DOI: 10.26314/GARR-Conf23-proceedings 05

IIT Dataverse: il repository dell'Istituto Italiano di Tecnologia per la conservazione e la condivisione dei dati FAIR della ricerca

Valentina Pasquale, Alessandro Bruchi, Ugo Moschini, Elisa Molinari, Daniele Rossetto, Francesca Cagnoni, Stefano Bencetti

Istituto Italiano di Tecnologia

Abstract. Nel riconoscimento dell'importanza di una gestione corretta e responsabile dei dati di ricerca, l'Istituto Italiano di Tecnologia ha implementato un data repository che consente ai ricercatori di documentare, conservare e condividere i dati che supportano le scoperte scientifiche dell'Istituto. Il repository è basato sul sistema open source Dataverse, adottato da numerose istituzioni a livello mondiale, che supporta i principi FAIR ed abilita il riuso dei dati per nuove ricerche. In questo articolo presentiamo il processo che ha portato all'adozione di questo strumento, la sua configurazione ed infrastruttura tecnica, i termini e le linee guida di utilizzo, ed infine i primi risultati e i possibili sviluppi futuri

Keywords. repository, Dataverse, research data management, FAIR data, storage

Introduzione

Negli ultimi anni, l'Istituto Italiano di Tecnologia ha dedicato risorse al consolidamento delle attività per la gestione dei dati della ricerca (Research Data Management), che comprende le varie fasi del ciclo di vita dei dati, dalla pianificazione alla conservazione, alla condivisione e al riuso. Tale supporto è una collaborazione delle Direzioni Organizzazione della Ricerca e ICT, con il contributo della Direzione Affari Legali.

In questo ambito, è stato implementato un repository per i dati della ricerca "definitivi", ovvero dati che sono selezionati, curati e pronti per la conservazione a lungo termine e/o condivisi in modo aperto in una forma definitiva o pseudo-definitiva. Lo strumento che è stato giudicato più idoneo per realizzare tale repository è il sistema Dataverse, un'applicazione web open source sviluppata da Harvard University nell'ambito delle Scienze Sociali e nata allo scopo di condividere, conservare, citare, esplorare ed analizzare i dati della ricerca. Questo sistema è facilmente configurabile e integrabile con altri sistemi informativi, supporta i principi FAIR (Wilkinson et al., 2016) per la gestione responsabile dei dati della ricerca ed è adottato da numerose Istituzioni a livello mondiale.

1. Identificazione dei requisiti e progetto pilota

A luglio 2019 è stato proposto un sondaggio, rivolto a ricercatori e tecnici di IIT (1547 destinatari), allo scopo di mappare le pratiche di Research Data Management (RDM) e il livello di consapevolezza in questo ambito. Il sondaggio era mirato all'individuazione

Firenze, 14-16 giugno 2023 - ISBN: 978-88-946629-2-4

dei principali punti critici e alla scelta di uno strumento che rispondesse alle esigenze dei ricercatori in tema RDM. Il sondaggio è stato basato in larga parte su quello condotto nel 2017 e 2018 presso TU Delft, EPFL, University of Cambridge e University of Illinois, con alcune modifiche e aggiunte per tenere conto delle specificità istituzionali (TU Delft Data Stewards et al., 2018). Sono state raccolte 557 risposte, dalle quali è emersa l'esigenza di dotarsi di un repository istituzionale per supportare la conservazione dei dati e la condivisione interna ed esterna con i collaboratori. Pertanto, abbiamo cercato di individuare uno strumento flessibile, open source e che offrisse una Application Programming Interface (API) sufficientemente sviluppata per configurare il sistema programmaticamente in modo da catalogare i dati sulla base dell'organizzazione dell'Istituto e definire uno schema di autorizzazioni e permessi sufficiente ad abilitare la condivisione selettiva. Lo strumento scelto, Dataverse, offre tutte queste funzionalità ed inoltre abilita la condivisione aperta dei dati in modo aderente ai principi FAIR, uno dei principali requisiti posto dagli enti finanziatori (es., la Commissione Europea). Pertanto, nell'aprile 2020 è stato avviato un progetto pilota di 8 mesi che ha coinvolto 8 Linee di Ricerca, 2 per ciascuno dei Domini nei quali si articola l'attività dell'Istituto: Computational Sciences, Life Technologies, Nanomaterials, e Robotics. Lo scopo del progetto pilota era testare lo strumento Dataverse proponendolo ai ricercatori, i quali sono stati intervistati successivamente alla fine del pilota per raccogliere suggerimenti sulla configurazione e su possibili miglioramenti.

