Noname manuscript No.
(will be inserted by the editor)

A Vision Towards Scientific Communication Infrastructures

On Bridging the Realms of Research Digital Libraries and Scientific Data Centers

Donatella Castelli · Paolo Manghi · Costantino Thanos

the date of receipt and acceptance should be inserted later

Abstract The two pillars of the modern scientific communication are Data Centers and Research Digital Libraries, whose technologies and admin staff support researchers at storing, curating, sharing, and discovering the data and the publications they produce. Being realized to maintain and give access to the results of complementary phases of the scientific research process, such systems are poorly integrated with one another and generally do not rely on the strengths of the other. Today, such a gap hampers achieving the objectives of the modern scientific communication, that is, publishing, interlinking, and discovery of all outcomes of the research process, from the experimental and observational datasets to the final paper. In this work, we envision that instrumental to bridge the gap is the construction of "Scientific Communication Infrastructures". The main goal of these infrastructures is to facilitate interoperability between Data Centers and Research Digital Libraries and to provide services that simplify the implementation of the large variety of modern scientific communication patterns.

Keywords Scientific Communication Systems · Data Infrastructures · Research Digital Libraries · Data Centers

1 Introduction

New high-throughput scientific instruments, telescopes, satellites, accelerators, supercomputers, sensor networks and running simulations are generating massive amounts of data. The availability of huge volumes of data is a big opportunity for scientists as it can revolutionize the way research is carried out and lead to a new data-centric way of thinking, organizing and carrying out research activities (Jim Gray's vision [1]). Such data-dominated e-Science has started to impact also on the scientific communication process (Towards 2020 Science report [2]). Research data are starting not to be exclusively understood as necessary sub-product of a scientific publication, but are increasingly regarded as first class citizens of the scientific communication, with their own identity and metadata, which can be discovered, accessed, validated, and possibly reused. In the modern scientific communication paradigm researchers should be able to publish intermediate and relevant products of the research process, i.e. raw data, secondary data, and publications, in a way that they are discoverable, meaningfully interlinked, and re-usable by others [3]. Researchers, funding agencies, and organizations require *modern scientific communication systems*, supporting all functionalities required to facilitate modern publishing practices in order to improve the quality and speed-up sharing and re-use of research outcomes.

The pressing community requirements gave life to several initiatives aiming at publishing data and/or interlinking them with other research outcome. The most prominent ones have to do with data citation practices, i.e. standards for metadata about data and persistent identifiers, and recognize the role of data as a primary research output; e.g. DataCite [4] and Dataverse [5]. Such initiatives leverage data publishing, discovery and reuse, and permit to reward researchers producing and sharing data. Although fundamental, these are not sufficient, as several cultural and technological barriers are still hindering the realization of modern scientific communication systems. On the one hand, data citation

D. Castelli, P. Manghi, C. Thanos
ISTI - Consiglio Nazionale delle Ricerche
Via Moruzzi 1, 56124 Pisa, Italy
E-mail: name.surname@isti.cnr.it

is still not a common best practice in many disciplines, which instead focus on metadata descriptions for re-use of datasets within the community. On the other hand, the technologies and the professionals traditionally involved in publication and data management find themselves far apart. Traditionally, scientific communication relies on publishers (i.e. journals), academic institutions, and research centers to support research communities with what we shall refer to as *Research Digital Libraries* (RDLs). Such systems provide the combination of technology (e.g. repository functionality, from search to peer-review systems) and organization (e.g. librarians, reviewers) required to assist the literature life-cycle, from drafting to publishing and dissemination. To cope with new requirements of data publishing and interlinking, RDLs should today integrate features which are typical of *Data Centers* (DCs), which are the organizational units providing the technology (e.g. data repositories, computing infrastructures) and organization (e.g. data managers, data curators) required by researchers to efficiently manage their data. Unfortunately, RDLs and DCs were devised to target complementary phases of the data research and publication process and their supporting systems, policies, and best practices are not conceived to facilitate their interoperability.

As a consequence, the realization of modern scientific communication systems must bear the cost of upgrading existing RDLs and/or DCs technologies to establish interoperability and deliver the expected functionalities. For example, some scientific journals made dedicated agreements with DCs or established dedicated data repositories in order to ensure their authors deposit peer-reviewed publications in the journal repository and the data they used or produced in the same experiment in a data repository; e.g. DRYAD repository [6] and its Joint Data Archiving Policy. In such cases, very often both RDLs and DCs are upgraded to keep references from publication to data and vice versa, exploiting known publication and data citation standards. In other cases, the integration might involve services of the *Research Infrastructures* (RI) [7] that generated the data. For example, a scientific communication system may provide data peer-review facilities, necessary to ensure quality of published data. Data analysis and validation may require exceptional computational power or highly specialized algorithms and workflows (e.g. PRIDE database [8]), which are out of the scope of traditional RDLs and typically offered by Research Infrastructures.

Software solutions can always be found. However, the resulting scientific communication systems tend not to be cross-discipline and cross-technology and in general may suffer from high costs of realization, maintenance, and extension to other functionalities. The purpose of this paper is to advocate the need for bridging RDL and DC realms by means of so-called *Scientific Communication Infrastructures* (SCIs). Such infrastructures should provide the services and tools necessary to integrate content and functionality from arbitrary RDLs, DCs and RIs in order to: (*i*) minimize the upgrade effort required by RDL and DC organizations to interoperate with the infrastructures, and (*ii*) minimize the effort for implementing advanced scientific communication applications by re-using RDLs, DCs and RIs functionalities. The enabling software of SCIs should be designed to be extendible, general-purpose, and component-oriented so as to facilitate its customization to different scenarios and support the evolution of such scenarios over time.

Outline The paper is organized as follows. Section 2 motivates and describes the effects of e-Science on scientific communication. Section 3 describes the current approaches to the construction of modern scientific communication systems. Section 4 reports on the cultural and technological issues arising in the realization of such systems. Finally, Section 5 presents our vision of future Scientific Communication Infrastructures as the organizational and technological means through which scientific communities will overcome such issues and fully address modern scholarly communication requirements.

2 Modern Scientific Communication

The research and publishing process is composed of the following phases: (*i*) a scientist produces, through research activity, primary, raw data; (*ii*) this data is analyzed to create secondary data; (*iii*) this is then evaluated, refined to be reported as tertiary information for publication; (*iv*) this then goes into the traditional publishing process and feeds publication repositories contained in RDLs, while primary data are archived into discipline-specific DCs. Top of Figure 1 illustrates the traditional scientific communication process and the different involvements of DCs and RDLs. DCs are designed to serve the needs of a community of scientists whose experiments and/or results are based on data acquisition and processing. They deal with aspects such as raw data acquisition and processing, production of secondary data, analysis and curation of data, data storage and preservation onto data repositories, data disposition etc. [1][7]. Once the results are finalized, researchers rely on RDLs to produce and publish literature and related data, i.e. technical reports, pre-prints, articles, PhD theses, hence effectively implementing the

scientific communication process. Literature, which may or may not be certified by a peer-review process, represents the only well-established means of research dissemination and only includes data as embedded information or as separate files of secondary data, uploaded in the same publication repository [9]:

- *Literature embeds secondary data.* The data are contained within (peer-reviewed) publications in RDLs, e.g. a table in a paper. This is the traditional publishing model where the publisher takes full responsibility for the publication of the article as well as for the aggregated data embedded in it and the way it is presented. The tight embedding of the data into the publication makes the data citable and retrievable only together with the publication. Besides, the re-usability of the data is limited. This model is not appropriate when large data sets are involved, as they do not fit the traditional publication format.
- *Literature comes with separate secondary data files.* The data resides in supplementary files added to the journal article, thanks to more advanced RDLs. The journal offers authors the service to add in supplementary files to their article any relevant material that is too big or that will not fit the traditional article format or its narrative, such as datasets, multimedia files, large tables, animations, etc.; e.g. Elsevier[1], SAGE[2]. This publishing model serves well the consumer of an article, which can possibly visualize supplementary material independently of the article itself, but carries issues such as the curation and preservation of such files as well as the ability to find and link them independently of the main publication. In addition, supplementary files are often constrained to given size thresholds and therefore confine the possibilities of data publishing to secondary data.

Today, the advent of data-driven science is forcing this scenario to change. All stakeholders in the research life-cycle, from funding agencies to scientists and hosting organizations, require that data must be validated, stored and preserved in the long-term, to be published and accurately described in order to enable discovery and re-use by other scientists [10]. Funding agencies aim at Return Of Investment (ROI) measurement,[3] and organizations, as well as researchers, at gaining credit [11][12]. Most importantly, scientists, who today can collaborate through e-Science (research) infrastructures via e-Research tools such as those offered by Virtual Research Environments [13], urge to include data in the scientific communication chain in order to improve its discoverability, interpretability, and re-usability. Such requirements are similar, parallel and interwoven with the one of publishing literature, which still represents the conclusive step of the research chain. To support modern data-driven science, raw data acquisition, secondary data production, drafting and publishing literature must all be different phases of an integrated scientific communication process. More specifically, researchers should be able to collaboratively produce and publish intermediate and relevant products of this process, i.e. raw data, secondary data, and literature, in a way that these are discoverable, possibly meaningfully (web) interlinked, and re-usable by others [14].

