D e-Infrastructures Area

Leonardo Candela, Akrivi Katifori and Paolo Manghi

1 Introduction

Quoting the e-Infrastructure home page[1] of the FP7 ICT Research Unit of the European Commission:

"The e-Infrastructures activity, as a part of the Research Infrastructures programme, focuses on ICT-based infrastructures and services that cut across a broad range of user disciplines. It aims at empowering researchers with an easy and controlled online access to facilities, resources and collaboration tools, bringing to them the power of ICT for computing, connectivity, storage and instrumentation. This allows for instant access to data and remote instruments, 'in silico' experimentation, as well as the setup of virtual research communities (i.e. research collaborations formed across geographical, disciplinary and organizational boundaries)."

In other words, e-Infrastructures support research infrastructures from the "virtual" perspective, by enabling community "actors" (researchers or their applications) to exchange their "resources" (research data and literature) by means of a controlled, regulated, digital environment. Specifically, researchers in the field of e-Infrastructure investigate solutions and methodologies enabling and facilitating the realization of e-Infrastructure platforms capable of supporting the activities of domain-specific research communities (e.g. Agriculture, ICT, Social Sciences). In general, e-Infrastructures can be considered as a combination of (i) established policies, standards and best practices and (ii) a set of technologies and tools, which together support an environment where researchers of a given domain can accomplish their daily activities in a collaborative and synergic fashion (Atkins et al., 2003; Ioannidis et al., 2005).

The main purpose of this chapter is to report how researchers investigating in the area of e-Infrastructures organize their activities of "data and publica-

[1] http://cordis.europa.eu/fp7/ict/e-infrastructure.

tion management" and themselves rely on research infrastructures to do so. Due to the early age of this field and its rather multidisciplinary computer science character, no well-established research infrastructure is available and researchers tend to follow "infrastructure-flavoured" solutions local to their organizations. As a consequence, the authors of this chapter (from the D-Lib research group at CNR, Italy and the MADGIK research group at the University of Athens, Greece) opted to approach this study by collecting a number of experiences from relevant stakeholders in the field in order to identify "local infrastructure" commonalities and "research infrastructure" desiderata.

We shall first elaborate on the strategy adopted to run this investigation, based on questionnaire-driven interviews to a number of representative organizations in the field. Subsequently, we shall present the specific case narratives, before finally drawing a summary of the current status and elaborate on possible future challenges.

2 Methodology and representativeness of the study

In order to investigate on the current status and future challenges of research infrastructures in the area of e-Infrastructure, we adopted a methodology based on questionnaire-driven interviews to experienced researchers in the field. The questionnaire[2] contains a structured set of the questions, which we perceived as crucial to gather the information necessary to gain in depth understanding of the research workflow lifecycle at the researcher's group or organization. Crucial is the distinction between literature and data, where issues such as management, exchange and Open Access are somewhat more cross-domain and mature for scientific publications and heavily domain specific and not as thoroughly investigated for research data. In the process, we collected a list of "community desiderata", intended as current issues and/or envisaged solutions which interviewees believed could contribute to improve the overall research activities of the community. The general outline of the questionnaire is the following, concentrating on four main question groups:

- **Research group profile:** general information about the research group, interests and available service and computing infrastructures.
- **Research data:**
  - **data and metadata typologies:** information on which kinds of research data the organizations deals with and which kind of metadata formats are used to describe research data.

[2] https://spreadsheets.google.com/viewform?hl=en&formkey=dGN4bnp1QWJONkdXZ3FRbEtmb2tlZ2c6MQ#gid=0.

        - data in this field are mostly *software* (source code), *software instances* (software in execution, also known as “process”), *benchmarks* (domain-specific research data collections or corpora used to validate software instances), *logs* (recorded history of actions or events, typically used to evaluate and monitor the activity of a software instance) and *statistics* (often derived from logs to evaluate software instance activities).
        - metadata can be “structured”, i.e. machine interpretable and consumable records/profiles or “unstructured”, i.e. documentation such as user manuals, specifications, installation guides, in any format (wikis, websites, document files).
    - **data lifecycle:** information on how data and metadata are produced, processed and stored.
    - **data management aspects:** information on aspects such as data and metadata versioning, provenance and preservation.
    - **data exchange:** information on how data and metadata are exchanged by researchers internally and externally to the organization.
    - **data and Open Access:** information on the awareness and status of application of Open Access principles to research data within the organization.
- **Literature**
    - **literature management:** information on the publication lifecycle established at the organization, from survey, drafting and publishing of literature.
    - **literature and Open Access:** information on the awareness and status of application of Open Access principles to publications within the organization.
- **Combination of literature and research data:** information on the awareness and status of application of literature and data interlinking within the organization.

Based on the questionnaire, we arranged interviews with a selection of key stakeholders in the European domain of e-Infrastructures. Our strategy has been that of selecting a set of organizations and individuals which are representative of wider classes of research institutions and companies, with respect to the size of the organization and research scopes. As e-Infrastructure is a rather new and multidisciplinary topic, the selection criteria cannot aim at providing a full coverage of the methodologies and research aspects carried out in the field. However, we believe the adopted perspective allows one to gain an adequate view of the European status for this novel research field.

More specifically, we approached research institutes (D-Lib Research Group, National Documentation Center and Greek Research & Technology Network (GRNET)), universities (MADGIK Research Group) and private companies (Agro-Know and Engineering R&D Unit on Clouds and Distributed Computing Infrastructures). In the following, section 3 Case narratives presents the information collected in the interviews, section 4 Current status synthesizes the interviews and reports on the current status on research infrastructures for e-Infrastructures, while section 5 Desiderata and future directions concludes the chapter elaborating on researchers desiderata and identifying future challenges to address them.

3 Case narratives

In the following sections we present the summary of the interview for each organization. For each case narrative, we provide:

- general information about the organization, which includes allocation of people over research activities and a description of its local service and computing infrastructures;
- a description of the organization research objectives and projects;
- a description of the organization's typical workflow in the production of literature and data.

3.1 D-Lib research group

3.1.1 General information

The D-Lib research group, led by Dr. Donatella Castelli, consists of around five researchers, 15 technicians and three administrative staff. It is part of the Networked Multimedia Information Systems Laboratory (NeMIS), which consists of 48 researchers and technicians conducting research and development activities on algorithms, techniques and methods for information modeling, access and handling, as well as new architectures and system services – P2P and Grid-based (Foster and Kesselman, 1999) – supporting large networked multimedia information systems. The NeMIS laboratory is in turn part of the Institute of Information Science and Technologies (ISTI) of the Italian National Research Council (CNR), which is organized in 16 laboratories and is committed to producing scientific excellence and playing an active role in technology transfer.

Organization of activities D-Lib group research activities are organized in two parallel tracks: research subjects and projects. Each research subject is managed by one researcher and is assigned a group of co-researchers and technicians to address prototypes and products releases; both researchers and technicians can be assigned to multiple branches. Each project is assigned to one researcher, who becomes responsible and ISTI representative for the project, and generally involves one or more research subjects. In order to serve the project needs, the project responsible is also in charge of coordinating the researchers in charge of the individual subjects to accomplish the project objectives.

Computing infrastructure In order to accomplish research and development tasks, researchers are equipped with personal workstations and can count on a shared computer infrastructure, offering a central processing unit (CPU) cluster equipped with a separate storage area network as described in Table D.1

Table D.1 D-Lib computing infrastructure

CPU	Cores: – 10 × dual AMD Opteron Processor 252 (no hvm) – 2 × dual Quad-Core AMD Opteron Processor 2356 – 2 × dual Six-Core AMD Opteron Processor 2427 – 2 × dual Quad-Core HT Intel(R) Xeon(R) CPU E5630 – 2 × single Dual-Core AMD Opteron Processor 1222 – 2 × single Quad-Core IntelQ6600 – + other miscellaneous hardware: total Total: 88 cores (104 cores considering hyper-threading) Total: 516 GB ram on the cluster
Storage	Protocol: SCSI, SAS, SATA Disks: 42 drives, raid1 pairs, effective 5.7 Tb
Storage area network	Protocol: AoE Disks: 16 sata drives, raid1 pairs, effective 7.2

3.1.2 Research objectives and projects

The team focuses on the following research and development activities regarding the realization of sustainable e-Infrastructures for research:

- foundations and data models of digital libraries;
- digital library management systems: design and realization of systems for the construction of digital library systems (Candela et al., 2008);

- data management and curation services: e.g. authority file management, bulk-data feature extraction and transformation, time-series management, compound objects management (DRIVER-II project[3]);
- design and development of frameworks (middleware): enabling large-scale data infrastructures (D-NET software Toolkit[4] and gCube Toolkit[5]);
- Cloud services: (Dikaiakos, Katsaros, Mehra, Pallis and Vakali, 2009), service on-demand frameworks providing abstractions over different Cloud platforms (VENUS-C project[6]);
- design and development of virtual laboratories or virtual research environments: in the context of large-scale data infrastructures (D4Science-II project[7]);
- foundation elements of "global" infrastructures and "ecosystems" of infrastructures: (see GRDI2020 project[8]).

