archive an electronic copy (eprint) of their published papers. Within the OAI framework several tools have been developed to allow the creation, management, and interoperability of open archives via the OAI-PMH[6] protocol. This efforts result in a wider dissemination, hence in a higher impact of research results[7], generating vast and predictable benefits both for authors and their institutions, from individual carrier advancing, to further research funding, from the enhancement of each institution's visibility to promoting general research progress[8].

The implementation of institutional eprints repositories in Italy is at its dawn. Nonetheless, eprint archiving experiences an unprecedented *momentum*[9] in the international context of open access to scientific publishing. New repositories (Data

---

[3] All links cited in this paper were last visited on August 2003

[4] BOAI, see: http://www.soros.org/openaccess/index.shtml

[5] OAI, see: http://www.openarchives.org/

[6] Open Archives Initiative Protocol for Metadata Harvesting, see: http://www.openarchives.org/OAI/openarchivesprotocol.html

[7] Steve Lawrence, *Free online availability substantially increases a paper's impact*, in *Nature*, 411, 6837, p. 521, 2001, http://www.neci.nec.com/~lawrence/papers/online-nature01/

[8] Stevan Harnad, *Promoting Self-Archiving of Institutional Research Output*, http://www.ecs.soton.ac.uk/~harnad/Temp/self-archiving.htm

[9] Peter Suber, *Momentum for eprint archiving*, http://www.topica.com/lists/suber-fos/read/message.html?mid=1607391538&sort=d&start=14

Providers) rise and grow at a rapid pace[10], and abroad several national or cooperative projects have been funded to enhance open archives visibility and provide common interfaces and platforms (Service Providers and portals) to research documentation[11].

The project, born from a feasibility study on scientific electronic publishing released in summer 2002[12], and inspired by the outcomes of a conference on scientific communication[13] organized by CILEA and the University of Milan in May 2003, aims to develop an Italian national platform providing a central access point to research papers collected by Italian Open Archive repositories. The amount of existing open source software packages developed in the international academic environment, together with CILEA's experience in designing and implementing advanced ITC facilities, will constitute the basis for the project.

In order to provide a significant impact to the project, it is necessary to gather a "critical mass" of quality data from open archives. For this purpose, part of CILEA's resources will be devoted to support the design, implementation, and running of independent Data Providers, whether institutional, disciplinary, or individual, and to promote Open Archives through various initiatives addressed to users and technicians.

Along the project presentation, main issues are unfolded in bar diagrams and UML (Unified Modeling Language) diagrams. See Appendix A for a detailed legenda.

2. Stakeholders

The project addresses the whole research community, from students to faculties, from PhD to senior researchers, both in the humanities and in the STM environment, and their institutions. By self-archiving their own papers in institutional archives, implemented either on site or at CILEA, scholars will benefit by an extended dissemination and impact of their research results, and students will experience a wider access to all literature relevant for their fields of study. Moreover, institutions will exploit archives as a display for their scientific production. The rise of a national platform will also provide advanced data search and retrieval tools, data aggregation, time stamping, protection against plagiarism, and long-term digital preservation. In addition, the availability of usage statistics and citation linking will provide measurement tools for research impact, that are useful at a higher level of research management, for assessment and funding purposes.

3. Platform architecture

The set up of a number of Data Providers by research institutions, on site or at CILEA, is necessary to provide significant services from the point of view of quantity. Their

---

[10] See the dinamic list maintained by the Open Archives Forum:
http://www.oaforum.org/oaf_db/list_db/list_repositories.php
[11] See for instance the United Kingdom and Nederlands projects, respectively the above mentioned ePrints-UK:
http://www.rdn.ac.uk/projects/eprints-uk/ and DARE: http://www.surf.nl/themas/index2.php?oid=18
[12] Valentina Comba, *AEPIC Academic E-Publishing Infrastructures - CILEA : progetto di editoria elettronica per la ricerca e la didattica*, http://eprints.rclis.org/archive/00000066/
[13] *Comunicazione scientifica ed editoria elettronica: la parola agli autori : l'utente-autore nel circuito della comunicazione scientifica: editoria elettronica e valutaizone della ricerca*, Milano, Italy, 20 maggio 2003, http://www.cilea.it/convegni/convegnoeditoria/presentazione.html