2. Configurazione dell'infrastruttura del sistema di produzione

Il repository IIT Dataverse è stato lanciato in produzione nel maggio 2021. Durante il primo semestre dell'anno abbiamo lavorato alla messa a punto del sistema, sia dal punto di vista tecnico, lavorando sull'infrastruttura hardware e sulla configurazione della parte applicativa, sia da quello normativo, redigendo i termini di utilizzo (https://short.iit.it/terms) e le linee guida per la creazione dei dataset (https://doi.org/10.48557/EFFP8S). Dataverse è stato installato su un'infrastruttura dall'architettura scalabile (Figura 1), basata su 3 virtual machines (VMware vCenter), per un totale di 12 CPU, 40 GB RAM, 200 GB HD e 10 TB di storage dati. Il sistema di produzione è replicato su un'istanza di staging, non accessibile dall'esterno e dotata di uno storage dati più limitato, che viene utilizzata per sviluppo e testing. Il sistema è attualmente aggiornato alla versione 5.10.1.

Da un punto di vista applicativo, IIT Dataverse è integrato con il sistema di autenticazione single sign-on dell'Istituto. Tramite API è stata automatizzata la creazione giornaliera dei dataverse associati alle Linee di Ricerca, dell'aggiunta degli utenti e dei loro permessi. IIT Dataverse è integrato con OpenAIRE, al quale invia i metadati necessari all'identificazione dei dataset e alla loro associazione con gli ORCID degli autori e i grant ID. Inoltre, è indicizzato in re3data.org e il nostro Istituto fa parte del Global Dataverse Community Consortium.

3. Linee guida e supporto

Dal punto di vista della formazione, è stata creata una pagina dedicata nella intranet di IIT che mette a disposizione materiale (presentazioni, linee guida stampabili, FAQ, video