The team has been involved in many EU-funded projects relevant to the topics of e-Infrastructures, namely:

- **FP6 projects:** DILIGENT (no. 004260, Scientific Coordinator) – see project description in 3.5 MADGIK research group – BELIEF (no. 026500) and DRIVER (no. 034047).
- **FP7 projects:** EFG (no. 517006), DRIVER II (no. 212147), D4Science (no. 212488), BELIEF II (no. 223759), D4Science-II, HOPE, VENUS-C, GRDI2020 and OpenAIRE.

Among these, the most relevant and still ongoing are:

- **DRIVER Targeted Project (IST FP6) and DRIVER II CP/CSA (INFRA FP7):**[9] DRIVER is a multiphase effort whose vision and primary objective is to create a cohesive, robust and flexible pan-European infrastructure for digital repositories. DRIVER has established a network of relevant experts and Open Access repositories. DRIVER-II aims to consolidate these efforts and transform the initial test-bed into a fully functional, state-of-the art service, extending the network to a larger confederation of repositories
- **OpenAIRE:**[10] OpenAIRE aims to establish and operate a data infrastructure for connecting EC FP7 projects with the scientific publications funded under such projects. The infrastructure allows the Commission

[3] http://www.driver-community.eu.

[4] http://www.d-net.research-infrastructures.eu.

[5] www.gcube-system.org.

[6] http://www.venus-c.eu.

[7] http://www.d4science.eu.

[8] http://www.grdi2020.eu.

[9] http://www.driver-repository.eu.

[10] http://www.openaire.eu.

and organizations participating to EC project to measure the impact of the Open Access mandates (Clause 39) across FP7 projects in several research areas. The group is responsible for the realization of the enabling layer of the infrastructure (core infrastructure services: e.g. information service, orchestration services) and for the data management and curation part.

- **D4Science CP/CSA (INFRA FP7) and D4Science II CP/CSA (IP FP7):**[11] D4Science and its continuation, D4Science-II, is a European e-Infrastructure project, co-funded by the European Commission's Seventh Framework Programme for Research and Technological Development. D4Science-II will develop technology and methodologies that will enable sustainable interoperation of multiple, diverse and heterogeneous data e-Infrastructures that have been established and are currently running autonomously, thereby creating e-Infrastructure ecosystems that can serve an expanded set of communities dealing with complex, multidisciplinary challenges whose solution is beyond reach with existing resources. Furthermore, D4Science-II will use the existing D4Science e-Infrastructure as a hub to bring and hold together several established scientific e-Infrastructures and, thus, set up a prototypical instance of such an e-Infrastructure ecosystem. The group is responsible for the realization of the enabling layer of the infrastructure (core infrastructure services: e.g. information service, orchestration services) and for the data management and statistics part.

Research data With respect to research data, the team produces open source software, software instances, technical websites, logs and test results, with related benchmarks. In particular, software is produced by adopting rigid programming policies, from development and testing to integration and production.

Researchers and technicians store their data relying on a local service infrastructure integrating tools such as TRAC (road maps and tickets), SVN (software versioning), BSCW (document and calendar sharing) and wikis, made available across several projects to a pool of "single sign-on" authorized users.

Software data, when possible, are searched and fetched from well-known software web sources (e.g. SourceForge, Google projects, Apache projects) and re-used as part of the resulting products. Similarly, the team may contribute to the open source community.

Software instances are also regarded as available and exchangeable research data. In this context, a software instance is a *service*, i.e. running instance of

[11] http://www.d4science.eu.

software accessible through the web, made available for access by authorized consumers through a service-oriented infrastructure.

Structured metadata formats for research data are mainly proprietary (e.g. software and software instances) and may change depending on the infrastructure implementation. For example, services are described by metadata properties (obliged to include the URL of the service) which enable its discovery based on given criteria and subsequence usage. Such metadata are typically proprietary and target the requirements of service consumption raised by the application domain. In other cases, for example documentation (see unstructured metadata below), metadata formats are imposed by the specific tool's default (e.g. BSCW for technical reports).

Unstructured metadata are continuously produced to support the software lifecycle (e.g. specifications, software documentation, user and installation manuals, websites) and to describe software results or applications (e.g. white papers, technical reports).

Desiderata: most of the software products (research data) in the literature are prototypes and therefore available only through organizations, groups or researchers' websites. As such, they cannot be easily discovered, located and re-used. A community e-Infrastructure serving the purpose of software and documentation sharing would ultimately benefit the community, by guaranteeing standard metadata descriptions, collaborative development and degrees of quality certification.

Literature The team is very active on publication production, as it considers it an important mean of dissemination. The survey phase of publication is typically carried out relying on known publication sources, such as Google, Google scholar, Wikipedia and publisher websites (e.g. Elsevier, ACM, IEEE) and less known but specific sources, such as the DRIVER infrastructure. The phase of publication drafting is typically carried out by physical meeting and multi-hand editing, using shared editors such as Google docs and file-sharing tools such as Dropbox, BSCW and email.

It is mandatory for researchers at ISTI to upload publications metadata and full text, with proper access policies, into the PUMA-ISTI repository.[12] Through PUMA, publications are made available to Google Scholar or other aggregators, such as the DRIVER infrastructure and BASE.

Desiderata: there is no web source focusing on scientific publications on e-Infrastructure research. Relevant results in the field are to be discovered with parallel searches across several websites and cumbersome refinement and skimming cycles, often by reading the article abstracts or full text. A community e-Infrastructure serving the purpose of sharing e-Infrastructure

[12] PUblicationMAnagement, http://puma.isti.cnr.it.

literature would ultimately benefit the community, by guaranteeing standard metadata descriptions and tailored focus.

Desiderata: there are no conferences or journals focusing on e-Infrastructure research. Only a few conferences, such as TPDL (formerly ECDL) or IFLA have “special tracks” dedicated to the topic. Most submissions of scientific publications are therefore sent to conferences and journals whose main topic “touches” that of this research, i.e. service-oriented architectures, digital libraries, knowledge management, Grid (Foster and Kesselman, 1999), etc. In some cases, conferences and journals specific to the application domain of a given e-Infrastructure may also accept submissions of “methodological” papers. The domain of e-Infrastructure has reached sufficient maturity to deserve special venues and classification in the computer science world.

Combining literature and data The group always refers from publications the websites of products cited in the narration. However, this practice follows common sense rather than given policies. Although it would be desirable in many cases, the team is not aware of any best practices or tools for managing or providing combinations of literature and data.

Open Access The team is well aware of Open Access mandates, as it works on projects such as DRIVER and OpenAIRE which are trying to advocate and promote its adoption across Europe and beyond. In particular:

- **research data:** data are stored within ISTI infrastructure and not made openly available to third party consumers, which on request can be granted access to the data, i.e. Open Access policies are figured out case by case. Exceptions are made for software data, which are open source (hence Open Access) and directly available from the product websites;
- **literature:** researchers, when having to choose between equivalent publication venues, tend to prefer those supporting Open Access policies. Unfortunately, most relevant forums in the fields often rely on publishers that do not support Open Access rights.

3.1.3 Research workflows

The typical research production workflow of the team consists of the following phases:

1. problem identification, based on experience and intuition;
2. survey of the literature and data (software, documentation, reports) to find similar or useful (i.e. reusable) resources and “certify” the validity of the intuition;

3. design of a solution, possibly reusing existing data (e.g. software);
4. production and maintenance of unstructured metadata (e.g. software documentation, installation guides, roadmaps, technical reports);
5. development of prototype;
6. definition of benchmarks and testing;
7. release of a product;
8. publication writing and publishing.

Such steps are accomplished by exploiting the local service and computing infrastructure available at ISTI in combination with the above mentioned web tools for discovery, collaborative production and sharing of literature and data.

3.2 Agro-Know

3.2.1 General information

Agro-Know Technologies[13] is a new research-oriented enterprise that focuses on knowledge-intensive technology innovation for agriculture and rural development. The company focuses on realization of systems and services for organization and delivery of agricultural knowledge, promoting the usage of semantic web technologies and Web 2.0 tools. It also explores their deployment and testing in application domains such as education and training, commerce and public administration.