users (authors of scholarly papers) register and self-archive an electronic copy (eprint) of their production (usually in the shape of a pre-print, submitted to a journal to be published) in the appropriate repository. These archives also allow the availabilty of access control tools and the possibility to include different materials, from dissertations to didactic or administrative documentation.
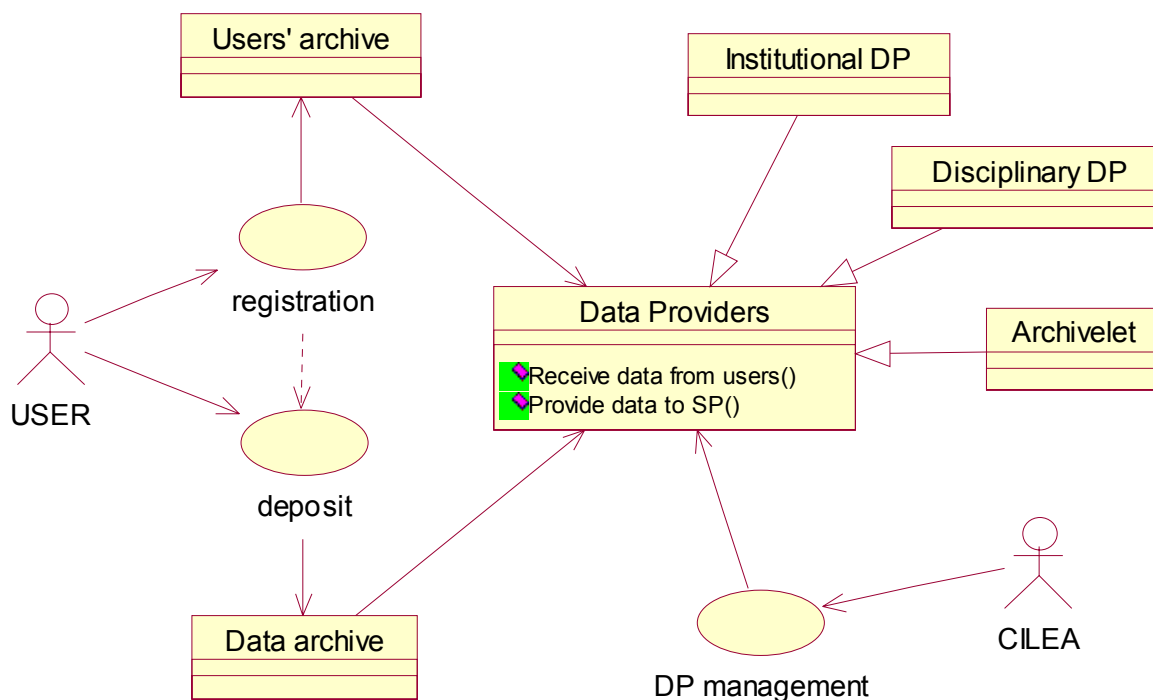

DIAGRAM 1 - DATA PROVIDERS

In order to achieve the project's main goal, that may be summarised in enhancing impact of research literature produced in Italy through its extended access, CILEA will organize promotional activities and contextually develop a national platform to provide a centralised access to scholarly literature archived in the already existing Italian open repositories.

