

Invenio, Zenodo e Archivi aperti

BY S. DAL PRA, A NOME DEL GLOS E S.N. INFN

OpenScience week, Genova, 05/11/2024



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. per soddisfare le esigenze di condivisione tra grandi collaborazioni scientifiche.

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

SW+Data. /cvmfs/ . . . filesystems open e accessibili a chi si installi un client.

Per dare sostegno alle pratiche FAIR si sono inoltre sviluppati gli **archivi aperti (Open Access Repository)**

INFN Open Access Repository

- <https://openaccessrepository.it>, installato e gestito a INFN Sez. Catania 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 **INSPIRE** e **arXiv** 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 rispettivo amministratore.
- Stiamo finalizzando la migrazione verso CNAF/SN, su Invenio RDM, **v12.0**
- Lavoro iniziato per migrare verso **v11.0**;
- la **v12.0**, rilasciata a fine Luglio 2024, supporta numerose features richieste, per cui stiamo adattando per migrazione diretta a v12.0

Zenodo, Invenio, InvenioRDM

Zenodo. è un *repository service*, quindi un archivio, con codice Opensource.

Invenio. Un *framework/toolbox*: una libreria software che fornisce strumenti per realizzare un servizio come zenodo. Sono state sviluppate molte implementazioni indipendenti basate su Zenodo → frammentazione

InvenioRDM. Collaborazione formata dalla comunità degli sviluppatori/utenti di Invenio, per realizzare una piattaforma omogenea.

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 INVENJO RDM platform. The header includes the logo, a search bar, and navigation links for 'Communities' and 'My dashboard'. The record is published on May 6, 2024, and is a 'Journal article'.

Enabling INFN-T1 to support heterogeneous computing architectures

Del Pra, Stefano¹, Spiga, Daniele², Boccali, Tommaso³, Chierici, Andrea⁴, Morganti, Lucia⁵, Sapunenko, Vladimir¹, Cesini, Daniele⁵, Rinaldi, Lorenzo⁵, Gregori, Daniele⁵, Cicala, Marco⁵

Affiliations:

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

Citation: Style: APA

Del 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 HTCondor-HPC system with the PRACE CINECA supercomputer center located 80km 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 TEXAROSSA 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

Viewing a PDF document titled 'epjconf_chep2024_11006.pdf'. The document is from 'EPI Web of Conferences 295, 11006 (2024)' and 'CHEP 2023'. The URL is 'https://doi.org/10.1051/epjconf/202429511006'.

Actions: 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 (v3) → Invenio RDM v12

- Normalmente InvenioRDM fornisce script di upgrade da una versione alla successiva.
- Nel nostro caso: v3.0 → v4.0 → ... → v12.0 non è, di fatto, praticabile.
- La versione di partenza è modificata v3.x da alcune customizzazioni e questo rende inapplicabili gli script ufficiali.
- Versioni recenti arricchiscono la semantica dei metadati e rendono obbligatori valori che prima non lo erano
- validazione metadati più rigorosa con le versioni nuove

È necessario *arricchire* i metadati prima di migrarli; alcune inconsistenze saranno inevitabili.

Analisi sui metadati di Invenio v3

Alcuni aspetti da considerare

- Campo unico per nome autore → in v12 campi dedicati `family_name`, `given_name`, entrambi obbligatori.
- affiliazioni autori definite a “mano libera” → moltissime rappresentazioni diverse per la stessa affiliazione.
- **ORCID** presente solo in pochissimi casi → in v12 va reso obbligatorio per autori INFN.
- **ROR** assente, o definito “impropriamente”.
- Campi malformati: per es. alcuni `doi` con spazi o aggiunte improprie, e così anche per altri metadati.
- Manca una “curation policy” per alcune communities → obbligatoria in v12