Agro-Know spun off from a group of researchers working in R&D projects in GRNET SA[14] (Greek Research & Technology Network) and today counts 15–20 employees, assigned to research and innovation, design and development activities.

Organization of activities Agro-Know is internally organized in three research teams of about five people, where one or two members are dedicated to software development. In parallel with the research teams, the company has a technical development team, led by one technical coordinator, whose purpose is to support cooperation and sharing of resources among the research teams.

Computing infrastructure The company supports an intranet connecting workstations of researchers and developers, plus common servers for file sharing. In many cases, research teams rely on computing infrastructures provided by the organizations they cooperate with or they work for.

[13] http://www.agroknow.gr.

[14] http://www.grnet.gr.

3.2.2 Research objectives and projects

The main e-Infrastructure research objectives of the company are:

- e-Infrastructures for agricultural research data;
- e-Infrastructures for museums of Natural History with extensive content on biodiversity, botany, etc.;
- e-Infrastructures for education;
- repository platforms adaptable to diverse application scenarios.

Agro-Know gives special emphasis in understanding the needs of the user communities they work with. They feel that a lot of interesting e-Infrastructures research issues can be identified through efficient observation of the user practices, the in-depth understanding of the problems they face and the support that the researchers need in their everyday work.

Among the projects that the Agro-Know team has been or still is involved are:

- **Organic.Edunet:**[15] a multilingual federation of learning repositories with quality content, which support the awareness and education of European Youth about topics related to Organic Agriculture and Agroecology;
- **Natural Europe:** an integrated effort to make knowledge residing in a vast array of Natural History Museums (NHMs) commonly accessible. Accessibility means that the impressive abundance of high-quality digital content is pedagogically structured and presented to the consumer in personalized and contextualized ways;
- **ARIADNE:**[16] an infrastructure of a distributed network of learning material repositories.

Research data Agro-Know creates and processes data such as software, system logs and analytics described by structured and proprietary metadata and by unstructured metadata (e.g. documentation, XML/RDF data models, websites).

The data are produced on the workstations (or private laptops) and then stored for sharing and exchange on the local computing infrastructure through version systems (e.g. Git).

Research data are often exported through project websites (e.g. software, technical reports).

Desiderata: privacy policies at different organizations have hindered reuse and publication of log-file data. An e-Infrastructure for research data in this area could also impose common protection policies and access protocols and ensure these are respected by participating organizations.

[15] http://portal.organic-edunet.eu.

[16] http://www.ariadne-eu.org.

Literature Researchers survey and share the literature through Google Scholar and Mendeley, but in general no collaborative tool (e.g. Google Doc-like) is used. Publications are mostly drafted on workstations (and private laptops), exchanged by email and eventually stored within the company's file server folder structure. However, when drafted in collaboration with external research teams, web tools such as BSCW, Dropbox, Google Docs and wikis may be adopted.

Desiderata: researchers find difficult to share their bibliography, i.e. to exchange their references in a meaningful and organized way. An e-Infrastructure for literature in this area could offer to researchers in the field services for ensuring controlled sharing of publications and bibliographies.

Combining literature and data Researchers at Agro-Know are not aware of publishers that allow the combination of literature and data nor of policies and best practices that would enable such combination.

Desiderata: although the benefits of this approach are evident, based on the experience at the company their application may encounter the issues of:

- metadata: standard representation formats for most of the data do not exist;
- privacy issues: in the case of log files the publication of the datasets may not be possible due to privacy laws/policies;
- unavailability: some data are not available for external referencing, i.e. not available through the internet, e.g. logs on a server.

Open Access issues Agro-Know supports and promotes Open Access. In particular:

- literature: researchers favour publishers supporting Open Access. When possible, publications are public in the project websites and also on the company website as a draft with a link to the editor site;
- research data: the software produced by the company are made available as open source and the educational material with a Creative Commons licence.

Desiderata: researchers believe it is crucial that funding agencies impose Open Access for the results of the projects they fund. As a side effect, this would push publishers at finding new business models.

3.2.3 Research workflows

The general workflow employed in each of the Agro-Know projects is the typical specification, design, development and documentation and evaluation cycle:

1. understanding and defining: describing the objectives of the project along with the partners that may be involved;
2. requirement analysis: producing requirement analysis documents by close interaction with the recipients of the technology to be delivered;
3. design: producing functional and architectural specifications of the technology to be developed, in strict collaboration with the recipients;
4. development: implementation of the technology based on the given specifications. Developers tend to re-use, adapt and customize core technology developed at Agro-Know and to re-use third-party open source software;
5. documentation: in parallel to development, researchers focus on technical reports or publications writing in collaboration with the technology recipients (e.g. user communities) and with project partners;
6. testing and deployment: after strict testing and evaluation, the technology is released and put into production. The underlying software is made available openly to the public, unless project copyright obligations are involved.

3.3 National Documentation Center (EKT)

3.3.1 General information

The National Documentation Centre[17] (EKT) is the Greek national infrastructure for scientific documentation, online information and support services on research, science and technology. The Centre was founded in 1980. It is integrated with the National Hellenic Research Foundation (NHRF) and is supervised by the General Secretariat for Research and Technology of the Ministry for Development.

EKT is both a major e-Infrastructures developer in Greece and one of the main providers for science and technology services and content, as it operates, among others, the Science and Technology digital library, including the digital library of Greek PhD theses.

Organization of activities EKT operates as partner of several projects and to each of them it assigns one coordinator supported by a research team. Research teams may share members and operate over more than one project. EKT elects one of the project coordinator as research supervisor of all projects, in order to maximize re-use of resources and collaboration.

EKT also undertakes close collaborations with external research teams. The most relevant experiences are with the institutes of the National Hellenic

[17] http://www.ekt.gr.

Research Foundation (the Pandektis project[18]) and with GRNET, in the context of GÉANT project.[19]

Computing infrastructure The EKT computing infrastructures is described in Table D.2

Table D.2 EKT computing infrastructure

CPU	– Virtualization platforms comprising 8 servers, 64 processing cores, 192 GB of memory in high availability configuration – 77 physical and virtual CentOS Linux, Redhat, Windows 2003 and Sun Solaris servers – 36 high-end 64-bit Intel Xeon, AMD Opteron and Solaris SPARC physical servers
Storage	– Storage Area Networks, coupled with 5 FC switches providing 83 TB of raw disk space – LTO3 and LTO4 tape libraries with 156 TB raw capacity
Storage area network	– Fully redundant IP network featuring no Single Points of Failure, Gigabit Ethernet end to end, redundant 1 Gbps firewall, border/core router configuration, VPN – Active Directory/LDAP infrastructure, high capabilities work stations, Gigabit Ethernet until the end user

3.3.2 Research objectives and projects

EKT research teams have expertise in the following research topics and activities:

- aggregation of heterogeneous resources;
- Open Access infrastructures;
- websites;
- digital library technologies;
- repository platforms;
- digitization;
- organizing national and international working groups for thematic studies to produce best practice or policy documents.

In particular, EKT participates and in some cases coordinates several research projects, both European and national. Those related to e-Infrastructures include:

[18] http://pandektis.ekt.gr.
[19] http://www.geant.net.

- **EuroRIs-Net and its continuation EuroRIs-Net+:** EuroRIs-Net is a coordination action supports the network of national contact points for Research Infrastructures;
- **OpenAIRE[20] project (EC FP7):** see project description in 3.1 D-Lib research group;
- **Pandektis:** Pandektis aimed to provide free access to 11 integrated and scientifically elaborated collections produced by the three humanistic Institutes of the National Hellenic Foundation for Research: Institute of Greek and Roman Antiquity, Institute of Byzantine Research and Institute of Neohellenic Research;
- **Argo:**[21] Argo aimed at realizing an environment which facilitates Open Access and search across bibliographical information resources available in Greece as well as abroad.

Research data Software is the main forms of data that EKT produces, together with unstructured metadata in the form of technical reports. Data and unstructured metadata are stored for internal sharing between the research teams in common file servers at EKT, with different access rights for different groups of users and over different projects. For software, a version control management system is used, as well as issue tracking (Mantis), while technical reports are drafted collaboratively as wikis.

Data exchange with groups of other organizations is accomplished mainly through the project websites.

Research Literature Researchers survey the literature through Google Scholar[22] and Scopus[23] and manage references using CiteULike.[24] For collaborative drafting they use SVN. Finally, preferred venues for publications are conferences such as TPDL (formerly ECDL) and IFLA and journals related with the topic of interest. Interestingly, some PhD theses have been followed in cooperation with universities and research centres.