The bottom of Figure 1 shows how modern scientific communication involves DCs as well as RDLs. It requires their interaction for establishing bi-directional links between data and literature as well as between data and data. Such process enables stakeholders to review the method of conducting the science as well as its final conclusions. It enables greater sharing, re-use and comparison of scientific results, reduces duplication of efforts, and insures against data loss because the additional, contextual and provenance information improves the repeatability and verifiability of the results. For example, data journals offer today manual peer-review of datasets, which entails lack of data certification quality [15]. Modern scientific communication should support systems providing workflows for automated data submission and analysis by interoperating with Research Infrastructure services capable of performing such validation. In addition, the integration of data and publications can produce significant benefits [1], since publications help the data to be better discoverable and interpretable, and provide the author better credits for the data; and reversely: the data add depth to the article and facilitate better understanding. Overall, such systems also impact on reading practices as they allow scientists to move beyond the paper to engage the underlying science and data much more effectively and to move from paper to paper, or between paper and reference data collection, with great ease, precision and flexibility [16].

2.1 Data citation standards and practices

In an attempt to deliver modern scientific communication systems, research in the area has already pro-

[1] Elsevier Supplementary Data, `http://www.elsevier.com/journals/vaccine/0264-410X/guide-for-authors#87000`

[2] SAGE Journals, Author Guide to Supplementary Files, `http://www.uk.sagepub.com/repository/binaries/doc/Supplemental_data_on_sjo_guidelines_for_authors.doc`

[3] For example JISC's "what we do": `http://www.jisc.ac.uk/whatwedo/programmes/-di_researchmanagement/managingresearchdata/research-data-publication.aspx`

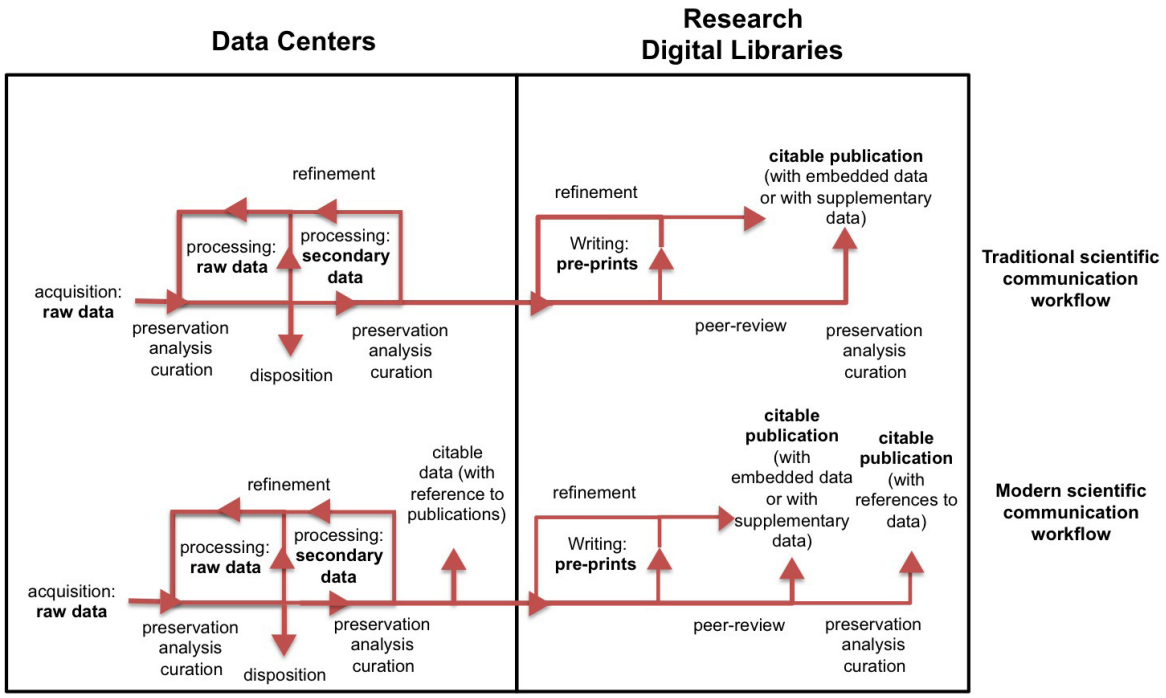

Fig. 1 Traditional vs Modern Scientific Communication

vided solutions to data publishing, discovery and re-use and interlinking with literature. Such solutions are more "infrastructural" and include metadata best practices for citing and reusing data from publications and vice versa. The main mechanism enabling the alignment and integration between data and publications in the scientific communication process is *data citation*. Data citation is the practice of providing a reference to data (or a dataset) intended as a description of data properties that enable discover, interlinking, and access to the data. As such, proper citation mechanisms rely on the assignment of persistent identifiers to data (hence on some entity guaranteeing the identifier and the data themselves will persist in the long-term), together with a description (metadata) of the data, which allows for discovery and, to some extent, re-use of the data. Several standards exist for citing data and practices vary across different disciplines and data repositories, supported by initiatives in various fields of applications. Their common objectives are to align data citation with that of publications, in order to support easier access to scientific research data on the Internet, increase acceptance of research data as legitimate, citable contributions to the scientific record, support data archiving that will permit results to be verified and re-purposed for future study, and give credit to the author and publisher of the data.

The *Dataverse Network* [12] is an initiative maintaining open source software for the installation and maintenance of a network of federated data repositories originally devised in the field of social sciences (other sciences have been targeted and others have on going requirement analysis). The software offers out-of-the-box facilities for long-term preservation, citation, and reuse of data according to standard practices and over data of several formats in a given domain. In particular, a running network, for each deposited dataset, requires a metadata description to be provided as means for data citation, hence discovery and re-use in the network. The metadata is "flat" and mandatorily includes title, authors, publishing year, distributor, a persistent identifier, and a Universal Numeric Fingerprint (UNF), i.e. a short, fixed-length string of numbers and characters that summarizes all the content in the data set, such that a change in any part of the data would produce a completely different UNF.

The *DataCite* initiative [4] forms an international consortium addressing the challenges of making data citable in a harmonized, interoperable and persistent way. In particular DataCite supports data centers by providing persistent identifiers for datasets, workflows and standards for data publication and journal publishers by enabling research articles to be linked to the underlying data. As such, unlike Dataverse, DataCite targets a wider audience and focuses on the minimal infrastructural aspects to enable cross-discipline best prac-

[4] Data Cite, `http://www.datacite.org`

tices for data citation. DataCite members must assign Digital Object Identifiers [5] (DOIs) [17] to their data sets and provide metadata descriptions responding to the DataCite metadata format specification [18]. DataCite mandatory metadata is a subset of the Dataverse mandatory fields (no property UNF) but it is "hierarchical" (e.g. creators can be more than one, have separate name separate from surname property, and may have a unique persistent identifier). On the other hand, the whole set of fields, including optional ones, is richer. For example, it includes properties to classify the data based on subject, format, typology, its access rights, language, and how it is interlinked with other datasets and publications. Many Data Centers (or simply data repositories) is today part of DataCite and follows its directives. For example, PANGAEA [6] is a system acting as an Open Access library whose goal is to archive and publish geo-referenced data from earth system research. The system guarantees long-term availability of its content through a commitment of the operating institutions in the domain. Data published in PANGAEA is described by DataCite mandatory fields and assigned a DOI by the infrastructure, but can include references to publications in the case data is kept as supplementary to such publications.

The *Organization for Economic Co-operation and Development (OECD)* constantly produces results of data processing that are widely cited and referred to from media and research journal papers. In order to provide the reader with in-depth reference to such resources, OECD provided a specification on how to formally cite their secondary data to facilitate their discovery and re-use [19]. The mandatory metadata fields proposed by the initiative are a superset of Dataverse's, completed with properties such as the abstract, periodicity, links to digital representations of the data (e.g. PDF, Excel), and copyright. As DataCite, no UNF property is considered, and other optional fields are available, including links to other dataset and country covered by the data.

3 Current trends in developing Scientific Communication Systems

Today's scientific communication is mainly driven by Research Digital Libraries (RDLs) whose technology (e.g. DSpace [20], Fedora [21], Greenstone [22]) supports the activities of research institutions and scientific journals. The objective of RDLs was traditionally that of supporting the processes of acquisition, organization, peer-review, preservation, and access to electronic scientific publications by implementing indexing, storing, searching, and retrieving techniques. In the last decade, as mentioned in the previous section, RDL technologies evolved into an attempt to cope with data publishing requirements, beyond the initial solutions of embedding data into publications and attaching supplementary files to publications. New Scientific Communication systems and tools have been realized, capable of indexing, storing, searching, retrieving and interlinking publications with datasets from DCs. Typically, organizations or research communities ended-up sustaining the cost of constructing such systems, investing in the development and maintenance of the relative software. These can be categorized in four broad categories:

- Journal publishers which support an RDL and invest in a "local" DC, typically consisting of one data repository, to support data publishing as mandatory to literature publishing;
- Research communities sustaining a shared DC (typically a data repository) and investing in RDL technologies in order to publish their data as it is traditionally done with literature.
- Research communities implementing data and literature publishing practices independently (hence operating RDLs and DCs) investing in the realization of technologies for the integration of their two worlds. The resulting systems may allow the author of publications and/or data to deliver the respective object to the proper technological support (respectively RDLs and DCs), or to create links between publication and data in order to enable better discovery practices.
- Research communities that, assuming data publishing practices are well-established, focus on "modern" RDL document models, where publications are intended as "information packages" somehow unifying data and publications into one navigable and/or machine re-usable object.

