The INFN Open Access Repository

Conceptual Design Report

R. Barbera, S. Bianco, M. Fargetta, M. Maggi,
D. Menasce, L. Patrizii, R. Rotondo,
in collaborazione con P. Lubrano

Abstract

This document presents the conceptual design of a long-term institutional Open Access Repository for the INFN. The motivations and the objectives as well as the state-of-the-art technologies available and some already existing examples are discussed. Both the current implementation and the proposed long-term solution are presented, together with the human and financial resources needed.

Contents

List of figures

List of tables

Acronyms and abbreviations

Acronym/Abbreviation	Full name
API	Application Programming Interface
AWS	Amazon Web Services
CERN	European Organization for Nuclear Research
COTS	Commercial Off-The-Shelf
CRIS	Current Research Information System
DESY	Deutsches Elektronen-Synchrotron
DMCI	Dublin Core Metadata Initiative
DOI	Digital Object Identifier
DR	Data Repository
EC	European Commission
EOSC	European Open Science Cloud
EPFL	Ecole Polytechnique Federale de Lausanne
EU	European Union
FAIR	Findable, Accessible, Interoperable, Re-usable
FNAL	Fermi National Accelerator Laboratory
FP4, FP5, FP6, FP7	4th, 5th, 6th, 7th Framework Programme of the European Union
GUI	Graphic User Interface
H2020	Horizon 2020 Programme of the European Union
HEP	High Energy Physics
HPC	High-Performance Computing
HTC	High-Throughput Computing
ICDI	Italian Computing and Data Infrastructure
INFN	Italian National Institute of Nuclear Physics
INSPIRE	The High Energy Physics information system
IT	Information Technology
LAN	Local Area Network
LHC	Large Hadron Collider
LOD	Linked Open Data
MARC21	MARC 21 Format for Bibliographic Data
NFS	Network File System
OADR	Open Access Document Repository
OAI	Open Archives Initiative
OAI-PMH	Open Archives Initiative - Protocol for Metadata Harvesting
OpenDOAR	Directory of Open Access Repositories
ORCID	Open Researcher and Contributor ID
PLAN S	Initiative for Open Access publishing
SCOAP³	Sponsoring Consortium for Open Access Publishing in Particle Physics
SLAC	Stanford Linear Accelerator Center
VQR	Valutazione della Qualità della Ricerca (Assessment of the quality of research)
VRC	Virtual Research Community
WAN	Wide Area Network
WLCG	Worldwide LHC Computing Grid

Executive summary

Dopo quattro secoli dalla prima rivoluzione scientifica, che vide l'affermazione del Metodo Scientifico induttivo e la nascita delle riviste scientifiche per la condivisione e la trasmissione della conoscenza, ed a circa un secolo dalla seconda, basata sull'affermazione definitiva del pensiero matematico e sull'adozione del Metodo Scientifico deduttivo per la formulazione delle grandi teorie (l'elettrodinamica classica di Maxwell, la relatività di Einstein e la meccanica quantistica), in questi ultimi dieci anni la scienza sta vivendo una sua terza rivoluzione, testimoniata da un approccio intrinsecamente multidisciplinare, dalla condivisione di tutti i prodotti della ricerca (non soltanto le pubblicazioni ma anche le moli sempre più grandi dei dati raccolti, il software, ecc.) e dall'adozione pervasiva delle tecnologie di Internet e del World Wide Web.

Termini come "Computational Science", "Data Science", "Science 2.0" e "Open Science" sono stati coniati per indicare un vero e proprio "paradigm shift" nel modo di fare ricerca scientifica e, conseguentemente, innovazione. Questo nuovo approccio ha creato iniziative internazionali quali cOAlition S, CODATA, GOFAIR, OpenAIRE, OpenDOAR, Research Data Alliance (RDA), ecc. che hanno fornito le basi a importanti decisioni politiche quali la creazione dell'European Open Science Cloud (EOSC) che avrà un ruolo di primo piano nel prossimo programma quadro dell'Unione Europea ("Horizon Europe") e si confermerà come "ombrello" delle attività di punta per l'implementazione definitiva e l'abilitazione dell'Area Europea della Ricerca.

A livello nazionale, da un punto di vista tecnico, il GARR sta coordinando il gruppo di lavoro sull'Italian Computing and Data Infrastructure (ICDI) creato dai rappresentanti di alcune tra le principali Infrastrutture di Ricerca e Infrastrutture Digitali italiane, con l'obiettivo di promuovere sinergie a livello nazionale, al fine di ottimizzare la partecipazione italiana alle attuali sfide europee in questo settore, tra cui EOSC, la European Data Infrastructure (EDI) e HPC.
Dal punto di vista politico, il MIUR sta prendendo coscienza di questa onda di cambiamento e ciò è testimoniato dalla recente istituzione, da parte del Capo Dipartimento per la Formazione Superiore e per la Ricerca, di tavoli tecnici per la redazione di un Position Paper dell'Italia su EOSC e di un Piano Nazionale per la Scienza Aperta. Inoltre, è in discussione al Senato la proposta di legge d'iniziativa del Deputato Gallo sulla materia dell'accesso aperto all'informazione scientifica.
In tutti i documenti in discussione in tali contesti, la creazione di archivi istituzionale è definita come una priorità.

Uno degli elementi abilitanti dell'Open Science è infatti un archivio digitale ad accesso aperto (o almeno controllato) nel quale conservare, etichettare e rendere disponibili, citabili e riusabili tutti i prodotti della ricerca. A tutt'oggi, l'INFN ha diversi database contenenti pubblicazioni ma non si è ancora dotato di un archivio istituzionale vero e proprio, **integrato sinergicamente all'interno del Sistema Informativo. Tale integrazione va coordinata in strettissima collaborazione con i responsabili del Sistema informativo. Occorre altresì un forte coordinamento con il Gruppo di Lavoro Valutazione per il quale l'archivio istituzionale rivestirebbe il ruolo di strumento essenziale.**

Al momento in cui questo documento viene redatto:

- nessun archivio di dati INFN è elencato nella nota piattaforma re3data.org (i dati dell'INFN sono "invisibil"i su una piattaforma che sta assumendo una grande rilevanza internazionale);
- l'archivio "general purpose" Figshare annovera 18 record dei quali almeno un coautore è affiliato all'INFN;
- l'archivio "general purpose" Zenodo include 147 record dei quali almeno un coautore è affiliato all'INFN: 75 pubblicazioni, 27 software, 19 presentazioni, 10 dataset, 9 poster, 5 video and 2 lezioni.

Questa situazione va modificata e l'INFN merita, al pari di altri importanti enti di ricerca in Italia e nel resto del mondo, di avere un proprio archivio istituzionale che soddisfi, tra gli altri, i requisiti di essere:

- aperto e basato su standard internazionali ampiamente adottati;
- capace di conservare tutte le tipologie di prodotti della ricerca;
- capace di estrarre e salvare al proprio interno prodotti della ricerca di personale INFN contenuti in altri archivi (es., arXiv, OpenAIRE, ecc.);
- capace di supportare il Gruppo di Lavoro Valutazione dell'Ente in occasione delle valutazioni periodiche della ricerca (VQR);
- capace di fornire valore aggiunto e supportare la Terza Missione dell'INFN;
- capace di aumentare la visibilità sia della ricerca che dei ricercatori dell'INFN (attraverso l'integrazione con ORCID);
- compatibile con i principi FAIR ("findable, accessible, interoperable, reusable");
- conforme ai requisiti ed alle linee guida definite da cOAlition S nel PLAN S (v. tabella 2 più avanti);
- conforme con EOSC e interoperabile con i servizi che fanno parte del suo catalogo;
- un elemento abilitante chiave di ICDI, specialmente per la cosiddetta "long-tail of science", sia all'interno che all'esterno dell'Ente.

Questo documento descrive il progetto di lungo termine di un archivio istituzionale sostenibile per l'INFN. L'archivio accoglierà e conserverà tutto il patrimonio documentale. La tecnologia proposta è in grado di gestire ogni tipo di file, incluso file di dati con gestione delle macchine virtuali, contenenti tutto il software necessario alla loro elaborazione. L'archivio proposto intende sostituire tutti i database di pubblicazioni e dati esistenti nell'INFN, archivi che consistono attualmente di database delle pubblicazioni, archivio delle note interne INFN e LNF, archivio foto/audio-video LNF, ecc..
La logica di uso sarà quella della massima semplicità. L'utente potrà depositare il prodotto una unica volta, non saranno necessarie altre manipolazioni da parte sua, ogni propagazione sarà gestita automaticamente dal sistema. Nel caso di preprint scientifici accettati da arXiv, l'utente avrà la possibilità di depositare su arXiv e il sistema copierà il documento sull'archivio istituzionale.

Motivazioni per un archivio istituzionale interno all'INFN
I database esistenti al momento:

- non coprono simultaneamente tutte le tipologie di documenti di interesse per l'INFN, come ad esempio la letteratura grigia, il software, i documenti di progettazione, i dati sperimentali, il portfolio dei brevetti, il materiale disseminativo multimediale, ecc.;

- non possono automaticamente trovare ed ingerire prodotti della ricerca con autori INFN che siano presenti anche su altri archivi (questo rende difficoltosa e manuale la composizione di bibliografie complete ed aggiornate con autori appartenenti all'INFN);
- non sono integrati con ORCID;
- non sono conformi con PLAN S;
- non sono conformi al catalogo EOSC né interoperabili con altri servizi di EOSC;
- non sono comprensivi;
- non forniscono un DOI;
- non sono interoperabili, senza non trascurabili investimenti in termini di risorse umane, con gli archivi istituzionali di altri enti quali CNR, INAF, INGV, ISS, ecc.
- non sono in grado di ospitare dati sperimentali prodotti dall'INFN in modalità FAIR, sia per il pubblico generico che per necessità scientifiche;
- non sono in grado di gestire la Data Preservation attraverso il meccanismo di Container e/o Virtual Machine (problematica differente dall'Open Data ma con molti punti architetturali e di gestione in comune).