The platform architecture will offer a common search and retrieval interface, disciplinary access by subject, and complementary services to different and independent repositories. The platform architecture is structured into two main objects: a collection of service providers and a portal. To sum up, a cluster of eprints archives, called Data Providers, which contain documents and associate metadata deposited by users at each institution, will provide the raw data, and a cluster of additional services, called Service Providers, will add value to these data, and rexpose them to users through a portal, that will allow results customisation.
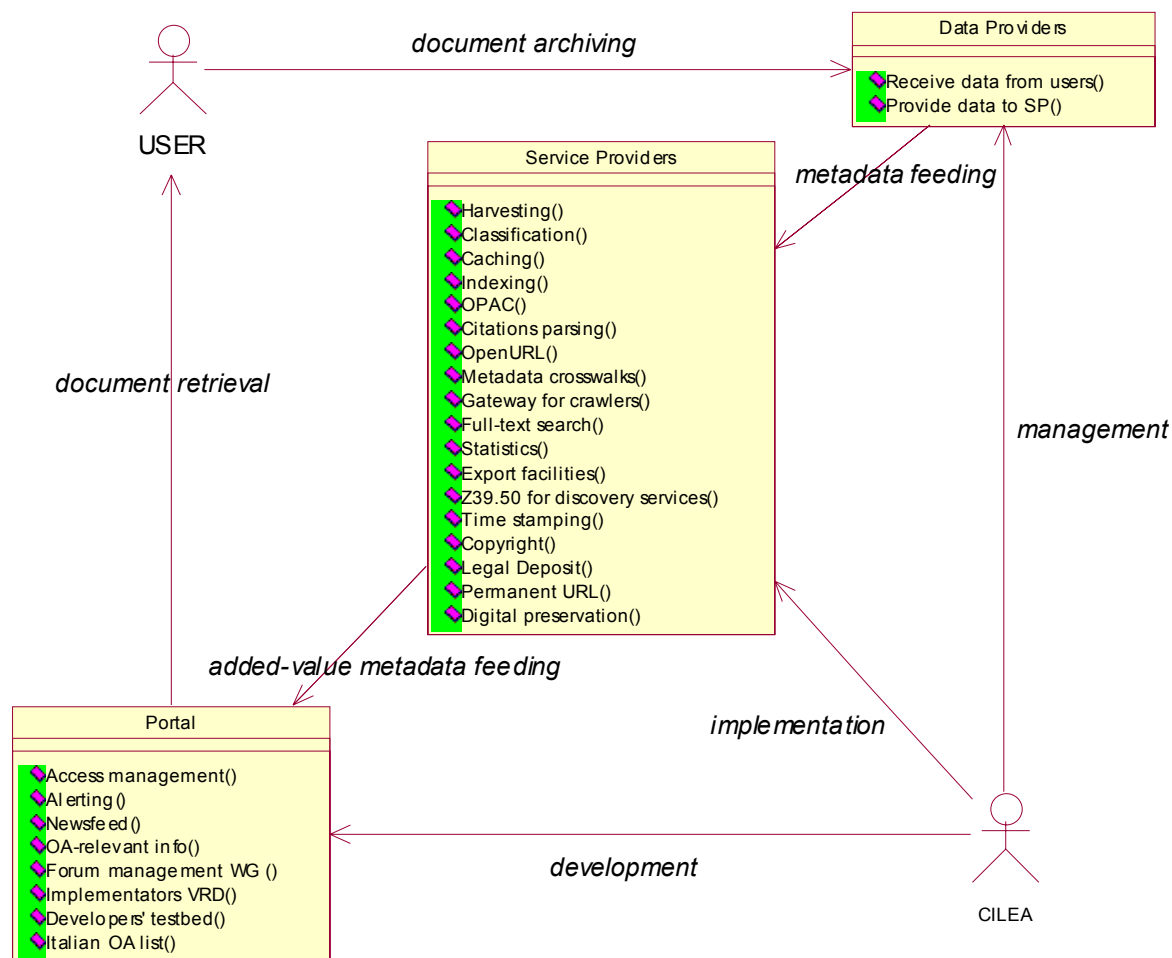
DIAGRAM 2 - PLATFORM ARCHITECTURE

3.1. Service Providers

Metadata and full text of research papers will be harvested from available and appropriate eprints archives, prefererably devoted to Italian research literature, using the most recent version of the OAI-PMH protocol, and then passed to external Web services that will enhance the records.

Several tasks will be performed at the utmost automation level:
- metadata attribution of a broad subject classification, based on the Italian official academic research areas[14];
- central caching and indexing functionalities;
- advanced searching, browsing, and retrieval capabilities;
- citations parsing in the document text to be used by citation services[15] and to extract machine-readable OpenURLs[16], possibly to be registered in an existing OpenURL resolution service;

---

[14] http://www.murst.it/atti/2000/alladm001004_01.htm.

