DOI: 10.26314/GARR-Conf17-proceedings-15

Data Management per la ricerca: un approccio metodologico

Paola Galimberti[1,4], Jordan Piščanc[2,4], Susanna Mornati[3,4]

[1]Università degli Studi di Milano, [2]Università degli Studi di Trieste, [3]4Science, [4]IOSSG

Abstract. Il contesto europeo pone la gestione dei dati della ricerca e la loro corretta archiviazione e conservazione per possibili ed eventuali utilizzi futuri al centro delle politiche sulla research integrity. La possibilità di riprodurre i risultati delle ricerche pubblicate è diventata infatti uno degli elementi essenziali nelle politiche della ricerca della Commissione Europea per poter restituire una credibilità alla scienza, da tempo minata da male pratiche e frodi scientifiche.
Queste tematiche hanno trovato fino ad ora poco riscontro in Italia, sia da parte dell'Ente finanziatore unico della ricerca (il MIUR) sia da parte delle istituzioni.
L'intervento descrive il percorso fatto bottom up da un gruppo di atenei per cercare di fornire ai propri ricercatori politiche e strumenti che avvicinino le nostre pratiche a quelle del resto dei paesi europei.

Keywords. Research Data Management, FAIR data

Introduzione

Il contesto europeo pone la gestione dei dati della ricerca e la loro corretta archiviazione e conservazione per possibili ed eventuali utilizzi futuri al centro delle politiche sulla research integrity. La possibilità di riprodurre i risultati delle ricerche pubblicate è diventata infatti uno degli elementi essenziali nelle politiche della ricerca della Commissione Europea per poter restituire una credibilità alla scienza, da tempo minata da male pratiche e frodi scientifiche.

Le iniziative legate alla gestione dei dati sono numerosissime e sono stati messi a disposizione una serie di toolkit: dalla roadmap sui research data della LERU ai risultati del progetto LEARN, dai Data Management Plan (cartacei o online) alle linee guida FAIR. L'idea è quella di mettere a fattor comune esperienze e strumenti che favoriscano, almeno in Europa, l'adozione di comportamenti coerenti e condivisi.
La necessità di collegare i risultati delle ricerche con i dati che ne stanno alla base è molto sentita nei Paesi "research intensive" dove sono stati organizzati servizi centralizzati sia dal punto di vista delle infrastrutture che da quello delle politiche, linee guida, supporto legale.

In Italia la necessità di indicazioni (tecniche e legali), di formazione (di figure esperte nel Research Data Management) e soprattutto di infrastrutture deputate alla raccolta archiviazione e conservazione dei dati, e' molto sentita dai ricercatori che spesso si trovano ad affrontare le richieste della Commissione Europea senza strumenti, senza indicazioni precise e senza infrastrutture di supporto, mentre il MIUR per ora non ha ancora dato

Conferenza GARR 2017 - The data way to Science - Selected Papers
Venezia, 15-17 novembre 2017 - ISBN: 978-88-905077-7-9

alcun segnale d'interesse rispetto a questo argomento così rilevante.

Nell'attesa che il Ministero elabori politiche e indicazioni sul Research Data Management, un gruppo di lavoro costituito da atenei in cui la sensibilità verso il tema della gestione dei dati della ricerca era maggiormente sentito ha intrapreso un percorso per la definizione di strumenti condivisi (e da condividere) in particolare un modello di policy sul Research Data Management, un modello di Data Management Plan, un'ipotesi di infrastruttura per la raccolta archiviazione e conservazione dei dati.

1. Dalla teoria alla pratica: un percorso di apprendimento

Dalle linee guida di OpenAire si evince che per soddisfare le direttive della Commissione Europea recentemente espresse anche nella EOSC declaration per quel che riguarda la gestione dei dati della ricerca ci si può anche avvalere di Data Repository generici quali Zenodo o FigShare. Numerosi sono anche i repository tematici di dati che possono essere usati per la gestione del ciclo di vita dei dati. Si ripresenta però la stessa situazione di quando, più di 10 anni fa, si iniziò a gestire l'accesso aperto alle pubblicazioni scientifiche e la loro archiviazione negli archivi istituzionali. Per usare archivi generici quali Zenodo o Repository tematici bisogna in ogni caso "accettare" i loro termini d'uso e sottostare ai loro limiti (ad esempio una dimensione massima di spazio per utente usando account gratuiti). Se poi i ricercatori usano archivi diversi per i propri progetti si rischia di avere una frammentazione della produzione dei dati generati.