La soluzione tecnologica proposta
- È basata su framework, quali Invenio e Zenodo, sviluppati, supportati e mantenuti dal CERN, garanzia di evoluzione sostenibile nel tempo.
- È conforme con tutti i più importanti standard bibliografici internazionali.
- È open source, paradigma centrale; soluzioni proprietarie non verrebbero accettate dalla comunità scientifica internazionale.
- Permette installazione e gestione locale, con uso dinamico ed "elastico" delle risorse.
- È conforme con la versione più recente delle linee guida di PLAN S; in particolare, con quelle riportate di seguito che non sono invece soddisfatte da un archivio molto usato nell'ambito della Fisica delle Alte Energie qual è arXiv:
    o *"High-quality article level metadata in standard interoperable non-proprietary format, under a CC0 public domain dedication. Metadata must include complete and reliable information on funding provided by cOAlition S funders (including as a minimum the name of the funder and the grant number/identifier)."*
    o *"Full text stored in XML, conforming to an open community standard, such as JATS."*
    o *"Linking to data, code, and other research outputs that underlie the publication and are available in external repositories."*

Implementazione nell'architettura documentale dell'INFN
Questo è certamente un punto critico: un'agile integrazione con il Sistema Informativo dell'INFN e col sistema documentale permetterebbe una notevole flessibilità nel fornire snapshot istantanei, su richiesta del management e/o del Gruppo di Lavoro Valutazione, circa la situazione corrente dei prodotti della ricerca dell'Ente. Affinché questa integrazione avvenga occorre omogeneizzare e coordinare lo sviluppo di tutte le component software del sistema integrato e questo richiede personale con esperienze in diversi ambiti: dal software puro, al web semantico, alle problematiche bibliotecarie e di "data stewardship" e una certa dose di conoscenza delle problematiche scientifiche.
L'attività relativa alla valutazione e in particolare alla VQR è strategica ed ha la massima priorità. **La soluzione tecnica scelta per l'archivio istituzionale è flessibile e consentirebbe, se ritenuto necessario, di integrarsi con l'esistente database delle pubblicazioni lasciandone inalterate le funzionalità e i collegamenti con gli altri applicativi del Sistema Informativo.** Questa possibilità è concettualmente simile a quella in

fase di implementazione nel CNR, dove l'esistente database (di soli metadati) PEOPLE sarà interfacciato con il nuovo repository Open Science basato su Zenodo.

Risorse finanziarie e loro profilo temporale

Data la recente approvazione del progetto IBiSCo (Infrastruttura per Big data e Scientific COmputing) da parte del MIUR, con orizzonte temporale fino ad almeno il 31/12/2032, e considerando un ragionevole profilo temporale d'utilizzo del servizio da parte del personale INFN dipendente e associato, il progetto dell'archivio istituzionale appare senz'altro sostenibile da un punto di vista infrastrutturale.

Da un punto di vista delle risorse umane necessarie, invece, le seguenti richieste sono considerate minimali, in aggiunta naturalmente a quelle che costituiscono già il Gruppo di Lavoro sull'Open Access che dovrebbero essere integrate da competenze di "data stewardship" e potrebbero all'occorrenza essere riorganizzate secondo un Management Board, un Editorial Board, un Help Desk, ecc.:

- fase di "commissioning" (2019-2020):
  - n. 1 tecnologo (con profilo di informatico o ingegnere informatico) con esperienza nella gestione di archivi digitali;
  - n. 1 tecnologo (con profilo di informatico o ingegnere informatico) per il supporto all'integrazione dell'archivio istituzionale con gli altri servizi del Sistema Informativo dell'INFN, con particolare riguardo a quelli che sono interessati dalla VQR;
  - n. 1 bibliotecario digitale;
- fase di normale operatività (dal 2021 in poi):
  - n. 1 tecnologo (con profilo di informatico o ingegnere informatico) con esperienza nella gestione di archivi digitali;
  - n. 1 bibliotecario digitale;

Oltre alle risorse umane elencate sopra, è considerato altresì necessario un budget per coprire le spese di mobilità (contatti, collaborazioni, partecipazione a workshop tecnici, eventi di outreach e di disseminazione, eventi di aggiornamento e formazione, ecc.) e consumo (costo annuale del prefisso DOI, metabolismo nelle sedi coinvolte, ecc.).

Alternative commerciali

Questo è un punto delicato. La base concettuale dei principi FAIR è che tutto il sistema di gestione e fornitura di letteratura Open Access e di Open Data sia, per l'appunto, "Accessible", nell'accezione *"as open as possible, as closed as neccessary"* del termine. Questo non necessariamente implica che lo sia anche il sistema software alla sua base, ma vincolare il principale asset di proprietà dell'ente (i propri prodotti della ricerca, siano essi pubblicazioni, progetti o addirittura dati) a software proprietario e commerciale pone evidenti rischi, tra i quali possiamo indicare, come i maggiori possibili:

- la sindrome del "single-vendor-lock-in": una volta basato il sistema su un software proprietario si rischia, in caso di controversia col produttore o di una sua uscita dal mercato, di non poter più garantire il futuro accesso ai dati stessi;
- sempre in questo contesto, ogni modifica futura richiesta al software (inevitabile in un mondo in continua evoluzione) richiederebbe un accordo col produttore con prevedibile lievitazione dei costi e possibili problemi circa i tempi di consegna degli aggiornamenti stessi;
- in ogni caso, la gestione del sistema non potrebbe fare a meno di personale INFN con le necessarie competenze e, se questo non avesse possibilità di intervenire sul software, si

potrebbero creare dei significativi colli di bottiglia nella gestione del sistema, in particolare in quella inerente agli Open Data: definire gli opportuni metadati che accompagnano ogni singolo prodotto e fornire un sistema che davvero permetta una fruizione efficace dei dati resi pubblici non può infatti prescindere dalle competenze di un esperto del settore specifico.

Durante la stesura del presente documento sono state acquisite delle informazioni sui costi di alcune soluzioni commerciali esistenti, sia a livello nazionale che internazionale, e dalla loro analisi, la soluzione "in house" è più conveniente. Alcuni dati sono forniti nel resto del documento.

1. Introduction and context

1.1 Introductory concepts and driving considerations

In the last 30 years or so, scientific computing has steadily evolved from centralized to a more distributed environment. This has been due to the concurrent availability of cost-effective COTS components and decrease of costs of LANs. In the first half of 90's, the emergence of cluster computing for HTC applications was confirmed and "farms" of computers with many-core processors, interconnected by low-latency networks, become the norm. This eventually extended to the domain of HPC to the extent that about 80% of the TOP500 machines built in the last ten years are based on a cluster architecture (1).

Furthermore, the steep decrease of costs of high-bandwidth WANs has fostered in the recent years the spread and the uptake of the Grid Computing paradigm and the distributed computing ecosystem has become even more complex with the emergence of Cloud Computing.

At the onset of the 21st century all these developments have led to the new concept of **e-Infrastructure** - defined as *"an environment where research resources (hardware, software and content) can be readily shared and accessed where necessary to promote better and more effective research; such environment integrate hard-, soft- and middle-ware components, networks, data repositories, and all sorts of support enabling virtual research collaborations to flourish globally"* (2).

Indeed, e-Infrastructures have been built over several years both in Europe and the rest of the world, to support diverse multi- and inter-disciplinary VRCs (3). There is a shared vision that e-Infrastructures will allow scientists across the world to do better (and faster) research, irrespective of where they are and of the paradigm(s) adopted to build them.

E-Infrastructure components can be key platforms to support the Scientific Method (4) the "knowledge path" followed in many aspects by scientists since the time of Galileo Galilei. With reference to Figure 1, Distributed Computing and Storage Infrastructures (local HPC/HTC resources, Grids, Clouds, long term data preservation services) are ideal both for the creation of new datasets and the analysis of existing ones while Data Infrastructures (including OADRs and DRs) are essential to evaluate existing data and annotate them with results of the analysis of new data produced by experiments and/or simulations. Finally, Semantic Web-based enrichment of data is key to correlate documents and data, allowing scientists to discover new knowledge in an easy way, and engage in a more robust scholarly discourse.
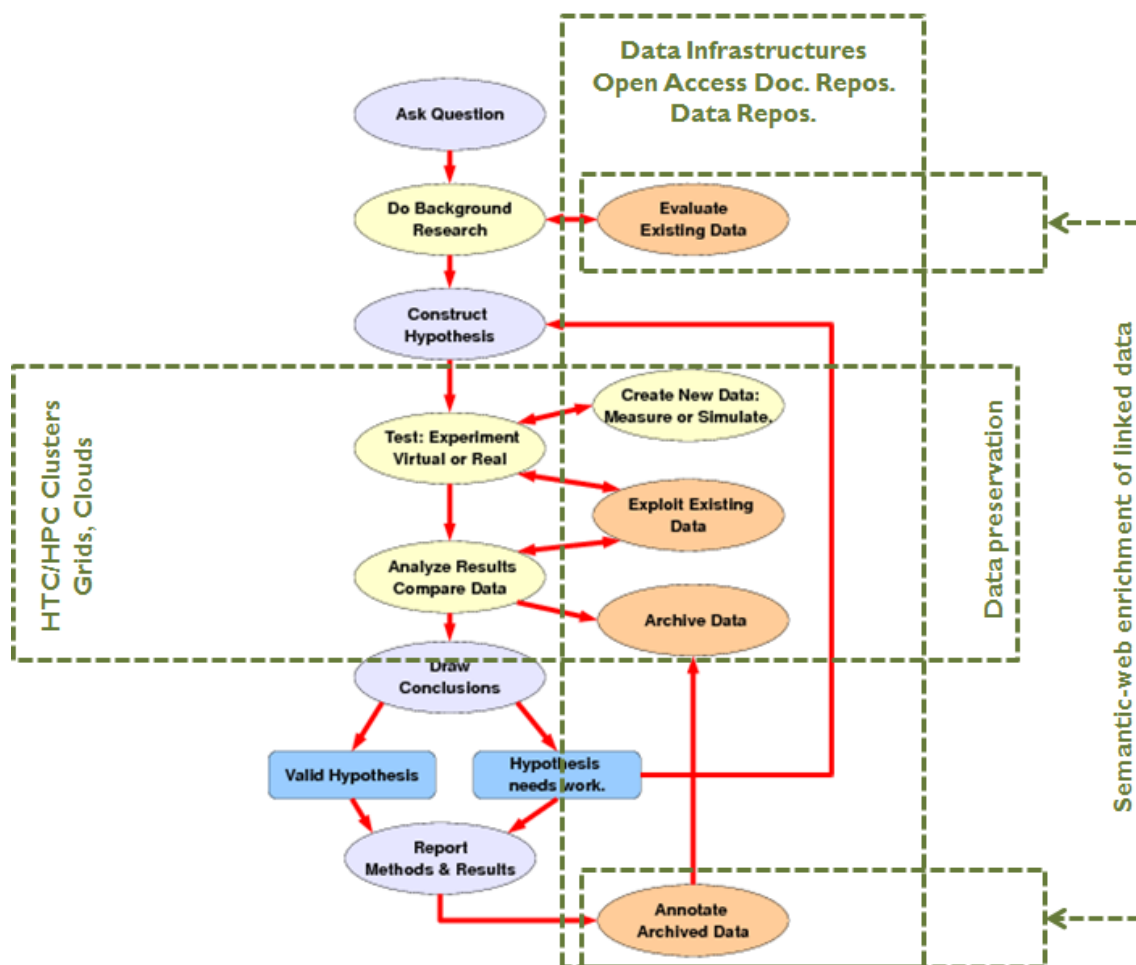
Figure 1 – Pictorial view of the Scientific method (the background figure comes from ref. *(4)*).