3.1 RDL Organizations supporting typical DC services: Making Related Data Available

Many scientific journals have started to require data valuable for the evaluation of an article to be deposited prior submission into a data archive or Data Center. Such journals generally rely on external data repositories (or Data Centers) which offer the storage and preservation capacity necessary to cope with size and long-term sustainability of deposited data [23]. The Joint

[5] Digital Object Identifier System, `http://www.doi.org`
[6] PANGAEA, `http://www.pangaea.de`

Data Archiving Policy[7] (JDAP) proposed by the DRYAD initiative[8] describes the requirement that data supporting publications must be publicly available (license CC0): "This policy was adopted in a joint and coordinated fashion by many leading journals in the field of evolution in 2011, and JDAP has since been adopted by other journals across various disciplines". In this case, journals subscribing to this policy rely on the DRYAD data repository [6], which was specifically devised and supported by the committed consortium of journals for this purpose. In their policy, DRYAD also adopts the DataCite approach and generates a proper DOI and metadata for all deposited material, making it discoverable and re-usable independently of the original publication. A similar service is offered by the data repository PANGEA introduced above, which offers storage for supplementary data for Elsevier articles at ScienceDirect.

3.2 DC Organizations supporting typical RDL services: Publishing Data

A recent new trend is that of data journals whose mission is to disseminate data by leveraging analytic precision and transparency, minimize replication of work, and disclose new research avenues. Researchers can submit to a journal their valuable qualitative dataset together with a description, i.e. a short publication. An example is the GigaScience journal[9] (supported by BGI Shenzhen and BioMed Central), which accepts "data notes" submissions relative to relevant datasets (license CC0) in the ambit of biological and biomedical research. Another interesting notion is the one of *data papers* [24], whose motivations are three-fold: (*i*) providing a citable publication to bring scholarly credit to the creators of the data, (*ii*) describing data in a human readable form to incentivize re-use, and (*iii*) enabling discovery of data by the research community; e.g. the Journal of Open Archeology Data[10], Global Biodiversity Information Facility (Pensoft)[11]. The journal organizes the logistic of the peer review of the data by selecting capable reviewers in the field. As in the case above, in the case of acceptance, the journal must ensure the long-term availability and preservation of the data and to this aim relies on external support. The data repository PANGAEA introduced above supports the Earth System Science Data (ESSD) journal[12], dedicated to publishing original research data in the field. The interesting novelty introduced by data journals is that of proposing a publishing process for data that resembles the one of publications. Data are not a supplement to a publication, but vice versa. Peer-review, aiming at measuring originality and quality of data, is applied to the data rather than to the publication, and its "blessing" is mandatory for the data to be published.

3.3 Community organizations integrating their DCs and RDLs

A further approach is that of integrating existing and autonomous RDLs and DCs by means of "gluing" or "embedded" technologies. The idea is to deploy and manage RDLs and DCs for their regular missions, but apply the necessary changes to make them interoperate and offer functionalities typical of modern scientific communication systems. A real example is that of the European Bioinformatics Institute (EBI), a non-profit academic organization that forms part of the European Molecular Biology Laboratory (EMBL). EBI supports a DC for research and services in bioinformatics, including databases of biological data such as nucleic acids, protein sequences and macromolecular structures. Another unit of EBI provides an RDL publication repository called UK PubMedCentral[13] (today changing to Europe PubMedCentral), which offers advanced functionality for linking biomedical literature to scientific data at EBI. To this aim, EBI extended the publication repository to include references to data stored at EBI Data Center and then realized services capable of: (*i*) interacting with the repository to mine biomedical literature (PDF files) and identify possible links to datasets [14] (e.g. proteins) and (*ii*) semi-automatically (prior data curator validation) materializing such links from literature to data and vice versa.

3.4 Research communities developing tools for "modern publications"

Scientific publications in both digital and physical form will likely never lose their role of communication means. However, literature publishing will inevitably change to

[7] Joint Data Archiving Policy (JDAP), `http://www.dryad.org/jdap`

[8] DRYAD Repository, `http://datadryad.org/`

[9] GigaScience journal, `http://www.gicasciencejournal.com`

[10] JOAD, `http://openarchaeologydata.metajnl.com`

[11] Global Biodiversity Information Facility, `http://www.gbif.org`

[12] Earth System Science Data Journal, `http://www.earth-system-science-data.net`

[13] UK PubMedCentral, http://ukpmc.ac.uk/

[14] What's it!, `http://www.ebi.ac.uk/webservices/whatizit/info.jsf`

address the evolving requirements of data-driven science and its supporting technologies [25]. Such a process is already ongoing and RDL technologies started supporting new conceptions of scientific publication, not merely with different business models, but also with different editorial and technical approaches [26]. These are typically based on "document models" where a publication is intended as a set of "information units", including text and datasets, images, videos, sound recordings, mathematical models, workflows, presentational material, and software packages meaningfully connected by relationships. Their principle is that of exploiting data identification, citation and linking technologies (see Section 2.1) together with metadata descriptions enabling different degrees of human and machine interpretation. In the literature, two major classes of publication models seem to emerge: structured publications and experiment-oriented publications. In the following we shall present them together with real-case instantiations.

Structured publications "Fine-grained" structured publications are intended as one textual information object structured in well-defined subparts, which may include sections, paragraphs, figures, tables, as well as images or web references to external sources and interactive applications. Their structure is designed to enable smart visualization of the publication through Web applications, i.e. navigation through its subparts, and browsing of links to external Web resources, such as remote data available through HTTP. Investigations on such kinds of publication models started a decade ago, e.g. OpenDLib data model [27], but were recently re-proposed as underlying models for Web 2.0 publications, such as the Article of the Future of Elsevier [28]. Other examples, are Utopia Documents [29] and SOLE documents [30]: Utopia Document is a novel PDF reader that semantically integrates visualization and data analysis tools with published research articles, via links to external objects (e.g. biochemical datasets[15]); similarly, SOLE is a tool for linking research papers with associated *science objects*, such as source codes, datasets, annotations, workflows, packages, and virtual machine images. Authors of SOLE are investigating the possibility of enabling re-use of datasets linked by a SOLE document via given services; in this case these documents would fall in the category of "experiment-oriented publications" explained below. Finally, *live publications* have recently emerged in the context of e-Science infrastructures and consist in textual publications (typically research reports) which embed data descriptions, tables, histograms, summaries, and statistics based on "live data", generated at access time and updated in the publication by the underlying infrastructure. A publication can therefore be "instantiated" in a given moment in time to describe current status/results for a given scenario. Examples of such publications can be found in the D4Science and iMarine infrastructures, serving respectively the communities of European Space Agency and FAO [33].

"Coarse-grained" structured publications are intended as "compound objects", i.e. sets of existing objects meaningfully interlinked and packaged to form one new digital object. Examples are *enhanced publications* [31] and *modular articles* [32]. An enhanced publication consists of an existing publication, e.g. a peer-reviewed textual article, enhanced with relationships to a number of existing objects, such as further publications (cited, similar, etc.) or datasets (used in experiments, resulting from experiments, etc.). Examples are research data that provides evidence of the research, its associated contextual and provenance metadata and the derived information, extra materials useful for clarification purposes, post publication data that could provide commentaries, and web resources. An enhanced publication encodes the structure of a graph rooted in an existing publication and connecting objects which can be distributed over several locations (typically identified by a persistent identifier, e.g. DOI). Similarly, a modular article mirrors the vision of Kircz, according to whom data sets, images, sounds, simulations and videos are part (i.e. modules) of the publishing environment, next to text. A module is defined as a uniquely characterized, self-contained representation of a conceptual information unit, aimed at communicating that information. Each type of information unit should be well defined and therefore be endowed with different sets of metadata, each set describing a different aspect of the information entity. A modular article consists of modules and Internet links between them into a coherent unit for the purpose of communication, but none of them is privileged like in the case of enhanced publications.

Experiment-oriented publications Such publications are inspired by structured publications, but generally contain, beyond digital objects, also information units whose purpose is enabling automatic reuse of their content [34]. Examples of such publications are Scientific Publication Packages, Research Objects, and executable papers. A *Scientific Publication Package* (SPP) [35] is a new information format that encapsulates raw data, derived products, algorithms, software, textual publications and associated contextual and provenance metadata. This new information format is fundamentally different from the traditional file-based formats.

