POLITECNICO DI MILANO
GRADUATE SCHOOL
OF BUSINESS

Gestione dei dati digitali della ricerca: linee guida per garantirne la produzione e conservazione sostenibile

Alida Isolani, Scuola Normale Superiore di Pisa
Leonardo G. Mezzina, Scuola IMT Alti Studi Lucca

TUTOR: Fabrizio Pedranzini, Politecnico Di Milano

Project Work
Master in Management dell'Università e della Ricerca
Master SUM VIII edizione

Gli autori ringraziano per il loro supporto, Fabrizio Pedranzini, Susanna Mornati, Marisol Occioni e Fabrizio Luglio.

Da Alida: un ringraziamento speciale a Aldo Tommasin, Larissa Zoni e Michele Fiaschi per avermi dato fiducia; grazie a Valentina Iacomino e Elisa Guidi perché hanno reso unica questa esperienza; grazie a Leonardo Mezzina perché lavorare insieme è stata una bellissima esperienza.

SOMMARIO

1. Abstract

Gli atenei sono sempre di più incentivati a rendere disponibili i risultati delle ricerche, in modo da massimizzare i benefici pubblici con lo scopo di alimentare un circuito virtuoso che, attraverso il libero accesso ai dati e alle pubblicazioni da parte della comunità scientifica e della cittadinanza, produce altra conoscenza.
La gestione dei dati di ricerca alla fine del ciclo di vita di un progetto di ricerca è sicuramente una problematica aperta che rende necessaria l'applicazione di soluzioni tecniche specifiche per tipologia di progetto. Essendo queste soluzioni tecniche peculiari dei singoli progetti e dei tipi di dato all'interno dei vari domini accademici, abbiamo pensato di contribuire alla soluzione utilizzando un approccio di tipo gestionale. L'obiettivo del nostro Project Work è quindi quello di investigare su come garantire la sostenibilità della produzione dei dati digitali della ricerca attraverso la definizione di un framework ed eventualmente di linee guida che disciplinino ad esempio l'utilizzo di specifici strumenti di ateneo. Il caso di studio preso in analisi riguarda i molti progetti di ricerca ormai conclusi della Scuola Normale che risalgono al periodo tra il 1999 e il 2013 e che hanno prodotto applicazioni web divenute obsolete, non più aggiornate né aggiornabili ma molto consultate e di elevato valore scientifico. In questo senso, la Scuola ha permesso di evidenziare le principali difficoltà e problematiche relative alla gestione e conservazione di dati prodotti dalle ricerche.
Il nostro lavoro introduce le problematiche relative alla gestione sostenibile dei dati della ricerca attraverso un processo incrementale, descrivendo dapprima il caso concreto della Scuola Normale (Capitolo 2).
Nel Capitolo 3 abbiamo analizzato le raccomandazioni della Comunità Europea in merito alla gestione sostenibile dei dati digitali delle ricerche descrivendo il caso dei Data Pilot in Horizon 2020 e il Data Management Plan.
A questo punto ci siamo domandati in quali fasi del processo della ricerca degli atenei è possibile inserire attività relative alla creazione e gestione sostenibile dei dati prodotti dalle ricerche e quindi nel Capitolo 4 è stato descritto il processo della ricerca e la catena del valore dei progetti. Ci siamo concentrati sull'aspetto della realizzazione di software specifici concludendo che nonostante a livello nazionale non ci sia molto, esistono alcuni repository di dati sviluppati proprio con l'idea di fornire strumenti per la conservazione sostenibile dei dati: Zenodo e Dataverse (Capitolo 5).
Nel Capitolo 6 abbiamo riportato i risultati delle interviste fatte a diversi atenei italiani per meglio comprendere quanto è sentita la questione relativa alla gestione e conservazione sostenibile dei dati prodotti dalle ricerche all'interno delle Università italiane.
Il Capitolo 7 descrive il nostro framework e le raccomandazioni che secondo noi sarebbe utile che ogni ateneo mettesse in campo per rendere sostenibile negli anni la produzione digitale della ricerca e il Capitolo 8 riporta le considerazioni conclusive e le possibili attività da sviluppare in futuro.

2. Introduzione al problema

Com'è cominciato tutto: il caso SNS

La Scuola Normale organizza la ricerca all'interno delle proprie strutture distribuendo le attività tra centri, laboratori e gruppi di ricerca. Queste strutture sono incentivate ai principi dell'accesso aperto alla letteratura scientifica e alla libera diffusione dei risultati delle ricerche prodotte al proprio interno.

Il nostro caso di studio muove esattamente dai dati digitali realizzati nell'ambito di diversi progetti di ricerca da alcune di queste strutture che negli anni hanno terminato la loro attività e hanno lasciato in eredità all'ateneo i risultati delle proprie ricerche. Gli studi svolti da queste strutture rientrano nell'ambito delle digital humanities: hanno infatti prodotto, tra le altre cose, applicazioni web di supporto alla catalogazione di beni culturali, archivistici e storico-artistici, edizioni critiche digitali, digital library, XML di testi rari e preziosi e sistemi di ricerca full text.

Esempi di alcuni prodotti (in realtà ve ne sono molti altri, ma questi sono stati il *casus belli*) realizzati dalle strutture di ricerca della Scuola sono i seguenti:

- **BiViO - Biblioteca Virtuale Online**
  - *url: bivio.filosofia.sns.it*
  - Anno di inizio 2006, incrementato con testi nuovi fino al 2012.
  - Finanziato con finanziamento.
  - Lo scopo è quello di orientare ricerche filosofiche, storiche, storico-artistiche, filologiche alla costituzione di una biblioteca virtuale on line, capace di offrire testi rari nelle edizioni e traduzioni più significative, rese consultabili da adeguati sistemi informatici, garanti di ricerche a vari livelli, dai più semplici, come frequenze di parole, ai più sofisticati in grado di interrogare il contenuto.
- **La biblioteca ideale di Giordano Bruno**
  - *url: bibliotecaideale.filosofia.sns.it*
  - Anno di inizio 2000.
  - Finanziato nell'ambito delle celebrazioni per il quarto centenario della morte di Giordano Bruno.
  - Il progetto ha previsto la digitalizzazione e la taggatura mediante una codifica XML nella versione TEI delle opere latine e volgari di Giordano Bruno e la sua biblioteca. La marcatura è stata pensata per rappresentare adeguatamente i diversi e complessi fenomeni presenti nel testo.
- **La Bibbia nel '500**
  - *url: bibbia.filosofia.sns.it*
  - Anno di inizio 2006.
  - PRIN 2006-2008.
  - Lo scopo del progetto è stato quello di permette la consultazione di edizioni significative della Bibbia, riprodotte in formato immagine e, in alcuni casi, rese disponibili in formato testo ricercabile, ma anche testi di filologia, opere di riformatori italiani e le documentazioni inerenti a importanti processi per eresia, in modo da offrire uno spaccato che, pur non potendo essere esauriente, si è proposto comunque di

ritrarre nei suoi tratti principali la vivacità di un processo di cambiamento del rapporto dell'uomo con il testo sacro della religione cristiana.

- **Imago Historiae**
  - *url: imagohistoriae.filosofia.sns.it*
  - Anno di inizio 2008.
  - PRIN 2006-2008.
  - Il fine è stato quello di realizzare una biblioteca digitale della storiografia italiana dell'età umanistico-rinascimentale. Sono state acquisite, in formato immagine, ad una risoluzione di 300 dpi, più di 50000 pagine, per un totale di 48 opere, visualizzabili seguendo un percorso tematico (Percorsi); o l'elenco delle opere (Opere), ordinate cronologicamente secondo la data di compilazione; o scorrendo la lista per autore e titolo; o ancora effettuando una ricerca sui metadati, di cui ciascuna unità catalografica trattata, o componente di unità, è stata corredata, secondo lo schema MAG 2.0.1. Ciascun testo è fornito di un proprio indice elettronico, conforme all'originale, che facilita l'accesso alla consultazione delle sue singole parti (tomi, libri, capitoli, lettere dedicatorie, ecc.) e di una scheda di introduzione, che lo contestualizza storicamente e che specifica le caratteristiche materiali del libro presentato.
- **Giorgio Vasari. Le vite - Edizione Giuntina e Torrentiniana**
  - *url: vasari.sns.it*
  - Anno di inizio 1999.
  - Finanziamenti interni.
  - Questo progetto ha previsto la marcatura XML TEI dei testi de Le vite, nell'edizione Giuntina e Torrentiniana e la realizzazione di un sito per permettere la ricerca testuale attraverso un motore di ricerca sviluppato in SNS.
- **La fortuna visiva di Pompei - Archivio di immagini e testi dal XVIII al XIX secolo**
  - *url: pompei.sns.it*
  - Anno di inizio 2002.
  - PRIN 2002, BRICKS 2004-2007, CARARE 2010-2013.
  - Le attività di ricerca e analisi condotte sono supportate da un archivio e da una biblioteca digitali, progettati con la duplice funzione di offrire ai ricercatori uno strumento per l'archiviazione e la gestione di informazioni complesse e risorse tipologicamente diverse (testi e immagini), e di rendere consultabile per tutti attraverso un sito web un patrimonio culturale e artistico, spesso raro e difficilmente accessibile.
- **Monumenta Rariora**
  - *url: mora.sns.it*
  - Anno di inizio 1999.
  - PRIN 2006, finanziamenti interni.
  - Il progetto ha portato alla realizzazione di un sistema informatico per il trattamento integrale delle opere e dei documenti riguardanti la fortuna della statuaria antica. Il sistema implementato consente la consultazione e la ricerca interrelata tra informazioni contenute in un data base e quelle desumibili dalla fruizione di testi contenuti in una digital library, trattati secondo procedure di Information Retrieval (motore di ricerca testuale).

I dati digitali realizzati nell'ambito dei progetti sono ancora molto utilizzati da accademici, studenti e ricercatori. Le statistiche di accesso infatti ci dicono che nel 2019, ad esempio, un

applicativo web come bivio.filosofia.sns.it ha avuto da solo oltre 800 visitatori diversi al mese e mediamente gli altri progetti sono acceduti da oltre 3000[1] utenti al mese.

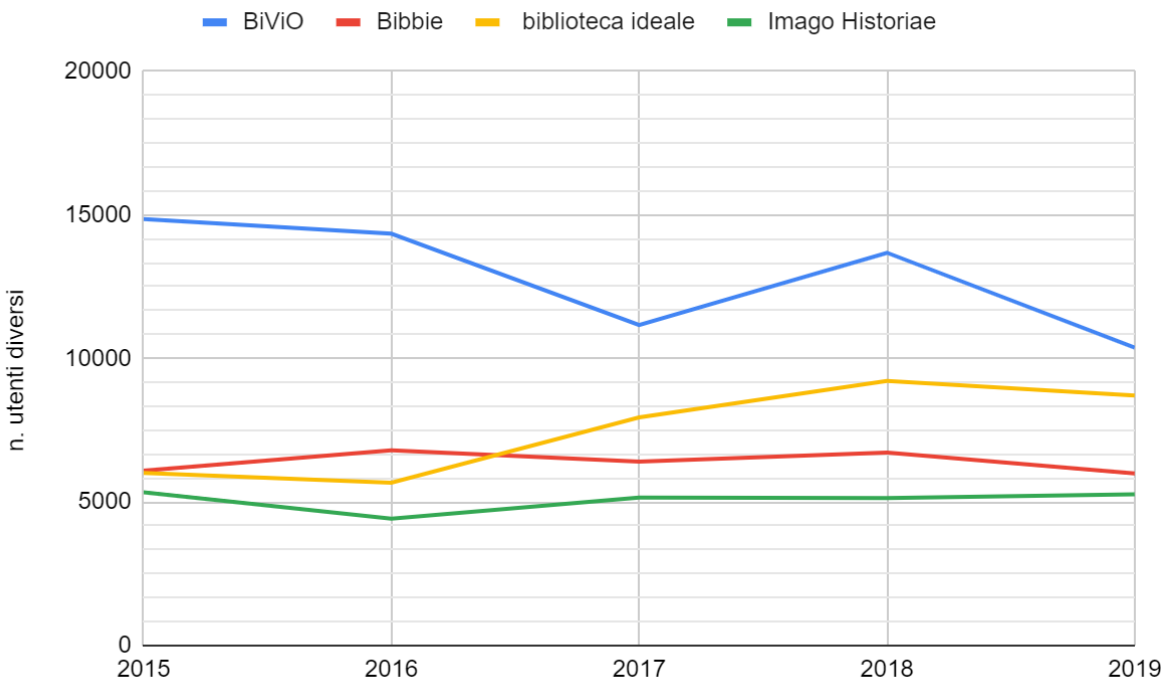

Figura 1 - Andamento degli accessi dal 2015 al 2019

Si può quindi intuire come questi lavori continuino negli anni ad avere una valenza scientifica rilevante ma a causa dell'impossibilità di aggiornamento legato anche all'utilizzo di tecnologie proprietarie che nel tempo sono divenute obsolete, risultano di difficile gestione e conservazione sollevando evidenti problemi di sicurezza informatica e di mancato allineamento con le normative attuali.

