

Openscience per INFN, INFN per Openscience

BY S. DAL PRA, A NOME DEL GLOS.

CCR WS, Palau, 24/05/2024

Open Science & FAIR Data

Open Science è un movimento culturale teso a rendere aperto ogni passo della ricerca scientifica (fonte: Wikipedia). I principi **FAIR** conseguono direttamente:

- **Findable**: I prodotti O.S. si possono individuare univocamente (→ doi)
- **Accessible**: Disponibili senza barriere di accesso (costi o dipendenze)
- **Interoperable**: Sono in formato aperto (leggibili tra tool differenti).
- **Reusable**: Sono ben descritti, e ben conservati (possibile uso, anche futuro).

INFN, Open Science, FAIR data

Strumenti chiave per l'Open Science erano già presenti e diffusi naturalmente in HEP/INFN da tempo prima che si codificassero ufficialmente i principi O.S.

SW. CNAF/SN: <https://baltig.infn.it> basato su GitLab, offre **Control Version**, che permette di individuare univocamente prodotti sw

SW+Data. /cvmfs/ . . . filesystems accessibili a chi si installa un client.

Per aderire alle pratiche FAIR sono nati gli **Open Access Repository**

INFN Open Access Repository

- <https://openaccessrepository.it>, installato e gestito a INFN-CT da fine 2014, è basato su zenodo/Invenio v3, col mandato (CDR 2019) di ospitare:
 - i. l'archivio delle Note INFN (che risale al 1955)
 - ii. la collana di proceedings Frascati Physics Series
 - iii. tutti i preprint arXiv con almeno un autore INFN.
 - iv. tutti i preprint INSPIRE con almeno un autore INFN.
- Ospita inoltre contenuti di altri EPR (es. ISPRA, MoU in preparazione).
- I contenuti sono organizzati in communities, che possono seguire i flussi di approvazione previsti dal proprio admin.
- È in corso la migrazione verso CNAF/SN, su Invenio RDM, v11.0
- Trepida attesa per la v12, ma per ora si lavora con v11.