[15] Pilot with Biochemical Journal, `http://www.biochemj.org/bj/424/3/`

The different information units must be specified and can either be included as references to a unique identifier or actual bit streams incorporated within the package. Tools are provided to the scientists that allow him or her to specify the precise components, including: data, mathematical functions, software specifications, and textual documents. The Scientific Publication Package, i.e. a compound digital object is represented as a PDF package. A *Research Object* [36] (MyExperiments.com) is a compound object obeying to some extent to the following properties (the "six R's"): re-playable, repeatable, reproducible, reusable, re-purposeful, and reliable. The vision behind such model is to replace traditional models of publications with others capable of "providing sharable, reusable digital objects that enable research to be recorded and reused" - which, fundamentally, is what Science and e-Research involves. Other approaches like Paper Machè [37] or SHARE [38] make use of virtual machines that provide an environment for publishing "executable papers". Such a virtual machine would include all required tools and the complete software setup, which is needed to reproduce and verify an experiment described in such papers. The virtual machine may also contain data, the required scripts and embedded code snippets to generate updated revisions of a paper and allow reviewers to trace back the steps and verify results of the authors.

4 Issues in Realizing Scientific Communication Systems

The solutions presented in Section 3 suffer from two main inter-dependent weaknesses that make them fail at satisfying the requirements of modern scientific communication processes. On the one hand the lack of data publishing best practices for DCs and the relative communities. On the other hand, the sustainability costs which organizations willing to realize scientific communication systems have to bear.

4.1 Barriers for Data Centers

Scientific communication is still framed too narrowly, typically focusing on the final result of the research and publication process, that is, the scientific article in RDLs. Indeed, Data Centers (DCs) mainly function as central services where researchers can both deposit data they have created and also find data they can reuse within their own work. In addition, they support researchers in preparing their data for wider presentation and reuse in particular, in the creation of appropriate metadata and bear the responsibility for the curation and long-term preservation of the data. Although new trends are emerging, DCs typically do not target publishing aspects of the data and suffer from a major lack of best practices and technologies in order to support a rigorous scientific communication process. This is not surprising, as data citation is far more complicated than citation of scientific publication. For example, data sets generally are not locatable and attributable in the same way as scientific publications, they are often versioned, and they are mostly not peer-reviewed, hence in the need of quality control [39]. More generally, most of the data is still "hidden" into data repositories at Data Centers (when not open to the Internet) or in scientists' hard disks.

Culture of sharing Despite the urging requirements of data-driven science, data citation is still not widely adopted in many areas due to cultural barriers. This trend, not only deprives scientific communication of relevant research outputs, but also hinders the adoption and uptake of new publications models, thereby hampering the effective implementation of modern science. A recent study, carried out in [4], has summarized the current status of data citation standards, instruction, and practices among the "breadth of academic research, through a content analysis of journal articles, style manuals, and journal guidelines". Interestingly, such aspects are benchmarked against a Data Citation Adequacy Index, which takes into account the usage of various data citation standards, in order to measure the efficacy of current practices. The results are not surprising and confirm scientists are not yet well acquainted with data citation practices; for example, the majority of citations make use of in-text data titles and authors and publishers of the dataset are often missing. The problem is mainly cultural, since shifting behavioral norms is a slow process, and requires all stakeholders, from librarians and repository managers to data managers, to understand and disseminate the benefits of data citation for researchers; especially on aspects such as data discovery and re-use and credits for authors publishing quality data.

Metadata structure and semantics When cultural barriers are not an issue, Data Centers often encounter another difficulty: data citation not only as a mean to discover the data, but also as a mean to re-use the data by a human or a machine. Metadata structure and semantics may not be limited to the high-level bibliographic-like description of data, but also include specific properties enabling discipline-specific (e.g. device-specific) re-use of the cited data. In this direction, several proposals have appeared in the literature.

We have seen how different initiatives tend to propose metadata descriptions whose structure and semantics may reach different depth of discipline or cross-discipline insights (e.g. INSPIRE directive[16]), be limited to data citation, bearing or not relationships with other data or publications, provenance information, authorship information, hence enabling different degrees of automatic interpretation and reuse [1][39][40]. Varying aspects are data granularity, data formats, data quality (parameters and measures), data re-use, data publishing policies (what data of a Data Center should be published), and data linking (what data should be made available within, be made supplemental to or be linked with publications). Identifying and investing in the right direction might be difficult in absence of well-proved trends and existing experiences. Similarly, keeping up with metadata trends and requirements entailed by the evolution of one discipline or the multi-disciplinary participations requires efforts [41] that might fall out of the scope of DCs and beneficiary scientific communities.

Exporting metadata When cultural and metadata format barriers are not an issue, Data Centers must commit to the technology required to export their dataset metadata. Several standard formats and protocols for exporting metadata about (modern) publications and datasets have been proposed and increasingly adopted in the DC and RDL realms. Among several initiatives, Linked Data [42][43], OAI-ORE [44], and OAI-PMH[17] are known representatives of methods for encoding and exporting metadata of objects for third party re-use.

Linked Data proposes a set of best practices for publishing and connecting structured metadata on the Web as a graph of interrelated objects encoded in RDF format. The adoption of Linked Data by an increasing number of data providers led towards the vision of the Web as a Global Data Space [45], i.e. a global data space containing billions of assertions relative to publications and datasets. Similarly, OAI-ORE defines standards for the description and exchange of "aggregations of Web resources", which are representations of graphs of web resources. The common goal of these standards is to expose metadata object descriptions (e.g. title, publisher and date of a dataset) and relationships between them (e.g. citedBy, partOf) as labeled graphs, together with structural information required to make it automatically accessible and interpretable by consumers. LinkedData SPARQL entry points and OAI-ORE aggregations expose data source metadata as searchable and navigable graph of objects respectively.

OAI-PMH was devised to support bulk-exports of XML metadata records describing the "resources" of a "repository". Although the protocol was conceived in the digital library context, its adoption went beyond this scenario and several dataset repositories and digital archives are today supporting it to expose discipline-specific metadata descriptions (e.g. DataCite, LIDO, EAD). OAI-PMH exposes a list of metadata descriptions whose granularity is expressed by the XML format. For example, a metadata record may encode the metadata of one object together with relationships to metadata descriptions of other objects; i.e. the records represents sub-graphs, rooted subsets of the aforementioned graph of objects.

Therefore, DCs must choose the protocol and implement the required export technologies. Such actions are often driven by community policies. DCs typically pick export formats and protocols guided by the existence of services capable of exploiting and rewarding their efforts. The scientific panorama is extremely heterogeneous on this respect, with some communities thriving with common solutions and others still unaware, uninterested, or not sufficiently motivated to invest in the direction of data publishing and interlinking with publications. For example, the Cultural Heritage community has a long history in sharing content, since disclosure and dissemination are intrinsic part of their mission. Libraries need to share their metadata descriptions to reduce redundant cataloguing work. Museums and archives hold more unique digital artifacts, but need to share vocabularies and authority files, e.g. events, people, topics, places, to collaboratively annotate their collections uniformly and facilitate discovery and interpretation. Moreover, persistent identifiers play a crucial role for digital objects and their descriptive concepts (e.g. vocabularies and authority files) to be uniquely referred and properly preserved into the future. Despite the "stumbling blocks" [46], the Cultural Heritage community has embraced the Linked-Data initiative (and the Linked Open Data project), where metadata sharing and accessibility, vocabulary and authority file sharing, and persistent identifiers are addressed by tools such as RDF*, SKOS, W3C Open Annotation [18] and many others. LinkedData as a publishing practice has brought real benefits and opportunities to the community, which has constructed around it technologies for exporting RDF datasets, collection and aggregation of RDF datasets, collaborative annotation of digital artifacts, generation of common on-

[16] Infrastructure for Spatial Information in the European Community, `http://inspire.jrc.ec.europa.eu`

[17] OAI Protocol for Metadata Harvesting, `http://www.openarchives.org/pmh`

[18] Open Annotation W3C community group, `http://www.w3.org/community/openannotation`

tologies and vocabularies, etc. [47]. However, the same story may not hold in other disciplines. In some cases the cultural barrier makes scientists perceive dataset sharing as harming (others may "steal" results) or a futile action [48]. In other cases, the lack of "community agreements and services" [49] makes the choice difficult to take and the trade-off "cost vs. uncertain benefits" heads off versus a non-choice; for example in the field of neuroimaging, the will to share datasets still finds both cultural and technological barriers [50].

4.2 Barriers for research community organizations

Realizing and maintaining Scientific Communication systems is an expensive activity for a research community and its organizations. In the four categories of solutions presented in Section 3, the first one described how an organization familiar with and operating an RDL needs to invest in the realization of a data repository, hence in a system providing at least minimal but expensive typical DC functionality. In the second case, the same scenario occurs but with an organization operating a DC data repository deciding to invest in the operation of a dedicated RDL [51]. In both cases, the delivery of such "integrated systems" has clear main drawbacks, namely software and system sustainability costs. The technological effort needed to achieve the objectives leads the organizations involved to operate beyond their usual areas of expertise. This is generally an expensive approach, involving software development and refinement costs, as well as personnel expenses. In the third case, the organizations already bear the cost of personnel and maintenance of RDLs and DCs, but still have to realize the software integrating such systems, which generally are not designed to interoperate with each other. Revising code and writing mediation services in order to interlace RDLs and DCs to support different phases of the same scientific communication process is again a non-trivial task. In summary, mainly due to the implementation and maintenance cost of such integrated systems, these three solutions are very pragmatic and tailored to the requirements they must address. As such, they tend to be "minimal" and "static", which means limited to the minimal functionalities required by the community and generally not designed to facilitate further integration of functionality.