- crosswalks for data conversion, to import/export from/to databases adopting different metadata standards;
- gateway service for web crawlers;
- full-text search capabilities;
- usage statistics;
- export facilities to build individual and/or institutional web pages for presentation and/or assessment purposes;
- implementation of a Z39.50 server in order to offer a configurable discovery service to be embedded in partner organizations' site resources, such as library catalogues;
- time stamping, copyright protection, legal deposit, PURL, and preservation issues will also be addressed.

Part of the aforesaid services may be embedded directly into each eprints repository tool.
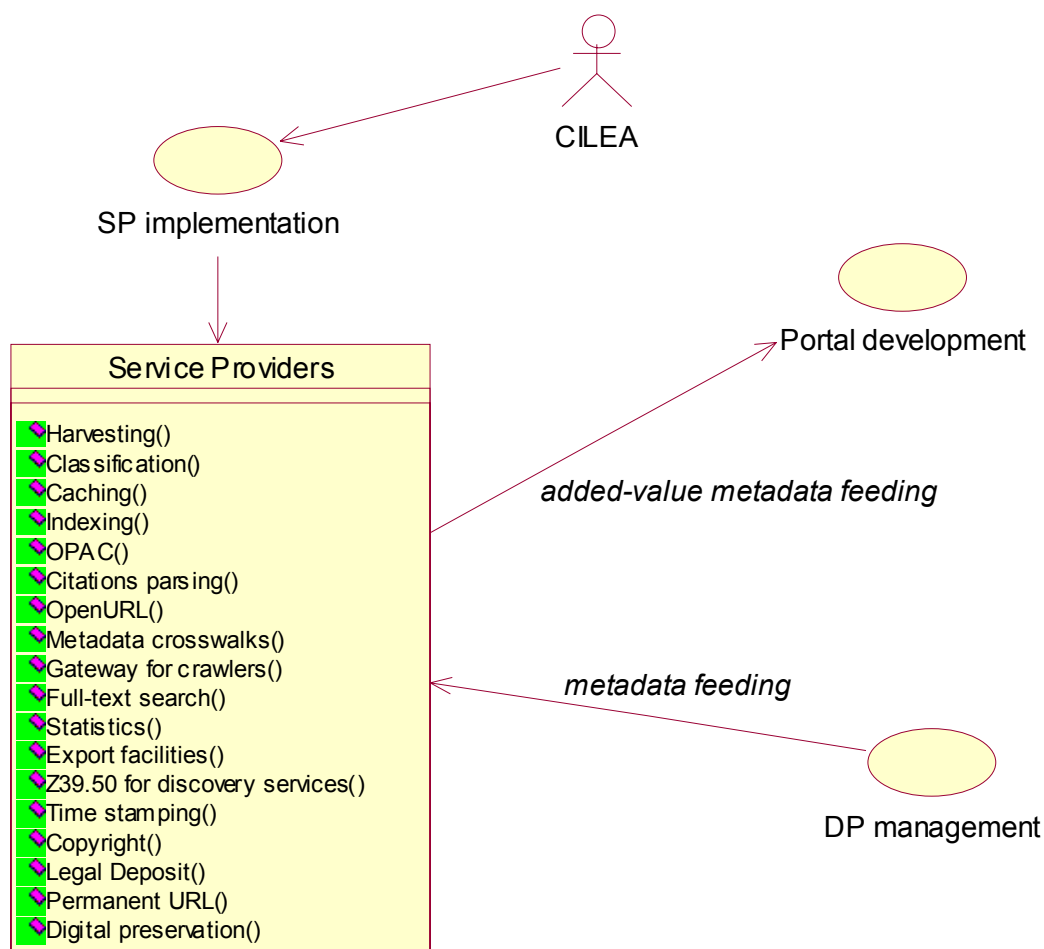

DIAGRAM 3 - SERVICE PROVIDERS

---

[15] Such as CiteSeer, also known as ResearchIndex, see: http://citeseer.nj.nec.com/
[16] http://library.caltech.edu/openurl/Public_Comments.htm