Publications are made available for web search and access through the Helios[25] repository, realized at EKT.

Combining literature and data Combining data and literature is considered a good practice at EKT. On the other hand, the absence of best practices and tools available to support it does not make it an option.

[20] http://www.openaire.eu.

[21] http://argo.ekt.gr.

[22] http://scholar.google.com.

[23] http://www.scopus.com.

[24] http://www.citeulike.org.

[25] http://helios-eie.ekt.gr/EIE.

Open Access issues EKT is one of the first organizations in Greece to actively adopt and promote Open Access and one of the first to sign the Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities. It is the creator and owner of the main website for Open Access in Greece,[26] which provides information on best practices, policies and existing repositories that have adopted Open Access, for example.

3.3.3 Research workflows

EKT adopts clearly defined procedures for research and development, specifically:

1. requirement analysis: producing requirement analysis documents by close interaction with the customers;
2. design: Producing functional and architectural specifications of the technology to be developed;
3. development: Implementation of the technology based on the given specifications, possibly reusing EKT software. Progress is monitored by the project coordinator and by the EKT research supervisor;
4. documentation and publications: In parallel to development, researchers focus on technical reports or publications writing in collaboration with the technology recipients (e.g. user communities) and with project partners;
5. testing and deployment: After strict testing and evaluation, the technology is released and put into production.

3.4 Greek Research & Technology Network (GRNET)

3.4.1 General information

GRNET SA[27] operates the Greek Research & Technology Network, according to the operating model described by the EU Research and Education Networks. It operates both at a national and international level and constitutes the setting for the development of innovative services for the members of the Greek research and education communities. GRNET SA connects more than 90 institutions, including all Greek universities and technical and research institutes, as well as the public Greek School Network, supporting more than 500,000 users all over the country. Moreover, it provides local interconnection services to the main Greek Internet providers, through the Greek Internet Exchange/GR-IX[28] infrastructure. GR-IX started operating

[26] http://openaccess.gr.

[27] http://www.grnet.gr.

[28] http://www.gr-ix.gr.

in 2008 and provides interconnection at Nx10 Gbps, enhancing the quality of internet service and infrastructure nationwide.

Organization of activities GRNET's technical personnel are organized in research groups, which in turn can be assigned to one or more projects. Researchers and developers can participate to and collaborate with several groups and projects, for both publication writing and software development activities.

Furthermore, GRNET collaborates via EC projects with major European institutes that work on infrastructures, such as CERN.

Computing infrastructure GRNET's computing infrastructure is presented in Table D.3. Occasionally, the activities may require Cloud (Dikaiakos et al., 2009) resources rental, to acquire CPU and data storage capabilities on demand.

Table D.3 GRNET computing infrastructure

CPU	– 26 servers – 512 cores
Storage	– 200 TB storage

3.4.2 Research objectives and projects

The main research topics at GRNET are:

- e-Infrastructures for research infrastructures;
- Grid solutions (Foster and Kesselman, 1999);
- service Cloud solutions;
- access to digital content.

Among the projects that GRNET has participated in are:

- **StratusLab:**[29] StratusLab is developing a complete, open-source Cloud distribution that allows Grid and non-Grid (Foster and Kesselman, 1999) resource centres to offer and to exploit an "Infrastructure as a Service" Cloud. It is particularly focused on enhancing distributed computing infrastructures such as the European Grid Infrastructure[30] (EGI).

[29] http://stratuslab.eu.

[30] http://www.egi.eu.

- **Organic.Edunet:**[31] Organic.Edunet had as its aim to facilitate access, usage and exploitation of digital educational content related to Organic Agriculture (OA) and Agroecology.

Research data GRNET researchers deal with research data such as software, virtual machine images/appliances, websites and system logs. These are often accompanied by unstructured metadata in the form of manuals, documentation and technical reports. Data and metadata are stored and archived in server storage devices private to the groups. Unstructured metadata are often in Latex and multi-hand drafted with the support of a version control system. Similarly, software is organized and managed through version control systems.

As for metadata, GRNET tends to use proprietary formats for software and currently is designing metadata standards for virtual machines in collaboration with external groups (Dublin Core model and RDF encoding).

Data exchange between members of the group and across several groups is made possible through wikis, which are used as structured and organized directories to the data files. In general, data are open for others to use, except when external collaborators require a non-disclosure agreement.

Desiderata: exchanging research data with external groups in different projects is made difficult by the adoption of different version control systems. An e-Infrastructure for this research community may offer services for storing and sharing research data based on common formats and policies to be adopted as standards by the community.

Research literature GRNET researchers focus more on software development than on publication writing. As such publication management is not accomplished through specific tools. When surveying and drafting Google and Mendeley might be used to search publications and manage references. Researchers mostly publish at conferences and journals.

Combining data and literature Combining data and literature would be considered very useful but is not yet an option as there are no best practices to follow or wide-spread tools available to support it.

Open Access issues GRNET is aware of the advantages of Open Access policies, but is not pursuing them actively. Specifically:

- literature: researchers do not prefer Open Access publisher to others and do not invest in buying Open Access licences;

[31] http://www.organic-edunet.eu.

- research data: software data are usually available as open source (e.g. Apache 2 licence) and technical reports are available with a Creative Commons licence.

3.4.3 Research workflows

The GRNET e-Infrastructures team focuses mostly on software development, less on publication writing, but does not implement strict development procedures. To achieve its objectives, the team exploits collaboration tools, both for software development and technical report writing, and adopts design and development methodologies that may vary from project to project.

3.5 MADGIK research group

3.5.1 General information

The Management of Data, Information, & Knowledge Group[32] (MADGIK), led by Prof. Yannis Ioannidis, is part of the Department of Informatics and Telecommunications[33] of the School of Sciences of the National Kapodistrian University of Athens. Research and development activities within the department cover a wide spectrum of information and communication technologies. The group has a rich and long experience in several topics of computer science including digital libraries (information integration and access, Grid-services, cultural heritage systems) and e-Infrastructures.

Organization of activities The MADGIK group counts around 40+ members, including five faculty staff, several R&D staff and students at all educational stages. Being active in research and development, it includes 15 full time technical people, organized in R&D project-dedicated teams, each led by team leaders and supervised by the scientific coordinator.

The group is in close collaboration with other groups of the same organization for publication writing and software development issues and has a strong and long tradition of cooperation with groups of other organizations.

Computing infrastructure The group has a local storage and computing infrastructure, consisting of personal workstations and shared servers in a local network, organized in virtual machines. In projects such as D4Science-II, part of this infrastructure joins a larger development and execution environment that consists of a cluster of 110 CPUs with 300 GB RAM and 15 TB of storage.

[32] http://madgik.di.uoa.gr.

[33] http://www.di.uoa.gr/en.

3.5.2 Research objectives and projects

The MADGIK group has the following general research objectives:

- databases and information systems: Data repositories, query optimization, personalization, intelligent databases, etc.;
- digital libraries;
- human computer interaction: user interface for databases, complex data visualization;
- scientific repositories: scientific experiment management, data repositories, workflow management.

It participates and has participated in a large number of national and European projects related to e-Infrastructures which include:

- **OpenAIRE project (EC FP7):**[34] (see project description in 3.1 D-Lib research group) the group focuses on designing and developing user interfaces and end-user functionality services, as well as on services for the integration of access statistics collected from European repositories.
- **DRIVER Targeted Project (IST FP6) and DRIVER II CP/CSA (INFRA FP7):**[35] (see project description in 3.1 D-Lib research group) the group focuses on end-user functionality services, such as user profiling, user recommendations and "generic" portals dynamically adaptable to match functional requirements of end-users of different communities;
- **D4Science CP/CSA (INFRA FP7) and D4Science II CP/CSA (IP FP7):**[36] (see project description in 3.1 D-Lib research group) the group focuses on optimized and distributed search services, as well as on highly configurable data transformation services.
- **DILIGENT Integrated Project (IST FP6):**[37] the main objective of DILIGENT (Castelli, Candela, Pagano and Simi, 2005) has been to create an advanced testbed for knowledge e-Infrastructure that will enable members of dynamic virtual e-Science organizations to access shared knowledge and to collaborate in a secure, coordinated, dynamic and cost-effective way.

Research data The group produces mostly research data in the form of software, software instances, benchmarks, experimental data, XML, system logs and websites. Software is stored and versioned through SVN services, in some cases shared with project partners.

[34] http://www.openaire.eu.

[35] http://www.driver-repository.eu.

[36] http://www.d4science.eu.

[37] http://diligent.ercim.eu.

