DOI: 10.26314/GARR-Conf21-proceedings-13

Infrastrutture digitali per i dati della ricerca: adozione e personalizzazione del Research Data Management System all'Università Milano-Bicocca

Bonaria Biancu, Alessandro Andretto, Paolo Brambilla

Università degli Studi di Milano-Bicocca

Abstract. BOARD (Bicocca Open Archive Research Data) è il Research Data Management System che l'Università di Milano-Bicocca mette a disposizione di docenti e ricercatori per il deposito e la pubblicazione dei dati della ricerca. Frutto di un'attenta analisi comparativa tra i prodotti esistenti sul mercato, da luglio 2020 la piattaforma è soggetta a intense implementazioni e personalizzazioni, non solo per soddisfare le esigenze dell'istituzione ma anche per contribuire a restituire un servizio valido all'intera comunità di prodotto

Keywords. Open Science, Research Data, Research Data Management, Principi FAIR

Introduzione

La policy sull'Open Science approvata nel novembre 2019 dall'Università Milano-Bicocca poggia su tre cardini: Open Access per le pubblicazioni scientifiche, Open data per i dati della ricerca, Open Infrastructure per l'utilizzo in modalità open delle infrastrutture di ricerca dell'Ateneo.

Per la gestione dei dati della ricerca, tra le prime azioni che l'Ateneo ha posto in essere per dare seguito a quanto stabilito nella policy, vi sono state l'assunzione di una figura dedicata di data steward e l'adozione di una piattaforma per la gestione, pubblicazione e disseminazione dei dati.

Per la scelta del Research Data Management System (RDMS) si è cercata una soluzione che contemperasse le esigenze concrete dei ricercatori dell'Ateneo, le indicazioni della policy sulla Scienza Aperta e la gestione tecnico-organizzativa del processo di pubblicazione degli open data. L'adozione del RDMS è stata preceduta da una lunga e articolata fase di analisi, nella quale sono state messe a confronto le principali piattaforme di Research Data Management attualmente presenti sul mercato. Assodato che tutte le piattaforme erano compliant ai principi FAIR, in fase di analisi è stato dato molto peso alla presenza di determinate caratteristiche tecniche, valutate come imprescindibili per la scelta finale.

Il risultato di questa approfondita indagine ha portato all'impiego e allo sviluppo di BOARD (Bicocca Open Archive - Research Data; Url: https://board.unimib.it/), il RDMS fornito a Bicocca in modalità Software as a Service da Elsevier.

1. BOARD: caratteristiche della piattaforma

La scelta del RDMS è stata determinata tenendo bene a mente alcune caratteristiche fon-

Online, 7-16 giugno 2021 - ISBN: 978-88-946629-0-0

damentali, considerate dei must have per la piattaforma dei dati di Bicocca.

È innanzitutto importante ribadire che la piattaforma è FAIR compliant: al dataset pubblicato vengono associati un DOI e un numero di versione, in modo da renderlo immediatamente citabile e da poter tenere traccia di eventuali, successive modifiche. Le 15 licenze a disposizione (preset sulle Creative Commons ma possibilità di selezionarne altre, sempre in ambito Open Access/Open Source) permettono inoltre di qualificare la disseminazione dei dati in modo corretto.

La pubblicazione dei dataset è soggetta a un processo di validazione e approvazione, necessario perché i data steward possano verificare la corretta implementazione dei principi FAIR e della policy di Ateneo sull'Open Science, nonché l'allineamento con quanto dichiarato in un eventuale Data Management Plan (DMP).