Esempio: Affiliazione autori

```
zenodo=# SELECT COUNT(*) n, creator_affiliation aff FROM aux_r00
WHERE LOWER(creator_affiliation) LIKE '%nfn%enov%' GROUP BY aff ORDER BY n desc;
```

```
n | aff
-----+-----
18 | INFN Sezione di Genova
10 | INFN, Sezione di Genova
 4 | INFN, Sezione di Genova Italy
 3 | INFN Genova, Italy
 3 | INFN, Genova
 2 | Universita degli Studi di Genova and INFN, ' GENOVA, Italy
 2 | INFN Genova
 2 | INFN - Genova, Italy
 2 | INFN - Sezione di Genova
 2 | Dipartimento di Fisica dell'Università di Genova and INFN-Sezione di Genova, Via Dodecaneso 33, 16146 Genova, Italy
 1 | Dipartimento di fisica & INFN Genova
 1 | INFN Sezione Genova
 1 | INFN Genova and Sapienza University of Rome
(13 rows)
```

- Questa “ricchezza” di rappresentazioni per uno stesso valore si riscontra naturalmente per tutte le affiliazioni.
- Durante l’analisi sono emersi dei disallineamenti con l’archivio ROR (es. alcuni gruppi collegati non presenti). Abbiamo colto l’occasione per emendare le voci ROR e sanare le discrepanze.
- Operazioni analoghe sono state compiute o sono in corso avanzato per altri metadati.
- Si considera quindi la possibilità, con la migrazione, di migliorare la qualità di alcuni metadati.

Procedura Extraction, Transform, Load

Esistono tools pronti disponibili (es. 1 e 2) ma non sono direttamente applicabili; dobbiamo ricorrere al procedimento generico: **Extraction**, **Transform**, **Load**, con l'aggiunta di un passo supplementare, di **Enrichment**.

Extraction. Metadati estratti via SQL queries sul database sorgente (**zenodo**). Per comodità tali dati estratti vengono inseriti in **tabelle ausiliarie** (alcuni dettagli a seguire)

Transform. Si adattano i metadati al formato necessario per l'inserimento in Invenio. È la parte piú complessa, laboriosa e destinata a rimanere in alcuni casi *incompleta* o *errata*:

1. semantica sorgente piú povera (es. **“name”**: **“Nome Cognome”**)
2. Invenio v12 richiede obbligatoriamente alcuni valori, che in v3 possono mancare.

Enrichment. Si aggiungono dati assenti (o si migliorano quelli presenti) su v3, spesso obbligatori su v12.

Load. si caricano i dati estratti in Invenio v12 attraverso le sue **api**. La procedura, articolata in piú fasi, rispecchia quel che farebbe un utente via dashboard:

- Draft upload** → Carica metadati, dati, genera e assegna un **pid** (Principal ID).
- Draft update** → Si associa il record a una community, si aggiungono eventuali altri metadati
- Files upload** → Si caricano i files associati (pdf, datasets ecc.).
- Review request and publish** (con registrazione del **DOI**, se necessaria).

Note su Extraction, Enrichment, Transform

- I metadati inseriti in origine sono spesso incompleti o inaccurati, in particolare per i dettagli anagrafici e di affiliazione degli autori.
- 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**.

- Per **uniformare le affiliazioni** abbiamo applicato una Convolutional Neural Network, allenata da M. Gattari. Questo ci permette anche di **associare il codice ROR**

```
>>> print(n,"updates",up,"records affected")  
469 updates 15710 records affected
```

- **Nota:** Molte affiliazioni sono inserite *multiple*, es. **Univ-X e INFN-X**. In questi casi la rete seleziona **INFN-X**.

Invenio e vocabolari La v12 supporta i **vocabularies**, che permettono di precaricare valori ORCID, ROR e associarli preventivamente agli autori. **Questo risolve il problema per nuovi inserimenti**. Parte del lavoro di Enrichment è usato anche per produrre i vocabularies.

I vocabularies sono soggetti a cambiare (es. nuovi nomi). Invenio v12 offre la possibilità di aggiornarli, che in v11 era molto più limitata.