Le azioni che sono state messe in campo dall'ateneo per permettere la sopravvivenza di questi applicativi non sono state incardinate all'interno del processo di produzione della ricerca e pertanto risultano isolate e non sistematizzate. Ciò significa che questi prodotti hanno bisogno di più respiro, ma niente è stato ancora fatto per i nuovi progetti di ricerca e quindi, inevitabilmente, i prodotti di queste ricerche affidano la propria sostenibilità nel tempo alla sensibilità dei singoli ricercatori anziché a un'azione strategica dell'ateneo.

Cosa intendiamo per dati digitali della ricerca

Il consorzio OpenAIRE (1) definisce i dati della ricerca come i fatti, le osservazioni, le immagini, i risultati di elaborazioni informatiche, le registrazioni, le misurazioni o le esperienze su cui argomenti, teorie, test o ipotesi della ricerca sono basati. I dati possono essere numerici, descrittivi, visuali o tattili, possono essere grezzi o lavorati e possono essere mantenuti in qualsiasi formato o media (2).

[1] Statistiche di accesso ai portali ottenute mediante AWStats: https://awstats.sourceforge.io/

La Queensland University of Technology considera dato ogni elemento del ciclo di vita della ricerca prodotto durante le fasi di pianificazione, archiviazione, conservazione, accesso, riutilizzo e smaltimento (3). Nella Figura 2 è riportato uno schema grafico dei dati generati durante il processo della ricerca.
Tutti i ricercatori lavorano con i dati, ma ciò che viene definito “dato” dipenderà dalla propria disciplina (4). In particolare:

- lo studioso di discipline umanistiche potrebbe riferirsi alle fonti o ai testi principali;
- se la ricerca riguarda le scienze sociali, si penserà in termini di risultati del sondaggio, interviste e statistiche;
- uno scienziato probabilmente considererà i risultati dei propri esperimenti e delle osservazioni.

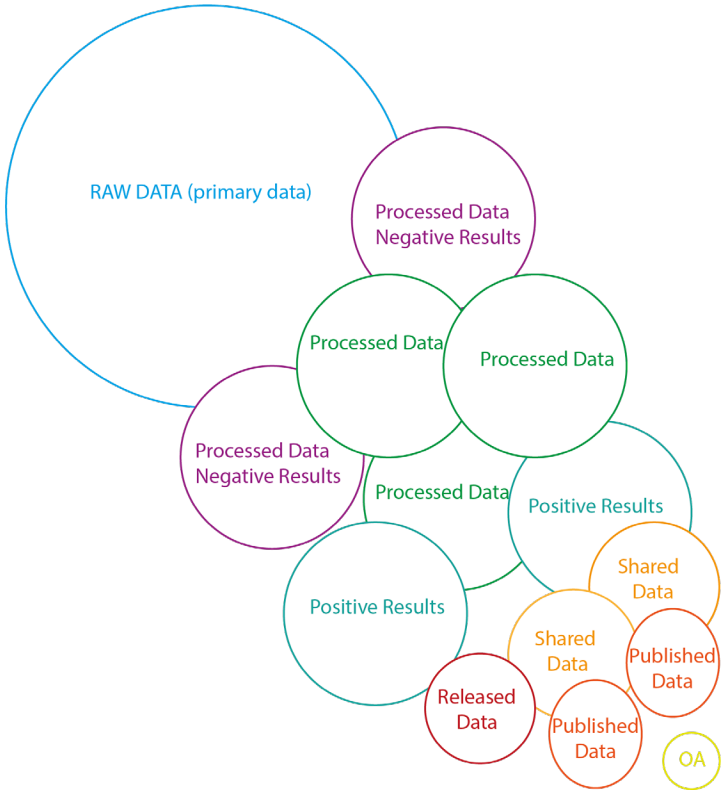

Figura 2 - Dati dal processo della ricerca (5)

Nel seguito del documento adotteremo la definizione di dati della ricerca fornita dal consorzio OpenAIRE. Tale definizione ci sembra abbastanza generica per tutti i tipi di dato della ricerca tuttavia il nostro focus sarà sui dati della ricerca che sono disponibili in formato digitale che chiameremo dati digitali della ricerca.

Cosa succede quando finiscono i progetti di ricerca?

Quello che risulta da recenti studi è che ogni anno grandi quantità di dati prodotti dalla ricerca e dalla sperimentazione vanno perduti. In un recente articolo (6), un team di scienziati canadesi ha studiato la probabilità che un set di dati di ricerca continui ad essere reperibile negli anni successivi alla pubblicazione.
Il report prende in considerazione degli articoli pubblicati tra il 1991 e il 2011 che contengono dati morfologici di piante o animali e che come metodologia di analisi utilizzano un metodo

statistico di classificazione degli individui denominato DFA. La base dati di articoli ricavata mediante il motore di ricerca Web of Science aveva un totale di 516 articoli. Per richiedere agli autori i dati analizzati all'interno degli articoli, sono stati utilizzati gli indirizzi di posta elettronica riportati negli articoli come punto di contatto. Tuttavia se nell'articolo non era specificato un indirizzo di posta elettronica o se la casella di posta elettronica risultava non funzionante gli autori del report hanno provato a cercare indirizzi di posta elettronica funzionanti anche online. La probabilità di riuscire a trovare indirizzi di posta elettronica funzionanti decrementa di circa il 7% all'anno.
Risultati interessanti del report sono:

- esiste una forte correlazione negativa tra l'età dell'articolo e la probabilità che i dati utilizzati siano ancora esistenti. Nella Figura 3, la curva ottenuta mediante una regressione di Poisson mostra come la probabilità che i dati siano ancora esistenti, probabilità condizionata dal fatto di aver ottenuto una risposta utile al messaggio di richiesta, decresce del 17% ogni anno. Le risposte negative sulla disponibilità dei dati includono il fatto che i dati siano stati persi (memorizzati su un dispositivo rubato), memorizzati da qualche parte in un luogo distante inaccessibile all'autore, oppure memorizzati su hardware datato come floppy disk o ZIP.

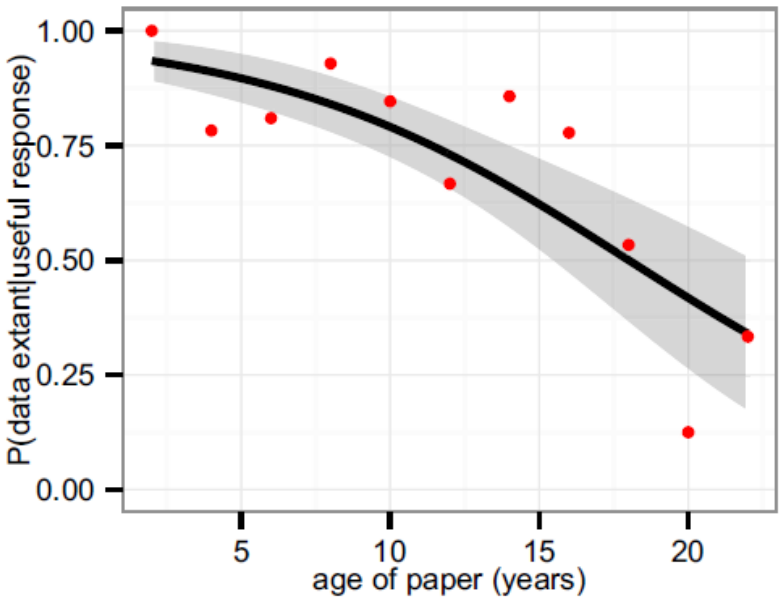

Figura 3 - Probabilità prevista che i dati siano esistenti col passare del tempo

- si potrebbe pensare che gli autori di articoli meno recenti rispondano meno facilmente, ma dai risultati invece si evince come la probabilità che si riceva una risposta o una risposta utile non dipende dall'età dell'articolo (Figure 4 e 5).

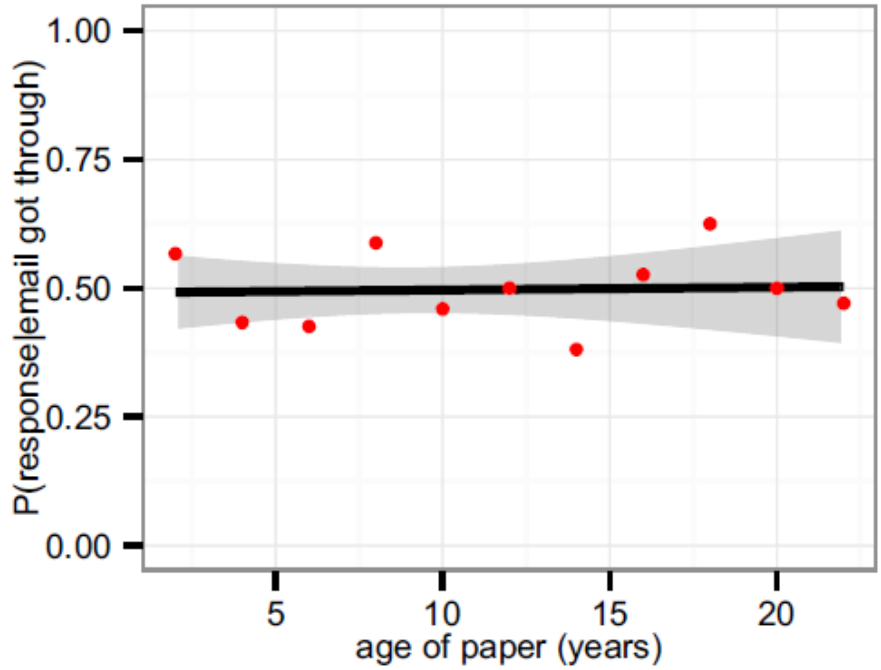

Figura 4 - Probabilità prevista di ricevere una risposta utile se gli autori son stati contattati con successo via e-mail