3.2. Portal

Via a search interface, a portal will provide a single point of access to the centralised data, and to resources relevant to open archive issues, together with a range of additional services, such as a user profile management system, tailored e-mail alerting services, aggregated newsfeed, and so on. The portal's web site will also host a list of Italian OA initiatives, a testbed of tools for developers, a VRD for implementators, and a forum for policy issues to be dealt with at a national level.
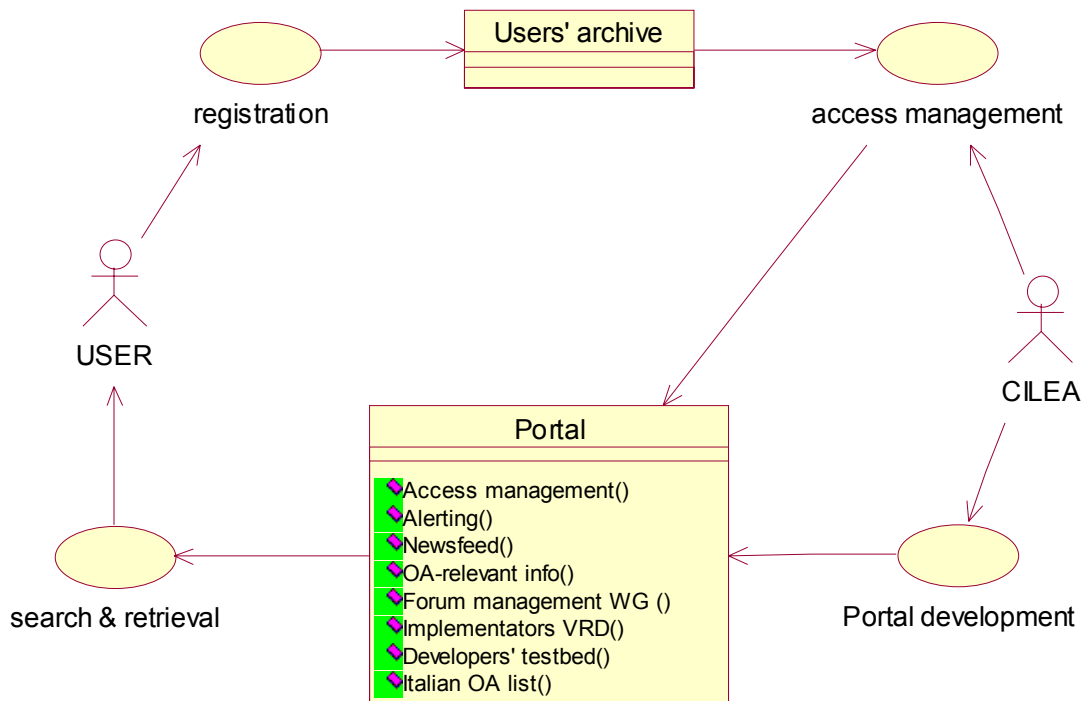

DIAGRAM 4 - PORTAL

4. Detailed plan and milestones

WP1 - Management, advocacy and marketing activities

This work package will develop along the whole course of events related to the project. It includes:
- project management;
- conferences presentations and seminars on open archives, copyright, peer-reviews processes and other related issues;
- organization of a national conference on Open Archives in Italy in 2004;
- management of a national Working Group, with the aim of providing guidelines on metadata sharing (including classification and indexing, and possibly authority lists), on copyright issues (including registration, submission, and self-archiving policies), on tools for research assessment, on advocacy within institutions;

- contacts with similar or related projects, in Italy or abroad;
- studies on users' behaviours and perspectives.

<u>Start date</u>: July 2003
<u>End date</u>: October 2004

<u>Related activities</u>: project design and evaluation, partecipation to national/international conferences on OA, partnership in national/international projects.

<u>Milestones</u>:
- Presentations: one seminar/presentation per month during nine months: autumn 2003 to mid-2004, one national conference in 2004.
- Deliverables: two progress reports on project advancement (February and July 2004), three Working Group reports on metadata, copyright, and assessment (mid-2004), two papers on users' communities, social sciences versus STM researchers (autumn 2004).

WP2 - Data Providers installation and support

This work package will install, configure, and run Data Providers at CILEA site and/or customer institutions. Support will also be provided to institutions or disciplinary groups for autonomous set up. It includes:
- set up of hardware and operative system environment for Data Providers;
- installation, testing, and evalutation of open source software packages;
- users' and implementators' technical training.

