INTELLIGENZA ARTIFICIALE

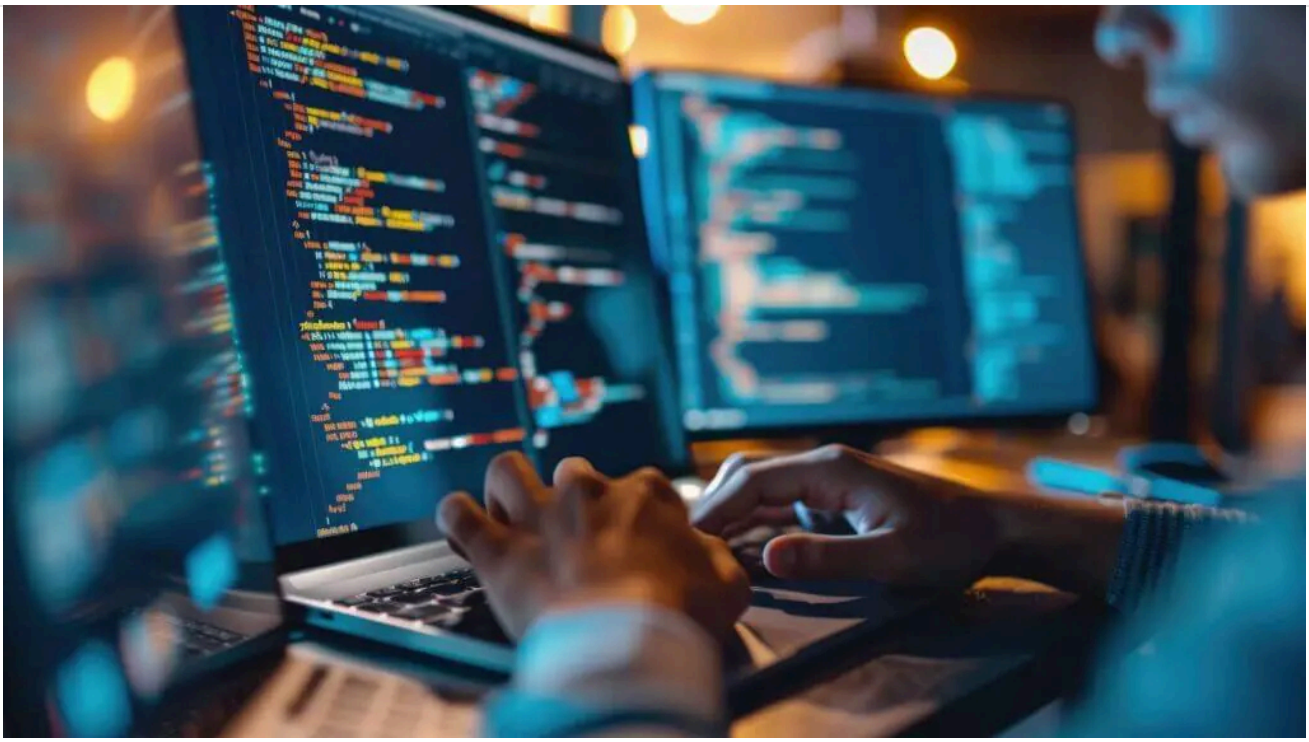
Formazione dei data steward: l'Università di Torino apre la strada in Italia

La gestione dei dati secondo i principi FAIR (Findable, Accessible, Interoperable, Reusable) diventa cruciale nell'epoca dell'Intelligenza Artificiale. Il corso universitario di aggiornamento per Data Steward, avviato a Torino, forma esperti capaci di pianificare la gestione dei dati in ambito scientifico, tecnologico, etico e legale, rispondendo alle esigenze di ricerca e industria