One of the cornerstones of the Scientific Method, which is a key driver through the knowledge path, is science reproducibility. In recent years, the issue of the reproducibility of scientific results has attracted increasing attention worldwide, both inside and outside scholarly communities, to which a recent Special Edition of Nature (5) is testament. As striking examples, Begley and Ellis (6) could not reproduce the results of 47 out of 53 "landmark" publications in cancer research and Casadevall et al. (7) have identified more than 2,000 articles listed in Pubmed (8) as retracted since the first identified article was retracted in 1977.

The problem goes well beyond the topic of cancer. In March 2012 a committee of the US National Academy of Sciences heard testimony that the number of scientific papers that had to be retracted increased more than tenfold over the last decade while the number of journal articles published rose only 44% over the same period (9). At the current rate, by 2045 there could be as many papers published as retracted.

Considering these findings, researchers and other scholarly observers have recently been proposing and conducting initiatives to help the scientific community to address the issue of reproducibility. Some of the most interesting ones are gathered under the umbrella of the Reproducibility Initiative< (10) jointly started by the lab-services start-up Science Exchange (11) and the open access journal PLoS ONE (12). Scientists can submit studies to Science

Exchange that they would like to see replicated. An independent scientific advisory board selects studies for replication and service providers are then selected at random to conduct the experiments. The results are returned to the original investigators, who can then publish them in a special issue of the open-access journal PLoS ONE and are awarded with a "certificate of reproducibility" for studies that are successfully replicated.

Although the initiative of Science Exchange is commendable, it is however limited to the health domain, authors must pay to have their results reproduced, and the choice of studies to be reproduced is entirely decided by the advisory board.

Furthermore, some very important considerations are in order.
1. As pointed out by C. Drummond (13), reproducibility and replicability are different concepts and *"replicability is not reproducibility"*.
   The "re-'s" of the Scientific Method go beyond replicability and re-producibility and indeed include both repeatability and re-usability (see Figure 2).
2. In the last 2-3 decades science has become more and more computationally intensive and computer simulations are actually "reconciling" the inductive and deductive approaches of the Scientific Method. Then:
   a. *"An article about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete software development environment, [the complete data] and the complete set of instructions which generated the figures"* (14).
   b. *"Scientific communication relies on evidence that cannot be entirely included in publications, but the rise of computational science has added a new layer of inaccessibility. Although it is now accepted that data should be made available on request, the current regulations regarding the availability of software are inconsistent. We argue that, with some exceptions, anything less than the release of source programs is intolerable for results that depend on computation. The vagaries of hardware, software and natural language will always ensure that exact reproducibility remains uncertain, but withholding code increases the chances that efforts to reproduce results will fail"* (15).
   c. *"The publication and open exchange of knowledge and material form the backbone of scientific progress and reproducibility and are obligatory for publicly funded research. Despite increasing reliance on computing in every domain of scientific endeavor, the computer source code critical to understanding and evaluating computer programs is commonly withheld, effectively rendering these programs "black boxes" in the research work flow"* (16).
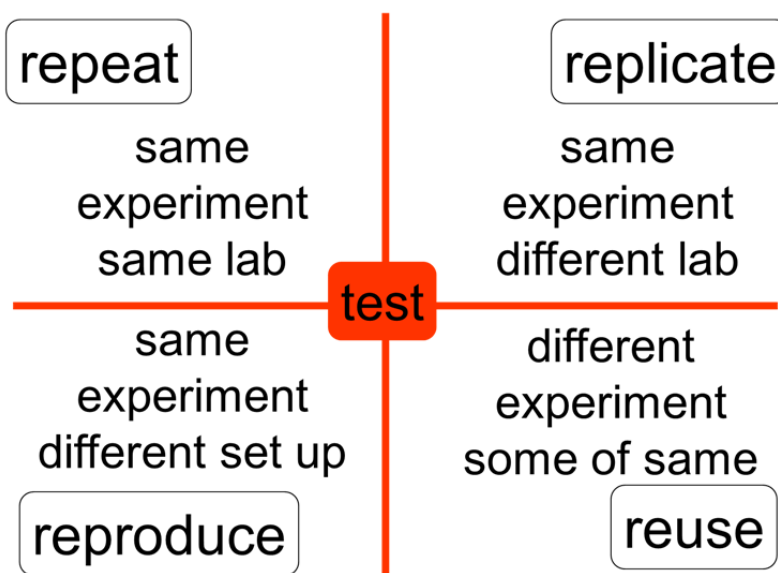
Drummond C Replicability is not Reproducibility: Nor is it Good Science, online
Peng RD, Reproducible Research in Computational Science *Science 2 Dec 2011: 1226-1227.*

Figure 2 - The "re-'s" of the Scientific Method.

For all the above, real science reproducibility should include full access to papers, datasets, data collections, algorithms, configurations, tools and applications, codes, workflows, scripts, libraries, services, system software, infrastructure, compilers, hardware, etc. In order to ensure all that, besides and beyond e-Science, the new concept of Open Science (also referred to as Open Knowledge) is emerging and its key enablers are pictorially depicted in Figure 3.
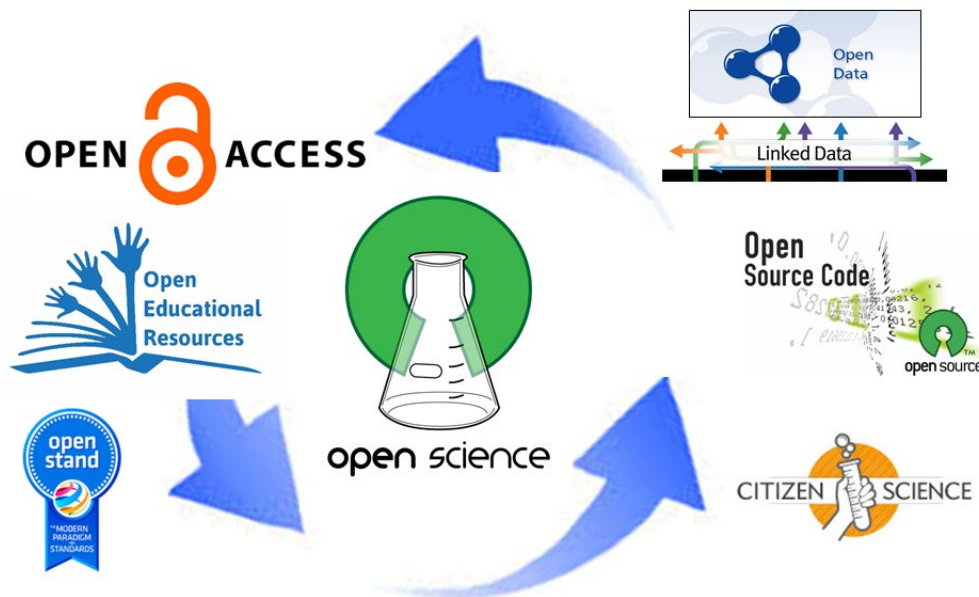

Figure 3 - Key enablers of Open Science.

According to a seminal book (17), Open Science *"refers to a scientific culture that is characterized by its openness. Scientists share results almost immediately and with a very wide audience"*.

Five schools of thought on Open Science have been identified so far (18), characterised by their central assumptions, the involved stakeholder groups, their aims, and the tools and methods used to achieve and promote these aims (see Figure 4). The **infrastructure school** is concerned with the technical infrastructure that enables emerging research practices on the Internet, for the most part software tools and applications, as well as computing networks. The infrastructure school regards Open Science as a technological challenge and focuses on the technological requirements that facilitate research practices, such as Grid and, more recently, Cloud Computing.
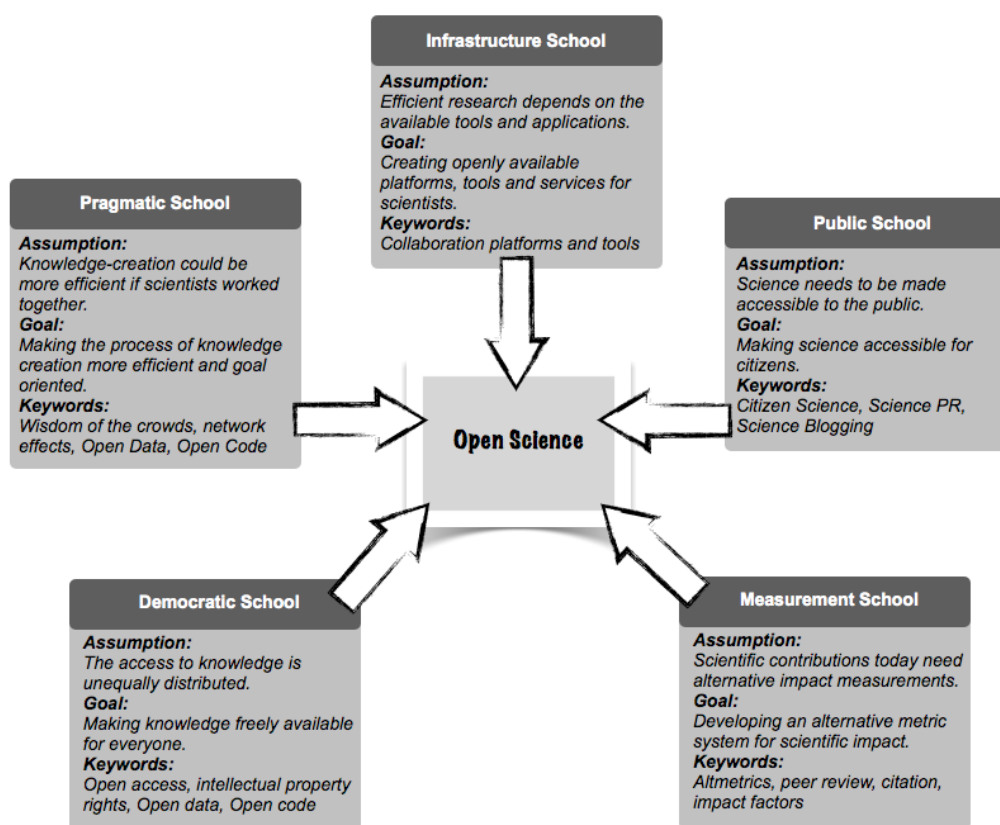

Figure 4 -The five schools of thought of open science (the figure comes from ref. *(18)*).