Si presenta perciò la necessità di istituire e creare anche per le gestione dei dati della ricerca una infrastruttura istituzionale che sia in grado di garantire un'identità e un presidio continuo fornendo altresì servizi a valore aggiunto. In questo caso la messa in produzione di un archivio per i dati della ricerca che gestisca tutto il ciclo di vita (dai dati grezzi, attraverso le diverse versioni, alla descrizione, fino al Data Management Plan) fino ad arrivare al dataset definitivo, risulta ben più onerosa.

Guardando alle esperienze già consolidate in altri paesi europei, dove ci sono buone pratiche e realtà già in atto, ci si rende conto che per gestire efficientemente tutto il ciclo di vita dei dati della ricerca diventa una scelta strategica "fare rete" e instaurare la collaborazione tra più organizzazioni (come ad esempio centri di ricerca e università), come nei Paesi Bassi dove 10 Università e diversi centri di ricerca hanno aderito e usano un Data Research Repository condiviso: DataVerseNL gestito dal DANS.

2. L'esperienza dell'università di Milano e di Trieste

In questa prospettiva le Università di Milano e Trieste insieme ad altri Atenei stanno lavorando a un progetto comune per sperimentare una soluzione di archivio dei dati della ricerca che possa rispondere alle esigenze di comunità scientifiche anche molto diverse fra di loro. Il gruppo di lavoro ha il compito di definire i requisiti di un sistema per la gestione dei dati della ricerca da parte di gruppi disciplinari diversi e di definire, monitorare e validare le attività di test coinvolgendo ricercatori e gruppi di ricerca provenienti da aree diverse. I requisiti saranno analizzati sia dal punto di vista dell'infrastruttura IT e delle risorse umane a supporto, sia dal punto di vista legale che di policy.

L'università di Milano è stata fra le prime in Italia ad adottare un archivio istituzionale per le pubblicazioni e il lavoro sui dati appare il naturale proseguimento di un'attività di supporto all'apertura e alla trasparenza dei risultati della ricerca che si è consolidata nel corso degli anni.

Trieste sarà Capitale Europea della Scienza ESOF 2020 e per l'Università di Trieste l'adozione di una solida soluzione di archiviazione dei dati della ricerca diventa di importanza strategica. L'Università di Trieste è anche coinvolta in UnityFVG, progetto di cooperazione tra gli Atenei del Friuli Venezia Giulia che prevede tra l'altro anche la condivisione delle risorse e delle best practices della ricerca. Sperimentare perciò una soluzione di archivio dei dati della ricerca porterebbe vantaggio anche agli altri Atenei della Regione e potrebbe coinvolgere un maggior numero di Ricercatori.
L'università di Milano ha inoltre coordinato il gruppo di lavoro che ha portato alla definizione di un modello di Policy per il research data management ora a disposizione dell'intera comunità italiana. Sia Trieste che Milano hanno aderito a IOSSG, il gruppo di lavoro che ha il compito di fornire supporto ai ricercatori italiani sulle questioni legate ai dati della ricerca.

Lo strumento scelto da Milano e Trieste per il progetto pilota è Dataverse.
Si è arrivati a questa scelta dopo aver svolto una serie di interviste nei dipartimenti che hanno messo in luce le esigenze principali dei ricercatori. Si sono considerati una serie di strumenti adatti a soddisfare tali esigenze e la scelta è caduta su Dataverse perché è un software open source, che può contare su una comunità di sviluppatori ampia e su un grandissimo numero di utilizzatori in tutto il mondo.