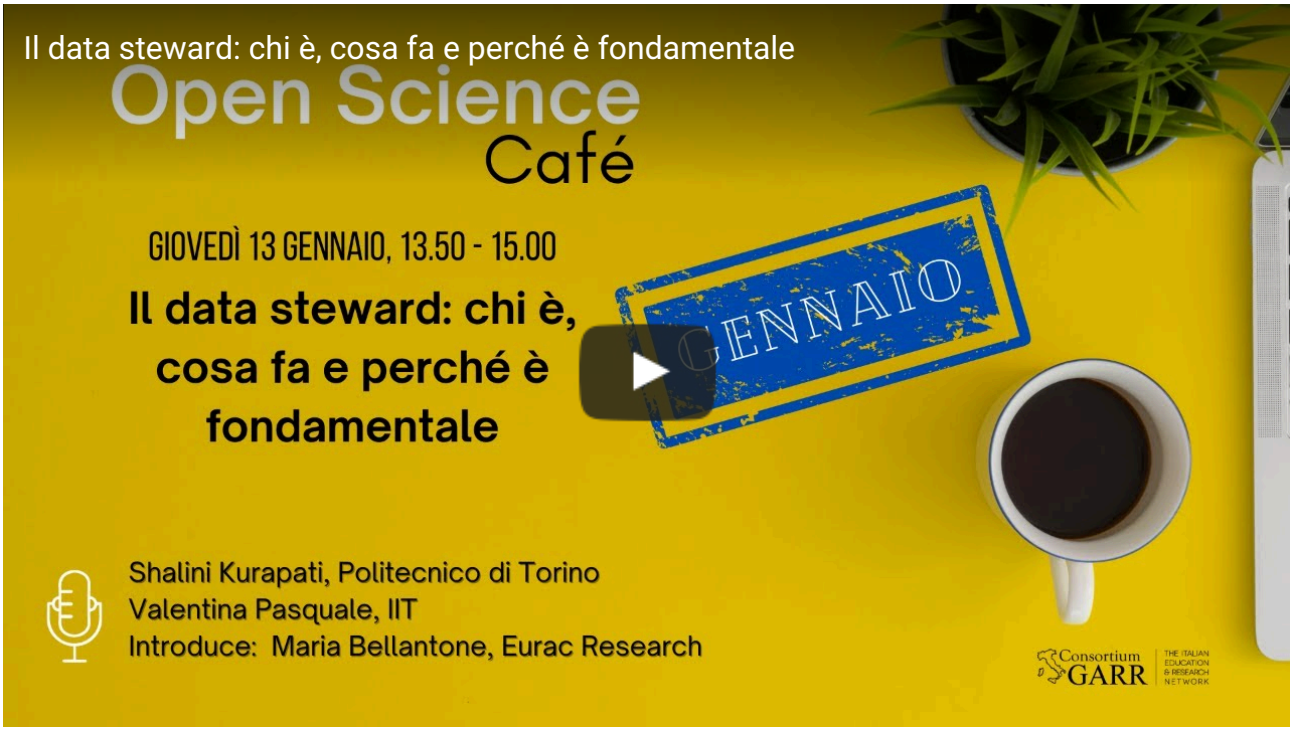
Pubblicato il 27 ago 2024

Marco Aldinucci
Dip. di Informatica, Università di Torino e CINI HPC-KTT

Rossana Damiano
Università di Torino



La gestione dei dati secondo i principi della **scienza aperta** riveste sempre maggiore importanza in ambito globale. In questo quadro, i concetti di riuso, accesso e interoperabilità (FAIR, ovvero Findable, Accessible, Interoperable, Reusable) hanno assunto un ruolo primario nella gestione dei progetti e delle infrastrutture di ricerca, tanto da essere spesso posti come vincolo alla concessione di finanziamenti [1].