The INFN very much supports the Open Science "paradigm" and has a strong focus on the application of the guidelines of the infrastructure school. The INFN is committed to setup an infrastructure for re-producible and re-usable science and one its key components is a standard-based Open Access Repository (OAR).

The present document outlines the conceptual design of a sustainable INFN OAR and it is organised as follows. Section 2 lists the objectives of the OAR. Existing technologies for and

examples of OARs are briefly described in Section 3. Section 4 summarises the current status and functionalities of the prototype of the INFN OAR, which is being developed since 2014, and highlights its strengths as well as its weaknesses. The proposed long-term solution for the INFN institutional repository is outlined in Section 5, while the resources needed are discussed in Section 6.

2. Objectives

In our vision, INFN deserves to/should have, as many other important research organisations in the world, an institutional repository that matches the requirements to be:

- Open and fully based on widely accepted standards;
- Able to store all kind of research outputs;
- Able to automatically retrieve and store INFN-authored research outputs contained in other repositories;
- Able to support the management during the periodic national research assessments (VQR);
- Able to add value and fully support INFN Third Mission;
- Able to increase the visibility of both INFN research and researchers (through the integration with ORCID);
- FAIR;
- Compliant with the requirements and guidelines of PLAN S (19);
- Compliant with the EOSC catalogue and interoperable with other EOSC services;
- A key enabler/component of ICDI (20), especially for the so-called "long-tail of science", both inside and outside the organisation.

In the rest of the present document we discuss the state-of-the-art technologies available to reach the above objectives and present the current implementation at INFN as well as the proposed long-term solution, including the human resources and the operational budget needed.

3. Technologies for and Examples of Open Access Repositories

3.1 Technologies

Open Access repositories are powered by Digital Asset Management Systems (DAMS), which are intertwined structures incorporating both software and hardware that take care of management tasks and decisions surrounding the ingestion, annotation, cataloguing, storage, retrieval and distribution of digital assets. Types of digital assets include, but are not exclusive to, photography, logos, illustrations, animations, audio-visual media, presentations, spreadsheets, documents (e.g. Word, PDF, etc.), and a multitude of other digital formats and their respective metadata.

There is a plethora of DAMS's available and some of the most common used in the Open Access domain are listed in Table 1.

DAMS	Home page	License
CKAN	http://ckan.org/	Free
CONTENTdm	http://www.oclc.org/contentdm.en.html	Commercial

Dataverse	https://dataverse.org/	Free
Digibib		
Digital Commons	http://digitalcommons.bepress.com/	Commercial (hosted service)
DigiTool	http://www.exlibrisgroup.com/category/DigiToolOverview	Commercial
DiVA-Portal	http://www.diva-portal.org	Free (hosted service)
dLibra	http://dingo.psnc.pl/dlibra/	Commercial
Drupal	https://www.drupal.org/	Free
DSpace	http://www.dspace.org/	Free
Earmas	http://www.earmas.net/	Free
EPrints	http://www.eprints.org/software/	Free
EQUELLA Repository	http://www.equella.com/	Commercial
ETD-db	http://scholar.lib.vt.edu/ETD-db/index.shtml	Free
Fedora	http://www.fedora-commons.org/	Free
Fez	http://apsr.anu.edu.au/currentprojects/fez06.htm	Free
Figshare	https://figshare.com	Commercial (hosted service)
Greenstone	http://www.greenstone.org/	Free
HAL	https://hal.archives-ouvertes.fr/	Free (hosted service)
Invenio	http://invenio-software.org/	Free
Islandora/Fedora	http://islandora.ca/	Free
intraLibrary	http://www.intrallect.com/solutions/managing_content/	Free
Mendeley	https://www.mendeley.com	Commercial (hosted service)
Mendeley Data	https://data.mendeley.com	Commercial (hosted service)
MyCoRe	http://www.mycore.de/	Free
Open Repository	http://www.openrepository.com/	Commercial (hosted service)
OPUS	http://www.kobv.de/entwicklung/software/opus-4/	Free
PURE	https://www.st-andrews.ac.uk/staff/research/pure/	Free (hosted service)
SciELO	http://scielo.org/php/index.php	Free (hosted service)
VITAL	https://www.iii.com/products/vital	Commercial
WEKO	http://weko.wou.edu.my	Free
XooNIps	http://xoops.org/modules/repository/	Free

Table 1 - List of the most used Open Access Digital Asset Management Systems.

Others, more business- and/or social-oriented, are listed in (21).

3.2   Examples of Open Access Repositories

As of today, a few thousands of either institutional or general purpose official OADRs and DRs exist worldwide. According to the global Directory of Open Access Repositories - OpenDOAR (22), most of them have been deployed and commissioned in the last 15 years or so (see Figure 5), i.e. since the Open Access movement has become a solid reality in the scientific and technological research landscape.
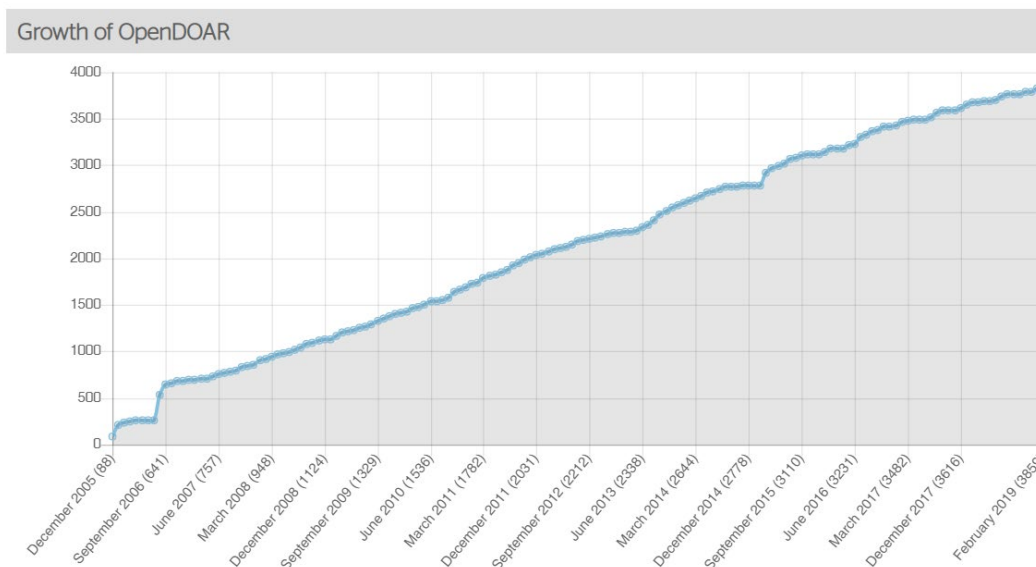
Figure 5 - Temporal growth of the number of repositories in OpenDOAR (source: OpenDOAR website).

The pie-chart of the DAMS technologies used to implement the OARs registered in OpenDOAR is reported in Figure 6.
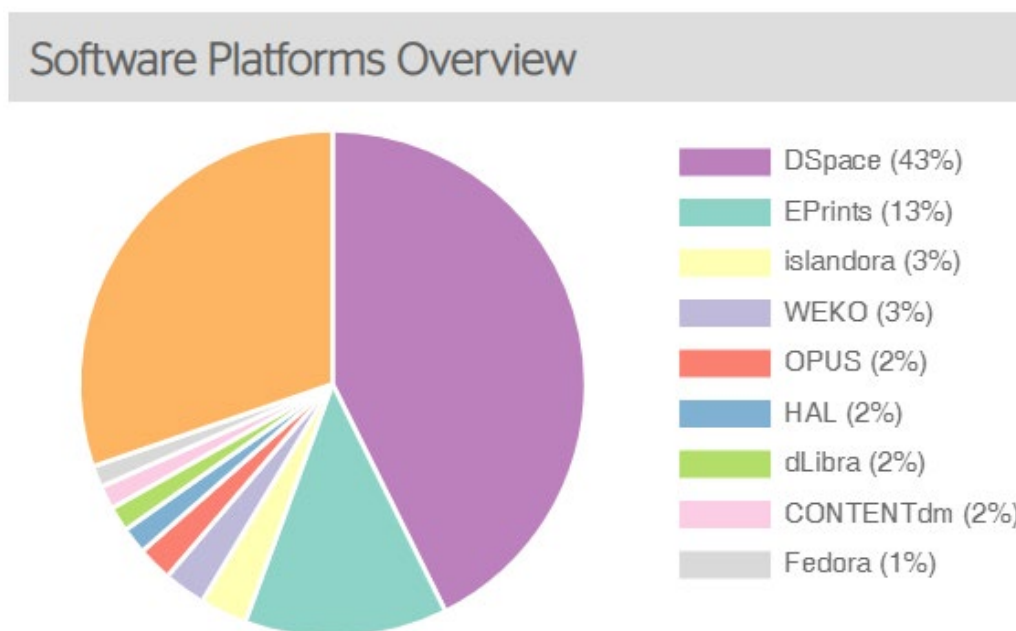

Figure 6 - Technologies used to implement the repositories stored in OpenDOAR (source: OpenDOAR website).