Tabella ausiliarie

Esempi di dati ausiliari usati per normalizzare o arricchire quando possibile informazioni mancanti tra i metadati dei record da migrare

```
zenodo=# SELECT ror,name,acronym FROM ror WHERE name LIKE '%INFN%' LIMIT 4;
```

ror	name	acronym
05eva6s33	INFN Sezione di Roma I	INFN-ROMA
05symbg58	INFN Sezione di Pisa	INFN-PI
02pq29p90	INFN Sezione di Catania	INFN-CT
015kcdd40	INFN Sezione di Napoli	INFN-NA

```
zenodo=# SELECT familyname,givenname,orcid,ror FROM aux_id_names_orcidror LIMIT 2;
```

familyname	givenname	orcid	ror
nome1	cognome1	0000-0003-0326-6368	049jf1a25
nome2	cognome2	0000-0001-9934-5081	025e3ct30

```
zenodo=# SELECT * FROM aux_md5_filepath LIMIT 2;
```

md5	filepath
0000037db4cdc9dd36e2c60787edaa69	./150/880/r/.../-fulltext.pdf
000011369674e792ca218179e1e6dca6	./134/669/r/.../record-json.json

E+T Si lavora su una copia locale di dati e metadati di **v3**. 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
Python		Json v12 records
create_users.py	←	rdm_init.sh
migrate_communities.py		
dump_records.py		
migrate_record.py		

- 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.
- In v3: ~ 97.5 Krecords, ~ 300GB prodotti. Non tutti vengono migrati

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
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"]
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/...672-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 'Stefano Dal Pra' is displayed with a red 'S' icon. Below the profile, there are tabs for 'Uploads', 'Communities', and 'Requests', with 'Uploads' being the active tab. A search bar for uploads and a 'New upload' button are present. The main content area shows a list of 10 draft records, sorted by 'Recently updated'. The first record is titled 'CovidStat project data' by Menasce, Dario; Mezzetto, Mauro; and Pedrini, Daniele. The second record is titled 'Ottimizzazione della risoluzione energetica per un fascio γ di backscattering di luce laser da un fascio di elettroni' by Preger, M. The left sidebar contains filters for 'Access status' (Metadata-only), 'Status' (Unpublished), and 'Resource types' (Publication, Dataset, Image).

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

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

S Stefano Dal Pra

Uploads Communities Requests

Search in my uploads... New upload

Access status

- Metadata-only 10

Status

- Unpublished 10

Resource types

- Publication 6
- Dataset 3
- Image 1

Help

[Search guide](#)

10 result(s) found Sort by Recently updated

Draft January 31, 2022 (v1) Dataset Metadata-only Edit

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

Draft November 12, 1985 (v1) Technical note Metadata-only Edit

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

Migrazione vs Deployment

Attualmente in uso tre istanze separate:

1. messa a punto procedura **migrazione e validazione complessiva** → S. Dal Pra
2. setup di aspetti particolari (**autenticazione Oauth, integrazione mail** ecc.) → S. Antonelli/SSNN.
3. Istanza con deployment di produzione (per verificare procedure di amministrazione dell'istanza: **start / stop / backup / recovery**, ecc.) → E. Cesarini / SSNN.

Quest'ultima è conforme all'architettura complessiva dei **Servizi Nazionali INFN**, quindi integra ridondanza geografica delle applicazioni e dello storage