Il gruppo di lavoro si avvale del supporto tecnico di 4Science, azienda specializzata nel fornire soluzioni open source per la ricerca, con cui si cercherà di tradurre in pratica le esigenze espresse nei due atenei.

Dataverse è un'applicazione web open source per condividere, conservare, citare, esplorare e analizzare i dati della ricerca, sviluppato dall'Institute for Quantitative Social Science della Harvard University. Permette di mettere a disposizione della comunità scientifica i propri dati, aumentandone la visibilità. Facilita inoltre il riutilizzo dei dati stessi e, di conseguenza, la replicabilità delle ricerche.

Tramite Dataverse, i ricercatori possono organizzare, condividere e conservare i propri dati, corredati di metadati descrittivi, possono gestirne le diverse versioni e attraverso citazioni formali ricevere credito (citazioni) e adempiere alle richieste dei finanziatori della ricerca. I dati restano sotto il controllo del ricercatore che decide cosa e quando rendere pubblico; la piattaforma è inoltre interoperabile con altre fonti di dati attraverso il protocollo OAI-PMH.

Dataverse mette a disposizione funzionalità avanzate per l'analisi dei dati tabellari, mediante l'integrazione con l'applicazione "TwoRavens". L'interfaccia per le analisi statistiche fornita da TwoRavens è utilizzabile da un'utenza con diversi livelli di competenze statistiche e si configura anche come possibile strumento didattico. Infatti mediante tale interfaccia è possibile visualizzare statistiche di base relative alle variabili che caratterizzano il dataset, effettuare analisi su un sottoinsieme di valori e testare modelli statistici.

TwoRavens è stato concepito proprio come uno strumento per aumentare il numero di utenti in grado di effettuare ragionamenti di tipo quantitativo, mettendo a disposizione funzionalità di analisi che non hanno necessità di grandi infrastrutture e un'interfaccia grafica mediante la quale effettuare le analisi.

Viene supportata anche l'analisi dei dati geospaziali (shapefile) che possono essere esplorati e manipolati attraverso l'integrazione con WorldMap, uno strumento per la visualizzazione e l'analisi di dati geospaziali, sviluppato dal "Center for Geographic Analysis" dell'Università di Harvard.

Attraverso un apposito plug-in, infine, è possibile collegare un repository Dataverse a una specifica rivista gestita attraverso OJS (Open Journal System), in modo da consentire agli autori di sottomettere, oltre all'articolo, anche i dataset ad esso collegati.

3. Conclusioni

Il progetto è appena iniziato ed è presto per poter fare un bilancio. Allo stato attuale sono molto chiare le esigenze di formazione e la richiesta di competenze specifiche all'interno degli atenei che possano supportare i ricercatori nella attività di gestione dei dati.
La sperimentazione che durerà circa sei mesi dovrà aiutare il gruppo di lavoro a capire se Dataverse sia lo strumento in grado di rispondere alle esigenze di trasparenza, ricercabilità, e riusabilità ormai diventate urgenti anche qui da noi.

Autori

Paola Galimberti paola.galimberti@unimi.it
Paola Galimberti si occupa di accesso aperto, qualità dei dati, cura gli strumenti a supporto della valutazione interna ed esterna della ricerca, di etica e integrità della ricerca.

Jordan Piščanc piscanc@units.it
Responsabile IT all'Università di Trieste degli Archivi Istituzionali e sistemi CRIS OpenstarTS, ArTS. Attivo da più di 10 anni nella community DSpace. Il focus principale della sua attività sono gli Open Archive e infrastrutture DSpace-CRIS/GLAM. Segue con molto interesse gli argomenti di Open Science ed è membro di IOSSG.

Susanna Mornati susanna.mornati@4science.it
Direttore Operativo a 4Science S.r.l., ha un'esperienza trentennale in sistemi informativi per la ricerca e la gestione di progetti complessi come l'implementazione di un nuovo Research Information Management System basato su DSpace-CRIS in oltre 60 enti e l'adozione nazionale di ORCiD nel 2015. Susanna ha una reputazione internazionale come sostenitrice dell'Open Access e Open Science ed è membro di IOSSG.