Altri punti qualificanti per la scelta del repository sono stati l'indicizzazione e la ri-pubblicazione dei dataset di Bicocca già presenti sulle principali piattaforme di disseminazione dei dati della ricerca: attività, queste, che rispondono all'esigenza di non interferire con le abitudini dei ricercatori, ormai invalse nei vari ambiti disciplinari. È necessario evitare di depositare più volte gli stessi dataset, sia per l'ovvio motivo di non moltiplicare i punti di accesso ai dati creando nel tempo inconsistenze, sia per non dover chiedere lo sforzo di ri-caricare dati già depositati su altre piattaforme. In BOARD sono così ricercabili i dataset già resi disponibili altrove (circa 1700 a settembre 2021), mediante la pubblicazione dei loro riferimenti e dei metadati principali, e un puntatore alle sorgenti di provenienza. Le risorse sono state inoltre organizzate in collezioni dipartimentali.

Per quanto riguarda il caricamento dei dati sul repository istituzionale, si è agito sulla personalizzazione del workflow di deposito, a partire dalle impostazioni relative ai connettori verso i servizi di storage forniti dai principali cloud provider. Essenziale è stato l'interfacciamento con i connettori di Google Drive e Microsoft OneDrive, dal momento che Bicocca ha sposato a livello istituzionale l'adozione sia della Google Suite sia di Office365. Ugualmente importante l'integrazione con Microsoft Azure, che risponde all'esigenza di offrire, a quei ricercatori che già utilizzano risorse e laboratori virtuali in cloud, una modalità semplice e immediata per trasferire i propri dati dall'infrastruttura virtuale di Azure a BOARD. Sono inoltre garantiti i connettori a DropBox e Box, ed è supportato il protocollo WebDav per il caricamento di file da server remoti.

L'accesso a BOARD è inoltre integrato con il sistema di autenticazione federata di Ateneo. Oltre ai comuni vantaggi di un accesso federato, questa integrazione permette di agganciare in automatico i profili degli utenti affiliati a Bicocca ed assegnare loro i privilegi del repository istituzionale (100 GB di spazio per dataset, metadati customizzati, affiliazione automatica all'istituzione e al dipartimento).

Infine, ai tempi in cui è stata svolta l'analisi comparativa, il servizio di Elsevier era l'unico, tra i vari competitor analizzati, che includesse uno spazio di archiviazione per la conservazione perpetua di dati e metadati. La long term preservation è garantita tramite un accordo con il DANS (Data Archiving and Networked Services) olandese e permette a Bicocca di evitare la completa dipendenza dal fornitore.

2. Integrazioni e sviluppi

L'adozione di un RDMS richiede un processo di miglioramento continuo, al fine di aumentare la qualità dei servizi già disponibili e di implementare nuove funzionalità con l'obiettivo di rendere il sistema funzionale ed interoperabile, mettendo al centro le esigenze dei ricercatori.

Proprio ai fini dell'interoperabilità è iniziato un percorso che avrà come obiettivo finale l'integrazione tramite web service REST con le fonti dati di Ateneo, in primis le anagrafiche utenti di IRIS (il CRIS di Bicocca). Un primo risultato evidente in questa direzione è la funzionalità di affiliazione automatica, che associa di default il contributor del dataset all'istituzione e al dipartimento di afferenza. Anche in questo caso è valso il principio di non interferire con le abitudini di pubblicazione dei ricercatori, abituati in IRIS a trovare già precompilata la propria affiliazione a livello dipartimentale. Sempre sull'interoperabilità, sono a disposizione le API (e la relativa documentazione) per interrogare i servizi e ottenere informazioni previa autenticazione.

Altro aspetto non secondario è rappresentato dai servizi complementari: è in sviluppo un sistema di reportistica che fornisce pubblicamente l'analisi dei dati di utilizzo di BOARD, mentre delle specifiche aree riservate permettono a utenti e amministratori di approfondire il grado di utilizzo e di disseminazione delle risorse depositate.

Il mantenimento di un alto standard di qualità dei metadati rimane un punto fondamentale: mentre per gli schemi da adottare si è cercato di attingere, per quanto possibile, allo standard Dublin Core, vi è una costante ricerca volta al miglioramento delle informazioni fornite. Tra i metadati personalizzati per Bicocca sono presenti le classificazioni ERC e SSD, mentre si sta lavorando per una possibile integrazione di tassonomie per la classificazione formale dei ruoli dei contributor al dataset.