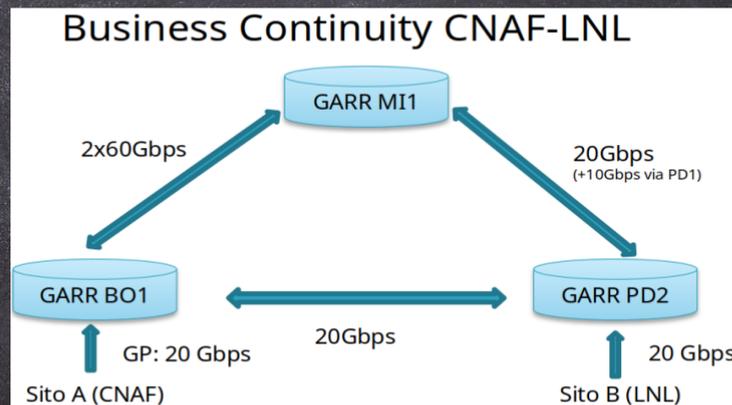
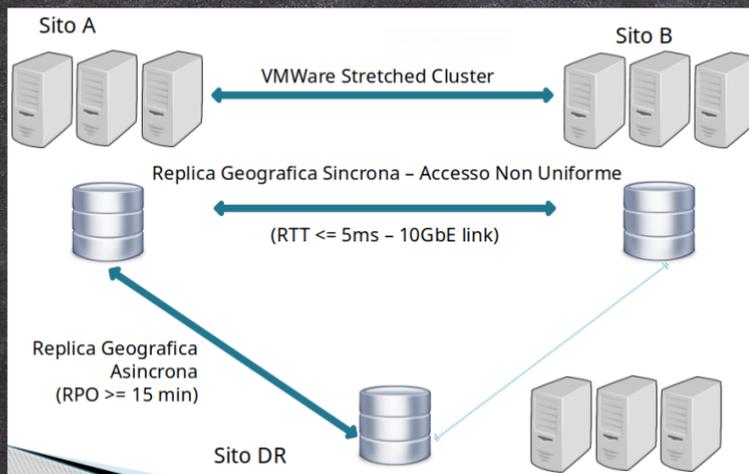
Note

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

→ si può chiedere aiuto attraverso un canale discord usato dagli sviluppatori

→ crea una ml invenio-admin@lists.infn.it per "scambio di esperienze" tra gestori di istanze Invenio RDM.

Ridondanza geografica e Business Continuity



- Un servizio software gestito dai SSNN (InvenioRDM v12 nel nostro caso) rimane attivo e disponibile anche in caso down del CNAF (Sito A) e l'istanza torna disponibile in modo trasparente al Sito B (INFN-LNL)
- I dati sono continuamente replicati geograficamente tra i due siti
- Oltre a questo si aggiungono i regolari backup

Ticketing system, code repository etc.

Utilizziamo una istanza INFN di GitLab per tenere il codice sviluppato, immagini container, configurazioni dell'istanza, documentazione interna e coordinare le attività tecniche in corso.

The screenshot displays the GitLab Issues interface for the 'Invenio / Invenio INFN / Issues' project. The top navigation bar includes a search bar and filters for 'Open 61', 'Closed 53', and 'All 114'. A sidebar on the left shows the project structure with 'Issues' selected. The main content area lists several issues, each with a title, ID, creation date, author, labels, and status.

Issue ID	Title	Created	Author	Labels	Status	Comments
#112	connessione v12 <--> opensearch	2 weeks ago	Stefano Dal Pra	deployment, opensearch, security	Closed	1
#111	CSP in avatar con proxy esterno	2 weeks ago	Ettore Cesarini	bug	Closed	2
#106	Migrazione v12: problemi migrazione irrisolti	3 weeks ago	Irene Piergentili	metadati, migration	Closed	2
#105	Inserimento metadato Conference/Meeting	3 weeks ago	Irene Piergentili	metadati, migration	Closed	2
#103	Inserimento metadato Software	3 weeks ago	Irene Piergentili	metadati	Closed	1
#101	Keywords non separate	1 month ago	Irene Piergentili	metadati	Closed	1

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
- verifiche finali sui contenuti migrati e correzione problemi ← siamo anche qui
- setup istanza di produzione ← e qui
- Todo: Inserimento note INFN mancanti (fattibile, solo questione di tempo)
- Ingestione da Inspire ← in corso
- Verifiche di robustezza / resilienza

Coordinamento attività

- task tecnici gestiti seguito con l'issue tracker di baltig.
- Coordinamento generale gestito a livello GLOS su ONLYOFFICE (GARR).

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, E. Cesarini, S. Bianco, S. Dal Pra, I. Piergentili, F. Marchegiani, S. Stalio, A. Bombini

Credits. R. Rotondo, S. Monforte, M. Gattari, A. Paoletti, S. Longo e i SSNN.

Contatti. openscience@lists.infn.it