Some notable examples of OADRs and DR are reported in the following list, with their pros and cons:
- arXiv (23)

- o *"Started in 1991, arXiv.org is a highly-automated electronic archive and distribution server for research articles. Covered areas include physics, mathematics, computer science, nonlinear sciences, quantitative biology, quantitative finance, statistics, electrical engineering and systems science, and economics. arXiv is maintained and operated by Cornell University with guidance from the arXiv Scientific Advisory Board and the arXiv Member Advisory Board, and with the help of numerous subject moderators."* The main limitation is the fact that only papers can be uploaded on the repository.
- CERN Open Data Portal (24)
  - o *"The CERN Open Data portal is the access point to a growing range of data produced through the research performed at CERN. It disseminates the preserved output from various research activities, including accompanying software and documentation which is needed to understand and analyse the data being shared."* The portal is currently focused basically on LHC data. The data of only one non-LHC experiment are available in the repository.
- figshare (25)
  - o *"figshare is a repository where users can make all their research outputs available in a citable, shareable and discoverable manner."* The main limitations are the fact that figshare is a private company and the 5 GB limit for the uploads done by single users, although they claim that single-user storage space is unlimited. At the time of writing, figshare includes 18 records where at least one co-author belongs to the INFN.

    Figshare also provides services for institutions and their pricing model is based upon the "research intensity" of an organisation as defined by the number of publications it has published in the last 5 years, according to data from Dimensions (26). Figures reported in Dimensions for INFN place it in the middle of figshare's five pricing tiers, corresponding to an annual subscription rate of 36,200 Euros for 100 TB of AWS storage. Extra fees are charged in the case migrations of records from other existing repositories (e.g., arXiv) should be needed.
- Mendeley (27) and Mendeley Data (28)
  - o *"Mendeley is a free reference manager and academic social network* [Ed, provided by Elsevier] *that can help you organize your research, collaborate with others online, and discover the latest research"*. Mendeley offers all the functionalities of a repository for publications and it is free for single users. Mendeley included Mendeley Institutional Edition (MIE), which offers premium packages and paid plans for institutions, but this option is not anymore available to new organisations.

    Besides Mendeley, Elsevier also provides Mendeley Data, which is *"a modular* [Ed. FAIR-compliant]*, cloud-based platform helping research institutions to manage the entire lifecycle of research data"*. It is free for single users only.
- INSPIRE (29)
  - o *"INSPIRE is an open access digital library for the field of high energy physics (HEP). It is the successor of the Stanford Physics Information Retrieval System (SPIRES) database, the main literature database for high energy physics since the 1970s."* As arXiv, only papers can be uploaded on the repository and it is "single-domain" driven and oriented.
- SCOAP^3 (30)

- - *"SCOAP3 is a one-of-its-kind partnership of over three thousand libraries, key funding agencies and research centers in 44 countries and 3 intergovernmental organisations. Working with leading publishers, SCOAP3 has converted key journals in the field of High-Energy Physics to Open Access at no cost for authors."* SCOAP3 is only devoted to already published papers and does not act a repository of preprint or other research outputs.
- Zenodo (31)
  - *"Zenodo is a general-purpose open-access repository developed under the European OpenAIRE program* [(32)] *and operated by CERN. It allows researchers to deposit data sets, research software, reports, and any other research related digital artefacts. For each submission, a persistent Digital object identifier (DOI) is minted, which makes the stored items easily citable. […] Zenodo is supported by CERN 'as a marginal activity', and hosted on the high-performance computing infrastructure that is primarily operated for the needs of high-energy physics"* (33)*.* Nevertheless, Zenodo is often mentioned as a reference example of FAIR-compliant digital repository and considered a flagship service of EOSC.

    At the time of writing, Zenodo includes 147 records where at least one co-author belongs to the INFN. These records include 75 publications, 27 software, 19 presentations, 10 datasets, 9 posters, 5 videos and 2 lessons.

At national level, slightly more than 130 digital repositories (mostly institutional) are currently registered in OpenDOAR (34). Some relevant examples are listed below:
- CNR Central Library (digital) (35)
  - It is the digital counterpart of the CNR Central Library established in 1927. It includes the multimedia "grey literature" of the Organisation but nothing is said on the FAIRness of the contents. More information is provided here (36).
- CNR Science & Technology Digital Library (37)
  - It was the main deliverable of a CNR-internal project funded by the Italian Ministry of Education, University and Research over the period 2014-2016. *"The objective of the project was to build a Digital Library making science and technology available to everyone, promoting its most widespread use"* (38). However, the latest news on the repository website date back since 2016 and the status of the service (both current and future) is not clear.
- "Polvere di Stelle" ["Stardust" in English] (39)
  - It the official entry point to the cultural heritage repositories and libraries of the Italian National Institute of Astrophysics.
- INAF Institutional repository (40)
  - It is the institutional repository of the Italian National Institute of Astrophysics based on well-defined policies (41), (42). It is managed by a team of about 10 people, 3 of which are librarians. At the time of writing the present document, the service is not open to users yet.
- IRIS (43)
  - A special mention is deserved by the Institutional Research Information System (IRIS), which is a CRIS developed by CINECA. IRIS is an IT platform that makes it easy to collect and manage data on research activities and outputs within an organization. Researchers, administrators and evaluators are given all the tools needed to monitor research results, enhance visibility and efficiently allocate available resources. IRIS is currently available for all Italian universities (44) and

some of them use it also as an Open Access repository for publications (not for data, though). Informal contacts with some IRIS adopters place the cost of a full IRIS license in the order of 100,000 Euros per year. As IRIS is developed following the requirements of all universities, release schedule seems not to be fast and new functionalities required by some adopters may require time to be added to the main code base.

4. Current Status

4.1   Approach and generalities

The INFN operates a prototype of an institutional OAR at the Division of Catania since 2014 (the current address is http://legacy.openaccessrepository.it). The requirements for a DAMS to be used within/for the organisation were the following:
- Open source;
- Distributed under an open license;
- Deployable on a local infrastructure (i.e., not a hosted service);
- Standard compliant;
- Well supported;
- Scalable, up to $O(10^6)$ - $O(10^7)$ resources (to begin with).

There are many comparisons of different DAMS's available on the web and we went through several of them (45), (46), (47), (48). Another element for the choice was to have the most direct possible know-how of the DAMS to be adopted, in order to have something solid but, at the same time, deployable in a very short amount of time. In this respect, INFN Catania has indeed a deep knowledge of and a long experience with Invenio (49) since 2005.

For this reason, the choice was Invenio (**version 1**, referred to in the following as Invenio 1) and the additional motivations for it were the following:
1. It is fully compliant with all most important library standards, such as, for example: DCMI (50), Marc21 (51) and OAI-PMH (52);
2. It is co-developed by a large international collaboration comprising institutes such as CERN, DESY, EPFL, FNAL, SLAC and used as institutional DAMS by tens of scientific institutions worldwide (53);
3. INSPIRE, SCOAP[3] and Zenodo repositories are based on Invenio;
4. The CERN Document Server (54) operates and manages since 2002 more than 0.6 million records in high-energy physics, covering articles, books, journals, photos, videos, and more. Furthermore, Zenodo currently hosts about one million records belonging to various disciplines.

The INFN OAR is the key enabler of a vision where a DAMS can allow researchers to deposit their science products, including FAIR data and open source software, and, at the same time, be part of an Open Science platform ensuring full reproducibility and re-usability.

The INFN OAR allows single researchers to upload their products but also connects to external archives to harvest products having at least one author belonging to INFN author. External archives can contain papers, such as arXiv, SCOAP[3] and OpenAIRE, data, such as EUDAT (55), re3data.org (56) and Zenodo, and software, such as GitHub (57).

Furthermore, INFN OAR allows each resource to be citable and discoverable, through DataCite (58) DOIs, and reproducible/re-usable, thanks to the connection to Science Gateways and to Grid and Cloud infrastructures as well as to local HPC facilities.

It is worth noting that INFN OAR allows federated login and it has been registered as a Service Provider of several Identity Federations, including the official Italian federation IDEM (59), the GrIDP "catch-all" one (60), co-managed by the INFN and the University of Catania, and the eduGAIN inter-federation (61), coordinated and managed by GEANT (62).

Once logged in, users can add their ID's and INFN OAR supports several credential providers, such as Google Scholar (63), GitHub, ORCID (64), ResearcherID (65), ResearchGate (66), and SCOPUS (67). In the case of ORCID, which is becoming a "de facto" standard for author disambiguation, when available ORCIDs can be shown as links next to the authors' names in the resources stored on INFN OAR.

Finally, records stored on the INFN OAR are linked to Altmetric (68). In case they have web, wiki and/or social citations the well know "donut" appears in the detailed view page of the record with a link to the Altmetric Explorer.

4.2   Types of resources

INFN OAR can store several types of digital assets: audio-video recordings, datasets, images, presentations, posters, publications and software. The software collection includes the possibility to store entire virtual machines containing all the code to reproduce and extend the analysis performed in a given paper with a given dataset. **Through the combination of DOIs and ORCIDs, all research outputs stored on the INFN OAR can be claimed from within and linked to the ORCID profiles of their author(s)**.

4.3   Certification and compliance

As already mentioned above, Invenio is fully standard-based and a lot of effort has been devoted to make INFN OAR officially certified and compliant with the most relevant and widely known Open Access initiatives.

As shown in Figure 6 and Figure 7, INFN OAR is an OAI-conforming repository and it is an official OpenDOAR data provider.

Figure 7 - INFN OAR conforming with OAI specifications.

Figure 8 - INFN OAR as an OpenDOAR data provider.

It is worth underlying that INFN OAR is also one of the officially certified OpenAIRE archives, compliant with version 3.0 of its guidelines (see Figure 8).

Figure 9 - INFN OAR as an OpenAIRE official archive.

When a resource is uploaded to/harvested in the INFN OAR, the project under which it was produced can be specified. INFN OAR includes the databases of FP4, FP5, FP6, FP7 and H2020 project as well as those of the National Operating Program of the Italian Ministry of Education, University and Research and of the INFN internal projects. **This makes the INFN OAR easily customisable as a CRIS**.