<u>Start date</u>: August 2003
<u>End date</u>: December 2003

<u>Related activities</u>: developers' training (external, internal, or self-training).

<u>Milestones</u>:
- set up of two Data Provider prototypes: December 2003

WP3 - Service Providers development and implementation

This work package will install, configure, and run several Service Providers at CILEA's site, and possibly develop open source additional software. It includes:
- set up of hardware and operative system environment for service providers prototype;
- installation, testing, and evalutation of open source software packages for every service provider separate functionality: harvesting, subject classification, caching, indexing, searching, parsing citations, OpenURL resolution, crosswalks, gateway for crawlers, full-text search, statistics, export, Z39.50 target for configurable discovery services, time stamping, copyright protection, legal deposit, PURL, and preservation.

<u>Start date</u>: September 2003
<u>End date</u>: April 2004

<u>Related activities</u>: developers' training (external, internal, or self-training).

<u>Milestones</u>:
- set up of a Service Provider prototype, with most functionalities: April 2004

WP4 - Portal design and development

This work package will design and implement a portal for Italian research literature at CILEA's site. It includes:
- search interface
- user profile management system
- tailored e-mail alerting services
- aggregated newsfeed
- list of Italian OA initiatives
- testbed of tools for developers
- VRD for implementators
- forum for national working group on policy issues

<u>Start date</u>: January 2004
<u>End date</u>: June 2004

<u>Related activities</u>: developers' training (external, internal, or self-training).

<u>Milestones</u>:
- set up of a portal prototype: June 2004

5. Time plan

Duration: activities will extend from May 2003 to October 2004.

Detailed schedule:

2003-05: CILEA Workshop on Scientific Communication & Electronic Publishing
2003-07: WP1 project design
2003-09: WP1 project assessment and WP2-WP3 first tests
2003-10: WP1 first presentation and WP2-WP3 further tests
2003-11: WP1 second presentation and WP2-WP3 implementations
2003-12: WP1 third presentation and WP2 prototypes release
2004-01: WP1 fourth presentation and WP4 beginning of design
2004-02: WP1 fifth presentation and first progress report on project advancement
2004-03: WP1 sixth presentation and national Conference - WP4 test
2004-04: WP1 seventh presentation and WP3 prototype release
2004-05: WP1 eighth presentation and WP4 implementation
2004-06: WP1 ninth presentation and release of three deliverables (WG reports) - WP4 prototype release
2004-07: WP1 second progress report on project advancement
2004-10: WP1 release of two papers on users' communities - WP2-3-4: release of production services.

CILEA Open Archives Platform - Project plan

```
        2003                    2004
         7    8    9   10   11   12   1    2    3    4    5    6    7    8    9   10
WP1  ████████████████████████████████████████████████████████████████████████████
WP2       █████████████████████████
WP3            ███████████████████████████████████████████
WP4                            █████████████████████████████
```
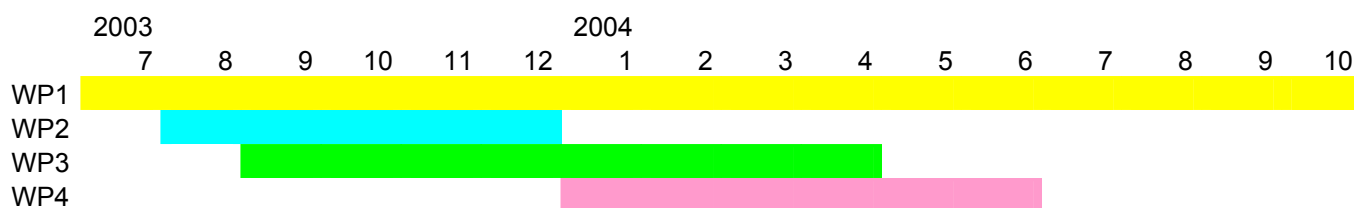
DIAGRAM 5 - TIME PLAN

Constraints: full success of the platform services may depend on the amount and quality of data gathered by Data Providers, which in turn depend on the institutional policies supporting the filling of archives.