Indice degli argomenti

L'Importanza dei principi FAIR nella gestione dei dati

I concetti base dell'approccio **FAIR**, che sono mostrati in Figura 1, descrivono un ecosistema di **sorgenti di dati** ("dataspaces") coordinate e globalmente utilizzabili dai sistemi di calcolo senza l'intervento umano ("machine actionable"). Sebbene questa visione sia maturata nell'ambito della ricerca, **la crescente domanda di dati di alta qualità per l'allenamento e la messa a punto di modelli di Intelligenza Artificiale** (AI) ha rapidamente amplificato l'interesse per la scienza aperta in diversi ambiti applicativi, dalla ricerca all'industria alla pubblica amministrazione e in tutte le aree della conoscenza: dalle scienze dure a quelle sociali, dalla medicina alle discipline umanistiche.

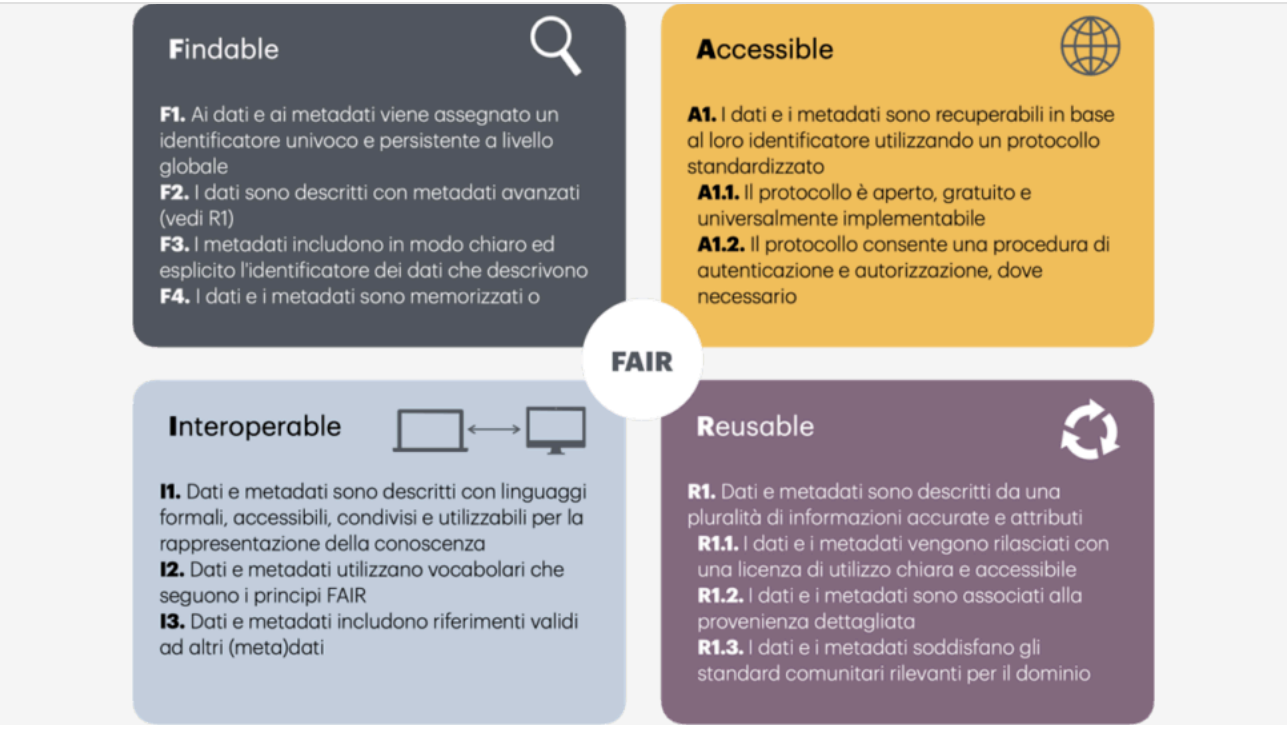


Figura 1: I principi FAIR in pillole.