Unstructured metadata, in the form of technical reports and project deliverables, are compiled (possibly in collaboration with other project partners) and exchanged through e-mail when edited. Tools such as Google Docs or common project wikis may be adopted for collaborative editing but are not the rule.

Depending on the domain of the e-Infrastructure to be delivered, domain-specific research data may be collected and used as benchmarks; e.g. images and raw scientific data, audio and video, publication full texts, big data, time-series. Interestingly, the work space resulting from the D4Science project is used for benchmark data storage and exchange by the group itself. This platform has been developed to support scientific research in general with environmental and maritime data as the use cases and allows data management and exchange through web user interfaces. Similarly, metadata formats of domain-specific research data may be regarded as benchmarks; examples vary from standard, e.g. Dublin Core, Darwin Core, SDMX, ISO for geographical data, to proprietary formats.

For software and software instances, custom metadata may be used, in agreement with the specific project requirements, which in turn depend on shared development policies. The use of custom metadata for software instances has been the result of user or system needs, as the standards were not defined or sufficient (e.g. an example is the need to record in the metadata service dependencies). Technical reports are rarely annotated with metadata but it is planned to make this annotation standard within the group in the near future.

Desiderata: after the end of an EC project, consortiums have an obligation to keep the resulting reports only for a few years. The EC project BELIEF provides a digital library where documents of past projects can be stored for future storage in time. However, it would be desirable if funding agencies, such as the EC, would provide a “place” (namely an infrastructure) where past and ongoing projects could store and retrieve their data outcomes, from software to technical reports and deliverables.

Research literature Scientific publications are exchanged through e-mail and rarely edited through collaborative tools, like Google Docs. In some cases, some of the authors may be reluctant to learn and use a new collaborative tool, so e-mail exchange is the more common practice.

In order to search for publications, tools like Citeseer and Google Scholar are more commonly used and, to a lesser extent, the DRIVER infrastructure. The group’s preferred publication forms are conferences, online and print journals and PhD and MsC theses.

Combining data and literature The group believes in the publication of data combined with literature, as a mean to verify the experimental results and conclusions of the publication. However, it does not implement those practices, due to the lack of standards and tools.

Open Access issues The group supports Open Access for publications and also adopts it, although not as a strict policy. The reason is that the top conferences and journals touching the fields typically do not implement Open Access business models.

Most of the research data and metadata are open but exceptions exist:

- software data: the group tends to adopt GPL licences and open source;
- unstructured metadata: technical reports and documentation are available openly on the wiki, except for the project managerial/financial ones;
- logs: service activity logs are used for debugging purposes and for measuring the usage of the infrastructure from several perspectives, including end-users and applications. As a consequence, logs can be released to third parties only after proper permissions, as they may be used to infer private information.

3.5.3 Research workflows

The group works on system design and development based on research findings and on relative scientific publications. When operating in the context of a project whose aim is to deliver an e-Infrastructure, the typical workflows consists of the standard phases of requirement analysis, design and implementation, by reusing, experimenting or devising research achievements and solutions of the research group. Design, development and testing of software are often carried out in cooperation with project partners, by sharing hardware and supporting tools. Research papers are often presenting a system or part of it, together with experimental results which prove its effectiveness or quality.

3.6 Engineering R&D Unit on Clouds and distributed computing infrastructures

3.6.1 General information

Engineering Group is Italy's largest systems integration group and a leader in the provision of complete IT services and consultancy. Engineering Group has about 6500 employees and 35 branch offices, throughout Italy, in Belgium and (outside the EU) in Brazil. The Engineering Group operates through seven business units: Finance, Central Government, Local Government and Healthcare, Oil Transportation and Services, Utility, Industry and Telecom, supported by an SAP transverse skills centre and by its Central Office for Research & Innovation, with researchers active in Italian and EU projects. Engineering was one of the first Italian companies to adopt the Quality standard ISO 9001 in the early 1990s. Since 1996 the company has adopted NATO standard AQAP 2110/160 certification. And recently the production units have been certified CMMI® level 3. The Pont Saint Martin Service Centre (PSM) provides to more than 100 Italian and international customers, 40,000 workplaces, 1000 remote connections, 10,000 electronic mail boxes and about 7000 SAP users. The R&D department is organized to work in strict cooperation with business divisions in order to facilitate knowledge and technology transfer.

The Engineering R&D Unit is involved in the NESSI[38] and NEM ETPs[39] initiatives and in a number of Grids (Foster and Kesselman, 1999) and Cloud (Dikaiakos et al., 2009) related initiative including VENUS-C (see 3.1 D-Lib research group), VisionCloud, Passive, TEFIS,[40] ERINA4Africa,[41] ERINA+,[42] ARISTOTELE[43] and D4Science-II (see 3.1 D-Lib research group). The Engineering team interviewed consists of 16 members.

Organization of activities

Research and development activities are managed by dedicated teams that are formed by taking into account the requirements of the specific activity and evolve during the activity itself, e.g. new members can be added or members having different expertise might replace previously allocated members. Members of the group partake to multiple activity teams. The overall goal is to maximize the use of human resources.

Computing infrastructure

The infrastructure supporting the activities of the interviewed group consists of 16 workstations (the policy is to have one workstation per group member) plus the computing resources listed in Table D.4. In addition to that, the team makes use of resources acquired through one or more Cloud infras-

[38] http://www.nessi-europe.com.

[39] http://www.future-internet.eu/news/view/article/the-cross-etp-vision-document.html.

[40] http://www.tefisproject.eu.

[41] http://www.erina4africa.eu.

[42] http://www.erinaplus.eu.

[43] http://www.aristotele-ip.eu.

tructures, including Windows Azure, Barcelona Supercomputing Center and Engineering Group data centre.

Table D.4 ENG computing infrastructure

CPU	12 Servers (bi processor – quad processor)
Storage	2 TB
Storage area network	1 SUN (1.7 TB)

3.6.2 Research objectives and projects

The Distributed Computing R&D group focuses on a number of research and development activities including:

- software configuration, build and testing;
- authorization, authentication and accounting in distributed infrastructures including service-oriented architectures (Lomow and Newcomer, 2005), Grid (Foster and Kesselman, 1999) and Cloud domains;
- Grid and Cloud computing (focusing on their exploitation in Real Business ENvironments).

The team has been involved in many EU-funded projects relevant to the topics of e-Infrastructures, namely:

- **D4Science-II:**[44] actually the third phase of a project started with the name of DILIGENT (Castelli et al., 2005) where the Engineering has been involved since the beginning. D4Science-II is developing an infrastructure enabling the interoperation of diverse infrastructures that are running autonomously, thereby creating ecosystems that can serve a significantly expanded set of communities. In this project, Engineering mainly works on the design and implementation of security-related solutions, focusing on interoperability aspects and takes care of the overall coordination of the integration, testing and distribution activity;
- **VENUS-C:**[45] an FP7 Research Infrastructures project, coordinated by the Engineering team is building open source facilities to provide an easy-to-use and service-oriented Cloud infrastructure. From a technical standpoint, Engineering leads research and technological development activities dedicated to Monitoring, Accounting and Billing while also contributing to activities related to Application Security. Engineering is also the lead partner to evaluate new business and sustainable models for scientific computing in close synergy with partners from enterprise as part of the activities pertaining to Communication and Sustainability;

[44] http://www.d4science.eu.

[45] http://www.venus-c.eu.

- **ERINA+:**[46] a project that is developing and applying techniques for measuring the socioeconomic impact of the project funded by the European Commission within unit F3 (Research Infrastructures) by enhancing and applying the socioeconomic methodology for the impact evaluation and assessment, already conceived and experimented during the ERINA study. Engineering is the coordinator of the project and is leader of the activities on the dissemination of project results;
- **ETICS 2:**[47] a project (the second phase) that developed an out-of-the-box software build and testing infrastructure, powered with a build and test product repository, and automatic collection of software quality metrics. Engineering was involved in tuning, improving and integrating the Grid Quality Certification Model (Meglio, Bégin, Couvares, Ronchieri and Takacs, 2008), with other established certification procedures and standards as well as developing and maintaining a web client to facilitate the interaction with the ETICS service.

Research data With respect to research data, the team mainly deals with software artefacts, project reports and technical documentation leading to websites, wiki pages and manuals. Unfortunately, although scientific paper production is encouraged, it is not frequent.

These research data are shared mainly among teammates by relying on tools that might depend on the activity the team is involved. Among these tools there is intranet, CSV and ETICS (for software artefacts) – which are exploited by all the teams – as well as tools like BSCW[48] and TRAC[49] – which are mainly used in the context of specific teams because are somehow a working practice imposed by the activity, e.g. they are imposed in a research project like D4Science-II.