Ultimo, ma non meno importante, è un generale approccio verso l'internazionalizzazione: BOARD è già registrato in registri europei di repository dei dati, come FAIRsharing (https://doi.org/10.25504/FAIRsharing.DDotIl) e Re3Data (http://doi.org/10.17616/R31NJMUX). In ottica EOSC, invece, l'obiettivo è di diventare data provider dell'infrastruttura europea.

3. Conclusioni: sostenibile e digitale

Guardando alle implementazioni realizzate fino ad oggi e agli sviluppi futuri, è chiaro come il soddisfacimento delle necessità degli utenti rimanga il punto centrale per lo sviluppo di BOARD. Questo approccio si è tradotto in soluzioni pratiche, frutto anche di un confronto iniziato a settembre 2020 con i ricercatori di Bicocca mediante la partecipazione a focus group creati per migliorare la user experience della piattaforma.

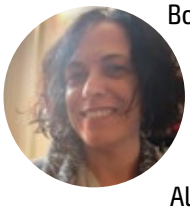
Proprio in ragione di questa attenzione costante verso i docenti e i ricercatori che dovranno effettivamente gestire e pubblicare i dati, si è prestata attenzione a soddisfare le esigenze di enti finanziatori e riviste, ormai orientati sempre più a richiedere di default il deposito dei dati della ricerca in repository qualificati.

La predilezione per l'opzione SaaS incontra il favore delle direttive AgID per le quali la soluzione di Elsevier è qualificata. Proprio questa qualificazione - voluta fortemente da Bi-

cocca - assicura che il servizio di Elsevier sia sviluppato e fornito secondo criteri di affidabilità e sicurezza, considerati necessari per i servizi digitali pubblici. In aggiunta, BOARD garantisce un ulteriore livello di sicurezza sulla disponibilità delle risorse grazie alle API, che permettono di esportare massivamente dati e metadati in qualsiasi momento.

Vale infine la pena di evidenziare un ritorno sugli investimenti che l'adozione di questa piattaforma ha fin qui generato: dal punto di vista dei costi diretti, gli sforzi implementativi del gruppo di lavoro di Bicocca, in collaborazione col team di Elsevier, hanno reso la piattaforma accogliente, aperta, strutturata ma flessibile ed interoperabile, e predisposta alla long term preservation. Di questo gioverà l'intera comunità di prodotto, che oggi trova una soluzione sicuramente più funzionale di quella presentata nel 2020. Dal punto di vista dei costi indiretti, non meno importanti, gli sforzi portati avanti mirano ad evitare inutili duplicazioni dei dati, in modo da ridurre quel caos informativo in cui è purtroppo molto facile incappare.

Autori

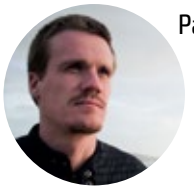


Bonaria Biancu bonaria.biancu@unimib.it

Bonaria Biancu lavora in università dal 2002. Si occupa di gestione dei sistemi per la ricerca e dei servizi web, e di digitalizzazione dei processi di business. Dal 2020 è componente del Consiglio di Amministrazione dell'Università Milano-Bicocca.

Alessandro Andretto alessandro.andretto@unimib.it

Alessandro Andretto è il responsabile dell'ufficio Sistemi Integrati per la Ricerca presso i Sistemi Informativi di Bicocca e si occupa della gestione del CRIS di Ateneo, Open Science, campagne di valutazione e business intelligence per la ricerca.



Paolo Brambilla paolo.brambilla2@unimib.it

Paolo Brambilla è Data Specialist presso l'Università di Milano-Bicocca. Si occupa principalmente di Open Science ed è amministratore della piattaforma BOARD. In precedenza ha lavorato presso diversi enti e istituzioni per lo sviluppo di politiche sui dati aperti.