4.4   Knowledge workflow and knowledge "nexi"

In the previous sub-sections, we have shown how, thanks to the adoption of important bibliographic standards, INFN OAR is both an Open Access Initiative conforming archive and an official OpenDOAR data provider, able to automatically harvest resources from different sources, including SCOAP[3], using RESTful API. It is also one of the official OpenAIRE archives, compliant with version 3.0 of its guidelines.

INFN OAR allows SAML-based federated authentication and it is one of the Service Providers of the eduGAIN inter-federation. Furthermore, it is also connected to DataCite for the issuance and registration of Digital Object Identifiers (DOIs).

But what makes INFN OAR different from other Open Access repositories is its capability to connect to Science Gateways and exploit Distributed Computing and Storage Infrastructures worldwide, including WLCG (69), EUDAT and EOSC (70) ones, to easily reproduce and re-use/extend scientific analyses.

Indeed, INFN OAR is a key enabler of the "share, validate, preserve, reproduce" workflow depicted in Figure 10 and allows to "walk through the knowledge path in a circular way".

Starting from the upper-left panel of the figure and going clockwise, either a researcher or a citizen scientist can:

1. Search and discover a research product, e.g. a scientific publication or an analysis object;
2. Be re-directed to an Open Access Repository where that product is stored as one of its resources;
3. Be-redirected to a Science Gateway, such as those implemented with the FutureGateway framework developed by INFN Catania, where that analysis can be reproduced or even extended;
4. Write a new paper about the new extended analysis;
5. Upload the new paper on the Open Access Repository, assign a DOI to it, and "connect" the new paper to the old one, to the needed dataset(s) and to virtual appliance containing the software to read and analyse it(them).

Upon completion of the virtuous cycle, new knowledge has been added to the existing one and both the new and the existing one are citable, searchable and discoverable.

Figure 10 - The knowledge workflow implemented by INFN OAR.

It is also important to underline that, using Semantic Web technologies, developed at the INFN Division of Catania in the context of past EU funded projects, the INFN OAR can be included in LOD federations and the "knowledge" contained in the records stored in the repository can be "browsed" and "navigated" and new "knowledge nexi" can be discovered (see Figure 11).
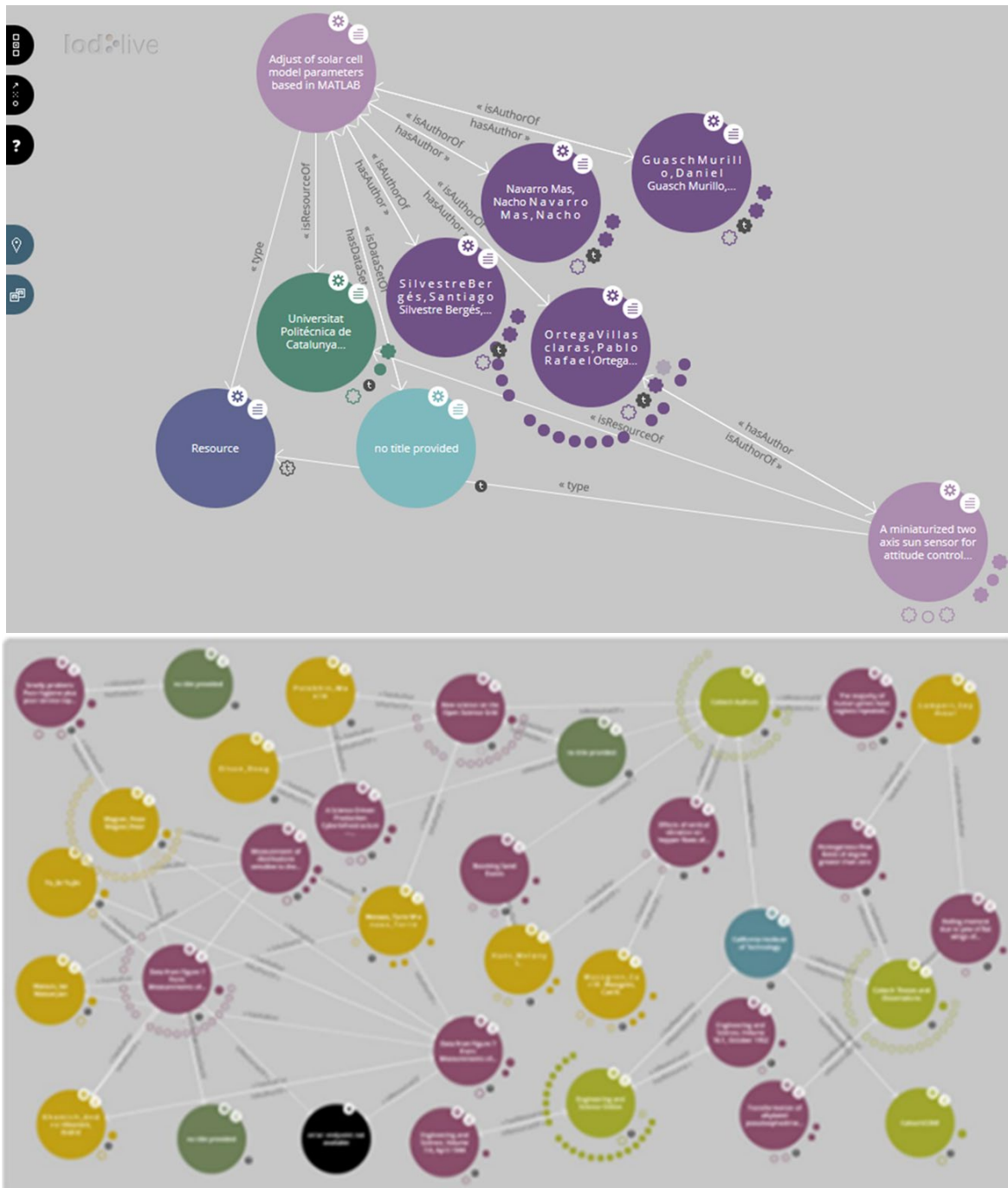

Figure 11 - Various-complexity knowledge nexi.

5. Proposed solution

Since the deployment of the INFN OAR based on Invenio 1, which took place in mid-2014, several important things have happened:

- Invenio 1 has reached its end of life and will not receive any further updates.
- Invenio 2, which was a transitional release never made widely public, combining the old Invenio 1 legacy code base with the new technology stack, paving the way towards the Invenio v3.0 release, reached end-of-life and will not receive any further updates.
- Zenodo was officially launched;
- Invenio v3.0 (referred to in the following as Invenio 3) was officially released on June 2018 (71).

Invenio 3 contains a long list of improvements with respect to Invenio 1 (as well as to Invenio 2) and its most important features are summarised here (72). Furthermore, Invenio 3 has an implementation roadmap (73) that really makes it an outstanding *"Open Source framework for __large-scale__ digital repositories"*, as advertised in its home page.

The great flexibility and customisability of Invenio 3 is one of its most important strengths but, at the same time, it is one of the hurdles in its wide adoption. Unlike Invenio 1, indeed Invenio 3 does not ship with a GUI for users and administrators, which has instead to be created on purpose using its very detailed API. Invenio 3 developers say that *"Invenio is like a Swiss Army knife of battle-tested, safe and secure modules providing you will all the features you need to run a trusted digital repository"* (74) and they actually delegate the development of the GUI to the managers and operators of the specific instance.

The developers of Zenodo have recently released their code on GitHub (75), including the GUI. So, the INFN Open Access Group has decided to:

- Update the back-end of the INFN OAR from Invenio 1 to Invenio 3;
- Customise the Zenodo GUI and adapt it as the front-end of the new version of the INFN OAR.

The clear advantage of the above decision is manifold:

- Improve the INFN OAR with leading-edge technologies in the domain of digital repositories;
- Include the possibility for digital assets, offered by the new Zenodo GUI, to be stored in the repository not only open access but also embargoed, restricted and closed;
- Exploit the concept of "communities", which is central in the Zenodo architecture, to cope with several aggregation of contents: per INFN division, per Scientific Committee, per project, per initiative, etc.;
- Create in-house competences in the operation and administration of large-scale FAIR-compliant archives;
- Hence, avoid any expenses for the outsourcing of the creation of a GUI for the INFN institutional repository.

The activities outlined above have already been undertaken and the new version of the INFN OAR is available at the URL www.openaccessrepository.it. **That is the proposed solution for a sustainable institutional repository for INFN**.

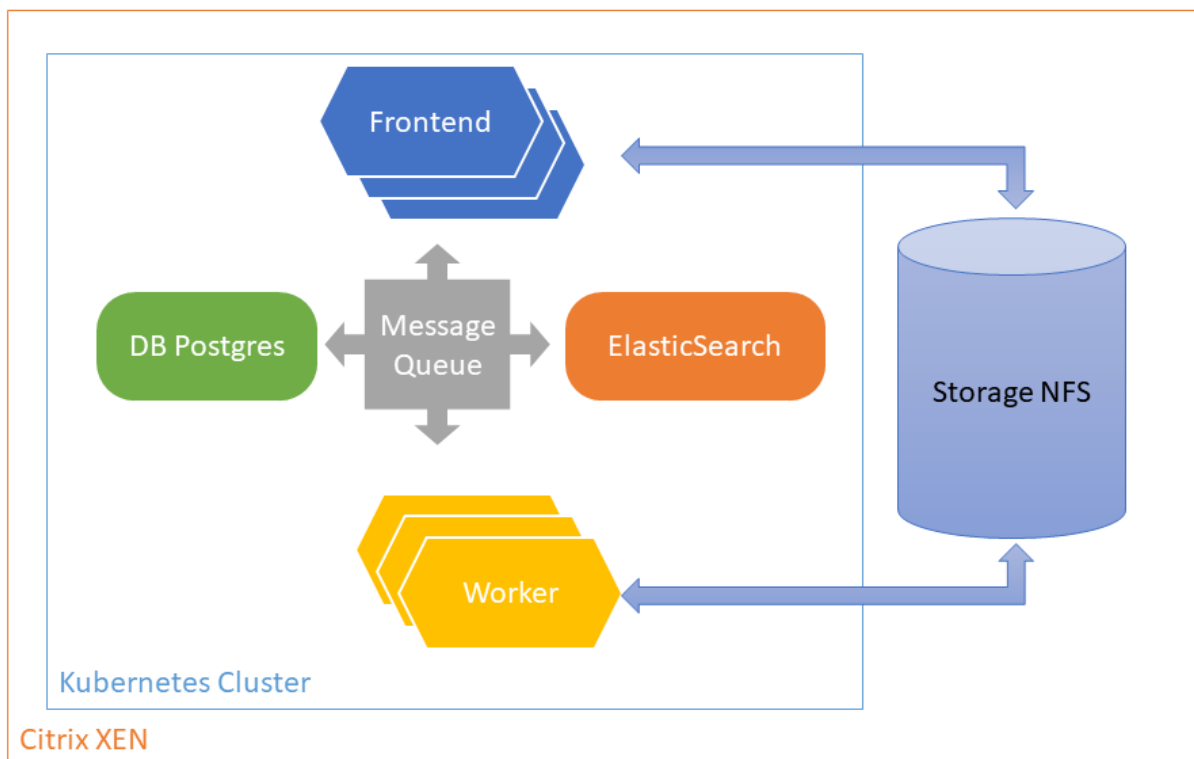The deployment layout of the new INFN OAR is sketched in Figure 12.