La natura *machine actionable* dei *dataspaces*, tuttavia, non si estende naturalmente alla definizione degli insiemi di dati e dei loro metadati, che devono essere progettati da umani che sono esperti della specifica area della conoscenza su cui insiste il dataspace. È opportuno ricordare che **le macchine non possono estrarre conoscenza di alta qualità da dati di bassa qualità**, secondo il ben noto principio *"garbage-in garbage-out"*.

Anche l'avvento dell'Intelligenza Artificiale Generativa ("GenAI") non cambia la sostanza. L'insieme dei modelli GenAI possono essere certamente strumenti utili nelle mani di chi colleziona nuovi dati per semplificare lavori routinari e sintetizzare riassunti, schemi e talvolta nuovi contenuti (talvolta di qualità discutibile [3]), ma più difficilmente potranno **generare nuova conoscenza**. Anzi, poiché la quantità di dati necessari per mettere a punto modelli GenAI cresce con la complessità dei compiti[1] che gli sono richiesti, la richiesta di dati necessari per la messa a punto dei modelli AI amplifica la richiesta di nuovi dataspaces in tutte le aree della conoscenza.

Data Steward: una figura professionale emergente

Per questo, mentre i compiti routinari nella gestione dei dati sono destinati ad essere sostituiti da sistemi AI più o meno sofisticati, la figura professionale esperto di dati (**"data steward"**) è destinata ad assumere un'importanza sempre maggiore nel flusso di lavoro –dalla pianificazione degli esperimenti, alla raccolta dei dati e la definizione dei metadati– che per essere accessibile dai moderni di sistemi di analisi di dati dovrà essere necessariamente machine actionable, quindi FAIR [4,5].

Il primo corso universitario per data steward in Italia

Nel marzo 2024 ha preso il via all'Università di Torino la prima edizione del **Corso Universitario di Aggiornamento Professionale (CUAP) per Data Steward**, il primo "master" per la formazione di data steward in Italia. Il corso risponde all'esigenza di enti di ricerca ed enti pubblici di incorporare nel proprio organico una nuova figura professionale che si può sommariamente descrivere come un tecnico esperto di dati in forma digitale in grado di comprendere la natura dei dati di una specifica area disciplinare e quindi in grado di dialogare con i ricercatori di quell'area, ma anche capace di pianificarne la gestione valutando gli aspetti tecnologici, etici e legali.

Struttura e durata del corso

Supportato nella sua interezza da un finanziamento della **fondazione Compagnia di San Paolo**, il corso consiste in 320 ore di lezione frontale (40 Crediti Formativi Universitari), erogati in presenza presso il Dipartimento di Informatica, che ne è organizzatore, nell'arco del periodo temporale che da marzo 2024 si estende a novembre 2024.

Risposta del mercato e profilo dei candidati

La selezione di accesso al corso, conclusasi all'inizio di marzo, ha rivelato **una buona risposta di interesse da parte del mondo della ricerca** e, più in generale, della gestione dei dati. A fronte dei 25 posti messi a bando, infatti, sono

pervenute 63 domande di iscrizione da parte di candidate e candidati con profili professionali e provenienze geografiche diverse. L'analisi delle candidature mostra un panorama diversificato per fasce di età, percorsi formativi pregressi, aree disciplinari e ambiti professionali, mostrando un interesse trasversale per la professione di data steward che si estende dai dipartimenti e centri di ricerca universitari ed enti di ricerca all'ambito della consulenza e della ricerca e sviluppo private.

Data Governance Act ora applicativo: così cambia l'economia digitale

22 Settembre 2023

di **Anna Cataleta**

Struttura del corso: formazione completa per data steward

Progettato con il fine di consolidare le competenze delle persone che, nell'ateneo torinese e nell'ecosistema della ricerca regionale, già svolgono il ruolo di data steward, ma anche di formare una nuova generazione di data

steward, il corso si articola in tre macroaree, ognuna consistente in moduli specifici per un totale di 17 insegnamenti specifici.