Approvazione contenuti INFN

Dettagli sul Disciplinare INFN, art. 5

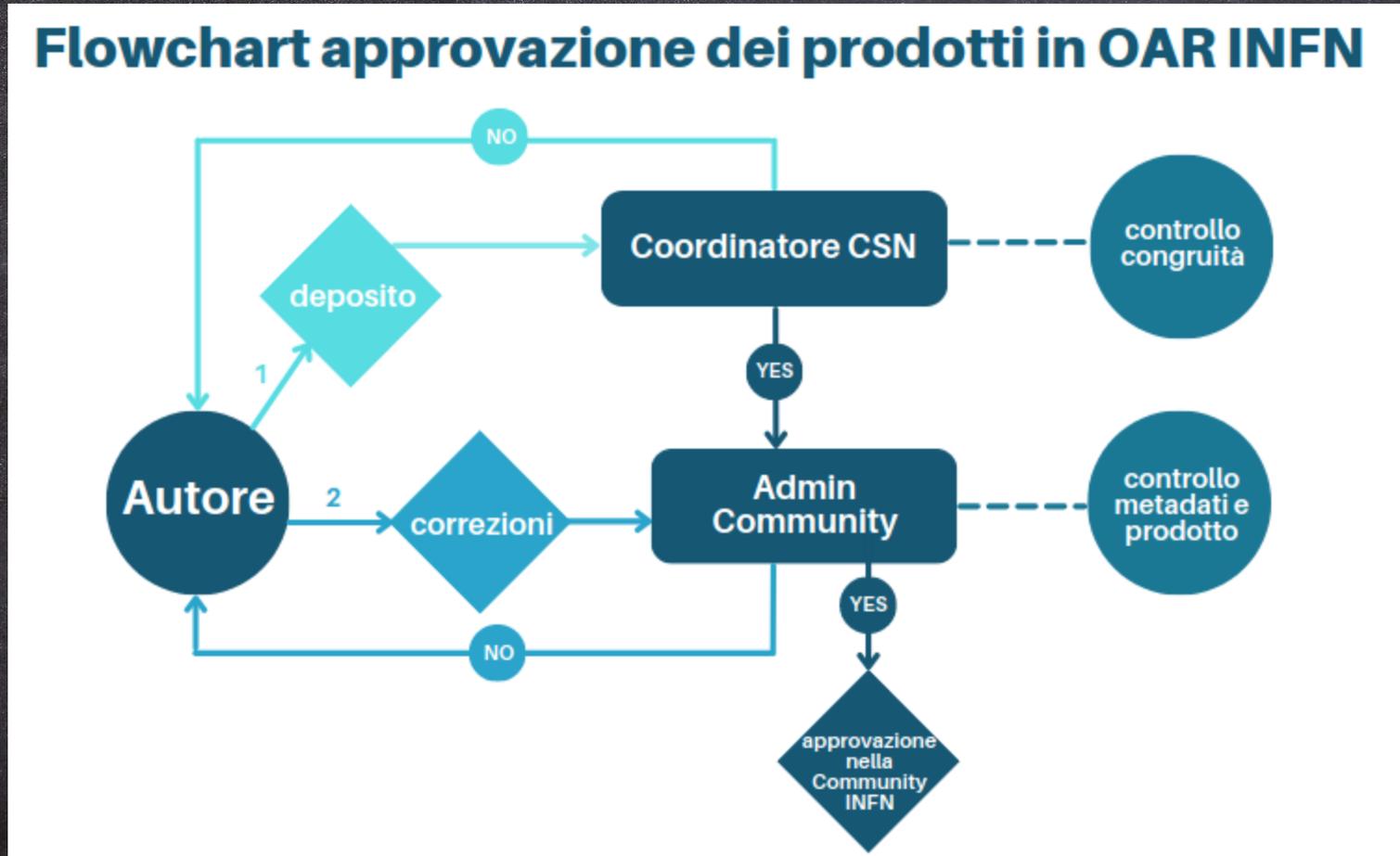


Table 1. openaccessrepository.it ospita diverse community; il diagramma rappresenta il flusso di validazione per la community INFN previsto per la nuova istanza.

Esempio Record pubblicato

The screenshot shows a record on the INVENIORDM platform. At the top, there is a search bar and navigation links for 'Communities' and 'My dashboard'. The record is published on May 6, 2024, and is version v1. The title is 'Enabling INFN-T1 to support heterogeneous computing architectures'. The authors listed are Dal Pra, Stefano; Spiga, Daniele; Boccali, Tommaso; Chierici, Andrea; Morganti, Lucia; Sapunenko, Vladimir; Cesini, Daniele; Rinaldi, Lorenzo; Gregori, Daniele; and Cicala, Marco. The affiliations include the National Center for Frame Analysis and various INFN sections. The citation is provided in APA style. A description of the INFN-CNAF Tier-1 center and its role in supporting heterogeneous computing architectures is included. A PDF file is attached, titled 'epjconf_chep2024_11006.pdf', which is a preprint from the EPJ Web of Conferences 295, 11006 (2024).

INVENIORDM Search records... Communities My dashboard

Published May 6, 2024 | Version v1 Journal article Open

Enabling INFN-T1 to support heterogeneous computing architectures

Dal Pra, Stefano¹; Spiga, Daniele²; Boccali, Tommaso³; Chierici, Andrea⁴; Morganti, Lucia¹; Sapunenko, Vladimir¹; Cesini, Daniele¹; Rinaldi, Lorenzo⁴; Gregori, Daniele⁵; Cicala, Marco⁵

[Hide affiliations](#)

1. National Center for Frame Analysis
2. INFN Sezione di Perugia
3. INFN Sezione di Pisa
4. INFN Sezione di Bologna
5. E4 Computer Engineering SPA

Citation

Style APA

Dal Pra, S., Spiga, D., Boccali, T., Chierici, A., Morganti, L., Sapunenko, V., Cesini, D., Rinaldi, L., Gregori, D., & Cicala, M. (2024). Enabling INFN-T1 to support heterogeneous computing architectures.

Description

The INFN-CNAF Tier-1 located in Bologna (Italy) is a center of the WLCG e-Infrastructure providing computing power to the four major LHC collaborations and also supports the computing needs of about fifty more groups - also from non HEP research domains. The CNAF Tier1 center has been historically very active putting effort in the integration of computing resources, proposing and prototyping solutions both for extension through Cloud resources, public and private, and with remotely owned sites, as well as developing an integrated HTC+HPC system with the PRACE CINECA supercomputer center located 8Km far from the CNAF Tier-1 located in Bologna. In order to meet the requirements for the new Tecnopolo center, where the CNAF Tier-1 will be hosted, the resource integration activities keep progressing. In particular, this contribution will detail the challenges that have recently been addressed, providing opportunistic access to non standard CPU architectures, such as PowerPC and hardware accelerators (GPUs). We explain the approach adopted to both transparently provision x86_64, ppc64le and NVIDIA V100 GPUs from the Marconi 100 HPC cluster managed by CINECA and to access data from the Tier1 storage system at CNAF. The solution adopted is general enough to enable seamless integration of other computing architectures at the same time from different providers, such as ARM CPUs from the TEXTAROSSA project, and we report about the integration of these within the computing model of the CMS experiment. Finally we will discuss the results of the early experience.