Figure 12 - Deployment layout of the new INFN OAR.

The new deployment of the OAR takes advantage of the cutting-edge technologies on service deployment and management. Unlike the previous version, Invenio 3 has a modular architecture, which better complies with current micro service architecture philosophy aiming at increasing scalability, availability and reliability of services. As a result, the new INFN OAR is deployed as a micro service swarm. The basic infrastructure is made of a virtual cluster, managed by Citrix XenServer (76), composed of 4 nodes: 1 cluster manager, 2 workers and 1 NFS share manager. Citrix allows migration and periodic snapshots of virtual machines, so hardware failures can be easily recovered with no or very limited impact on running services. Citrix stores snapshots for several days so it is always possible to restart from a certain working state.

Virtual machines are nodes of a Kubernetes cluster (77), a container orchestrator aimed at simplifying deployment and management of containerized applications. In this context, a container can be associated to a single microservice. Kubernetes manages the microservice composing Zenodo application and it is responsible for the migration between the workers. Furthermore, it is also responsible for scaling stateless micro services up by increasing the number of instances and balancing the traffic among them.

Looking at the figure above, the Zenodo application consists of 5 micro services: frontend, worker, database (DB), ElasticSearch and message passing. In the current deployment, all services are stateless apart from DB and ElasticSearch, which are instead stateful. Since those services store their state in a NFS share, they can easily migrate whereas scalability is not

managed by the orchestrator in order to avoid inconsistences at service level. However, both EleasticSearch and Postgres DB support horizontal and vertical scalability so it is possible to configure these services in multiple containers whenever needed. Kubernetes deploys stateless services in the available workers balancing the traffic and, in case of overloading, other workers can seamlessly be added to run additional containers.

Additionally, NFS provides a data storage for all the documents, media and data uploaded to INFN OAR. Continuous backup of the storage prevents data losses.

Currently, all micro services are running as single instances. We have successfully executed several reliability tests to verify no data are missed in case of failures and to prove the ability of Kubernetes to migrate containers and re-configure the network whenever needed. We will soon perform availability tests as well to identify the correct size of microservice pools and workers and how these have to dynamically scale to cope with the expected number of users.

The Zenodo software that powers the INFN OAR has been modified with respect to the official release and the changes are made available in a dedicated software repository on GitHub, which is referenced in the INFN OAR itself (78).

The main differences are:

Authentication: the INFN OAR supports authentication with SAML so users can access using identity providers, such as INFN AAI, and the creation of local account has been suppressed. The new INFN OAR is already a service provider of both the Italian federation IDEM and the inter-federation eduGAIN.

- Theming: the official Zenodo software has many aspects of the theme hard coded so, wherever possible, we made these configurable with specific variables for an easier customisation in the future.
- Bug fixing: the migration from the old version of the OAR to the new one has made evident some bugs, which have already been communicated to the Zenodo team. Not of them have been incorporated in the official code though, mainly because they have been deemed not relevant for the CERN instance. All fixes have of course been included and documented in the INFN OAR code base.

Besides the activity on the customisation and generalisation of the official Zenodo code, a big effort has been spent to create deployment files and containers. A Zenodo container with the modified code and running with Apache httpd and Shibboleth for the authentication has been created and made available on Docker Hub (79), together with the file to generate the container. Every commit in the docker files and other associated code triggers the building of a new container. In the future, this update will be integrated with our cluster to have a full continuous delivery/integration system compliant with the DevOps (80) paradigm.

Finally, the Kubernetes deployment files for all microservices in the cluster have been created and made available on GitHub to easily replicate the installation.

Before closing the section, it is worth underlining that, while the deployment and commissioning of the new version of the INFN-OAR is being finalised, a preliminary study of its compliance with the guidelines of PLAN S has been carried out. Results are summarised in Table 2 and relate with the version of the guidelines released by cOAlition S on the 24th of April 2019 (yet to be officially published).

Basic mandatory conditions for all publication venues (version of 24.04.2019)	arXiv	INFN-OAR	Notes
All scholarly publications on the results from research funded by public or private grants provided by national, regional and international research councils and funding bodies, must be published in Open Access Journals or on Open Access Platforms or made immediately available through Open Access Repositories without embargo.	YES	YES	
The journal/platform must provide on its website a detailed description of the editorial policies and decision-making processes. In addition, at least basic statistics must be published annually, covering in particular the number of submissions, the number of reviews requested, the number of reviews received, the approval rate and the average time between submission and publication.	Not mentioned in the website	YES	
The journal/platform must offer authors/institutions the option of copyright retention at no extra cost. Licenses to publish must not restrict the right of the author/institution to make the VoR or the AAM of the article open access immediately upon publication, under an open licence.	YES	YES	
The journal/platform must either enable authors to publish with immediate and permanent open access (without any kind of technical or other form of obstacles) under an open licence, or to deposit the AAM (Author's Accepted Manuscript) or VoR (Version of Record) in an Open Access repository at no extra cost. In either case, no embargo period can be applied (including for early view versions, i.e. online VoR before inclusion in an issue).	YES	YES	
The self-archiving policy of the venue must be registered in SHERPA/RoMEO.	NO	YES	INFN-OAR to be configured
Mandatory technical conditions for all publication venues			
Use of persistent identifiers (PIDs) for scholarly publications (with versioning, for example in case of revisions), such as DOI (preferable), URN, or Handle.	YES: only handles (for papers only)	YES: DataCite DOIs	

Deposition of content with a long-term digital preservation or archiving programme (such as CLOCKSS, Portico or equivalent).	Not mentioned in the website	YES	INFN-OAR to be configured
High-quality article level metadata in standard interoperable non-proprietary format, under a CC0 public domain dedication. Metadata must include complete and reliable information on funding provided by cOAlition S funders (including as a minimum the name of the funder and the grant number/identifier).	NO (no grant information can be provided at submission time)	YES	IMPORTANT: INFN-OAR can become a Current Research Information System (CRIS)
Machine-readable information on the Open Access status and the license embedded in the article, in standard non-proprietary format.	YES (only for papers)	YES (for all research outputs)	
Strongly recommended additional criteria for all publication venues			
Support for PIDs for authors (e.g. ORCID), funders, funding programmes and grants, institutions, and other relevant entities.	YES (only for ORCID)	YES	IMPORTANT: INFN-OAR can become a Current Research Information System (CRIS)
Availability for download of full text for all publications (including supplementary text and data) in XML and conforming to an open community standard, such as JATS.	YES (only the full text)	YES	
Direct deposition of publications (in XML, conforming to an open community standard, such as JATS, and including complete metadata as described above) by the publisher into author designated or centralised Open Access repositories that fulfil the Plan S criteria.	NO	YES	
OpenAIRE compliance of the metadata.	Not mentioned in the website	YES	
Linking to data, code, and other research outputs that underlie the publication and are available in external repositories.	NO	YES	
Openly accessible data on citations according to the standards by the Initiative for Open Citations I4OC.	YES	YES	
Mandatory criteria for repositories			
Use of persistent identifiers (PIDs) for the deposited versions of the publications (with versioning, for example in case of revisions), such as DOI (preferable), URN,	YES: only handles (for papers only) and no	YES	

Criteria			
or Handle.	versioning available		
High quality article level metadata in standard interoperable non-proprietary format, under a CC0 public domain dedication. This must include information on the DOI (or other PID) both of the original publication and the deposited version, on the version deposited (AAM/VoR), and on the open access status and the licence of the deposited version. Metadata must include complete and reliable information on funding provided by cOAlition S funders (including as a minimum the name of the funder and the grant number/identifier).	NO	YES	
Machine readable information on the Open Access status and the license embedded in the article, in standard non-proprietary format.	YES	YES	
Continuous availability (uptime at least 99.7%, not taking into account scheduled downtime for maintenance or upgrades)	Not mentioned in the website	YES	INFN-OAR to be configured
Helpdesk: as a minimum an email address (functional mailbox) has to be provided; a response time of no more than one business day must be ensured	YES	YES	
Strongly recommended additional criteria for repositories			
Manuscript submission system that supports both individual author uploads and bulk uploads of manuscripts (AAM or VoR) by publishers.	YES (subject to limitations to avoid denials of service)	YES	
Full text stored in XML, conforming to an open community standard, such as JATS.	NO	YES	
Support for PIDs for authors (e.g. ORCID), funders, funding programmes and grants, institutions, and other relevant entities.	YES: only handles (for papers only) and no versioning available	YES	

Table 2 - Compliance of INFN OAR with PLAN S guidelines.

6. Person-power and operational budget needs

At present, INFN OAR is deployed on the cloud infrastructure of the Division of Catania and this configuration is also envisaged for the next foreseeable future. Massive scale-up in the number of records stored in the repository should not pose any relevant problems at least for the next 4-5 years due to the expected forthcoming approval of the IBiSCo project, which has been submitted in response to a call of the National Operating Program 2014-2020 for the strengthening of research infrastructures in Italy.