The metadata collected depend on the tool/software they are conceived for, e.g. the metadata equipping software artefacts designed for ETICS are based on ETICS specifications. There is no metadata standard that the team is requested to use but those resulting from the tools they rely on to perform their activities.

Desiderata: the team is discussing the benefits and drawbacks in making the research data they produce publicly available, although they are regulated by policies. This holds mainly for software artefacts. On one hand, this practice is conceived to be a good practice leading to enhancement of organization visibility and business; on the other hand, it is conceived to be a

[46] http://www.venus-c.eu.

[47] http://etics.web.cern.ch/etics.

[48] http://public.bscw.de.

[49] http://trac.edgewall.org.

“dangerous” practice because of the risk of reducing the organization’s competitiveness. The desiderata are to have facilities for enhancing the visibility of the data that guarantee visibility of policies regulating data access and provenance.

Literature With respect to literature, the production of scientific papers is limited while the consumption is encouraged. Engineering team mainly relies on known publication sources, such as Google, Google Scholar and publisher websites (e.g. Elsevier, ACM, IEEE). As regards paper production, the team relies on “standard” editing tools (namely Microsoft Word) and file-sharing facilities, e.g. the intranet, email attachment, Dropbox.

Desiderata: because of the limited activity, there are no major desiderata but the overall team is interested in having a seamless access to all the literature. In particular, this seamless access should simplify the discovery of the so-far produced literature on a specific topic.

Combining literature and data With respect to linking data and literature, it is common to provide the paper with the URL(s) of the software artefacts the paper is documenting or is related to. In addition to that, it is quite common to have websites/web pages dedicated to document software artefacts.

Desiderata: the mechanisms for linking data and software artefacts should be strengthened. In addition to a simple link, a bunch of metadata should be either explicitly added or dynamically derived with the goal to enrich the paper with characteristics of the software artefact, such as the licences, technical requirements and software dependencies. These metadata should be machine oriented as to promote the implementation of tools benefiting from these data.

Open Access With respect to Open Access, there are no established policies within the group. Open Access strategies aiming at enhancing research and development results are encouraged. However, it should be possible to define fine-grained access policies.

3.6.3 Research workflows

The typical research production workflow of the team is pragmatic and quite standard since it is mainly oriented to produce new software artefacts. It includes the following phases (this is a simplistic view, the phases are organized in loops where decisions taken at certain points can be reconsidered thus leading to multiple iterations):

1. requirement analysis: producing requirement analysis documents by close interaction with the customers;
2. problem characterization and analysis;
3. survey: of existing tools (off-the-shelf solutions) and approaches that can be (re-)used in the context of the problem domain;
4. design: of a technical solution resolving the specific problem by promoting the (re-)use of existing technologies and standards;
5. implementing and testing: of the envisaged solution;
6. release: of the software artefact with the related documentation.

4 Current status

From the analysis of the interviews, it appears that researchers in the field of e-Infrastructure follow similar research workflow patterns, mostly in the direction of producing software data (to be used in the construction and maintenance of production infrastructure systems) and relative publications. The e-Infrastructure community, however, has not reached common agreements on policies, standards and best practices in the production of research data and literature. Depending on their focus (e.g. companies and research institutions), organizations and research groups tend to grow their own research infrastructures, based on proprietary best practices, policies, data formats, etc., in order to enable their researchers to collaboratively discover, produce, store, share and publish online both research data and literature. Typically, as illustrated in Figure D.1, such infrastructures are obtained as combination of:

- **local service and computing infrastructures:** examples are hardware (e.g. machine clusters), services such as SVN and TRAC for software data versioning and development and repository systems for literature storage and publishing;
- **web infrastructure:** as many other computer science research communities, the e-Infrastructure community makes heavy usage of the plethora of online tools for literature drafting (e.g. Google Docs, discovery, e.g. Google Scholar, BASE, OAIster, DRIVER) and sharing (e.g. SourceForge, Google projects, Apache projects, Dropbox). Among such online tools are included also local infrastructures which offer web access to their literature and data, e.g. institutional repositories, open source SVN systems.

Due this “local” approach, the e-Infrastructure research community has not established standards for data formats and classification or metadata for data resources, nor either policies and rules for interlinking data and literature.

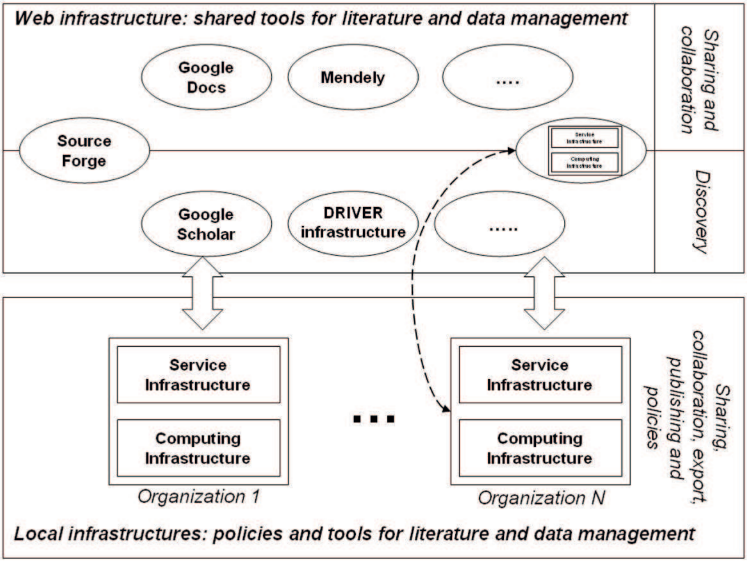

Figure D.1 Current status of research infrastructures for e-Infrastructure research

Overall, the research community has not grown a shared research infrastructure and, as a consequence, an e-Infrastructure, from both the organizational (i.e. policies, standards and best practices) and technical (i.e. services) point of views. Through such e-infrastructure, research data and literature in the field could be (possibly openly) collected, shared, exchanged and linked to each other, based on well-established participation and access policies and standard formats. Although researchers agree on the potential benefits that such infrastructure would bring, no plan in this direction is being undertaken. The reasons for this are many: for example the existence of practical and powerful online tools, reluctance to change methodologies, lack of funds and logistics and the youth of the discipline.

The unavailability of a common e-Infrastructure leads to two main drawbacks:

- **interoperability costs:** whenever organizations need to cooperate in the production of data and literature, for example within collaborative research projects, they have to bear a cost of interoperability of content (e.g. data and metadata exchange) and of learning new cooperation tools (e.g. file sharing, publication drafting, software versioning).
- **hardly reachable data and literature resources:** in order to discover and identify data and literature of interest to the field, researchers

need to access and search the plethora of web sources available for publication and data sharing (“aggregators”, e.g. Google Scholar, DRIVER, SourceForge), but also websites of organizations (e.g. to find software products and documentations), often reachable through generic searches on “The Web” (e.g. Google, Yahoo).

The following sections summarize the results gathered through the interviews, trying to cover all aspects of the typical e-Infrastructure research workflows and identifying the possible improvements that would derive by the establishment of a common research infrastructure.

4.1 Research data

4.1.1 Data types and metadata

Research data typologies are:

- **software:** intended as programming language code or the results of code compilation, such as installation packages;
- **software instances:** intended as software running on a machine (e.g. web services), often described by a so-called “profile” (Grid terminology) and therefore discoverable and reusable for interaction or “orchestration” by authorized applications;
- **benchmarks:** intended as collections of data available through any kind of storage support (e.g. file system, DBMS) and used for testing purposes. Typically their format, size and storage support vary depending on the application domain and can included videos, images, table data, database tables, files and folders;
- **logs:** intended as recorded histories of actions or events, typically used to evaluate and monitor the activity of a software instance. Their storage modes and formats vary, ranging from databases to text files;
- **statistics:** intended as qualitative or qualitative measures often derived from logs analysis to evaluate software instance activities (e.g. number of requests to a software instance in a given period).

Regarding metadata typologies, in general, data come with metadata information in order to make it available for discovery and re-use within and outside the local infrastructures.

Generally, structured metadata (produced in the form of records/profiles which are interpretable by a machine), can obey to proprietary or standard formats depending on the typology of data. In some cases, as for software and software instances data (e.g. D-Lib research group), proprietary metadata structures are introduced to be able to describe domain-specific properties of the data (e.g. dependencies of software packages). Metadata standards are also adopted (e.g. Dublin Core for technical reports), often imposed by the tools integrated in the local service and computing infrastructures (e.g. repository platforms, BSCW).