Files

epjconf_chep2024_11006.pdf

EPJ Web of Conferences **295**, 11006 (2024) <https://doi.org/10.1051/epjconf/202429511006>
CHEP 2023

Edit
New version
Share

Versions

Version v1	May 6, 2024
------------	-------------

Details

Resource type
Journal article

Export

JSON Export

Oltre a link ORCID e Affiliazione ROR, link a nuove versioni e prodotti collegati.

Migrazione Zenodo → Invenio RDM v11

Gli upgrade tools 1 e 2 non sono applicabili al nostro caso. Si applica il generico procedimento **Extraction, Transform, Load**

Extraction. Metadati estratti via SQL queries sul database sorgente (**zenodo**). Tra questi figura la **posizione virtuale** dei dati (si ricorre a hash **md5**). È necessario creare in **zenodo** delle tabelle ausiliarie

Transform. Si adattano i metadati al formato necessario per l'inserimento in Invenio. È la parte più complessa, laboriosa e destinata a rimanere non *completa*:

1. semantica sorgente più povera (es. **“name”**: **“Nome Cognome”**)
2. Invenio richiede obbligatoriamente alcuni valori, che in zenodo possono mancare

Load. si caricano i dati estratti in Invenio. La procedura via api è articolata in tre fasi e rispecchia quel che si farebbe via dashboard:

- i. **Draft upload** → Carica metadati, dati, genera e assegna un **pid**.
- ii. **Draft update** → Si associa il record a una community
- iii. **Review request and publish**

Note su Extraction + Transform

- I metadati inseriti in origine sono spesso incompleti o inaccurati, in particolare per i dettagli anagrafici e di affiliazione degli autori.
- Es. **Affiliazione**: “INFN-BO”, “INFN Bologna”, “Ist. Naz. Fisica Nucl., Sezione di Bologna” sono tutte possibili, e questo vale per **tutte** le sedi.
- Rimedio (parziale). Per arricchire i dati sorgente ricorriamo a diversi repo esterni:
 - ORCID, ROR, + metadati autori da VQR (thanks A. Paoletti, GLV/DSI). Quando si trova un match si possono associare i metadati mancanti al record originale.

Coordinamento attività

- task tecnici gestiti seguito con l'issue tracker di baltig.
- Attività più generali gestite a livello GLOS su ONLYOFFICE (GARR).

E+T Si lavora su una copia di zenodo (metadati) e un rsync dei dati. Si aggiungono in zenodo metadati ausiliari, poi:

DB zenodo		Tableaux aux
CREATE TABLE aux_r0 AS (SELECT . . .)	←	Campi di interesse
↓	←	Join External Info
aux_r1	←	+ORCID +ROR,...
↓	←	Join MD5 files
aux_r2	←	AGGR by auth, files
↓	←	1 row = 1 record
aux_r3	←	metadati pronti

- aux_r3 contiene metadati ~ pronti per inserimento draft in Invenio. Lato python rimane la trasformazione conclusiva in Json
- aux_r0 è “definitiva”: se arrivano ulteriori metadati si ricreano le successive.
- ~ 97.5 Krecords, ~ 300GB prodotti