Finally, in the fourth case, organizations must implement systems and tools for accessing publications and datasets as exported by RDLs and DCs to support the implementation of the modern publication models. Re-using and combining the metadata "graphs" (see previous section) exported by DCs and RDLs requires the realization, installation and maintenance of adequate "aggregative" systems. These are capable of interpreting the structure and semantics of the data sources (known schemas, vocabularies, etc.), fetch content according to the relative protocols and formats, and map such content onto the physical representation (e.g. triple stores, relational databases, column stores) of a common data model, i.e. structure, semantics. For example, in the Cultural Heritage, where LinkedData is becoming a new trend, several systems have been proposed. One of them is Semantic MediaWiki [52][53], which allows researchers to collaboratively create research corpus out of a set of aggregated LinkedData digital library resources; others are approaches based on distributed RDF queries [54][55]. Other examples are metadata aggregation infrastructures, such as Europeana [19], which collect Cultural Heritage XML metadata descriptions from archives and libraries and attempt to interconnect them to generate richer information corpora. National examples of aggregations are those of NARCIS[20], the gateway to scholarly information in the Netherlands, and Swedish ScienceNet[21] [56], the national scholarly communication infrastructure, which delivers CRIS-like functionalities[22] for the purpose of measuring national research impact (Current Research Information Systems [57]). The software solutions powering such systems suffer from two main drawbacks:

- Their re-usability in other contexts is possible only if the underlying "bottom up" assumptions remain the same (e.g. export and search protocols, metadata formats, vocabularies);
- They are conceived to integrate content in order to generate content, and not to be extended with new functionalities or to integrate existing functionalities, as it is typically the case in different application domains.

The resulting technologies are more general-purpose (e.g. Semantic MediaWiki [53]), but still focused on one technological setting, e.g. LinkedData exports, and deliver community specific services. These issues make them hard to re-use in alternative scenarios, where communities may have not opted for the same technological solutions. As a consequence, such communities are forced to bear the cost of realizing aggregative systems and tools from scratch, by integrating existing products and complementing missing functionalities with new code [58].

[19] Europeana, `http://www.europeana.eu`
[20] NARCIS, `http://www.narcis.nl`
[21] Sweden ScienceNet, `http://www.sciencenet.se`
[22] EuroCRIS, The European Organization for International Research Information, `http://www.eurocris.org`

5 Scientific Communication Infrastructures

Although RDLs and DCs were conceived to serve complementary and non-interoperable tasks of the research process, data and literature publishing requirements in the data-driven science are today demanding them to interoperate. Stakeholders in the research life-cycle (e.g. scientists, funding agencies, organizations) require advanced systems for tracking and identifying links between data and publications, contextualizing them with funding information and author identities, measuring research impact, etc. In the previous section, we highlighted how the implementation and maintenance of modern scientific communication systems fully addressing such requirements is hindered by lack of data publishing practices, technological issues (e.g. interoperability, lack of general-purpose software), and relative sustainability costs. While cultural issues and best practices are being and will be advocated by research communities and by funding agencies to eventually find standards and agreements [4][59][60], a lot of work has to be done in the direction of developing discipline-agnostic technologies capable of facilitating the realization of modern scientific communication systems. The "moving target" effect, being sciences in continuous evolution, and the discipline-specific requirements lead to realization of technology that is "hard" to maintain in the long term and to re-use in different contexts.

e-Science and e-Research trends are strongly advocating for a future where most research data, from raw to secondary, will have to be stored in discipline-specific DCs, and publications deposited in RDLs whose organizations have well-established policies, trained personnel, and sustainability plans to operate such systems. Such trend suggests that the best and more sustainable way to build modern scientific communication systems should be based on an economy-of-scale approach. Accordingly, communities should operate RDLs and DCs dedicated to their original duties and rely on scientific communication systems for the integration of RDLs and DCs so as to address modern dissemination needs. In the following we shall describe our vision towards the realization of scientific communication systems as peculiar cross-discipline research infrastructures, namely *Scientific Communication Infrastructures* (SCIs). In this process we shall present an abstract architecture for such infrastructures, mention the technologies that are today inspired by similar goals, and refer to the real case of the OpenAIRE infrastructure [23] [61] as an example of an embryonic scientific communication infrastructure.

[23] OpenAIRE project and infrastructure, `http://www.openaire.eu`

5.1 An Architecture for SCIs

The main challenge in the construction of modern scientific communication systems regards interoperability with and between RDLs and DCs, independently of their underlying technologies and the disciplines they serve. To serve all their actors, such systems should equally be able to interoperate with Research Infrastructures (RIs), whose functionalities produce and manage data (and indirectly publications), and with so-called *Entity Registries* (ERs), intended as services for maintaining "authority files" of relevance to scientific communication, e.g. authors (VIAF, ORCID, FOAF), funding schemes and projects (CRISs). In Section 3 we observed that existing solutions are mainly conceived to serve one technological domain (i.e. a class of applications based on the same technological approach) or, in some cases, one given discipline scenario (i.e. targeted application or service). In other words, they are not conceived having in mind re-usability and extendibility of software across domains and technologies. The software enabling modern scientific communication systems should instead incarnate such architectural principles, thus offer services for mediating with any kind of data source, manipulating content of arbitrary formats, and facilitate the integration of any functionality services. Such services should:

- Minimize the effort required to integrate content from DCs, RDLs and ERs: "you can take data as it is made available by data sources";
- Minimize the effort required to construct discipline-specific scientific communication workflows: "you can re-use and combine the functionalities in your DCs, RDLs, RIs, and ERs".

SCIs are Scientific Communication Systems satisfying such principles. In the literature their philosophy resembles the vision promoted by Virtual Research Environments (VREs) [13]. VREs are systems providing an integrated environment supporting the collaborative work of a community of researchers (e.g. myExperiment [36], OurSpaces[24]) by sharing a set of resources (e.g. data sources, tools, services, workflows). Example of functionalities researchers may expect from VREs are: authentication, collaboration, resource transfers, functionality over resources, customizability of functionality, re-use of resources, publishing resources, discovering resources, ownership awareness of resources, provenance and access tracking, etc.. SCIs follow as similar approach and provide designers and developers with tools facilitating the dynamic run-time construction and management of *SCI applications* out of con-

[24] OurSpaces, `http://www.ourspaces.net`

tent and functionality from a pool of *SCI resources*, i.e. RDLs, DCs, RIs and ERs. SCIs provide *mediation services* that encapsulate "SCI functionality" within "running services", and *enabling services* that allow for the construction of SCI applications as "service workflows", i.e. sequences of RDL, DC, RI and ER functionalities. Such abstractions offer the flexibility necessary to support and foster the implementation of discipline-specific and cross-discipline forms of scientific communication. This vision goes in the opposite direction with respect to the realization of the integrated systems described in Section 3, but adopts them as real-case scenarios to be served by SCI applications.

Figure 2 illustrates a SCI abstract architecture. The architecture comprises four main functional layers, i.e. *enabling*, *mediation*, *content*, and *application*, and is intended to offer the services to interoperate with and combine functionalities from a set of RDLs, DCs, RIs, and ERs. In the following we describe the core functionalities of such layers providing some concrete examples. The list is not comprehensive, as this would contradict the principle of "extendibility" of SCIs.

Mediation Layer The layer includes services required by the SCI to interact with external systems, such as RDLs, DCs, ERs, and RIs. Systems may offer functionalities via heterogeneous APIs allowing to fetch and feed content, process content, etc. Mediation services should "encapsulate" such functionalities into SCI services whose APIs, data exchange formats, policies follow SCI internal rules and enable interoperation (e.g. combination into workflows). Once integrated, external systems and relative functionalities become "registered resources" of the infrastructure, hence available for discovery and use into applications. For example, a special mediating service may be designed to encapsulate the LinkedData SPARQL entry point of DCs in order to make their content available as a bulk-list of metadata records from an OAI-PMH provider. Such a service should be configurable with a given RDF-XML mapping, possibly implement caching facilities, and support SCI proprietary APIs to exchange its records with other SCI services. More typically, mediation services offer functionality to access content from content resources via standard interfaces, such as OAI-ORE, ODBC, SRW, and to deposit content onto such resources, e.g. deposit a publication onto an RDL (e.g. SWORD project [62]) or a dataset onto a DC. Finally, the layer includes services for the encapsulation of advanced RI functionalities, for example to acquire the results of discipline specific processing workflows, run within the RIs, over content provided by the SCI itself; Figure 2 illustrates the example of a functionality for the analysis of dataset quality.