6. Resources

Effort: 24 man-month (FTE) best case, >36 man-month (FTE) worst case

Persons involved:
- WP1: Susanna Mornati (1/2 FTE: project management), external consultants
- WP2-3-4: Susanna Mornati (1/2 FTE: project design & requirements/specifications, customers' support), Zeno Tajoli (1/2 FTE: software configuration, Perl, metadata), Daniela Dragone (1/2 FTE: DB administration, Java), Nilde De Paoli (1/4 FTE: web site design and maintenance), other internal resources (1/4 FTE: system & development support, portal design & implementation).

Software requirements: open source software, continuously developed, upgraded and supported by foreign academic institutions, is available for the purpose. Some development may be needed to adapt or enhance existing functionalities, or design and implementation of new functionalities for the Italian context or due to users' requests.

Hardware requirements: existing hardware facilities at CILEA are enough to support the testing and starting phases. An extra server will become necessary in the production phase. Additional hardware may be required in case of a dramatic extension of services and their customers.

Technology and knowledge developed along the project may be reused in related projects on electronic publishing.

7. Sustainability

CILEA's activities have different funding models. To accomplish the activities planned in this project, funding may be obtained partly from CILEA's members and customers, partly from the Italian Ministry of Education, University and Research (MIUR). This project, which is to be considered of national interest, will address institutions that rely on CILEA's services and expertise, and at higher institutional level MIUR and CRUI.
Sponsorships and partnership with other funded projects may also be taken into account. Since this project shares goals and means with several initiatives started in other countries, it is appropriate to foresee partecipation of CILEA in international

projects within European Union funding programmes, sharing tasks with partner institutions and extend goals for higher level achievements.

## 8. Evaluation

Two progress reports on project advancement will be delivered during its duration, and a Working Group will be constituted with participants from different academic and research institutions involved. Usage statistics will be produced, both for fulfilling users' needs and in order to evaluate the impact of the project on the Italian research community.

Appendix A: diagrams legenda

UML (Unified Modeling Language) diagrams have been used to present the project. The meaning of main components is explained using general, non-technical expressions, as follows:
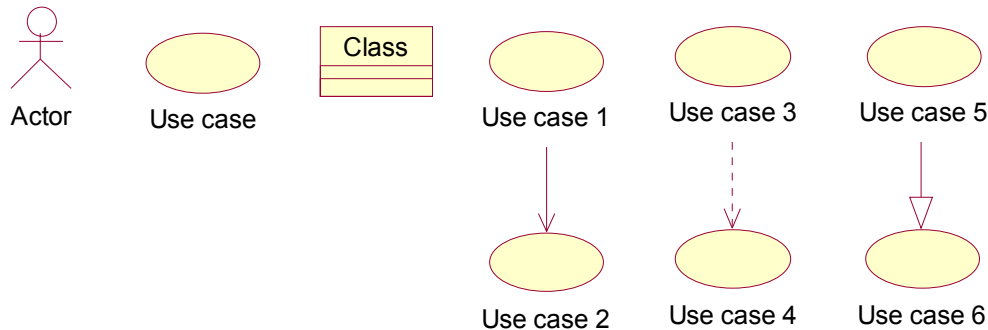
Actor: collection of roles, consistent with each *use case* interfaced.
For instance, *actors* are (see DIAGRAM 2):
1) scholarly authors toward archiving in institutional repositories;
2) CILEA toward development, implementation and management of services;

Use case: sequence of actions accomplished by a system, forming a consistent scenery toward any *actor* interfaced.
For instance, *use cases* are (see DIAGRAM 4):
1) the sequence of data search & retrieval toward users;
2) the portal development toward CILEA.

Class: collection of object sharing a series of technical features.
For instance, *classes* are (see DIAGRAM 2):
1) Data Providers;
2) Service Providers.

Association: semantic relationship among *actors*, *classes* and *use cases*.
For instance, *associations* are (see DIAGRAM 1):
1) unidirectional relationship between user and registration (arrow with continuous line), conveying data flow direction;
2) dependency relationship between registration and deposit (arrow with dotted line), conveying operations chronology;
3) generalisation relationship between institutional DP and generic DP (arrow with a triangular head), conveying that the first is a subset of the second.