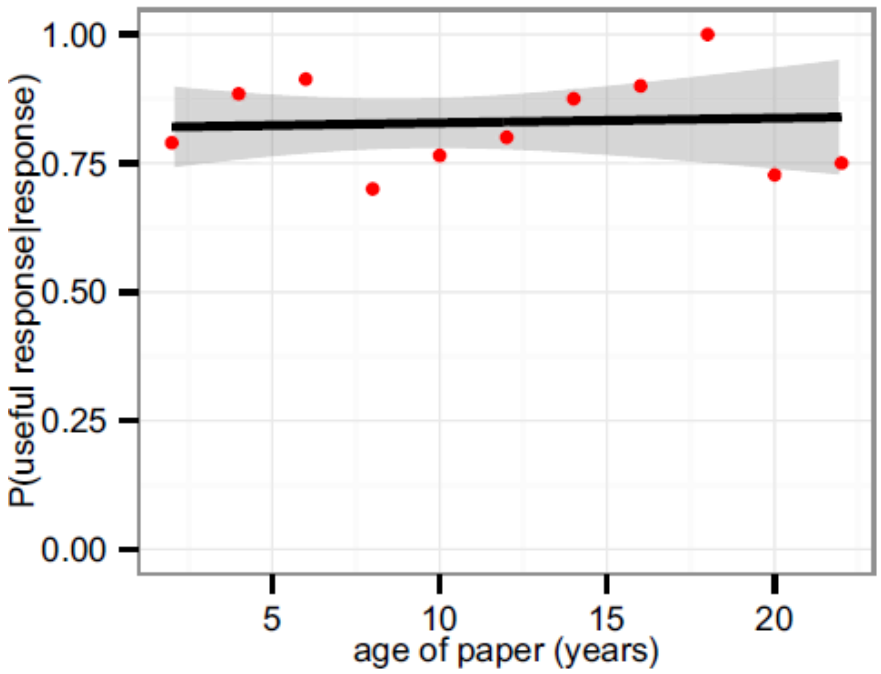

Figura 5 - Probabilità prevista di ricevere una risposta utile se si riceve una risposta

Sulle motivazioni personali che spingono o meno un autore a condividere i risultati della propria ricerca è possibile consultare (7) che si concentra sui dati digitali della ricerca nel campo delle scienze sociali. Gli autori sulla base della bibliografia esistente costruiscono delle ipotesi con peso negativo o positivo all'attitudine di condividere dei dati. Tali ipotesi sono:

- H1: i possibili benefici percepiti di carriera potrebbero influenzare positivamente l'attitudine degli autori di condividere i dati.
- H2: i possibili rischi percepiti alla carriera potrebbero influenzare negativamente l'attitudine degli autori di condividere i dati.
- H3: lo sforzo percepito a rendere disponibili i dati potrebbe influenzare negativamente la propensione degli autori alla condivisione dei dati.
- H4: lo sforzo percepito a rendere disponibili i dati potrebbe influenzare negativamente gli autori di condividere i dati.
- H5: la propensione stessa alla condivisione dei dati da parte degli scienziati potrebbe influenzare positivamente gli autori alla condivisione dei dati.
- H6: la percezione di avere subito a disposizione un repository dove inserire i dati potrebbe influenzare positivamente gli autori alla condivisione dei dati.
- H7: la percezione della pressione imposta da enti finanziatori potrebbe influenzare positivamente gli autori alla condivisione dei dati.
- H8: la percezione della pressione imposta dalle riviste potrebbe influenzare positivamente gli autori alla condivisione dei dati.
- H9: la percezione della pressione di tipo amministrativo potrebbe influenzare positivamente gli autori alla condivisione dei dati.

Sulla base di queste 9 ipotesi è stato preparato un survey inviato a 2.285 partecipanti e sono state ricevute 361 risposte. Le risposte ricevute tenevano conto di personale con demografia variabile, ad esempio: personale tenured (51%), personale in tenure track ma non tenured (16%) e personale non in tenure track (22%). Il risultato dell'analisi è riportato in figura, le ipotesi non supportate dai dati sono quelle collegate con linee tratteggiate.

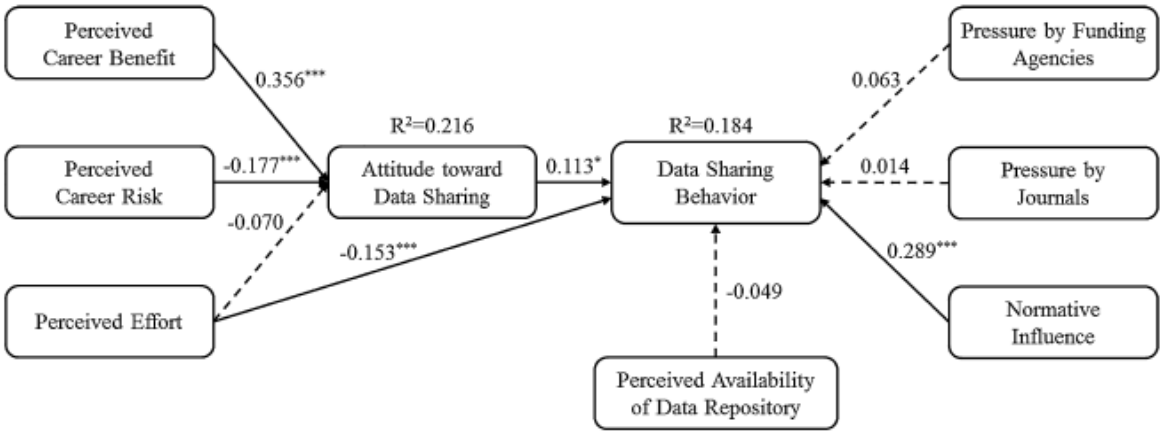

Figura 6 - Framework delle ipotesi supportato dal survey

Le motivazioni degli studiosi di scienze sociali a condividere i dati sono quindi guidate da motivazioni personali come il fatto di percepire un beneficio per la carriera, oppure un rischio e dalle motivazioni personali alla condivisione. Un altro aspetto fondamentale sembra essere l'influenza amministrativa. Interessante è il fatto che la disponibilità un data repository non venga visto come un ostacolo alla condivisione dei dati.

Il risultato di questi due studi è molto interessante: il primo ci dice che i dati della ricerca subiscono un declino col tempo, il secondo che la disponibilità a condividere dati è principalmente un fatto di natura personale (come evidenziato anche dal risultato sulla probabilità di ottenere una risposta ai messaggi di posta elettronica Figure 4 e 5 che rimane

costante col tempo) e può essere migliorato se imposto da una regolamentazione a livello amministrativo.
Infine, il fatto di non avere a disposizione un data repository di ateneo non sembra essere un ostacolo alla condivisione. Immaginiamo che ciò dipenda dalla disponibilità di molti mezzi per condividere e dall'esistenza di diversi data repository aperti.
Queste considerazioni ci suggeriscono che un'azione che miri a rendere organica e sistematizzata la produzione e conservazione dei prodotti digitali della ricerca avrà la necessità di far leva da un lato sugli aspetti amministrativo-organizzativi dell'ente e dall'altro sul fronte della sensibilizzazione del personale di ricerca verso le tematiche dell'accesso aperto e del riuso.

3. Le raccomandazioni della comunità europea sulla gestione dei dati

In questo capitolo vedremo ciò che suggerisce la comunità europea in merito alla gestione dei dati digitali della ricerca, descrivendo in particolare le particolarità dei progetti Open Research Data Pilot e il concetto di Data Management Plan, per il quale descriveremo altri modelli di template principali oltre a quello messo a disposizione dalla Comunità Europea. Lo scopo è quello di delineare i tratti comuni di un documento strutturato e abbastanza generico, utile alla gestione dei dati digitali della ricerca.

Il caso dei Data Pilot Horizon 2020

La comunità europea definisce dei progetti nell'ambito del framework H2020 denominati Open Research Data Pilot (8) abbreviato ORD con l'obiettivo di rendere i dati di ricerca aperti, "*as open as possible, as closed as necessary*".
Il *Pilot* ha due pilastri principali: sviluppare un piano di gestione dei dati (*Data Management Plan*) e, se possibile, fornire un accesso FAIR ai dati.
Il coordinatore di un progetto ORD deve quindi assicurarsi di:

1. creare e mantenere aggiornato un Data Management Plan (DMP);
2. depositare i dati della ricerca in un repository;
3. garantire che terze parti possano accedere, utilizzare, riprodurre e diffondere liberamente i dati della ricerca;
4. rendere chiaro quali strumenti siano necessari per usare i dati grezzi e validare i risultati della ricerca oppure fornire direttamente questi strumenti.

Nell'ambito dell'iniziativa ORD, il consorzio OpenAIRE offre supporto ai ricercatori relativamente alla gestione dei dati della ricerca e alla stesura del DMP sia attraverso siti web che rendono disponibile documentazione e FAQ mediante webinars con slide e schede informative.
Il consorzio OpenAIRE (1) sostiene e promuove, attraverso un'organizzazione umana e digitale, la gestione efficiente dei dati, l'*open access* e l'*open science* in genere. Questo sforzo è allineato con il piano della commissione europea di creare un *Open Science Cloud* a livello europeo (*European Open Science Cloud, EOSC* (9)).
EOSC è organizzata come rete di infrastrutture di ricerca federate, per favorire l'accesso e l'uso dei B*ig Data* nelle scienze e nei servizi pubblici.
La commissione europea all'interno del framework Horizon 2020 incoraggia l'accesso aperto e il riutilizzo dei dati digitali della ricerca secondo il principio *findable, accessible, interoperable and reusable*, indicato con l'acronimo FAIR.
Riprendiamo per *completezza la definizione introdotta nel Capitolo 1:*
si definiscono dati della ricerca come i dati nella forma di fatti, osservazioni, immagini, risultati dell'elaborazione di programmi, registrazioni, misure o esperienze sui quali vengono basati: un argomento, una teoria, un test, un'ipotesi o qualunque altro output della ricerca.
I dati possono essere numerici, descrittivi, visuali o tattili. Possono essere dati grezzi, puliti o il risultato di una elaborazione e possono essere memorizzati in qualunque formato o mezzo. Il nostro focus sarà sui dati che sono disponibili in formato digitale che chiameremo ***dati digitali della ricerca****.*

Con gestione dei dati (Data Management) si intendono tutti i processi e le attività necessarie per gestire i dati attraverso il ciclo di vita della ricerca al fine di renderli disponibili agli utenti e a scopi di ricerca presenti e futuri. Chiaramente i dati digitali della ricerca variano tra i diversi domini accademici e sono necessarie differenti attività per la gestione dei diversi dati. Tuttavia tutti i domini accademici hanno in comune il fatto che, per rendere i dati riutilizzabili, questi devono essere sempre accompagnati da metadati.
I metadati sono informazioni sui dati di ricerca; il loro scopo è di descrivere i dati in modo tale da abilitare altri nella ricerca. Si dovrebbero utilizzare informazioni strutturate per descrivere lo scopo, l'origine, i riferimenti temporali, le locazioni geografiche, il creatore, le condizioni di accesso e i termini di utilizzo delle collezioni di dati. Esistono già questi standard per diverse discipline (10) tuttavia se non sono disponibili standard, è possibile utilizzarne uno come il Dublin Core (11) e lo schema Datacite (12). Oltre ai metadati, per riuscire a utilizzare al meglio i dati digitali della ricerca potrebbe essere necessario fornire ulteriore documentazione e strumenti per l'accesso e l'analisi dei dati: dovrebbe pertanto essere archiviato l'intero pacchetto comprensivo di dati, metadati, documentazione e strumenti software.
In breve, la raccomandazione fondamentale, è quella di archiviare tutto il necessario per permettere di replicare uno studio, incluse le versioni intermedie e processate dei dati, quando queste versioni sono frutto di analisi della ricerca. Si potrebbe omettere di archiviare le versioni intermedie dei dati quando rigenerare la versione intermedia è meno dispendioso che archiviarla. Tuttavia questa omissione deve essere documentata insieme ai metodi necessari per ricreare i dati mancanti.
La commissione europea infine incoraggia gli autori a detenere i propri diritti (diritti d'autore e proprietà intellettuale) e a garantire licenze adeguate (eventualmente anche agli editori) utilizzando il meccanismo di licenza offerto da Creative Commons (13). Le licenze Creative Commons sono infatti è un utile alleato dal punto di vista legale per fornire accesso aperto nel senso più ampio possibile.