Esempio di record Estratto e pre-trasformato (aux_r3)

```
zenodo=# SELECT * FROM aux_r3 WHERE communities LIKE '%covid%' LIMIT 1;
```

```
id          | 000217c4-e4bb-4134-bf4e-5ce8630b2fe5
created     | 2020-05-26 10:00:52.543768
doi         | 10.15161/oar.it/23690
user        | marco.fargetta@ct.infn.it
title       | CovidStat project summary plots
description | This record contains the daily update summary plots of the data of the CovidS...
publication_date | 2020-05-25
resource_type | image-plot
sets        | ["user-infn", "user-covidstat-infn"]
communities | covidstat-infn
access_right | open
keywords    | ["COVID-19", "CovidStat project", "FAIR data", "Open Science"]
creators    | [{"name": "Menasce, Dario", "orcid": "0000-0002-9918-1686",
    "familyname": "menasce", "givennames": "dario", "affiliation": "INFN Milano Bicocca"},...]
jfiles      | [{"key": "italia_sommario.png", "size": 84049, "type": "png", "bucket": "488076ea-18f7-4102-81fc-82c46ff736df",
    "file_id": "089e0de7-3b11-49a8-addb-405478f72672", "checksum": "md5:361a8a76a981b12ed8084496dd7af18c",
    "version_id": "512d8281-d3b3-44d0-95cf-7bd61380c062"},...]
version_id  | 6
owner       | marco.fargetta@ct.infn.it
filelist    | [{"key": "italia_sommario.png", "md5": "361a8a76a981b12ed8084496dd7af18c",
    "path": "./236/90/r/2020-05-26T10:01:01.222608+00:00/data/files/089e0de7-3b11-49a8-addb-405478f72672-italia_sommario.png"}]
```

Invenio Drafts dashboard

The screenshot shows the Invenio Drafts dashboard for user Stefano Dal Pra. The interface includes a top navigation bar with the INFN logo, a search bar, and links for Communities and My dashboard. The user's profile is visible, showing the name Stefano Dal Pra and a search bar for uploads. The main content area displays a list of 10 draft records, sorted by 'Recently updated'. The first record is titled 'CovidStat project data' and is a Dataset type, uploaded on May 17, 2024. The second record is titled 'Ottimizzazione della risoluzione energetica per un fascio γ di backscattering di luce laser da un fascio di elettroni' and is a Technical note type, also uploaded on May 17, 2024. The left sidebar contains filters for Access status, Status, Resource types, and Help.

https://zenodo-dev.infn.it/me/uploads?q=&l=list&p=1&s=10&sort=updated-desc

INFN Search records... Communities My dashboard stefano.d...

Stefano Dal Pra

Uploads Communities Requests

Search in my uploads... New upload

10 result(s) found Sort by Recently updated

CovidStat project data
Menasce, Dario ; Mezzetto, Mauro; Pedrini, Daniele
This record contains the daily updated data of the CovidStat project. CovidStat is a project carried out by the CovidStat Working Group at INFN, whose creation was promoted within the Italian National Institute of Nuclear Physics with the aim of making a statistical analysis of the data provided daily by the Civil Protection on the spread of the...
Uploaded on May 17, 2024

Ottimizzazione della risoluzione energetica per un fascio γ di backscattering di luce laser da un fascio di elettroni
Preger, M.
La distribuzione energetica di un fascio di fotoni di alta energia ottenuti dallo scattering a $\sim 180^\circ$ di fotoni laser su di un fascio di elettroni di un anello dipende essenzialmente: 1) dall'angolo solido definito dal collimatore sul fascio γ ; 2) dalla distribuzione in energia degli elettroni; 3) dalla distribuzione degli angoli delle traiettor...
Uploaded on May 17, 2024

Access status
 Metadata-only 10

Status
 Unpublished 10

Resource types
> Publication 6
 Dataset 3
> Image 1

Help
[Search guide](#)

Deployment

Invenio RDM mette a disposizione due varianti:

Containerized install. → consigliata per valutazione; comoda per conoscere il sistema, ma considerata meno adatta in produzione. Segue questa via **LNGS (PNRR LNGS-FUTURE, WP3)**