Content Layer The layer includes services providing functionalities for content storage, processing, and provision. The services should offer different kinds of storage facilities, i.e. physical data models, and offer a variety of services to manage such content. For example, storage services may encapsulate relational databases (MySQL, Postgres), triple stores (Neo4J, Sesame), column stores and NOSQL databases (HBase, Cassandra, MongoDB, BIGDB, CouchDB), full-text indices (Apache Solr, ElasticSearch), and many others. Examples of content processing services are bibliometrics and statistics services for measuring research impact; deduplication services, necessary to delivery precise statistics, maintenance and merge of authority files, etc.; ontology services, to store, manage, and share ontologies within SCIs; transformation and cleaning services, capable of filtering metadata of a given format to generate metadata of an output format; mining services, capable of processing text or other digital content, in order to infer information to enrich or fix metadata information. Finally, provision services should be capable of interacting with storage services in order to expose their content via standard APIs. All content layer services, which should of course offer their full potential via proprietary APIs, should also offer SCI APIs to exchange their content with other services and from workflows.

Application Layer The layer includes services for constructing SCI applications out of running content services and mediation services. To this aim, SCI administrators are provided with tools for the construction and execution/orchestration of "applications", intended as combinations of end-user tools, i.e. portals, and (possibly inter-depending) workflows. Examples of typical applications are:

- Tools and workflows for data deposition policies, which give end-users one single entry point for publishing literature and related datasets by transparently exploiting available DCs and RDLs (see Figure 2;
- Workflows for data peer-review, which exploit RI data analysis services to perform the validation required after submission of data into a DC repository;
- Workflows for inferring relationships between datasets and publications, which process content from RDLs, DCs, and ERs to identify semantic relationships between such objects;
- Tools for managing modern publication models, which provide scientists with functionality to browse through

the objects residing in RDLs and DCs to support authoring, retrieval and navigation, visualization, and publishing of modern publications.

Enabling Layer The layer includes commodity services, which should minimally support the operation of a running SCI in terms of registration and orchestration of resources and authorized access to such resources. For example, a registry service for the registration of functionalities of different kinds from different resources. The registry keeps the "resource map" of the SCI and is the place where other services can discover the functionality services they need among those made available by the content layer and the mediation layer. An orchestration service, for example, may execute workflows in the application layer by discovering which services may accomplish at best its expected processing steps. Authorization and authentication services implement service-to-service and user-to-service access policies, to ensure end-users and applications do not violate agreements with the available resources. Other enabling services may be: subscription and notification services, to offer asynchronous communications between services; message exchange and delivery queues, in the style of ESBs (Enterprise Service Bus [63]), etc.

5.2 Towards the Realization of Scientific Communication Infrastructures

Lately, several research efforts in the field of research infrastructures and e-infrastructures [64] have led to the realization of software (often called "enabling software") for the construction and deployment of data infrastructures (e.g. D-NET [58], Cezary et al. [65], gCube [66]). For example, the *D-NET Software Toolkit* [58] was specifically devised to enable the construction of workflows by integrating third-party services with a set of highly configurable D-NET data management services. D-NET services are capable of storing, processing, and providing access to data according to several physical data models, logical data models, metadata formats, and standard access APIs. D-NET has been used to power the OpenAIRE infrastructure (Open Access Infrastructure for Research in Europe), realized and maintained by the homonymous project [67], to become the European Scholarly Communication Infrastructure. OpenAIRE's mission is to promote and measure the impact of Open Science and Open Access by means of a modern scientific communication system. The project has delivered a data infrastructure capable of collecting and interlinking content from RDLs (i.e. OA and non-OA publication repositories), DCs (i.e. research data repositories), and CRIS systems (i.e. funding information from European Commission and National funding schemes). Moreover, it supports advanced metrics to measure impacts of Open Access mandates and funding over research. The infrastructure populates a graph of (metadata of) objects spanning across all research disciplines and countries, with the major objectives of: (*i*) providing enhanced access to the graph for end-users and third-party systems, (*ii*) experimenting automatic inference of semantic relationships between different object typologies (e.g. datasets and publications), (*iii*) de-duplicating publication metadata, and (*iv*) construction and refinement of "enhanced publications". To this aim, D-NET offers a suite of services that cover the layers shown in Figure 2. In particular, mediation and enabling layers allow for the integration and access to content resources and for the encapsulation of RI functionalities, which are then combined to form OpenAIRE SCI applications. Examples of the latter are relationship inference functionality, which are deployed at RDL sites to parse article PDFs without violating copyrights; on-line key-word inference services, supported by the EBI institute (see Section 3.3); DataCite DOI dereference, etc..

On the other hand, D-NET covers only a portion of the possible interactions with DCs, RDLs, RIs and ERs. It focuses on storage and processing of metadata as XML files and their possible encoding onto relational databases (Postgres), full-text indices (Apache Solr) and column stores (HBase and Hadoop). For example it misses services for collection and processing of LinkedData, or services for long-term preservation of digital objects. This is to say that enabling software for SCIs may vary depending on the services they offer, the common data exchange APIs they are willing to impose, the kind of resources they are targeting, etc. In general, they can grow in functionalities depending on the scenarios and the domains they will serve. In the future, we expect the growing needs for scientific communication systems will push the scientific communities to adopt and extend such technological solutions, and encourage researchers in e-Science and e-Research to investigate into the realization of enabling software for SCIs.

6 Conclusions and Future Issues

A lot of work needs to be done. The idea of enabling a "global scientific communication infrastructure", unifying and giving access in a systematic, discipline-specific, authorized, and reusable way to the whole outcome of world's research, must rely on common practices and standard ways to engage SCIs themselves into larger

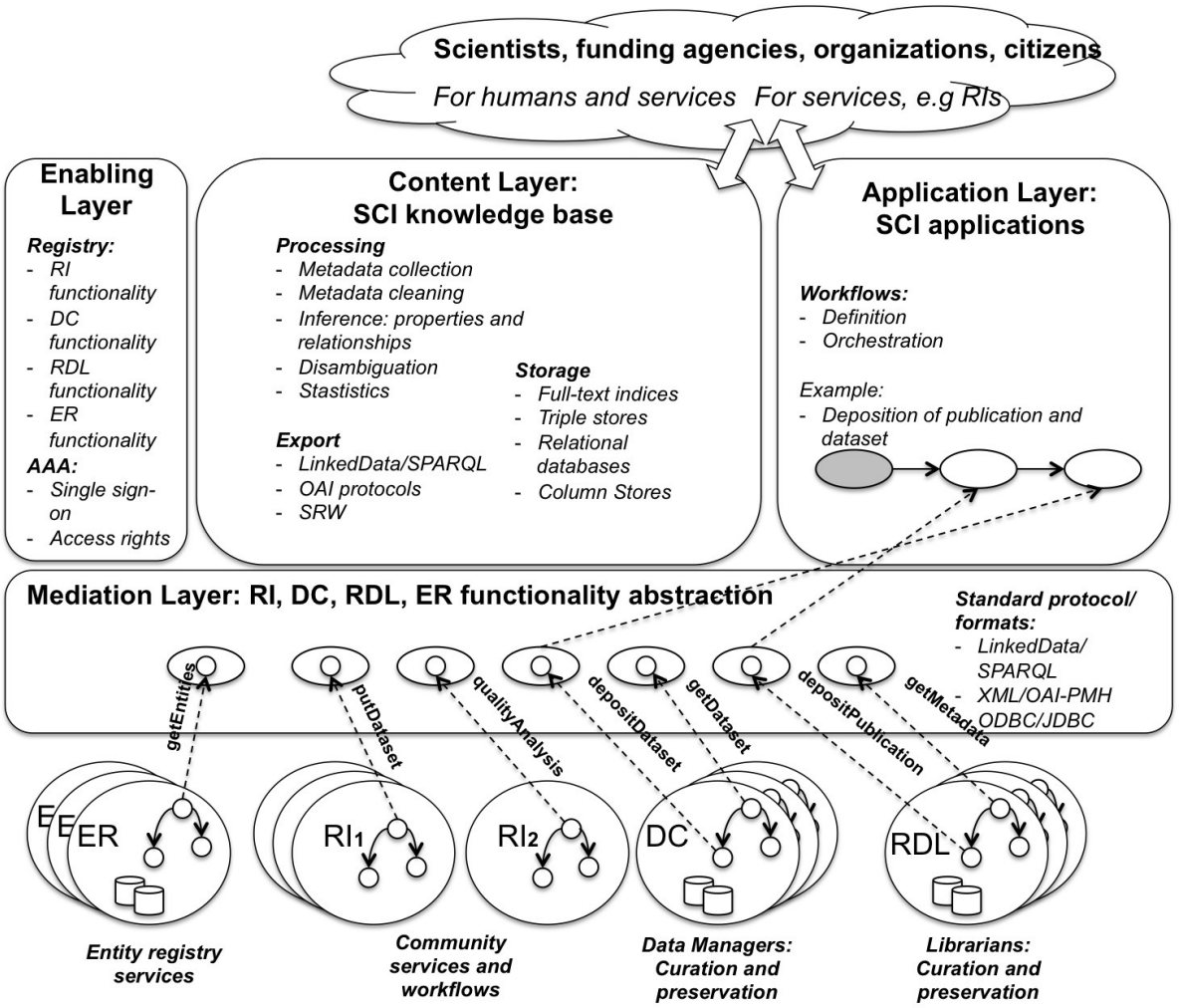

Fig. 2 Scientific Communication Infrastructures: a high-level architecture

eco-systems, i.e. infrastructures of infrastructures. However, existing solutions, although successful, are experimenting with the concepts underlying enabling software for SCIs. The relative communities and groups of scientists are still in the process of proposing new ideas rather than focusing on common solutions. Some of such solutions can partly be shared with those research communities targeting recommendations for the construction of research infrastructures. The Research Data Alliance[25] (RDA) and the e-Infrastructure Reflection Group[26] (e-IRG) , as well as other projects and initiatives world-wide, represent community efforts to achieve common best practices, standards, architectures, data models, and possibly services in the construction of research infrastructures. Other aspects, such as data models for modern publications, services, application patterns for scientific communication processes, are instead very specific to the realization of SCIs. We are convinced that these problems will offer a wide range of research opportunities and will become the focus of studies in the years to come.