Researchers also heavily rely on unstructured metadata in such forms as roadmap specifications, functional and architectural specifications, policy specifications, guidelines, usage and installation manuals, software documentation and technical reports. Documentation is made available in various standard formats, such as file formats PDF, docx, Latex, DocBook or through web formats, such as Web 2.0 wikis and more "traditional" websites.

4.1.2 Data management aspects

Organizations provide a wide range of data storage and export solutions, whose adoption depends on the typology of data and on the Open Access policies adopted. When asked, interviewees confirmed that they do not implement literature or preservation policies and no desiderata have been suggested in this direction.

Storage Organizations' local infrastructures are equipped with version control management systems (e.g. SVN, Git) and issue trackers (e.g. TRAC, RedHat Issue Tracker), through which they manage software data. Similarly, technical reports and benchmarks are stored using standard document management tools, such as repository platforms (e.g. DSpace, ePrints, Fedora, PUMA) and sometimes version control systems. Typically, such tools are under the control of the organization and to authorized users and applications.

Production Some organizations have adopted a systematic approach and have grown a local service and computing infrastructures where data can be managed across several projects and research activities, under controlled access policies. In other cases, such tools are deployed as independent instances, dedicated to the research activities of the case. In some cases, Cloud technology is exploited, in order to outsource the cost of temporary or high peaks of storage and computing power demand. For example, this may be useful when testing highly distributed algorithms to be run on Grid-oriented (Foster and Kesselman, 1999) research infrastructures. Typically, Cloud CPU rental enables the arbitrary growth of CPU or storage demand (especially, peaks of demand) at a cost that is lower compared to the one of purchasing and maintaining the machines required to run the same tests.

Collaboration Researchers collaborate in the production of software, unstructured metadata and benchmarks by exploiting the functionality offered by the tools available to them through the local service and computing infrastructures and through online tools such as Google Docs for technical reports.

Export and policies Research data, both data and metadata, are shared and published by means of local services, such as SVN or organization/project websites. Research data are subject to confidentiality and protection policies that depend on the organization and, within the organization, on the typology of data and on the project or research undertaken. The trend is for companies in the field to be reluctant on openly sharing the data they produce for business (e.g. Engineering). Such data are generally accessible within the boundaries of the organization and sometimes not available outside to the owning research group. On the other hand, research institutions tend to publish and disseminate their results through all possible means, to promote and give visibility to the results of their activities. In general, when disclosed to the world, the usage of software and unstructured metadata may be restricted according to standard licensing schemes and non-disclosure agreements.

4.2 Literature

e-Infrastructure researchers follow a typical literature lifecycle, made of phases of: (i) survey and analysis of the literature and (ii) drafting and publishing of an article, of course prior to submission, reviewing and acceptance to a venue, such as a conference, or a journal. Both phases are largely affected by the interdisciplinary nature of e-Infrastructure research, which is placed somewhere in between service-oriented architectures/infrastructures, Grid infrastructures, digital libraries, multimedia storage, information retrieval, big-data (NOSQL solutions) and the specific functionalities of the research field for which e-Infrastructures are necessary.

Survey and analysis There is no dedicated online literature source for e-Infrastructure research. Researchers rely in general-purpose online aggregators, such as Google Scholar, Citeceer, the DRIVER infrastructure (see 3.1 D-Lib research group), BASE,[50] OCLC-OAIster,[51] Scopus[52], publishers websites, such as Springer, Elsevier and ACM or the Web, with Google, Yahoo and other search engines typically used by the majority of computer science researchers. Similarly, some of them also exploit online tools such as Mendeley and CiteUlike[53] to share their favourite reading lists.

[50] http://www.base-search.net.

[51] http://www.oclc.org/oaister.

[52] http://www.scopus.com.

[53] http://www.citeulike.org.

Drafting As many researchers in computer science, articles are written exploiting online free tools for collaborative editing and file sharing, such as emails, Google Docs, Dropbox and SVN servers (e.g. for Latex articles).

Publishing Due to the interdisciplinary nature of the research field, only a few venues specific to e-Infrastructures are available, e.g. some tracks on Theory and Practice for Digital Libraries conference (formerly ECDL) and IFLA (International Federation of Library Associations and Institutions) conference series. As a consequence, articles in the field end up being submitted in journals and conferences related with digital libraries, service-oriented architectures, Grid and discipline-specific venues, those for which e-Infrastructures are constructed (e.g. biology, cultural heritage, grey literature). Some organizations from academia, research and industry also support and fund PhD and MsC theses.

4.3 Linking literature and research data

Researchers do reference their software and unstructured metadata from their publications by means of URLs indicating project website or downloadable files and as bibliography references. Moreover, data such as table data and graphs are placed/embedded within the publications text or, when too large for the publication body, as an appendix. This attitude reveals the awareness of the benefits of pointing readers to actual evidence of the results, but also shows the necessity of a more structured approach. In this process of linking publications and data, both writers and readers follow their intuition and not agreed-on rules, e.g. how to point to data, how to describe data properties and provenance. A more structured approach would enable better evaluation of the quality of the publication, avoid falsified data and enable discovery and re-use of the data, for example in order to improve previous scientific results. Interviewed researchers generally agreed on the benefits of such a combined approach for publication and expressed the need for both policies and tools to support its diffusion.

4.4 Open Access

It appears that most organizations are aware of the existence of the Open Access initiatives and agree with their mission and goals. In fact, many of them also actively promote it among their own researchers and in other communities (e.g. D-Lib group, EKT, MADGIK group). This is typical for research and academic institutions, whose interests are the dissemination of their achievements through Open Access literature and open source software data and unstructured metadata (e.g. technical reports). On the other hand, many organizations have interests which conflict with the consequences of Open Access especially on the side of software data (in this case Open Access translates in open source). In this context, Open Access may have the undesirable side effect of disclosing technology to third-party organizations, thus potentially reducing the possibility to sell it to customers (e.g. Engineering).

4.4.1 Literature Open Access issues

In general, although many organizations in the field are supporting and promoting Open Access, it seems that none of them has imposed Open Access policies as obligatory to its researchers. This choice has mainly to do with the lack of Open Access publishers linked with relevant conferences or journals in the field, i.e. those giving more value and thus visibility to research results, and with the high costs of purchasing gold Open Access licences from them (e.g. "Open Choice" publishing model from Springer).

Organizations store their publications in local repository platforms or websites in order for third-party organizations and researchers to follow their activities and get hold of the actual documents (for Open Access material) or to reach the toll-gate sources from which these can be requested. Since such sources are reached by online aggregators such as Google Scholar, DRIVER, etc. e-Infrastructure literature can be considered today discoverable through accurate and selective search activities.

Overall, no e-Infrastructure-specific literature sources are available on the web and researchers are required to tentatively search for publications in the field across online collections pertaining to several research domains.

4.4.2 Data Open Access issues

Open Access for data depends on the typology of data and on the specific policies of the organization involved.

For software data, Open Access, namely open source, is always considered a possibility and generally ruled by means of specific software licences, from GPL, Apache and non-disclosure agreements. Organizations make software available through product websites, local software repositories and sometimes through shared open source software repositories, such as SourceForge.

For software instance data, Open Access translates in open interaction with the APIs of running software. However, this is rarely the case. API access policies are often controlled through authentication and authorization protocols or, more simply, through white lists and black lists of IP addresses.

For unstructured metadata (e.g. technical reports, specifications), Open Access is a common practice, although often decided on a case by case basis. Organizations make available their unstructured metadata through product websites and local repository platforms, which are often aggregated, i.e. web crawled or OAI-PMH harvested (Lagoze and de Sompel, 2001), by online search engines, such as Google Scholar.

For benchmarks and log kind of data, Open Access policies are not frequently applied for a number of reasons. In some cases, these are simply not perceived as resources possibly reusable by the community. In others, they are produced in proprietary formats and may therefore result not interesting or not be easily re-used by third-party consumers. Finally, as for web log files or benchmarks obtained by protected information, there may be privacy issues that prevent such data to be openly disseminated.

Overall, e-Infrastructure data are available from the individual organization stores, websites and repositories, given these are made accessible from the Web and not only within organization intranets. This well-established attitude makes research data in the field hard to expose and discover, hence to re-use or reference by researchers.

5 Desiderata and future directions

The interviewees also suggested a number of desiderata on which aspects of e-Infrastructures could/would improve the current research workflows. In the following, such ideas are collected and presented according to the structure of the questionnaire: research data, literature, linking data and literature and Open Access. Finally, these are combined to figure out how an e-Infrastructure for e-Infrastructure research that meets such desiderata may impact on and benefit the overall community.