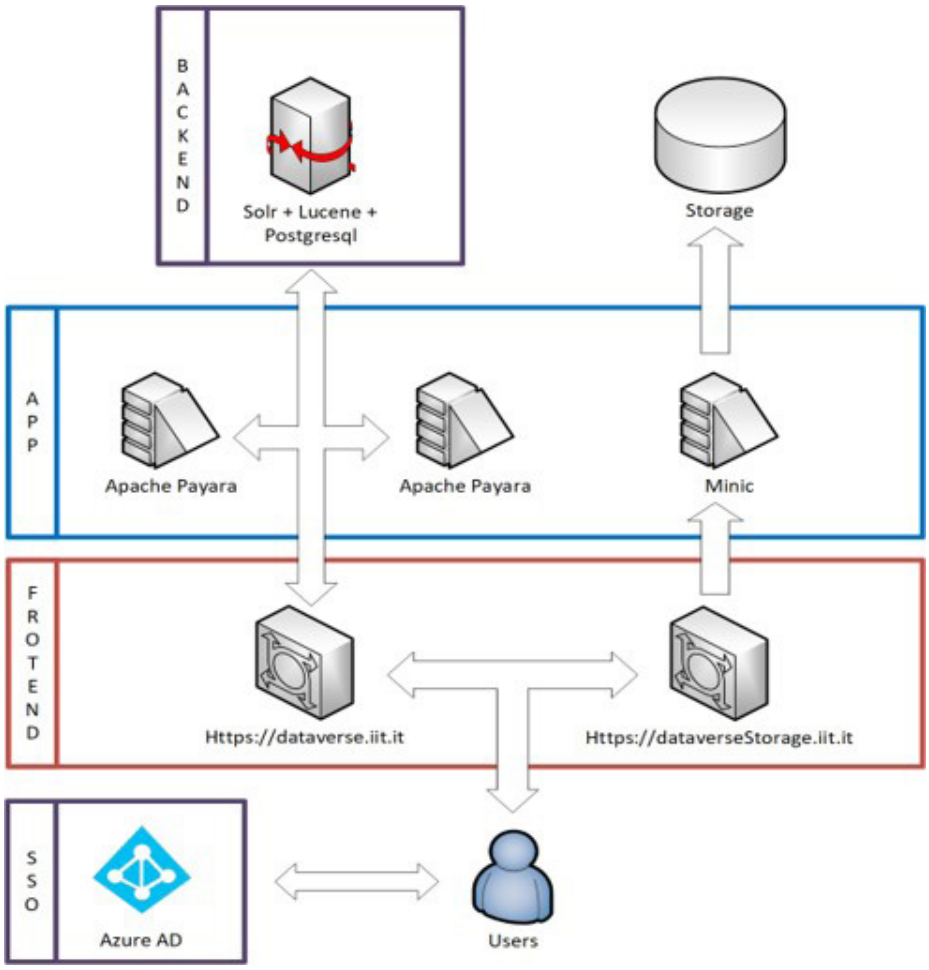


Fig. 1
Infrastruttura tecnica del repository IIT Dataverse.

dimostrativi, etc.) per assistere i ricercatori nel deposito dei dati secondo i principi FAIR. Nei mesi immediatamente successivi al lancio è stata abilitata la pubblicazione dei dataset sotto la supervisione e con il supporto del team RDM. È stata messa a punto una procedura di data curation, la quale prevede che i ricercatori attivino una "richiesta di pubblicazione dataset" attraverso un form e sottomettano il dataset per revisione. La richiesta viene presa in carico da un data steward che revisiona il dataset prima della pubblicazione sulla base di una checklist, in parte ispirata alla CURATED checklist della Data Curation Network e alla Curator Guide del repository DataverseNO. Eventuali richieste di aggiornamento o modifica del dataset vengono riportate al ricercatore, il quale si fa carico di implementarle al fine della pubblicazione. I termini di utilizzo del sistema sono conformi alle policy e procedure di IIT, in particolare 1) per la sicurezza delle informazioni (cyber-security), che definisce e regolamenta il trattamento dei dati sulla base della classe di rischio informatico associata, e 2) per la gestione dei dati personali secondo le leggi vigenti. Per scelta IIT Dataverse non è un repository per la condivisione dei testi delle pubblicazioni, che sono esplicitamente esclusi dalla tipologia di prodotto che è possibile caricare e condividere in questo sistema.

4. Conclusioni

Allo stato attuale (settembre 2023), IIT Dataverse contiene 30 dataset di ricerca, di cui 16 unpublished e 14 pubblici (escludendo dataset creati dal supporto RDM). Si rileva che, nonostante le periodiche campagne di formazione e sensibilizzazione interna, il numero di dataset pubblici rimane limitato in ordine di grandezza rispetto alla produzione di pubblicazioni scientifiche su journal dell'Istituto (> 1000/anno). È necessario considerare che IIT Dataverse non è l'unico strumento usato dai ricercatori di IIT per la condivisione dei dataset di ricerca, che sono pubblicati anche tramite Zenodo, Figshare e altri repository disciplinari (es., eBrains) già a disposizione della comunità scientifica. IIT Dataverse non viene presentato ai nostri ricercatori come l'unica soluzione disponibile per la condivisione aperta. Tuttavia, ci si pone come obiettivo di promuoverne ed incrementarne l'utilizzo, principalmente per la conservazione a lungo termine e il riutilizzo del patrimonio di dati di ricerca dell'Istituto, migliorando il supporto alla compilazione dei metadati e rendendolo uno strumento sempre più integrato all'interno dei flussi di lavoro.
Inoltre, IIT si è recentemente dotato di una nuova infrastruttura di storage centralizzata scale-out che ha l'obiettivo di aumentare la capacità di gestione dei dati, un loro uso sempre più efficace e che apre alla possibilità di applicare modelli di artificial intelligence ai dati stessi. IIT Dataverse sarà quindi integrato nel ciclo di vita del dato e verrà dotato di maggiore capacità, per garantire una migliore gestione di dataset di grandi dimensioni, e di termini di utilizzo aggiornati che abiliteranno anche la condivisione in forma restricted.