Il Data Management Plan

Introduzione

Ci sono molti vantaggi nella gestione e condivisione dei dati:

- è possibile trovare e comprendere i dati quando devi usarli;
- c'è continuità se il personale del progetto abbandona o si uniscono nuovi ricercatori;
- si evitano inutili duplicazioni, ad esempio ri-raccolta o rielaborazione dei dati;
- i dati sottostanti le pubblicazioni vengono mantenuti, consentendo la validazione dei risultati;
- la condivisione dei dati porta a una maggiore collaborazione e promuove la ricerca;
- la propria ricerca diventa più visibile e ha un impatto maggiore;
- altri ricercatori possono citare i tuoi dati in modo da ottenere credito.

In questo senso un documento comune in cui i ricercatori specificano quali dati saranno gestiti e condivisi all'interno di un progetto, ci può aiutare: questo è il *Data Management Plan* (DMP). Il DMP è uno strumento fondamentale per definire e pianificare i dati e gli strumenti che saranno utilizzati, forniti e condivisi in un progetto. È utile sia nel caso in cui i ricercatori usino nuovi dati sia nel caso di riuso di dati esistenti, specialmente quando gli enti di ricerca coinvolti sono diversi e fisicamente distribuiti su più territori.

Per ogni progetto di ricerca dovrebbe essere fornito un singolo DMP e non un DMP per ogni insieme di dati.
Nei progetti ORD la prima versione del DMP dovrebbe essere fornita entro i primi sei mesi del progetto come un deliverable ufficiale. In questa prima fase potrebbe essere coinvolto lo staff di supporto dell'ente con ruolo di *Principal Investigator* del progetto, ad esempio, per definire il data repository in cui conservare i dati per i futuri accessi.
Richiedendo il DMP nelle fasi iniziali del progetto, i finanziatori si assicurano che tutte le precauzioni e gli strumenti per memorizzare i dati siano stati presi in considerazione. Il DMP è un documento vivo che si evolve e ottiene maggiore precisione e consistenza durante l'arco di vita del progetto. Il DMP dovrebbe essere aggiornato almeno a metà e a fine progetto per renderlo definitivo con la versione dei nuovi dati; può comunque essere utile controllare il DMP durante l'arco di vita dell'intero progetto per verificare che gli obiettivi, dal punto di vista della gestione dei dati, siano stati ben formulati e se necessario rivederli.

Il Data Management Plan in dettaglio

Il DMP non è un documento che ha una struttura fissa e a seconda dei casi potrebbe o meno contenere differenti informazioni. Sono disponibili online diversi template per il DMP che è necessario compilare accuratamente quando si richiede un finanziamento dai diversi enti finanziatori. Come detto le informazioni del DMP sono utili per qualunque tipo di progetto ma per capire meglio quali informazioni siano necessarie, abbiamo scelto di descrivere in questa sezione i tre modelli che ci sembravano più significativi in ordine al dettaglio richiesto. Abbiamo notato che anche i modelli di DMP resi disponibili dai vari enti finanziatori vengono spesso revisionati e aggiornati.

Template Wellcome Trust

Il primo modello che abbiamo esaminato è quello richiesto *Wellcome Trust* (14) .
Wellcome Trust è un ente di beneficenza con sede a Londra; è stato fondato nel 1936 grazie all'eredità lasciata dal magnate americano dell'industria farmaceutica Sir Henry Wellcome, con lo scopo di finanziare la ricerca per migliorare la salute umana e animale. In particolare viene richiesta la compilazione di quattro macro-sezioni:

1. *data and software outputs*;
2. *research materials*;
3. *intellectual property*;
4. *resources required*.

Vediamo ora il dettaglio di ogni macro-sezione.

Template Wellcome Trust: *Data and software outputs*

In questa macro-sezione viene richiesto di indicare:

- **i dati e gli output software generati dalla ricerca**, descrivendo brevemente i tipi di dati e i software generati dalla ricerca, quali dati e software sono importanti per altri ricercatori e potrebbero essere condivisi, i formati e gli standard di qualità dei dati che saranno applicati per abilitare la condivisione efficiente di dati e software. Se non è possibile condividere i dati (per esempio per ragioni etiche o commerciali) in questa sezione ne dovrebbe essere specificato il motivo. I dati dovranno essere condivisi in conformità con standard riconosciuti qualora essi esistano, e in maniera di

massimizzare il collegamento dei dati e la condivisione. Il sito FAIRsharing (15) rappresenta una raccolta completa di standard di rappresentazione dei dati. È necessario fornire nei metadati informazioni dettagliate e quanto più possibile standard per permettere che il dataset sia ricercabile e interpretabile da altri utenti. Il software deve essere condiviso in maniera da permettere agli altri di essere usato effettivamente con appropriata e proporzionata documentazione di utilizzo. Anche i dati che non portano a nessuna scoperta dovrebbero essere condivisi per facilitare il supporto a nuove scoperte.

- **quando si intende condividere i dati**. In questa sezione deve essere specificata la finestra temporale di condivisione di dati e software, eventualmente specificando un periodo di esclusività. Come minimo i dati dovrebbero essere disponibili il giorno della pubblicazione ma è fortemente incoraggiata con le opportune precauzioni, la condivisione di dati con altri ricercatori nel periodo prima della pubblicazione.
- **come rendere disponibili dati e software.** In questa sezione va specificato come potenziali utenti possano scoprire, accedere e riutilizzare dati e software con tutte le condizioni associate
- **i limiti alla condivisione di dati e software.** Potrebbe essere necessario specificare in questa sezione limiti di utilizzo dei dati per preservare eventuale proprietà intellettuale. Le restrizioni devono essere minimizzate il più possibile e formulate chiaramente.
- **come i dati e il software saranno preservati**. In questa sezione va specificato come i dati e il software che abbiano valore di lungo termine siano preservati e mantenuti dopo la fine del finanziamento. Questa sezione è di particolare importanza quando si decide di non voler utilizzare un repository ben definito e di voler memorizzare i dati localmente.

Template Wellcome Trust: *Research materials*

Questa macro-sezione è relativa ai materiali di ricerca, intesa come produzione di materiale fisico risultato della ricerca.

Template Wellcome Trust: *Intellectual property*

La terza macro-sezione è invece relativa alla proprietà intellettuale ed è richiesto di descrivere:

- **che tipo di proprietà intellettuale sarà il risultato della ricerca.** In questa sezione è richiesto di prevedere se potenzialmente durante la ricerca possano nascere nuove proprietà intellettuali, così come potenziali scoperte e brevetti.
- **come si ritiene di proteggere eventuali proprietà intellettuali.** La pubblicazione di dettagli relativi ad un'invenzione può essere problematica, quindi è richiesto di chiarire come le varie pubblicazioni non interferiscano con l'uso di eventuali brevetti.
- **come le proprietà intellettuali saranno usate per pubblici benefici.** I profitti dalle nuove proprietà intellettuali dovrebbero essere secondari, l'obiettivo principale è il bene pubblico.

Template Wellcome Trust: *Resources required*

L'ultima macro-sezione è relativa alle risorse richieste. Descrizione di **figure professionali e competenze richieste. Costi relativi a memorizzazione e computazioni e costi relativi agli accessi.** Relativamente agli accessi è necessario descrivere costi relativi all'accesso ai

dati durante il periodo di vita del progetto, il costo di preparare i dati per gli utenti e per gli altri ricercatori. Per ultimo **costi relativi al deposito dei dati e del software** presso data repository strutturati per dominio accademico o data repository non strutturati (es. Zenodo) e di discutere eventuali problematiche di mantenimento di dati e software oltre il periodo di vita utile del progetto.

ERC Data Management Plan Template

Un secondo modello che abbiamo ritenuto significativo è quello pubblicato dall'European Research Council (ERC) (16), denominato appunto ERC Data Management Plan Template; dal sito è possibile notare che l'ultima versione del modello è aggiornata al 12 aprile 2017.
Il modello è suddiviso in cinque sezioni e un sommario e nell'introduzione viene specificato che "le sezioni [...] dovrebbero descrivere come avete intenzione di rendere i dati di progetto FAIR (Ricercabili, Accessibili, Interoperabili e Riusabili)".
Ognuna delle cinque sezioni di cui è composto il documento, dovrebbe essere completata con un livello di dettaglio appropriato al progetto.

ERC Data Management Plan Template: sommario

La prima parte del template è un sommario nel quale è richiesto di inserire riferimenti e nomi dei dataset (differenti dataset possono essere specificati all'interno di un singolo DMP), origine e dimensione attesa dei dati generati o collezionati, i tipi di dato e i formati.

ERC Data Management Plan Template: descrizione delle sezioni

La sezione 1 è intitolata *Rendere i dati ricercabili* ed è richiesto di specificare per ogni dataset, metadati e identificatori persistenti e univoci come per esempio i DOI (17).
La sezione 2 è intitolata *Rendere i dati accessibili in maniera aperta* ed è richiesto di descrivere:

- quali dati saranno ad accesso aperto e nel caso di dataset non accessibili come mai non viene fornito accesso;
- dove saranno depositati ed eventualmente specificare in quale repository i dati, i metadati associati, la documentazione e l'eventuale codice;
- come i dati possono essere acceduti e se eventuali software, strumenti o metodi per l'accesso sono forniti insieme ai dati.

La sezione 3 è intitolata *Rendere i dati interoperabili* ed è richiesto di descrivere standard e vocabolari specifici utilizzati per codificare dati e metadati in un certo settore disciplinare.
La sezione 4 è intitolata *Incremento del Riutilizzo dei dati* ed è richiesto di descrivere quali dati rimarranno riutilizzabili e per quanto tempo; se è previsto un eventuale embargo; che tipo di licenza specifica è prevista per i dati; procedure specifiche per garantire la qualità dei dati.
L'ultima sezione, la sezione 5 è intitolata *Allocazione di risorse e sicurezza dei dati* ed è richiesto di fare una stima dei costi di progetto per rendere i dati ad accesso aperto e una stima dell'eventuale valore di mantenere i dati ad accesso aperto per un lungo periodo; procedure di backup e recupero dei dati; eventuale trasferimento di dati sensibili per la memorizzazione sicura su repository per preservarli e curali dove con cura dei dati si intendono appunto le attività di gestione necessarie per mantenere i dati di ricerca a lungo termine in modo che siano disponibili per il riutilizzo e la conservazione.

DCC Template

Il terzo e ultimo modello che abbiamo ritenuto significativo è quello denominato "DCC Template" fornito dal Digital Curation Centre.
Il Digital Curation Centre (DCC) è nato nel Regno Unito per fornire a livello nazionale linee guida sui processi di preservazione digitale (assicurarsi che il valore dell'informazione digitale continui a rimanere accessibile ed usabile nel tempo) e cura digitale (la selezione, la manutenzione, la collezione e l'archiviazione dei dati digitali).
Per facilitare la compilazione del template è disponibile un applicativo web denominato dmponline (18). Una volta compilate tutte le sezioni richieste dal sito, è possibile salvare il proprio DMP per editarlo in qualunque momento e stamparlo direttamente in formato PDF. È inoltre disponibile un repository di DMP di esempio scaricabili e consultabili (19). Abbiamo trovato molto interessante il DMP intitolato "*Example poor DMP with pointers*" che mostra un esempio di cattiva compilazione di un DMP presentando i più comuni errori di compilazione e commentando in maniera puntuale ed esaustiva come risolverli.
Il modello è diviso in sette sezioni e in ogni sezione è richiesto di rispondere a domande puntuali.

DCC Template Sezione 1: Raccolta dati

- Che tipo di dati saranno raccolti o creati?
- Come questi dati saranno raccolti o creati?

DCC Template Sezione 2: Documentazione e Metadata

- Che tipo di documentazione e quali metadati saranno forniti insieme ai dati?

DCC Template Sezione 3: Conformità etica e legale

- Come saranno gestite eventuali problematiche a livello etico?
- Come saranno gestite le problematiche del diritto d'autore e di diritti di proprietà intellettuale?

DCC Template Sezione 4: Memorizzazione e Backup

- Come i dati saranno memorizzati e come sarà effettuata la copia di backup durante il periodo di durata della ricerca?
- Come sarà gestito l'accesso ai dati e la sicurezza dei dati?