Acknowledgements The authors wish to thank Maria Bruna Baldacci who provided valuable advice to the writing of this paper. This work is partially supported by the European Commission as part of the project OpenAIREplus (FP7-INFRA-2011-2, Grant Agreement no. 283595).

References

1. J. Gray, A Transformed Scientific Method. In The Fourth Paradigm: Data Intensive Scientific Discovery, Redmond, WA: Microsoft, 2009.
2. Towards 2020 Science, Report of the 2020 Science Workshop, Venice, 30 June-1 July 2005. Microsoft Corporation, 2006. `http://research.microsoft.com/en-us/um/cambridge/projects/towards2020science`
3. P. Lord, A. Macdonald, e-Science Curation Report prepared for the JISC Committee for the Support of

[25] Research Data Alliance, `http://rd-alliance.org`
[26] e-Infrastructure Reflection Group, `http://www.e-irg.eu`

Research. The Digital Archiving Consultancy Ltd., 2003. http://www.jisc.ac.uk/uploaded_documents/e-ScienceReportFinal.pdf
4. Mooney, H, Newton, MP. (2012). The Anatomy of a Data Citation: Discovery, Reuse, and Credit. Journal of Librarianship and Scholarly Communication 1(1):eP1035. http://dx.doi.org/10.7710/2162-3309.1035
5. M. Altman and G. King A Proposed Standard for the Scholarly Citation of Quantitative Data. D-Lib Magazine March/April 2007.
6. Hollie C. White, Sarah Carrier, Abbey Thompson, Jane Greenberg, and Ryan Scherle. 2008. The Dryad data repository: a Singapore framework metadata architecture in a DSpace environment. In Proceedings of the 2008 International Conference on Dublin Core and Metadata Applications(DCMI '08). Dublin Core Metadata Initiative 157-162.
7. Leonardo Candela, Akrivi Katifori, and Paolo Manghi. e-infrastructures. In: Christian Meier zu Verl and Wolfram Horstmann, editors, Studies on Subject-Specific Requirements for Open Access Infrastructures, pp. 125-164. Universitatsbibliothek Bielefeld, 2011
8. Attila Csordas, David Ovelleiro, Rui Wang, Joseph M. Foster, Daniel Ros, Juan Antonio Vizcano, and Henning Hermjakob PRIDE: Quality control in a proteomics data repository. Database 2012: bas004 doi:10.1093/database/bas004 published online March 20, 2012
9. S. Reilly, W. Schallier, S. Schrimpf, E. Smit, M. Wilkinson, Report on Integration of Data and Publications. Opportunities for Data Exchange (ODE), 2011.
10. Callaghan, Sarah and Donegan, Steve and Pepler, Sam and Thorley, Mark and Cunningham, Nathan and Kirsch, Peter and Ault, Linda and Bell, Patrick and Bowie, Rod and Leadbetter, Adam and Lowry, Roy and Moncoiffe, Gwen and Harrison, Kate and Smith-Haddon, Ben and Weatherby, Anita and Wright, Dan (2012), Making Data a First Class Scientific Output: Data Citation and Publication by NERC's Environmental Data Centers. International Journal of Digital Curation, 7(1), 107-113
11. Towards better access to scientific information: Boosting the benefits of public investments in research. Communication from the Commission to the European Parliament, The Council, the European Economic and Social Committee and the Committee of the Regions. COM(2012) 401 final. Brussels, 17..7.20112, http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2012:0401:FIN:EN:PDF
12. JISC an data. Data centres: their use, value and impact. A Research Information Network report, September 2011. http://www.jisc.ac.uk/news/stories/2011/09/~/media/-Data%20Centres-Updated.ashx
13. Alexander Voss, Rob Procter, (2009) Virtual research environments in scholarly work and communications, Library Hi Tech, Vol. 27 Iss: 2, pp.174 - 190
14. Borgman, C. L. (2012). The conundrum of sharing research data. Journal of the American Society for Information Science and Technology, 63(6): 1059-1078. http://dx.doi.org/10.1002/asi.22634
15. Pampel, H. , Pfeiffenberger, H. , Schfer, A. , Smit, E. , Prll, S. and Bruch, C. (2012): Report on Peer Review of Research Data in Scholarly Communication, hdl:10013/epic.39289
16. A. Renear, C. Palmer, Strategic Reading, Ontologies, and the Future of Scientific Publishing, Science, Vol. 325, 2009.
17. Natasha Simons, Implementing DOIs for Research Data, D-Lib Magazine,, Volume 18, Number 5/6, May/June 2012. doi:10.1045/may2012-simons
18. Joan Starr and Angela Gastlis, CitedBy: A Metadata Scheme for DataCite, D-Lib Magazine, January/February 2011, Volume 17, Number 1/2, doi:10.1045/january2011-starr
19. Green, T (2009), We Need Publishing Standards for Datasets and Data Tables, OECD Publishing White Paper, OECD Publishing. doi:10.1787/603233448430
20. M. Smith, M. Barton et al.. DSpace: An Open Source Dynamic Digital Repository. D-Lib Magazine, 9(1), 2003. http://www.dlib.org/dlib/january03/smith/01smith.html
21. S. Payette, C. Lagoze. Flexible and Extensible Digital Objects Repository Architecture (FEDORA). In: Research and Advanced Technology for Digital Libraries, Proc. of the Second Conference on Digital Libraries, ECDL 98, Crete, Greece. Springer Lecture Notes in Cpmputer Science), pp.41-59.
22. I.H. Witten, D. Bainbridge. How to Build a Digital Library. Elsevier, 2002
23. De Schutter, Erik, Data Publishing and Scientific Journals: The Future of the Scientific Paper in a World of Shared Data, Neuroinformatics Journal, 2010, Humana Press Inc., pp.151-153, Vol. 8, Issue 3, Doi:10.1007/s12021-010-9084-8
24. Vishwas Chavan and Lyubomir Penev, The data paper: a mechanism to incentivize data publishing in biodiversity science, BMC Bioinformatics 2011, 12(Suppl 15):S2. doi:10.1186/1471-2105-12-S15-S2
25. Shotton, D. (2009). Semantic publishing: the coming revolution in scientific journal publishing. Learned Publishing, 22(2), 85-94.
26. David Tempest (Universal Access Team Leader, Elsevier UK), Journals and Data Publishing: Enhancing, Linking And Mining. DCC Research Data Management Forum 8: Research data management - engaging with the publishers, Southampton, 29-30 March 2012
27. D. Castelli, P. Pagano, OpenDLib: A Digital Library Service System, In Proc. of the 6th European Conference on Research and Advanced Technology for Digital Libraries (ECDL'02)
28. Aalbersberg, I. J., Heeman, F., Koers, H., and Zudilova-Seinstra, E. (2012). Elsevier's Article of the Future enhancing the user experience and integrating data through applications. Insights: the UKSG journal, 25(1), 33-43.
29. Attwood TK, Kell DB, McDermott P, Marsh J, Pettifer SR, Thorne D. Utopia documents: linking scholarly literature with research data. Bioinformatics. 2010 Sep 15;26(18):i568-74. doi:10.1093/bioinformatics/btq383.
30. Quan Pham, Tanu Malik, Ian Foster, Roberto Di Lauro, Raffaele Montella, SOLE: Linking Research Papers with Science Objects, In Provenance and Annotation of Data and Processes. Lecture Notes in Computer Science Volume 7525, 2012, pp 203-208. Doi:10.1007/978-3-642-34222-6_16
31. S. Woutersen-Windhouwer, R. Brandsma, P. Verhaar, A. Hogenaar, M. Hoogerwerf, P. Doorenbosch, E. Durr, J. Ludwig, B. Schmidt, B. Sierman, Enhanced Publications, edited by M. Vernooy-Gerritsen, SURF Foundation, Amsterdam University Press, 2009
32. J.G. Kircz, New Practices for Electronic Publishing New Forms of the Scientific Paper. In: Learned Publishing, Vol. 15, No 1, January 2002
33. L. Candela, D. Castelli, P. Pagano, M. Simi, From Heterogeneous Information Spaces to Virtual Documents. In Digital Libraries: Implementing Strategies and Sharing Experiences: 8th International Conference on Asian Digital