Aree informatico-archivistica e etico-legale

Le prime due aree, Informatico-archivistica e Etico-legale, hanno lo funzione di riallineare le competenze in ingresso dei e delle partecipanti rispetto alle **nozioni di base sottese al funzionamento delle piattaforme di rappresentazione e archiviazione dei dati della ricerca** (formati, architetture cloud, ambienti virtuali, strumenti semantici, paradigma FAIR) e alle basi etico-legali su cui poggia la loro raccolta e condivisione (DRM, tutela della privacy, regolamenti, integrità della ricerca), senza trascurare la formazione sui principi e le basi della Scienza Aperta (vantaggi e strumenti per aprire la ricerca).

Una particolarità dell'impianto didattico del corso consiste nell'inserimento nell'area informatico-archivistica di **un nucleo di 12 CFU di informatica** (su un totale di 17 CFU di ambito informatico) dedicati allo studio dei meccanismi e degli standard che costituiscono il presupposto per la condivisione dei dati, con la finalità di creare, piuttosto che specifiche competenze pratiche, la consapevolezza degli strumenti necessaria per compiere scelte progettuali e gestionali informate.

Area dei casi di studio

La terza area, dedicata ai casi di studio, illustra gli standard di comunità, le buone pratiche e le infrastrutture negli ambiti specifici delle scienze umane, delle scienze sociali, delle scienze della vita e delle scienze dure.

Corpo docente e collaborazioni internazionali

Per riflettere la specificità dei moduli del corso i docenti sono stati reclutati, oltre che tra il personale di Unito attivo nell'ambito della **Scienza Aperta** (ricercatori, docenti e personale tecnico-amministrativo impegnato sui progetti e presso le

direzioni dell'Ateneo), anche tra il personale di centri di ricerca, quali **CNR, GARR e IIT in Italia e in Europa,** e liberi professionisti che svolgono attività di consulenza sulla gestione dei dati della ricerca. In totale**, i 40 crediti del corso sono erogati grazie al coinvolgimento di 45 docenti italiani ed europei,** di cui alcuni impegnati su contenuti di tipo seminariale in moduli di due o quattro ore, per i quali sono state privilegiate le sinergie con iniziative e infrastrutture attive a livello europeo (EOSC, GO FAIR, Skills4EOSC, Elixir e altre).

Sistema di tutoraggio

A complemento della didattica frontale, si è scelto di includere nelle attività del corso l'apporto di un gruppo di tutor che assistono gli studenti nelle prime due aree, secondo una modalità a sportello e appuntamenti concordati che si estende per 120 ore. L'orario è stato organizzato in modo da **favorire le persone già occupate**, concentrando le lezioni nelle giornate di giovedì e venerdì, e avvalendosi della piattaforma Moodle per la condivisione dei materiali didattici e l'interazione tra docenti e studenti oltre l'orario di lezione. Infine, è stata predisposta un'installazione della piattaforma Harvard Dataverse presso il Centro **HPC4AI** (High-Performance Computing for Artificial Intelligence at the University of Turin) per il training dei partecipanti al corso in una ambiente *sandbox*.

Project work e premio per la FAIRificazione

Per incentivare l'applicazione dei concetti e delle pratiche apprese nel corso, il project work, è stato sostituito con la possibilità di concorrere a un premio per la FAIRificazione di un dataset esistente: per candidarsi al premio i partecipanti al corso dovranno formare dei gruppi che presentino un progetto di FAIRificazione coerente con i principi e gli strumenti studiati durante il corso, a coronamento ideale del loro percorso di formazione.