DCC Template Sezione 5: Selezione e salvaguardia

- Quali dati sono considerati di valore a lungo termine e quindi dovrebbero essere mantenuti, condivisi e/o preservati?
- Che tipo di piani ci sono per preservare i dati a lungo termine?

DCC Template Sezione 6: Condivisione dei dati

- Come saranno condivisi i dati?
- Sono richieste restrizioni sulla condivisione dei dati?

DCC Template Sezione 7: Responsabilità e risorse

- Chi sarà responsabile per la gestione dei dati?

- Quali risorse sono necessarie per ottenere i risultati previsti dal DMP?

Analisi dei punti in comune

Dai tre modelli esaminati si può creare una vera e propria check list delle sezioni essenziali che un DMP dovrebbe avere. Similmente a quanto definito dall'Italian Open Science Support Group (20) una check list essenziale di partenza potrebbe essere:

1. Informazioni con i riferimenti amministrativi del progetto.
2. Informazioni dedicate alla definizione dei dataset e metadata.
3. Informazioni relative alla sicurezza e alla confidenzialità dei dati.
4. Informazioni relative alla condivisione e all'accesso ai dati.
5. Informazioni relative al ciclo di research data management e alla cura dei dati.

Compilando queste informazioni indipendentemente che si utilizzi un modello template o un'applicazione web come guida, dovrebbe essere facile transitare da un modello ad un altro eventualmente aggiungendo le varie informazioni mancanti. Per questo si potrebbe definire una sezione vera e propria che contiene un elenco di responsabili per le varie sezioni, che potrebbero essere coinvolti nel caso vengano richieste informazioni addizionali nella compilazione di uno specifico modello.

Lo sforzo quando si richiede la compilazione di un DMP, dovrebbe essere quello di non farla sembrare come l'ennesimo aggravio amministrativo, magari proponendo dei corsi di formazione che ne evidenziano i benefici anche dal punto di vista di chi fa ricerca.

4. Produzione e conservazione dei prodotti digitali della ricerca

In questo capitolo analizzeremo i processi della "filiera della ricerca" per comprendere se sia possibile (o necessario) includere nella filiera le attività relative alla gestione sostenibile dei dati digitali. Approfondiremo inoltre gli strumenti tecnologici di cui gli atenei si dotano, o che potenzialmente avrebbero facilmente a disposizione in qualità di consorziati del Cineca (21), durante le fasi di produzione e conservazione dei dati della ricerca. In questo modo la nostra idea è di fornire una fotografia della situazione attuale delle Università italiane relativamente alla gestione e conservazione dei dati.

La filiera della ricerca secondo Cineca

La "filiera della ricerca"[2] così come descritta in *U-GOV Ricerca - Il sistema per la gestione della ricerca in ateneo* (22), sintetizza i diversi processi che portano alla creazione del valore rappresentato dai risultati dell'attività di ricerca in un percorso relativamente lineare nel flusso e circolare nella capacità di re-investimento dei risultati.
Il diagramma di Figura 7 rappresenta appunto questa visione d'insieme dei processi facenti parte della filiera.

Le risorse e le competenze

Le attività di ricerca prendono le mosse dall'insieme del patrimonio di competenze e di risorse che gli atenei hanno costituito nel corso delle proprie attività istituzionali. Questo patrimonio è il *capitale* che, se reso operante, funge da motore propulsivo della catena del valore del *sistema Ateneo*.

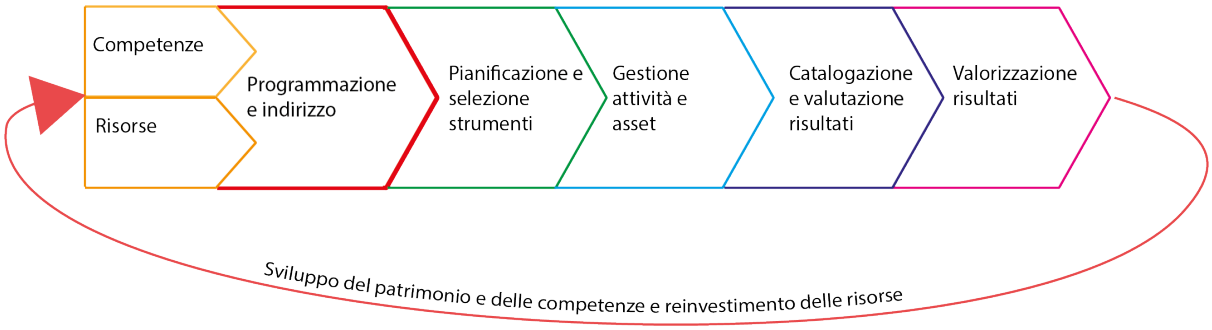

Figura 7 - Diagramma dei processi della filiera della ricerca secondo Cineca

Programmazione e indirizzo

Questo processo è collegato strettamente al perseguimento del mandato istituzionale per la ricerca scientifica e deve permettere quindi la definizione delle linee principali su cui

[2] La "filiera della Ricerca" è l'elemento centrale della più articolata "filiera della Conoscenza" e si colloca tra l'Alta Formazione e processi di Innovazione nella filiera produttiva che ne costituiscono rispettivamente il punto di ingresso e di uscita. L'Università è l'attore protagonista dei primi due passi e interlocutrice privilegiata del mondo imprenditoriale nel trasferimento dell'innovazione verso il mercato. (cfr. L.Nicolais, G.Festinese, Ricerca e innovazione, 2006, p.13 e succ.)

convogliare le risorse disponibili e degli obiettivi attesi in modo da agevolare la valutazione dei risultati.

Pianificazione e selezione strumenti

Sulla base degli obiettivi strategici gli attori possono procedere nella pianificazione dell'impiego delle risorse e delle modalità migliori per il raggiungimento degli stessi. Questo processo prevede la definizione di budget economici e di allocazione delle risorse umane e strumentali alle linee di ricerca, nonché l'identificazione degli strumenti operativi. L'idea è quindi che ciascuna struttura, sulla scorta della dotazione iniziale di competenze e risorse, potrà decidere quali attività pianificare verificandone contestualmente la sostenibilità economica e funzionale in relazione ai risultati attesi dal piano generale.

Gestione attività e asset

La visione d'insieme e gli obiettivi generali una volta declinati in attività operative entrano nei processi gestionali dell'Ateneo. Le attività di ricerca scientifica si realizzano in particolare con lo sviluppo di progetti che rappresentano il motore di questa area funzionale.

Catalogazione e valutazione dei risultati

Obiettivo principale di questo processo è l'alimentazione del catalogo dei prodotti volta a consentire le attività di verifica dei risultati dei processi di ricerca e di valutazione del grado di raggiungimento degli obiettivi. Allo stesso tempo il catalogo costituisce anche l'indice principale della base di conoscenza e diventa il punto di accesso ad essa per tutti i portatori di interesse.

Valorizzazione dei risultati

La valorizzazione del patrimonio di competenze e di risorse (sia umane che strumentali), ottempera sia ai fini istituzionali sia alla necessità di reperimento di ulteriori risorse economiche attraverso la costruzione di un'offerta verso il mercato della conoscenza.
Questo processo traccia le attività, gli strumenti, le risorse e le competenze interne e favorisce l'apertura e la diffusione dei dati e delle informazioni per la pubblicazione e la valorizzazione dell'eccellenza.

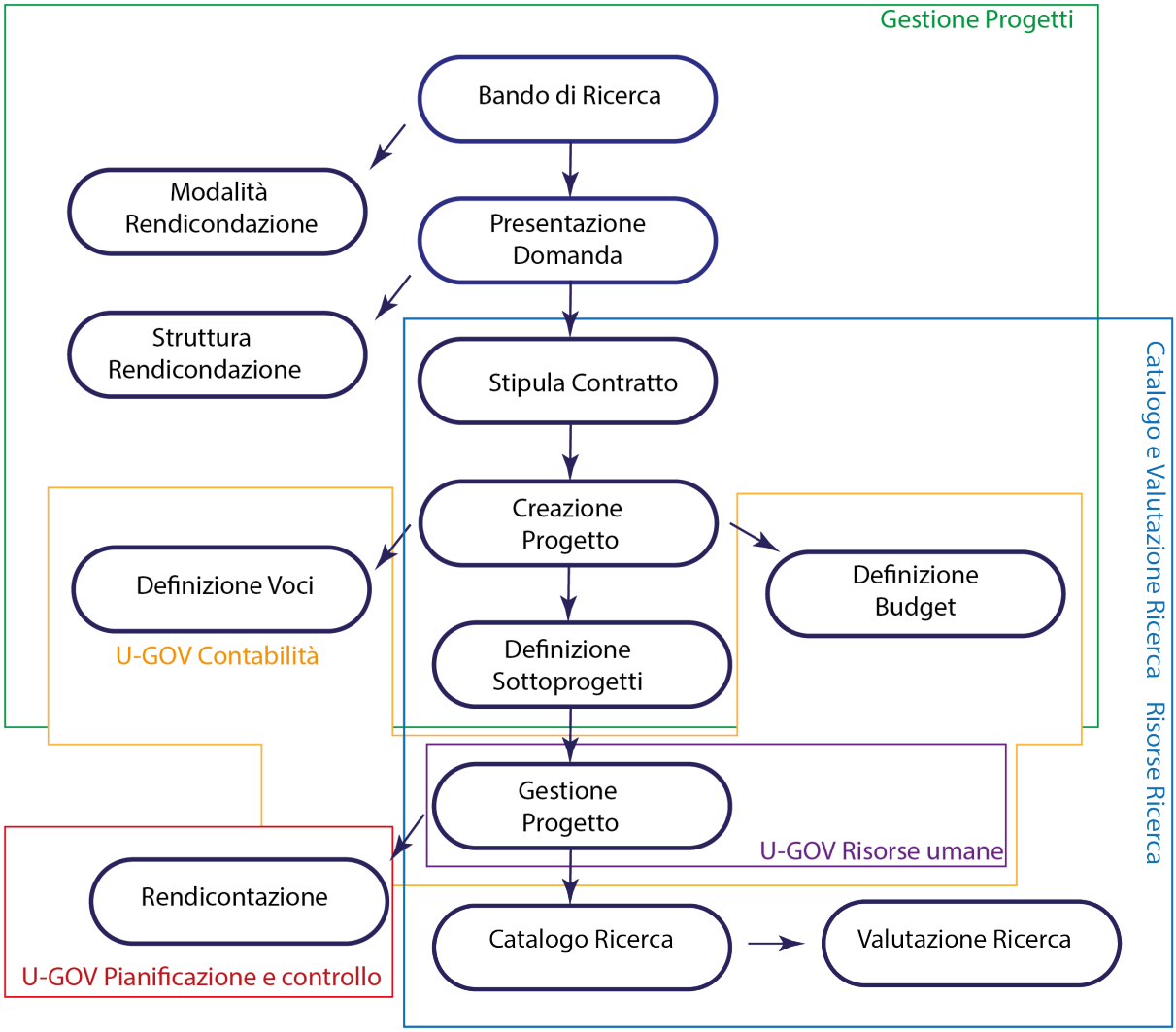

Figura 8 - Flusso di processo e applicativi

Cineca si è posto l'obiettivo di identificare i processi che in maggior misura traggono giovamento nell'essere supportati da un sistema informativo unificato e più generale possibile che possa gestire e monitorare le attività di ricerca a livello di ateneo.
L'approccio di Cineca si è quindi concentrato sulla realizzazione di un'astrazione della filiera della ricerca che permettesse l'utilizzo dei medesimi applicativi in tutti gli atenei con evidenti vantaggi in termini di manutenzione e gestione. In Figura 8 è possibile vedere quali applicativi della suite U-GOV coprono le fasi della ricerca, dalla pubblicazione del bando di finanziamento fino alla conclusione del progetto.
Questo punto di vista, per molti aspetti risultato vincente, ha quindi intenzionalmente, tralasciato tutte le peculiarità delle diverse discipline e in particolare, relativamente alla filiera della ricerca, sono state escluse le fasi legate alla gestione e conservazione dei dati prodotti dalla ricerca proprio perché hanno requisiti specifici e diversificati in base alla disciplina.
A conferma di ciò, c'è stata anche la nostra intervista telefonica con Fabrizio Luglio, Project Manager di Cineca che si occupa di IRIS, il quale ha appunto motivato l'esclusione della gestione dei dati prodotti dalla ricerca dall'analisi della filiera proprio a causa dell'elevata diversificazione nella loro gestione. Come ulteriore indagine abbiamo inserito nel survey inviato agli atenei e descritto nel capitolo successivo domande per capire se gli atenei

utilizzassero altro software applicativo che non sia direttamente sviluppato dal consorzio Cineca. La nostra indagine è principalmente volta a capire se i flussi dei processi descritti nella Figura 8 siano realmente comprensivi di tutto.