Libraries, ICADL 2005 (Bangkok, Thailand, 12-1 December 2005). Proceedings, pp. 11 - 22. Edward A. Fox, Erich J. Neuhold, Pimrumpai Premsmit, Vilas Wuwongse (eds).
34. C. Lynch, Jim Gray's Fourth Paradigm and the Construction of the Scientific Record. In: T. Hey, S. Tansley, and C. Tolle, Eds., The Fourth Paradigm , Microsoft Corporation, 2009, pp.177-183.
35. J. Hunter, Scientific Models A user oriented Approach to the Integration of Scientific Data and Digital Libraries, VALA 2006, Melbourne, February, 2006.
36. S. Bechhofer, D. De Roure, M. Gamble, C. Goble, I. Buchan, Research Objects: Towards Exchange and Reuse of Digital Knowledge. In: Proc. The Future of the Web for Collaborative Science (FWCS 2010), Raleigh, NC, USA. http://www.w3.org/wiki/HCLS/WWW2010/Workshop
37. Brammer, G. R., Crosby, R. W., Matthews, S. J., and Williams, T. L. (2011). Paper Mch: Creating Dynamic Reproducible Science. Procedia Computer Science, 4, 658-667. doi:10.1016/j.procs.2011.04.069
38. Van Gorp, P. and Mazanek, S. (2011). SHARE: a web portal for creating and sharing executable research papers. Procedia Computer Science, 4, 589-597. doi:10.1016/j.procs.2011.04.062
39. McCallum, I., Plag, H.P., and Fritz, S. 2012. Data Citation Standard: A Means to Support Data Sharing, Attribution, and Traceability. In EGU General Assembly Conference Abstracts. Abbasi, A. and Giesen, N., Vol. 14, Series EGU General Assembly, http://adsabs.harvard.edu/abs/2012EGUGA..1413029M
40. Schfer, A. , Pampel, H. , Pfeiffenberger, H. , Dallmeier-Tiessen, S. , Tissari, S. , Darby, R. , Giaretta, K. , Giaretta, D. , Gitmans, K. , Helin, H. , Lambert, S. , Mele, S. , Reilly, S. , Ruiz, S. , Sandberg, M. , Schallier, W. , Schrimpf, S. , Smit, E. , Wilkinson, M. and Wilson, M. (2011): Baseline Report on Drivers and Barriers in Data Sharing
41. Joo Rocha da Silva, Cristina Ribeiro, Joo Correia Lopes, Semi-automated Application Profile Generation for Research Data Assets, Metadata and Semantics Research Communications in Computer and Information Science Volume 343, 2012, pp 98-106
42. Berners-Lee. T. (2006b). Linked Data. Archived on December 1st, 2006. http://web.archive.org/web/20061201121454 http://www.w3.org/DesignIssues/LinkedData.html
43. Bizer, C., Heath, T., and Berners-Lee, T. (2009). Linked Data - The Story So Far. International Journal on Semantic Web and Information Systems (IJSWIS), 5(3), 1-22. doi:10.4018/jswis.2009081901
44. Carl Lagoze and Herbert Van de Sompel, The OAI Protocol for Object Reuse and Exchange. http://www.openarchives.org/ore
45. Heath, T. and Bizer, C. (2011). Linked Data: Evolving the Web Into a Global Data Space. Synthesis Lectuers on the Semantic Web: Theory and Technology, 1:1, 1?136. Morgan & Claypool. Retrieved from http://linkeddatabook.com/editions/1.0/#htoc9
46. Ed Summers, Linking Things on the Web: A Pragmatic Examination of Linked Data for Libraries, Archives and Museums. 2013. Library of Congress. arXiv:1302.4591
47. Eero Hyvnen: Publishing and Using Cultural Heritage Linked Data on the Semantic Web. Morgan & Claypool, Palo Alto, CA, USA, October, 2012.
48. Studies on Subject-Specific Requirements for Open Access Infrastructure, Meier zu Verl C, Horstmann W (Eds) (2011), Bielefeld: Universittsbibliothek. DOI:10.2390/PUB-2011-1
49. Sean Bechhofer, Iain Buchan, David De Roure, Paolo Missier, John Ainsworth, Jiten Bhagat, Philip Couch, Don Cruickshank, Mark Delderfield, Ian Dunlop, Matthew Gamble, Danius Michaelides, Stuart Owen, David Newman, Shoaib Sufi, Carole Goble, Why linked data is not enough for scientists, Future Generation Computer Systems, Available online 19 August 2011, ISSN 0167-739X, dpi:10.1016/j.future.2011.08.004.
50. Breeze, Janis L., Jean-Baptiste Poline, and David N. Kennedy. Data sharing and publishing in the field of neuroimaging. Giga Science 1.1 (2012): 1-3.
51. Parsons, M. A., R. Duerr, and J.B. Minster (2010), Data Citation and Peer Review, Eos Trans. AGU, 91(34), 297, doi:10.1029/2010EO340001.
52. Christoph Schindler, Cornelia Veja, Marc Rittberger, and Denny Vrandecic. 2011. How to teach digital library data to swim into research. In Proceedings of the 7th International Conference on Semantic Systems (I-Semantics '11), Chiara Ghidini, Axel-Cyrille Ngonga Ngomo, Stefanie Lindstaedt, and Tassilo Pellegrini (Eds.). ACM, New York, NY, USA, 142-149. DOI=10.1145/2063518.2063537 http://doi.acm.org/10.1145/2063518.2063537
53. A Krtzsch, Markus, Vrandecic, Denny, and Volkel, Max. 2006. Semantic MediaWiki. In The Semantic Web - ISWC. LNCS 4273, Springer Berlin Heidelberg. Pages 935-942, doi:10.1007/119260
54. Bastian Quilitz and Ulf Leser, Querying Distributed RDF Data Sources with SPARQL (2008). In Proceedings of The Semantic Web: Research and Applications.Lecture Notes in Computer Science Volume 5021, pp 524-538. doi:10.1007/978-3-540-68234-9_39
55. Zeng, K., Yang, J., Wang, H., Shao, B., and Wang, Z. (2013). A Distributed Graph Engine for Web Scale RDF Data. Proceedings of the VLDB Endowment, 6(4).
56. Johansson, . and Ottosson, M. O. (2012). A national Current Research Information System for Sweden. In e-Infrastructures for Research and Innovation: Linking Information Systems to Improve Scientific Knowledge Production (pp. 67-71). Agentura Action M.
57. Asserson, Anne, K. Jeffery, and Andrei Lopatenko. CERIF: past, present and future: an overview. Proceedings: Gaining Insight from Research Information. 6th International Conference on Current Research Information Systems, Kassel, Germany. 2002.
58. Paolo Manghi, Marko Mikulicic, Leonardo Candela, Donatella Castelli, and Pasquale Pagano. Realizing and Maintaining Aggregative Digital Library Systems: D-NET Software Toolkit and OAIster System. D-Lib Magazine, 16(3/4), March/April 2010
59. Developing Data Attribution and Citation Practices and Standards. An International Symposium and Workshop. August 22-23, 2011. US CODATA and the Board on Research Data and Information, in collaboration with CODATA-ICSTI Task Group on Data Citation Standards and Practices, http://sites.nationalacademies.org/PGA/brdi/PGA_064019
60. British Library, Datacite, Jisc. Workshop report - Describe, disseminate, discover: metadata for effective data citation, published by C. Wilkinson, 24 july 2012, http://www.datacite.org/node/67
61. Paolo Manghi, Lukasz Bolikowski, Natalia Manola, Jochen Shirrwagen, and Tim Smith. Openaireplus: the european scholarly communication data infrastructure. D-Lib Magazine, 18(9-10), September October 2012
62. Allinson, Julie, Sebastien Franois, and Stuart Lewis. Sword: Simple web-service offering repository deposit. Ariadne 54 (2008): 2.

63. Schmidt, M-T., et al. The enterprise service bus: making service-oriented architecture real. IBM Systems Journal 44.4 (2005): 781-797.
64. GRDI2020 Consortium. Global Research Data Infrastructures: The Big Data Challenges. GRDI200 Final Roadmap Report, February 2012. http://www.grdi2020.eu/Repository/FileScaricati/e2b03611-e58f-4242-946a-5b21f17d2947.pdf
65. Cezary Mazurek, Marcin Mielnicki, Aleksandra Nowak, Maciej Stroinski, Marcin Werla, and Jan Weglarz. Architecture for aggregation, processing and provisioning of data from heterogeneous scientific information services. In Robert Bembenik, Lukasz Skonieczny, Henryk Rybinski, Marzena Kryszkiewicz, and Marek Niezgodka, editors, Intelligent Tools for Building a Scientific Information Platform, volume 467 of Studies in Computational Intelligence, pages 529-546. Springer Berlin Heidelberg, 2013. ISBN 1978-3-642-35646-9. doi:10.1007/978-3-642-35647-632.
66. Candela, L.; Castelli, D., and Pagano, P. gCube: A Service-Oriented Application Framework on the Grid ERCIM News, 2008, 48-49
67. Paolo Manghi, Natalia Manola, Wolfram Horstmann, and Dale Peters. An Infrastructure for Managing EC Funded Research Output. International Journal on Grey Literature (TGJ), 6(1), Spring 2010, http://www.openaire.eu/it/about-openaire/publications-presentations/doc_details/189-an-infrastructure-for-managing-ec-funded-research-output