From the infrastructure point, the INFN OAR is sustainable. From the person-power perspective, instead, besides the members of the INFN Open Access Group, that could be re-organised into a Management Board, an Editorial Board and a Help Desk, the profiles listed in Table 3 are needed in addition.

Short-term: 2019-2020 – Commissioning phase	
FTE	**Profile**
1	Technologist (computer scientist/software engineer) with experience in digital repositories' operation and management.
1	Technologist (computer scientist/software engineer) to integrate OAR with the other services of INFN Information System.
1	Digital librarian.
Long-term: 2021 onwards – Normal operation phase	
FTE	**Profile**
1	Technologist (computer scientists/software engineers) with experience in digital repositories' operation and management as well as in data science.
1	Digital librarian.

Table 3 - Person-power and profiles needed for the operation of the INFN OAR.

Besides the human resources listed above, a budget is also deemed necessary to cover mobility (contacts, participation to technical workshops, outreach and dissemination events, training events, etc.) and consumables' costs.

References

1. **TOP500.** [Online] 2018. http://top500.org/statistics/overtime/ (Select Category = Architecture, choose Type = Systems Share, and then click on Submit to generate the graph).

2. **C(2012) 4890 final.** *COMMISSION RECOMMENDATION of of 17.7.2012 on access to and preservation of scientific information.* Brussels : European Commission, 2012.

3. **Andronico G., et al.** E-Infrastructures for International Cooperation. *Computational and Data Grids: Principles, Applications and Design (N. Preve Ed.).* s.l. : IGI Global - DOI: 10.4018/978-1-61350-113-9, see also www.igi-global.com/book/computational-data-grids/51946, 2011.

4. **B., Lawrence.** Scientific Method. [Online] http://home.badc.rl.ac.uk/lawrence/blog/2009/04/16/scientific_method.

5. **Nature.** Challenges of irreproducible research. [Online] http://www.nature.com/nature/focus/reproducibility/.

6. *Drug development: Raise standards for preclinical cancer research.* **Begley G. C., Ellis L. M.** Nature 483 (p. 531-533), 2012.

7. *Misconduct accounts for the majority of retracted scientific publications.* **Fanga F. C., Steenc G. R., Casadevall A.** 42 (p. 17028-17033), s.l. : Proceedings of the National Academy of Sciences of the United States of America, Vol. 109.

8. **PubMed.** [Online] http://www.pubmed.org.

9. **Reuters.** [Online] http://www.reuters.com/article/2012/03/28/us-science-cancer-idUSBRE82R12P20120328.

10. **Science Exchange.** Validating key experimental results via independent replication. [Online] http://validation.scienceexchange.com.

11. —. Change Science (TM). [Online] https://www.scienceexchange.com.

12. **PLoS ONE.** [Online] https://journals.plos.org/plosone/.

13. *Replicability is not reproducibility: nor is it good science.* **C., Drummond.** Montreal, Quebec, Canada : Proc. Eval. Methods Mach. Learn. Workshop 26th ICML, http://goo.gl/7f8WX9., 2009.

14. *WaveLab and Reproducible Research.* **Buckheit J. B., Donoho D. L.** s.l. : Lecture Notes in Statistics, 1995, Vol. 103 (p. 55-81).

15. *The case for open computer programs.* **Ince D. C., Hatton L., Graham-Cumming J.** Nature 482 (p. 485-488) doi:10.1038/nature10836, 2012.

16. *Shining Light into Black Boxes.* **Morin A, et al.** Science 336 (pp. 159-160), doi:10.1126/science.1218263, 2012.

17. **Many Authors.** *Opening Science - The Book.* http://book.openingscience.org, DOI: 10.1007/978-3-319-00026-8 : Bartling S., Friesike S. (Eds.), 2014.

18. **Fecher. B., Friesike., S.** Open Science: One Term, Five Schools of Thought. *Opening Science – The Book.* DOI: 10.1007/978-3-319-00026-8. : Bartling S., Friesike S. (Eds.), 2014.

19. **cOAlition S.** About Plan S. [Online] https://www.coalition-s.org/.

20. **ICDI.** Home page. [Online] https://www.icdi.it/.

21. **Capterra.** [Online] https://www.capterra.com/digital-asset-management-software/.

22. **JISC.** OpenDOAR. [Online] http://v2.sherpa.ac.uk/opendoar/.

23. **arXiv.** arXiv. [Online] http://www.arxiv.org/.

24. **CERN.** CERN Open Data Portal. [Online] http://opendata.cern.ch/.

25. **figshare.** figshare. [Online] https://figshare.com/.

26. **Science, Digital.** Dimension home page. [Online] https://www.dimensions.ai/.

27. **ELSEVIER.** Mendeley home page. [Online] https://www.mendeley.com.

28. —. Mendeley Data home page. [Online] https://data.mendeley.com/.

29. **INSPIRE.** INSPIRE. [Online] https://inspirehep.net.

30. **SCOAP3.** SCOAP3. [Online] https://scoap3.org.

31. **Zenodo.** Zenodo. [Online] https://zenodo.org/.

32. **OpenAIRE.** OpenAIRE. [Online] https://www.openaire.eu/.

33. **Wikipedia.** Zenodo. [Online] https://en.wikipedia.org/wiki/Zenodo.

34. **OpenDOAR.** Browse by Country and Region - Italy. [Online]
http://v2.sherpa.ac.uk/view/repository_by_country/it.html.

35. **CNR.** CNR Central Library. [Online] https://bice.cnr.it/.

36. —. Central Library brochure. [Online] https://bice.cnr.it/images/Brochure_inglese_copia.pdf.

37. —. Science & Technology Digital Library. [Online] http://stdl.cnr.it/it/.

38. —. Science & Technology Digital Library - What it is. [Online] http://stdl.cnr.it/en/the-
project/science-technology-digital-library/what-it-is.

39. **INAF.** Polvere di stelle. [Online] http://www.beniculturali.inaf.it/.

40. —. Open Access in INAF. [Online] https://openaccess-info.inaf.it/.

41. —. La Policy Open Access INAF. [Online] https://openaccess-info.inaf.it/policy-inaf.

42. —. Le politiche del repository. [Online] https://openaccess-info.inaf.it/policy-repository.

43. **CINECA.** IRIS - INSTITUTIONAL RESEARCH INFORMATION SYSTEM. [Online]
https://www.cineca.it/en/content/iris-institutional-research-information-system.

44. **U-GOV.** List of IRIS Installations (updated 18.04.2018). [Online] https://wiki.u-
gov.it/confluence/pages/releaseview.action?pageId=67639048.

45. **Repositories Support Project.** Repository software survey, November 2010. [Online]
2010. http://www.rsp.ac.uk/start/software-survey/results-2010/.

46. **Docslide.** DIGITAL LIBRARY REPOSITORY: INVENIO VS DSPACE. [Online] 2014.
https://docslide.us/download/link/digital-library-repository-invenio-vs-dspace.

47. **National Technical Library.** Comparison of Selected Software Systems for Creation of
Digital. [Online] http://wiki.lib.sun.ac.za/images/9/91/Nrgl-ir-survey.pdf.

48. *Digital libraries: Comparison of 10 software.* **Andro M., Asselin E., Maisonneuve M.**
https://www.tandfonline.com/doi/abs/10.1080/14649055.2012.10766332?journalCode=ulca20&,
s.l. : Library Collections, Acquisitions & Technical Services, 2013, Vol. 36.

49. **Invenio.** Invenio. [Online] https://invenio-software.org/.

50. **Dublin Core Metadata Initiative .** Dublin Core Metadata Initiative . [Online]
http://dublincore.org/.

51. **Library of the Congress.** MARC 21 Bibliographic Data. [Online] 2018.
http://www.loc.gov/marc/bibliographic/.

52. **Open Archives Initiative.** Open Archives Initiative - Protocol for Metadata Harvesting.
[Online] https://www.openarchives.org/pmh/.

53. **Invenio.** Invenio instances around the world. [Online] https://invenio-
software.org/showcase/.

54. **CERN Document Server.** CERN Document Server. [Online] http://cds.cern.ch/.

55. **EUDAT.** EUDAT. [Online] http://www.eudat.eu/.

56. **DataCite.** re3data.org. [Online] https://www.re3data.org/.

57. **GitHub.** GitHub. [Online] http://www.github.com/.

58. **DataCite.** DataCite. [Online] http://datacite.org.

59. **GARR.** IDEM GARR AAI. [Online] https://www.idem.garr.it/.

60. **GrIDP.** GrIDP. [Online] https://gridp.garr.it/.

61. **eduGAIN.** eduGAIN. [Online] http://www.edugain.org/.

62. **GEANT.** GEANT. [Online] https://www.geant.org/.

63. **Google.** Google Scholar. [Online] https://scholar.google.com/.

64. **ORCID.** ORCID. [Online] https://www.orcid.org/.

65. **Clarivate Analytics.** ResearcherID. [Online] http://www.researcherid.com/.

66. **ResearchGate.** ResearchGate. [Online] https://www.researchgate.net/.

67. **Elesevier.** SCOPUS. [Online] https://www.scopus.com/.

68. **Altmetric.** Altmetric. [Online] https://www.altmetric.com/.

69. **CERN.** WLCG. [Online] http://wlcg.web.cern.ch/.

70. **European Commission.** European Open Science Cloud. [Online]
https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud.

71. **Invenio.** Invenio v3.0.0 Released . [Online] 7 June 2018.
https://inveniosoftware.org/blog/invenio-v300-released/.

72. —. Invenio Features. [Online] https://invenio-software.org/#features.

73. —. Invenio Roadmap. [Online] https://invenio-software.org/roadmap/.

74. —. Invenio Documentation Home Page. [Online] https://invenio.readthedocs.io/en/latest/.

75. **Zenodo.** Zenodo Organisation. [Online] https://github.com/zenodo.

76. *Citrix web site.* **[Online] https://www.citrix.com.**

77. *Kubernetes web site.* **[Online] https://kubernetes.io/.**

78. **[Online] https://www.openaccessrepository.it/record/21260.**

79. **Zenodo container.** *Docker hub.* **[Online] https://hub.docker.com/r/infnct/zenodo.**

80. **Atlassian. DevOps: Breaking the Development-Operations barrier. [Online]**
https://www.atlassian.com/devops.