La nostra filiera della ricerca

In seguito alle considerazioni di Fabrizio Luglio ci siamo domandati come poter modificare la filiera della ricerca per includere nei processi la produzione, gestione e conservazione dei dati prodotti dalla ricerca. Riteniamo infatti che sia fondamentale ripensare la filiera della ricerca alla luce delle attività necessarie a consentire ai dati prodotti di sopravvivere alla ricerca stessa per poter così divenire patrimonio dell'ateneo e in generale della ricerca.
A nostro parere i processi coinvolti nella produzione, gestione e conservazione dei dati digitali della ricerca sono quelli iniziali di "Programmazione e indirizzo" e di "Pianificazione e selezione strumenti", e quelli finali di "Catalogazione e valutazione dei risultati" e "Valorizzazione dei risultati".
In particolare:

- "Programmazione e indirizzo" è coinvolto a livello di definizione di linee guida e best practice per una produzione, gestione e conservazione sostenibile dei dati digitali. Nello specifico le raccomandazioni della comunità europea relativamente alla gestione dei dati descritte nel precedente capitolo (es.: utilizzo di un DMP) potrebbero rientrare nelle azioni da compiere per rispondere alle strategie di ateneo in materia di open data e valorizzazione della ricerca.
- "Pianificazione e selezione strumenti" dovrà prevedere una fase di scelta e messa in opera degli strumenti informatici dedicati alla produzione, gestione e conservazione dei dati digitali. Anche in questo processo ci vengono in aiuto le raccomandazioni della comunità europea che indicano sia repository di tipo general purpose, come Zenodo (23) o Dataverse (24) (necessaria un'installazione locale), sia repository specifici per la singola disciplina (25) (26).

Nei due casi precedenti non riteniamo che sia necessario modificare la filiera della ricerca proposta da Cineca, in quanto la natura del processo rimane invariata: ci siamo limitati a esplicitare delle attività e degli strumenti.
La riorganizzazione dei processi avviene invece, a nostro parere, a livello di "Catalogazione e valutazione dei risultati" e "Valorizzazione dei risultati", dove abbiamo deciso di suddividere i due macro-processi nei seguenti:

1. Catalogazione e valutazione dei risultati: institutional repository
2. Catalogazione e valutazione dei risultati: disponibilità interna
3. Catalogazione e valutazione dei risultati: disponibilità esterna
4. Valorizzazione dei risultati: institutional repository
5. Valorizzazione dei risultati: disponibilità interna
6. Valorizzazione dei risultati: disponibilità esterna

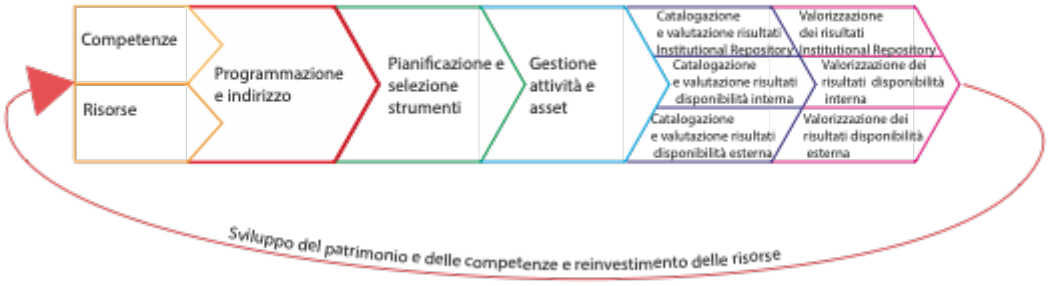

Figura 9 - La nostra versione della filiera della ricerca

Analizziamo adesso uno per volta i nuovi processi definiti.

Catalogazione e valutazione dei risultati: institutional repository

Corrisponde a quanto definito da Cineca, infatti anche nel nostro caso l'obiettivo principale di questo processo è l'alimentazione del catalogo dei prodotti per consentire le attività di verifica dei risultati e di valutazione del raggiungimento degli obiettivi.

Catalogazione e valutazione dei risultati: disponibilità interna

Questo processo raccoglie tutte le attività necessarie a garantire la condivisione dei dati digitali prodotti durante le fasi della ricerca all'interno del proprio ateneo di riferimento, quindi, ad esempio, tra diversi gruppi di ricerca.
Idealmente un vero e proprio laboratorio virtuale che permette ai gruppi di ricerca interni all'ente di sperimentare direttamente con i dati digitali.
Per garantire la disponibilità interna possibili soluzioni riguardano l'adozione a livello di ateneo di repository come Dataverse oppure di repository disciplinari. Repository come Dataverse dovrebbero essere preferiti a strumenti di conservazione più generici come Google Drive o Dropbox in quanto, a nostro parere, inducono un corretto comportamento di gestione. Dataverse che sarà analizzato nel capitolo successivo richiede ad esempio la creazione dei metadati, permette di sottoporre i dataset a revisione e offre la possibilità di versionare i dati (27).

Catalogazione e valutazione dei risultati: disponibilità esterna

Per gli atenei che hanno politiche relative alle modalità di divulgazione dei dati prodotti dalla ricerca, questo processo descrive le attività necessarie alla conservazione dei dati digitali attraverso strumenti ad-hoc per garantire la fruizione e la reperibilità da parte dei ricercatori, ma in generale da chiunque, all'esterno dell'ente.
Un possibile repository general purpose per offrire all'esterno dell'Università i dati prodotti è Zenodo, di cui vedremo più avanti caratteristiche e statistiche d'uso da parte degli atenei italiani. Ovviamente anche Dataverse o i repository disciplinari possono divenire mezzi per rendere disponibili all'esterno i dati digitali della ricerca.

Valorizzazione dei risultati: institutional repository

Anche in questo caso, con questa denominazione, ci riferiamo al processo così come definito da Cineca.

Valorizzazione dei risultati: disponibilità interna

Questo processo riguarda la valorizzazione dei dati prodotti dalla ricerca all'interno dell'organizzazione stessa. In particolare si concentra sulle attività, gli strumenti, le risorse e le competenze per garantire che, nel lungo periodo, i dati prodotti continuino ad essere reperibili e riusabili tra gruppi di ricerca.

Valorizzazione dei risultati: disponibilità esterna

Questo processo riguarda la valorizzazione dei dati prodotti dalla ricerca per soggetti che sono all'esterno dell'organizzazione e che potenzialmente vorrebbero accedere ai dati. Analogamente a quanto descritto nella sezione precedente, qui ci si concentra sulle attività, gli strumenti, le risorse e le competenze per garantire che, nel lungo periodo, i dati prodotti continuino ad essere reperibili e riusabili da tutti i potenziali interessati.

Infine, vorremmo sottolineare come una catena del valore modificata solo nella parte finale come quella riportata nella Figura 10 risulti incompleta: è necessario far partire le attività fin dal processo "Catalogazione e Valutazione dei risultati" per poter gestire le tre code separate anche per quanto riguarda il processo di valorizzazione.

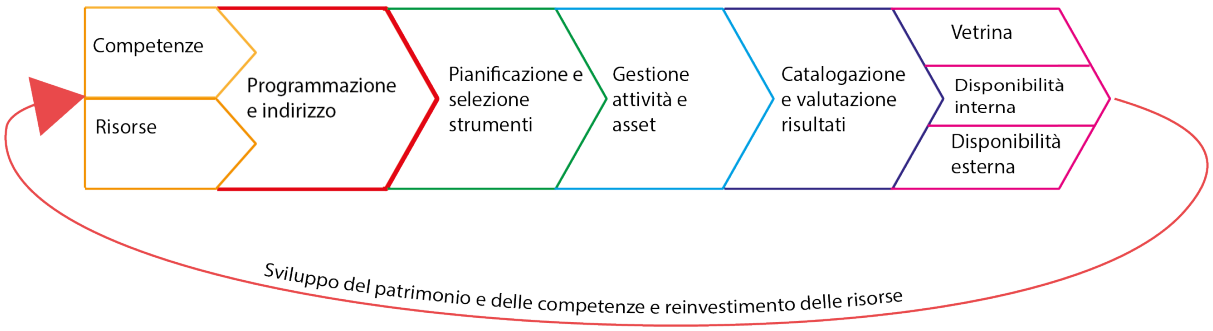

Figura 10 - Filiera incompleta

A conclusione riteniamo quindi necessaria a livello organizzativo, sia l'adozione di uno strumento come il DMP per ogni progetto di ricerca (anche quelli finanziati a livello interno dall'ente) sia l'adozione di un data repository di ateneo, preferibilmente un Dataverse di ateneo. Queste adozioni portano naturalmente a pensare la modifica della filiera da quella inizialmente proposta in Figura 7 a quella in Figura 9 con le tre code separate a distinguere Institutional Repository, Disponibilità Interna e Disponibilità Esterna (cioè i dati digitali della ricerca vengono visti dagli enti di valutazione, all'interno dell'ente e all'esterno dell'ente). Per completare quindi la nostra versione di seguito la nuova versione della Figura 8.