Riferimenti bibliografici

Wilkinson M., Dumontier M., Aalbersberg I. et al. (2016), The FAIR Guiding Principles for scientific data management and stewardship, Sci Data (3), 160018. https://doi.org/10.1038/sdata.2016.18

TU Delft Data Stewards, EPFL Library Research Data Team, Krause J., Lambeng N., Andrews H., Boehmer J., Cruz M., van Dijck J., den Heijer K., van der Kruyk M., & Teperek M. (2018), Quantitative assessment of research data management practice (Version 3) [Data set], Zenodo. https://doi.org/10.5281/zenodo.3380332

Autori

Valentina Pasquale valentina.pasquale@iit.it

Valentina Pasquale è Research Data Management Specialist all'Istituto Italiano di Tecnologia, dove si occupa dei servizi di supporto per la gestione dei dati della ricerca e la scienza aperta. Dal 2019 al 2021 è stata co-chair della rete di Data Stewardship Competence Centers in GO FAIR. Prima di occuparsi di dati della ricerca, ha ottenuto un dottorato in "Humanoid Technologies" all'Università di Genova e ha lavorato per diversi anni nel campo delle Neuroscienze Computazionali.

Alessandro Bruchi alessandro.bruchi@iit.it

Alessandro Bruchi è laureato in Ingegneria Elettronica e specializzato in progettazione microelettronica presso l'Università di Padova. Ha lavorato come consulente presso Engineering SPA

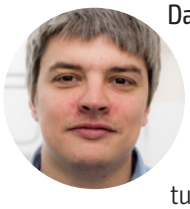
su progetti bancari e assicurativi, acquisendo esperienza in sistemi, reti e Linux. Dal 2007 lavora all'Istituto Italiano di Tecnologia e dal 2015 è responsabile dell'ufficio Cybersecurity, competente per la protezione delle infrastrutture e dei dati.

Ugo Moschini ugo.moschini@iit.it

Ugo Moschini, dopo aver lavorato all'Agenzia Spaziale Europea su algoritmi di compressione dati, ha conseguito un Dottorato in Informatica dall'Università di Groningen in analisi e classificazione di dataset di astronomia. Dal 2017 lavora nel Data Analysis Office dell'Istituto Italiano di Tecnologia su attività di analisi, visualizzazione e modellazione di dati riguardanti la produzione scientifica dell'istituto, con attenzione anche alle pratiche di Research Data Management e Open Science.

Elisa Molinari elisa.molinari@iit.it

Elisa Molinari ha conseguito un dottorato in Bioingegneria e Bioelettronica al DIST di Genova, progettando e sviluppando una piattaforma hardware e software per la gestione di dati neuroscientifici. Ha poi ottenuto un postdoc all'IIT e collaborato con università e l'ospedale pediatrico Gaslini nel campo dell'analisi di immagini biomedicali. È adesso Lead Data Analyst all'IIT di Genova e si occupa del coordinamento e dello sviluppo di sistemi di monitoraggio dell'attività di ricerca.



Daniele Rossetto daniele.rossetto@iit.it

Daniele Rossetto Casel si è laureato in Informatica presso l'Università di Genova nel 2004 e dal 2005 al 2009 ha lavorato presso lo stesso Ateneo come amministratore di rete e di sistema. Dal 2010 lavora presso l'Istituto Italiano di Tecnologia e dal 2014 ricopre il ruolo di responsabile dell'ufficio Gestione Infrastrutture, supportando l'Istituto e gli utenti su tutte le tematiche di infrastruttura IT.

Francesca Cagnoni francesca.cagnoni@iit.it

Francesca Cagnoni ha conseguito un dottorato in Immunologia Clinica e Sperimentale ed una specializzazione in Allergologia e Immunologia Clinica all'Università di Genova. Successivamente ha lavorato in un'azienda biotech prima di passare al supporto alle attività scientifiche. Oggi è Direttore del Research Organization Directorate di IIT, dove coordina un team trasversale che comprende gli uffici di Data Analysis, Projects, Research Agreement, e Outreach and Digital Production.

Stefano Bencetti stefano.bencetti@iit.it

Stefano Bencetti si laurea in Ingegneria Elettronica presso l'Università degli Studi di Genova e inizia la sua esperienza lavorativa nel settore ICT nel 1997 lavorando presso il Dipartimento di Matematica e il Dipartimento di Scienze dell'Informazione dell'Ateneo genovese. Dal 2009 è Direttore ICT presso l'Istituto Italiano di Tecnologia.