Un primo bilancio del corso

Nonostante non sia ancora possibile valutare in maniera definitiva l'efficacia del corso, dato che le lezioni termineranno a novembre 2024, è possibile tracciare un primo bilancio positivo facendo leva sulla **frequenza e sulla partecipazione delle/dei discenti**. Un bilancio maggiormente accurato dell'impatto del corso sulla creazione di nuove figure di data steward nell**e realtà pubbliche e private i**nteressate dal corso e sulle politiche dei dati in queste realtà è ancora prematuro, così come l'eventuale riprogettazione, anche parziale, del corso e dei suoi moduli per future riedizioni, sperabilmente estese al livello nazionale.

L'impatto dei data steward sulle infrastrutture di ricerca europee

Specificatamente, crediamo che un corso per la formazione di data steward, già sperimentato e costruito sul concetto europeo di microcredentials e forte della collaborazione con **il progetto Skills4EOSC** [6] finalizzato a produrre materiale didattico nell'ambito della Scienza Aperta, possa costituire le fondamenta per strutturare azioni al livello nazionale di ben più ampia portata sia nell'ambito delle infrastrutture cloud e HPC che nell'ambito dell'Intelligenza Artificiale. Infatti, nella torrida estate 2024, sono ai blocchi di partenza due manifestazioni di interesse per azioni Europee di grande dimensione e importanza a cui sarà necessario rispondere organizzando gli attori in nodi nazionali.

Ruolo dei Data Steward nelle iniziative EOSC

La prima è la manifestazione di interesse per costituire la nuova federazione di nodi EOSC (European Open Science Cloud) infrastrutture [7] che ha fra i suoi obiettivi anche la realizzazione di una infrastruttura federata al livello europeo in grado di ospitare dataspace FAIR provenienti da alter infrastruttura di ricerca più tematiche (nell'ambito della microbiotica, della medicina, del clima, etc.).

Importanza per le AI Factory europee

La seconda intende realizzare un certo numero di **AI Factory al livello europeo**, cioè ecosistemi costruiti intorno a un centro nazionale di supercalcolo che hanno l'obiettivo di creare e sostenere una vera e propria filiera dell'AI, dalla ricerca, al trasferimento tecnologico alla creazione di nuove imprese [8], che avranno necessità di personale formato per gestire i dati, cioè i data steward.

Note

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

[2] Yu-Cheng Tsai. Demystify Transformers: A Guide to Scaling Laws. SageAI: AI and Machine Learning for financials, planning, and beyond. Apr, 2024. **https://medium.com/sage-ai/demystify-transformers-a-comprehensive-guide-to-scaling-laws-attention-mechanism-fine-tuning-fffb62fc2552**

[3] Hicks, M.T., Humphries, J. & Slater, J. ChatGPT is bullshit. Ethics Inf Technol 26, 38 (2024). **https://doi.org/10.1007/s10676-024-09775-5**

[4] Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021). Datasheets for datasets. Communications of the ACM, 64(12), 86-92.

[5] Guizzardi, G. (2020). Ontology, ontologies, and the "I" of FAIR. Data Intelligence, 2(1-2), 181-191.

[6] Skill4EOSC EU project. Skills for the European Open Science Commons: creating a training ecosystem for Open and FAIR science. Last accessed Jul. 2024. **https://www.skills4eosc.eu/**

[7] EOSC Association: Advancing Open Science in Europe. Last accessed Jul 2024. **https://eosc.eu/**

[8] EuroHPC AI Factories. Last accessed Ju. 2024. **https://eurohpc-ju.europa.eu/eurohpc-joint-undertaking-amends-work-programme-incorporate-new-ai-factory-pillar-2024-07-26_en**

[1] Espressa con il numero di parametri. Per il Large Language Model (LLMs), Google DeepMind riporta una richiesta di dati almeno 20:1 rispetto al numero di parametri (tokens:parameters), Facebook riporta un fattore 215:1 per il modello LLama-3 [2].

Valuta la qualità di questo articolo

A

Marco Aldinucci

Dip. di Informatica, Università di Torino e CINI HPC-KTT

Seguimi su

D

Rossana Damiano

Università di Torino