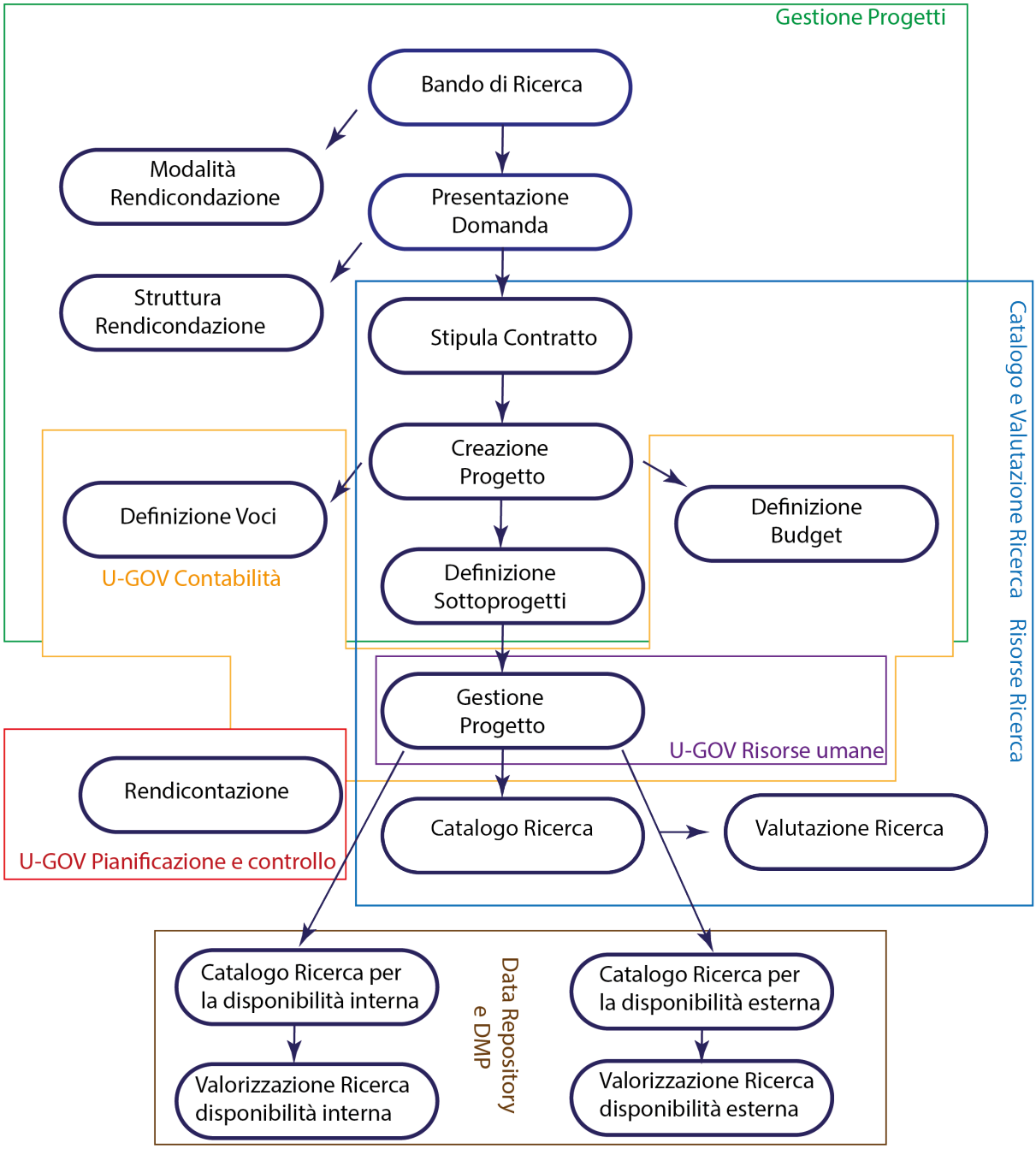

Figura 11 - Flusso ripensato di processo e applicativi

I processi che abbiamo genericamente etichettato come Data Repository e DMP tengono conto di tutte quelle attività utili a catalogare e a valorizzare il data repository dal punto di vista della disponibilità interna ed esterna, ovviamente sotto la guida del DMP.

5. Zenodo e altri data repository open: analisi e considerazioni

Per meglio analizzare la questione della conservazione a lungo termine dei prodotti della ricerca (cfr.: Catalogazione e valutazione dei risultati: disponibilità esterna) e verificare in che misura all'interno delle Università italiane i ricercatori si servono di sistemi di data repository general purpose, abbiamo deciso di estrarre in maniera automatica le statistiche d'uso di Zenodo investigando su diverse ipotesi.
Zenodo è un data repository aperto fornito gratuitamente dal CERN che offre a tutti i ricercatori una fetta della sua facility per la gestione dei *Big Data* in nome del movimento *Open Science*. Un grande incentivo per i ricercatori ad utilizzare Zenodo è la possibilità di ottenere un *DOI* (17) universale per eventuali citazioni. Come descritto dalle direttive del consorzio *OpenAIRE* le scelte da seguire in ordine di preferenza per la scelta di un data repository sono:

1. usare un data repository esterno ben stabilito all'interno di un certo dominio di ricerca che preserva e raccoglie i dati con standard riconosciuti per una certa disciplina;
2. se disponibile usare un repository data istituzionale;
3. utilizzare Zenodo;
4. utilizzare il REgistry of REsearch data Repository (28) come motore di ricerca per altri repository;

In ogni caso il data repository dovrebbe essere affidabile e rispettoso di alcune linee guida comuni come quelle sviluppate da "*Data Seal of Approval and World Data System*" (29).
Per completezza si riporta la terminologia sul deposito delle sole pubblicazioni ad accesso aperto all'interno del programma H2020 per quanto riguarda la scelta del modello di pubblicazione.
Esistono due tipologie di pubblicazione ad accesso aperto:
Green OA: è previsto che gli autori depositino (auto-archivino) la versione finale revisionata del manoscritto in un repository che rende possibile l'accesso aperto, di solito dopo un periodo di embargo per permettere di ammortizzare i costi della pubblicazione (mediante sottoscrizioni o download a pagamento).
Gold OA: accesso aperto immediato fornito dall'editor, i costi di pubblicazione sono coperti di solito dall'editore (open access journal) e gli articoli sono immediatamente disponibili alla data di pubblicazione.

Analisi dati di utilizzo di Zenodo

Navigando (30) è possibile consultare la documentazione delle API che permettono la gestione delle collezioni di dati automatizzata all'interno di Zenodo. In aggiunta, Zenodo come tutti i repository open offre la possibilità di *harvesting* dei record depositati mediante il protocollo OAI-PMH 2.0[3]; il punto di ingresso al protocollo è disponibile all'indirizzo (31).
Ci sono tuttavia delle restrizioni sia nell'utilizzo delle API REST che del protocollo OAI-PMH 2.0:

[3] Il protocollo OAI-PHP 2.0 è un'interfaccia di accesso ai metadati dei record contenuti in un repository che si basa su protocolli standard, richieste su canale HTTP e risposte in XML incapsulate in formati personalizzabili come per esempio elementi Dublin Core.

- le API sono limitate alle collezioni di dati depositate da un certo utente ed autenticato mediante un token rilasciato una tantum dall'apposito cruscotto di gestione;
- le possibilità di ricerca dei record mediante protocollo OAI-PMH 2.0 sono ridotte. Ad esempio il verbo *ListRecords*(tipo dello standard) permette di filtrare esclusivamente per:
    a. campi "from" e "until" che permettono di specificare un intervallo di tempo selettivo
    b. campo "set" che permette di specificare degli insiemi di record precedentemente classificati in set organizzati ad albero.

Aspetti implementativi del sistema di estrazione dati

I "set" per l'*harvesting* OAI-PMH 2.0 definiti su Zenodo sono scelti localmente dall'utente e non sono utili per ricercare dati a livello globale. La ricerca accessibile dal sito web mediante il campo "search" di Zenodo invece sembra utilizzare tutte le potenzialità di Elasticsearch[4] (32). Tuttavia, essendo Zenodo costruito e pensato per essere usato interattivamente, non risulta sufficiente richiedere una singola pagina web ad un certo URL per ottenere a schermo i risultati. Queste considerazioni ci hanno portato a decidere di interrogare Zenodo simulando un browser con funzionalità Javascript abilitate che fosse anche comandabile mediante script. Per la nostra indagine, onde evitare i falsi positivi anche a costo di sacrificare delle occorrenze, abbiamo pensato di interrogare il repository ricercando le Università italiane esclusivamente attraverso il loro nome ufficiale. Per questo alcune Università, come anche Scuola IMT Alti Studi Lucca e Scuola Superiore S. Anna Pisa, non compaiono nell'elenco finale poiché gli utenti che hanno depositato su Zenodo non hanno utilizzato il nome ufficiale dell'Università, ma ad esempio abbreviazioni e sigle.
Per ogni Università abbiamo registrato il totale dei record inseriti e il numero di record che sono catalogati come semplice pubblicazione. Il totale dei record depositati su Zenodo in cui sono coinvolte le Università è di 2.045. Nella Figura 12 è riportato il risultato della nostra prima indagine; il grafico è ordinato in ordine decrescente di utilizzo.

[4] Il motore di ricerca di Zenodo è costruito utilizzando la tecnologia Elasticsearch

Figura 12 - Diffusione di Zenodo tra gli atenei

Successivamente abbiamo voluto investigare l'andamento delle pubblicazioni su Zenodo negli ultimi anni per cercare di capire se il trend di utilizzo è in crescita o meno. Il risultato dell'andamento tra il 2015 e il 2019 delle dieci Università che maggiormente utilizzano Zenodo è riportato nel grafico di Figura 13. L'intervallo di anni tra il 2015 e il 2019 corrisponde all'80% dei record (1655/2045) individuati.

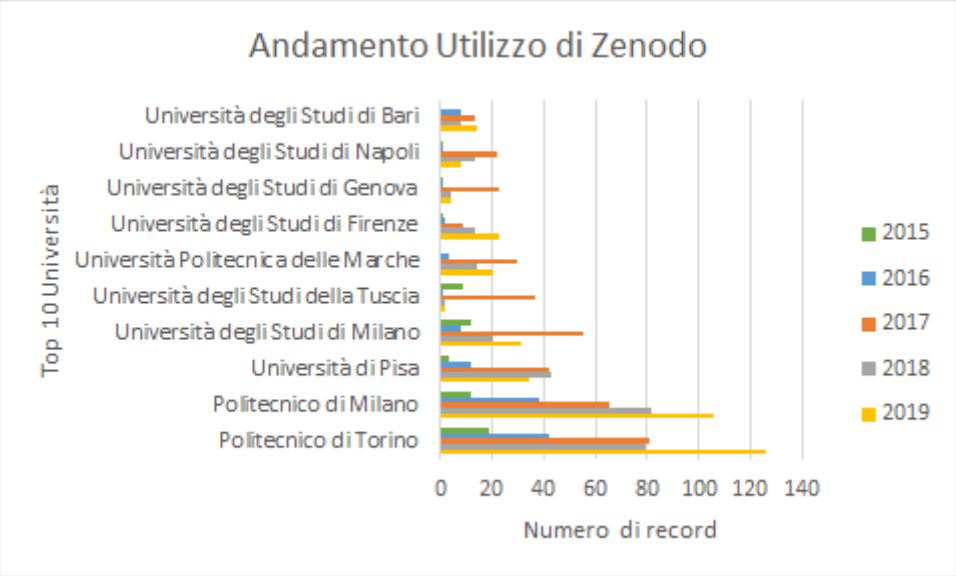

Figura 13 - Trend di utilizzo di Zenodo negli anni

Analisi delle Keywords

Per ogni record su Zenodo è possibile inserire un elenco di *keyword*. Nella Figura 14 è riportata la word cloud delle *keyword* presenti in ognuno dei 2.045 record depositati. Prima di stampare il grafico, l'elenco delle parole è stato preprocessato da uno script che ha eseguito la moltiplicazione delle occorrenze. Per esempio se la keyword "Animalia" compariva in due diverse ricerche per due differenti Università con occorrenze 18 e 5, lo script ha generato 23 occorrenze della parola "Animalia" nell'elenco finale.