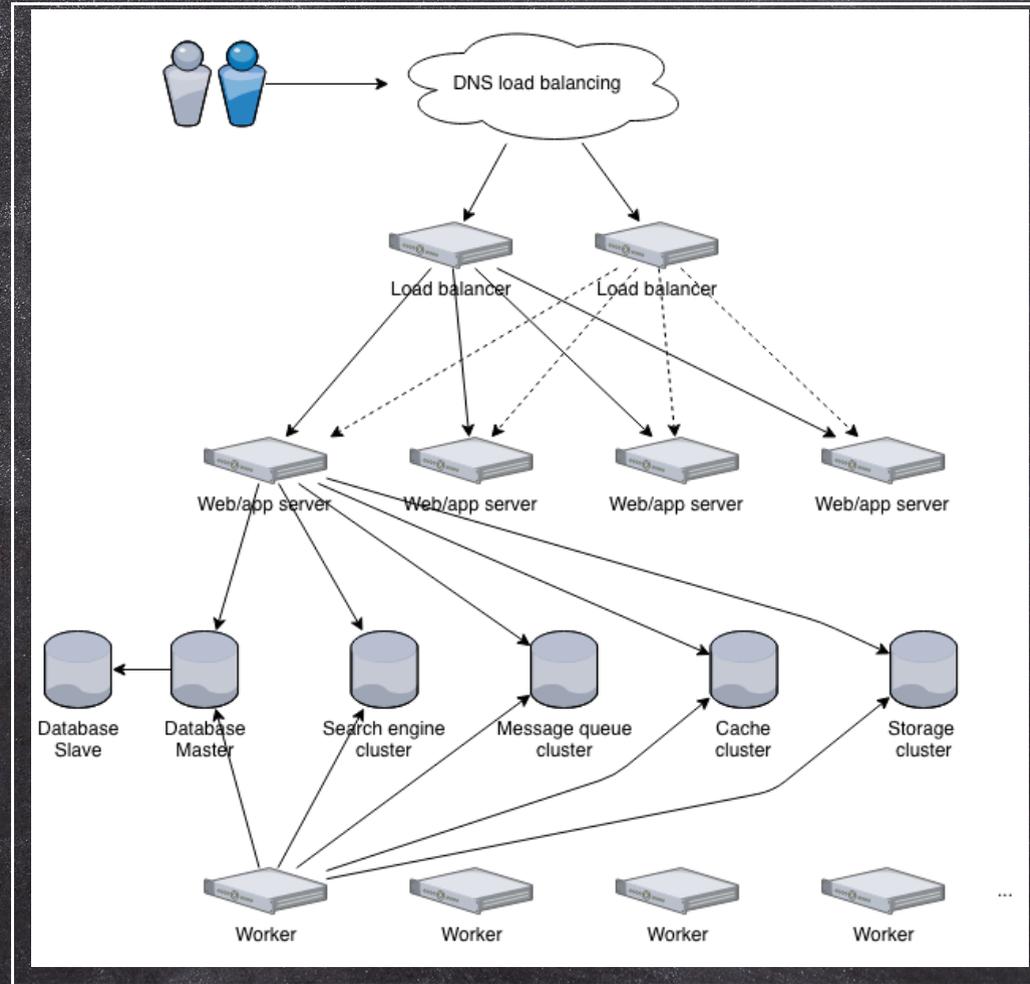
Local Install. → consigliata per produzione; presenta una serie di difficoltà rispetto al caso precedente, principalmente per una questione di “documentazione rarefatta”. Al CNAF seguiamo questa opzione.

→ **Nota:** il team Invenio NON ha risorse dedicate al supporto, concentra gli sforzi sul development

Configurazioni per accesso utenti

Invenio offre supporto per accesso OAuth2; l'istanza del CNAF è stata testata con Keycloak, AAI. In corso integrazione per accesso via ORCID. Pensiamo di escludere accesso via github (autenticazione carente).

Deployment Layout



Fonte: <https://inveniordm.docs.cern.ch/develop/architecture/infrastructure/>

Dove siamo

- Istanza di test <https://zenodo-dev.infn.it>, accesso a pochi “beta tester” via certificato.
- Procedura di migrazione completata.
- Applicare la procedura (piú e piú volte) ← **noi siamo qui**
- Todo: verifiche finali sui contenuti migrati (se ne mancano vanno cercati in zenodo e trattati a parte, e cosí per altri casi particolari)
- Todo: Inserimento note INFN mancanti (fattibile, solo questione di tempo)

Path to Production

- Finalizzare configurazione (in particolare: storage e scalabilità orizzontale) e dimensionamenti
- Verifiche di robustezza / resilienza

Next: Upgrade a Invenio **v12**. Di rilascio imminente da qualche tempo, offre features richieste, e **script di upgrade v11.0 → v12** (ma non **12b.x → 12.0**)

Next2: Storage dedicati (per documenti, per datasets)

Acknowledgments

Management Board.

M. Pallavicini (pres. G.E.), S. Bianco, M. Maggi, L. Patrizi, L. dell'Agnello
(Comitato ex art. 8, Disciplinare accesso ai prodotti)

Team. S. Antonelli, S. Bianco, S. Dal Pra, I. Piergentili, A. Bombini, F. Marchegiani, S. Stalio,

Credits. R. Rotondo, S. Monforte

Contatti. openscience@lists.infn.it