5.1 Research data

Controlled data sharing In general, e-Infrastructure researchers are willing to share their data so that they can reach and consume data produced by others. Sharing policies may range from open source licences and toll-gated copyrights to non-disclosure agreements, but the (marketing) principle is that data resources should be reachable and potentially accessible by researchers interested in them. For example software, unstructured metadata, benchmarks and logs should be always discoverable and reachable through community-oriented web tools, together with a metadata description of their degree of Open Access.

Data unreachability on the web In many cases, researchers find it hard to reach data they might need outside the boundaries of their organizations. For example, this is the case for software and unstructured metadata when these are not published on shared repositories such as SourceForge or exposed through repository platforms and product websites to be then crawled by web search engines. Researchers need to agree on best practices and policies for data publication and require community-specific tools for leveraging discovery of their data according to such policies.

Lack of metadata description standards In those cases where research data are available through web tools (e.g. software through Apache projects), the relative metadata properties are not peculiar to e-Infrastructure resources. This makes it hard for researchers to distinguish and identify the resources they require. Researchers need standards for data descriptive metadata and for data unique identifiers (e.g. DOIs, web handles).

Service and computing infrastructure sharing Typically software is developed, tested and integrated on local service and computing infrastructures featuring adequate CPU and storage quotas. Maintenance of services and hardware leads to high sustainability costs, hardly affordable by many communities. These costs could be reduced by adopting e-Infrastructures for sharing computational resources across multiple organizations according to a combination of service Cloud (Dikaiakos et al., 2009) and Grid resource sharing (Berman, Fox and Hey, 2004). This economy of scale approach would maximize the usage of resources and therefore minimize the overall cost of maintaining very large infrastructures and realizing complex e-Infrastructure software.

5.2 Literature

Lack of common classification schemes for literature The community calls for a clean classification scheme of the research field, in order to organize its scientific production and facilitate its discovery.

Lack of services for sharing literature e-Infrastructure literature is not easily discoverable through well-known web publications sources, mainly due to its interdisciplinary nature. The community calls for common services enabling the collection and discovery of publications in the field.

5.3 Linking literature and research data

Researchers realize the advantages of interlinking publications with research data in a meaningful way, from reusability of data to more effective validation of the results. To this aim, they need to agree on common policies for specifying references to data from within a publication text or from within the publication metadata. This work should be realized in conjunction with the definition of standards for metadata and unique identifiers for data resources.

5.4 Open Access

As in other research fields, e-Infrastructure researchers realize the importance of Open Access for both data and literature. On the other hand they are also aware of (i) the "certification of excellence" implied by peer review mechanisms, which often lead to retention of copyrights, and (ii) the return-of-investment principles behind the production of data for business. Hence, as for other research fields, to enforce Open Access, researchers need innovative business models.

5.5 A research infrastructure for e-Infrastructure researchers

The researchers' desiderata presented in the previous section seem to converge to the realization of an e-Infrastructure providing policies and services for sharing and collaboratively constructing research data and literature resources in the field of e-Infrastructures. As illustrated in Figure D.2, such an e-infrastructure would be complementary to the current local infrastructures. The combination of the two layers would give life to an effective research infrastructure for e-Infrastructure researchers. This would be spontaneously maintained by organizations willing to benefit from its services, based on well-known economy-of-scale principles. Its benefits would derive from a combination of organizational and technological efforts:

– **Organizational**
  - promote standards and policies for data and literature exchange (formats) and description (metadata);
  - promote standards and policies for interlinking research data and literature;
  - investigate on new business models capable of reaching the right compromise between publishers business and open access policies, without compromising the evaluation and publication process of research results.

– **Technological**
  - services for safely sharing and curating research data and literature in the field;
  - services for discovering and interlinking research data and literature in the field;

- services for collaboratively constructing research data and literature by reusing existing resources.

Investigations and studies on how communities could gradually move towards the realization of these objectives in a collaborative and synergic fashion are being undertaken in the EC project OpenAIRE. Experimental solutions in interlinking of research data and research literature have been realized in the EC project DRIVER-II, e.g. enhanced publications (Woutersen-Windhouwer, Brandsma and Hogenaar, 2009) and will be implemented in the EC project OpenAIREplus (to be started in December 2011).

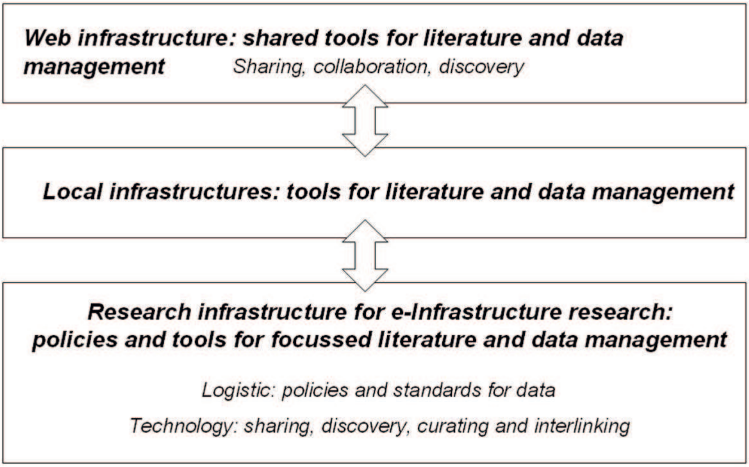
Web infrastructure: shared tools for literature and data management
Sharing, collaboration, discovery
Local infrastructures: tools for literature and data management
Research infrastructure for e-Infrastructure research: policies and tools for focussed literature and data management
Logistic: policies and standards for data
Technology: sharing, discovery, curating and interlinking

Figure D.2 Challenges: future research infrastructure for e-Infrastructure researchers

It is hard to envisage or quantify the cost for organizations willing to work in synergy to realize and maintain such infrastructure, as well as the cost of those organizations willing to join in a second stage, in order to benefit of its services. Certainly, as it happened in the past with other research infrastructures, the initial spark should come for a strongly motivated community, whose history and vision justifies common objectives, goals and risks. Although the e-Infrastructure community is probably the one which can at best realize this goal, its history is still in an early stage and such motivation is likely largely missing today.

6 List of figures

7 List of tables

8 Bibliography

Atkins, DE, Droegemeier, KK, Feldman, SI, Garcia-Molina, H, Klein, ML, Messerschmitt, DG, Messina, P, Ostriker, JP, & Wright, MH. *Revolutionizing Science and Engineering Through Cyberinfrastructure*. 2003.

Berman, F, Fox, G, & Hey, A. *Grid Computing: Making the Global Infrastructure a Reality*. John Wiley & Sons, 2003.

Candela, L, Castelli, D, Ferro, N, Ioannidis, Y, Koutrika, G, Meghini, Pagano, CP, Ross, S, Soergel, D, Agosti, M, Dobreva, M, Katifori, V, & Schuldt, H. The DELOS Digital Library Reference Model - Foundations for Digital Libraries. Version 0.98. February 2008.

Castelli, D, Candela, L, Pagano, P, & Simi, M. *DILIGENT: A DL Infrastructure for Supporting Joint Research*. 2nd IEEE-CS International Symposium Global Data Interoperability – Challenges and Technologies, 2005, 56-69, Society, I. C. (Ed.)

Dikaiakos, MD, Katsaros, D, Mehra, P, Pallis, G, & Vakali, A. Cloud computing: distributed internet computing for IT and scientific research. *Internet Computing, IEEE* 2009, 13, 10–13.

Foster, I & Kesselman, C. *The Grid: Blueprint for a New Computing Infrastructure.* Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1999.

Ioannidis, Y, Maier, D, Abiteboul, S, Buneman, P, Davidson, S, Fox, E, Halevy, A, Knoblock, C, Rabitti, F, Schek, H, & Weikum, G. Digital library information – technology infrastructures. *International Journal on Digital Libraries*, 2005, 5, 266–274.

Lagoze, C & de Sompel, HV. *The Open Archives Initiative: Building a Low-barrier Interoperability Framework.* Proceedings of the first ACM/IEEE-CS Joint Conference on Digital Libraries. ACM Press, 2001, 54–62.

Lomow, G & Newcomer, E. *Understanding SOA with Web Services*
Addison Wesley Professional, 2005.

Meglio, AD, Bégin, ME, Couvares, P, Ronchieri, E, & Takacs, E. ETICS: the International Software Engineering Service for the Grid. *Journal of Physics: Conference Series* 2008, 119, 042010.

Woutersen-Windhouwer, S, Brandsma, R, & Hogenaar, A. *Enhanced Publications: Linking Publications and Research Data in Digital Repositories.* Amsterdam University Press, 2009, 212.