Figura 14 - Word cloud con le keyword maggiormente utilizzate

La word cloud della Figura 14 ci ha dato un'idea della tipologia di dati memorizzati su Zenodo. Abbiamo quindi deciso di approfondire ulteriormente l'analisi utilizzando un modello di machine learning per classificare la tipologia di record; la scelta è ricaduta su Word2Vec. Word2vec (33) è una semplice rete neurale artificiale a due strati progettata per elaborare il linguaggio naturale, l'algoritmo richiede in ingresso un corpus che usa per analizzare la distribuzione semantica delle parole nel testo in input. Il corpus che abbiamo utilizzato è quello disponibile per il download denominato *word2vec-google-news-300* con un database di tre milioni di parole e la relativa rete neurale allenata grazie a Google News. Il listato del nostro algoritmo è riportato in Figura 15. Osservando la word cloud abbiamo scelto di studiare la somiglianza semantica delle *keyword* presenti nei record depositati dalle Università con le parole "Mathematics"," Engineering"," Humanities"," Biology"," Medicine"," Scientific"," Social". L'algoritmo utilizza la funzione di similarità tra due parole della rete neurale word2vec e nel caso di *keyword* composte, come per esempio "data mining" somma il peso della relativa similarità. La funzione di similarità restituisce anche similarità negative, quindi abbiamo scelto di accettare la somiglianza se si raggiungeva un certo valore di *threshold* che abbiamo fissato empiricamente a 0.14. I valore di *threshold* è stato scelto in maniera empirica osservando alcuni output. Se più somiglianze superano il valore di *threshold* abbiamo promosso quella con similarità maggiore.

```
import gensim.downloader as api
from gensim.parsing.preprocessing import *
wv = api.load('word2vec-google-news-300')
CUSTOM_FILTERS = [lambda x: x.lower(), strip_tags, strip_punctuation,strip_short]
classification=['Mathematics','Engineering','Humanities','Biology','Medicine','Scientific','Social']
keywords_array = []
with open('keywords') as my_file:
    for line in my_file:
        keywords_array.append(preprocess_string(remove_stopwords(line),CUSTOM_FILTERS))
print(keywords_array)
keywords_array = [x for x in keywords_array if x]
for keyword in  keywords_array:
    sommamax=-1
    cont=0
    for classi in classification:
        somma=0
        cont=0
        for subkeyword in keyword:
            try:
                somma=somma+wv.similarity(classi, subkeyword)
                cont=cont+1
            except KeyError:
                pass
        if (cont>0):
            if (sommamax<somma):
                sommamax=somma
                classimax=classi
    if (sommamax==-1):
        print('%r\t%r' % ("non presente nel modello", keyword))
    elif (sommamax<0.14):
        print('%r\t%r' % ("nessun match", keyword))
    else:
        print('%r\t%r\t%.2f' % (classimax, keyword, sommamax))
```

Figura 15 – Listato dell'algoritmo di classificazione delle keyword

Riportiamo qualche esempio di similarità:

Esempio 1
'Animalia' e 'Mathematics' hanno score di similarità 0.11
'Animalia' e 'Engineering' hanno score di similarità 0.09
'Animalia' e 'Humanities' hanno score di similarità 0.14
'Animalia' e 'Biology' hanno score di similarità 0.24

Esempio 2
'Aircraft' e 'Mathematics' hanno score di similarità 0.11
'Aircraft' e 'Engineering' hanno score di similarità 0.33
'Aircraft' e 'Humanities' hanno score di similarità 0.09
'Aircraft' e 'Biology' hanno score di similarità 0.09

Esempio 3
Mathematics' e 'experimental', 'tests' hanno score di similarità 0.21
'Engineering' e 'experimental', 'tests' hanno score di similarità 0.22
'Humanities' e 'experimental', 'tests' hanno score di similarità 0.09
'Biology' e 'experimental', 'tests' hanno score di similarità 0.30
'Medicine' e 'experimental', 'tests' hanno score di similarità 0.36
'Scientific' e 'experimental', 'tests' hanno score di similarità 0.37

L'algoritmo è stato lanciato su un totale di 527 keyword differenti, solo 141 (21%) keyword non erano presenti nel modello utilizzato. Il 10% delle keyword invece non aveva somiglianza con le parole che abbiamo scelto. La situazione è riassunta in Figura 16.

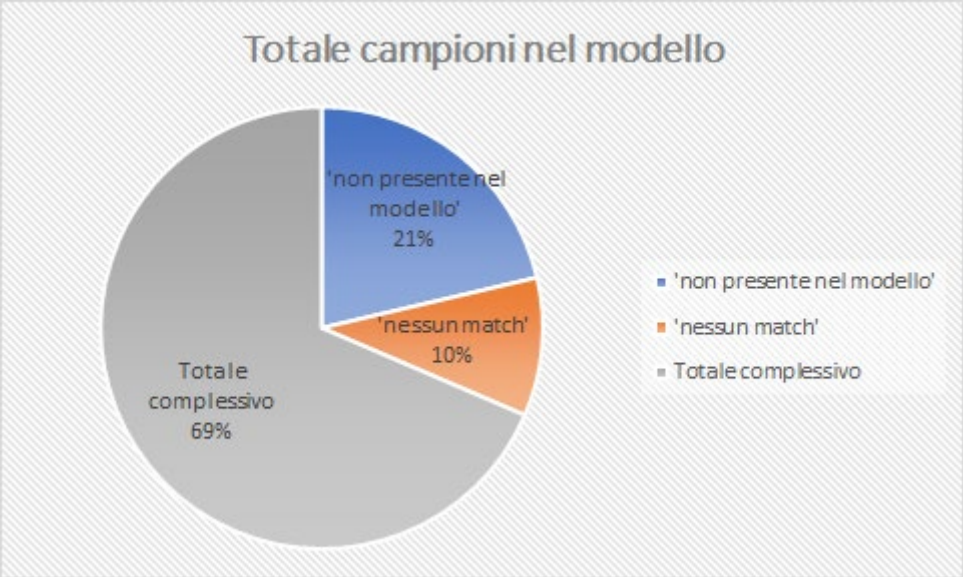

Figura 16 - Distribuzione dei risultati dell'algoritmo di classificazione delle keyword

Le restanti 453 keyword (69%) è stato classificato correttamente come in figura.

Figura 17 - Distribuzione delle keyword

Il 69% del totale ci sembra un campione significativo per poter trarre delle conclusioni. Per ottenere una copertura maggiore andrebbe allenata la rete neurale con termini specifici come 'basidiomycota' o 'marchantiidae' magari utilizzando proprio il testo degli articoli. Alcune keyword non riconosciute invece erano scritte in italiano ed il modello utilizzato non ha la lingua Italiana.

Considerazioni sull'utilizzo di Zenodo

Dalla Figura 13 è possibile notare come l'anno di maggiore utilizzo sia stato mediamente il 2017, con una tendenza in crescita per il 2019 (77 record per il 2015, 180 per il 2016, 474 per il 2017, 413 per il 2018 e 511 per il 2019). Un'ipotesi potrebbe essere che il 2017 è stato l'anno in cui si è diffusa maggiormente la soluzione IRIS di Cineca come repository istituzionale. Risale infatti al 29 novembre 2017 l'evento organizzato da Cineca e denominato IRIS day (34). Ricordiamo che la quasi totalità delle Università pubbliche ha scelto di utilizzare IRIS come Institutional Repository per le diverse possibilità che offre di estrarre al volo le metriche richieste dagli enti di valutazione e per il fatto che importa in maniera automatica i record già depositati sul loginmiur (35) dei docenti.
In generale, anche la Figura 12, ci suggerisce che, ad esclusione delle prime dieci Università più virtuose, l'uso che viene fatto di Zenodo non sia frutto di una strategia di ateneo volta alla conservazione dei prodotti della ricerca ma piuttosto legato a iniziative isolate. L'analisi semantica delle keyword infine (Figura 17) mostra come Zenodo sia utilizzato maggiormente in ambito scientifico e non in campo medico o delle scienze sociali. Questo potrebbe significare che gli autori di queste ultime discipline rendono disponibili i propri dati utilizzando canali differenti. Il deposito su Zenodo e su eventuali altri repository non è affatto organizzato e probabilmente rappresenta una soluzione temporanea nell'attesa che ad esempio a livello di ateneo possa rendersi disponibile una soluzione più strutturata basata ad esempio su Dataverse. Come abbiamo visto nel capitolo precedente la visione imposta da Cineca del processo della ricerca sembra essere diffusa quasi ovunque ma non è sufficiente anche per gestire i dati digitali.

Altri data repository

Navigando il sito web OpenAIRE|EXPLORE (36) è possibile trovare un catalogo di repository indicizzati dal consorzio OpenAIRE mediante harvesting di data repository compatibili col protocollo OAI-PMH (37) (Open Archives Initiative Protocol for Metadata Harvesting). Tra questi il più utilizzato come data repository è Dataverse. Dataverse è un'applicazione web per condividere, conservare, citare, esplorare e analizzare i dati digitali della ricerca. È stato sviluppato dall'*Institute for Quantitative Social Science* della Università di Harvard e permette di mettere a disposizione della comunità i propri dati, aumentandone la visibilità, facilitandone il riutilizzo e la replicabilità delle ricerche. Tra le funzionalità chiave di Dataverse sottolineiamo:

- citazione dei dati permanente;
- metadati, relativi alla citazione, relativi al dominio dei dati e relativi ai file;
- *faceted search* cioè la ricerca dei dati guidata e basata sui filtri basati sulla classificazione dei dati;
- controllo degli accessi;
- CC0 (nessun diritto d'autore) (13) / termini d'uso/ restrizioni/ guestbook;

- ruoli e permessi (se necessario le policy di accesso a dati e metadata possono essere più restrittive);
- *versioning* e verifica integrità. Verifica dell'integrità dei file: UNF (Universal Numerical Fingerprint) per i dati tabellari, MD5 *checksum* per le altre tipologie di file;
- impaginazione dei dati tabellari (I dati tabellari vengono convertiti in un formato adatto alla conservazione);
- API per l'integrazione con le riviste.
- *data visualizations* e analisi.

I dati in formato tabellare possono essere esplorati mediante Two Ravens, un'interfaccia per l'analisi statistica integrata con Dataverse, utilizzabile da parte di un'utenza con vari gradi di competenza statistica (anche come strumento didattico). Mediante Two Ravens è possibile sia visualizzare statistiche di base relative alle variabili, sia operare analisi su un sottoinsieme di valori che testare modelli statistici. Non c'è necessità per chi utilizza Two Ravens di scaricare i dati, le informazioni sono presentate in maniera riassuntiva e le analisi effettuate lato server. In questo modo ad esempio è possibile esplorare grosse quantità di dati senza trasferirli localmente sul pc dell'utente per essere processati e gestire allo stesso tempo la protezione dei dati mostrando solo statistiche riassuntive, senza dare accesso effettivo ai dati grezzi. L'obiettivo è anche quello di fornire ad un grande numero di utenti la possibilità di effettuare ragionamenti di tipo quantitativo con funzionalità di analisi che non hanno necessità di grandi infrastrutture e un'interfaccia grafica intuitiva mediante la quale effettuare le analisi e i ragionamenti.
Un repository Dataverse è un contenitore che ospita diversi Dataverse. Ogni Dataverse può contenere altri Dataverse o dataset e ogni dataset contiene metadati descrittivi e file (tipica struttura ad albero in cui le foglie sono i dataset e i nodi sono i Dataverse). Un dataset è un contenitore di dati, documentazione, codice e metadati descrittivi. I file in determinati formati (es. Stata, SPSS, R, Excel (xlsx) and CSV) possono essere importati come "dati tabellari". I dati tabellari caricati vengono estratti e archiviati in un formato facilmente leggibile e adatto alla "*Digital Preservation*". I metadati che descrivono il documento sono conservati separatamente in un database relazionale, in modo che siano accessibili in maniera efficiente. Se si ha un ruolo di Contributor si possono sottoporre i dataset a revisione. Gli amministratori o i curatori riceveranno una notifica in merito alla revisione del dataset e il Contributor riceverà a sua volta una notifica sia in caso di "Pubblicazione" sia in caso di "Rinvio all'autore". All'utenza generica è possibile navigare tra i diversi Dataverse, dataset e file pubblicati. Per vedere ciò che non è pubblicato, invece, è necessario essere autorizzati da un amministratore del Dataverse. Infine mediante plugin è possibile rendere compatibile Dataverse con i maggiori cataloghi e sistemi di *harvesting*.
A differenza di Dataverse, Eprints (38) e dSPACE (39) invece sono due repository che sono nati principalmente per essere utilizzati come Institutional Repository. dSPACE è utilizzato ad esempio come base per l'Institutional Repository IRIS fornito da Cineca. Eprints invece era utilizzato prima del passaggio ad IRIS come Institutional Repository dalle maggior parte delle Università ed ora quella stessa installazione è stata riutilizzata per la memorizzazione di dati digitali della ricerca (come rilevato nel survey qualche ente utilizza lo stesso IRIS per la memorizzazione dei dati). Mediante la personalizzazione della tipologia di pubblicazioni è infatti possibile configurare dSPACE ed Eprints per memorizzare dati generici. Tuttavia crediamo che questa struttura sia troppo rigida e Dataverse rappresenti ad oggi lo stato dell'arte dei data repository; basti pensare che oltre alle caratteristiche peculiari sopra descritte

è possibile effettuare con facilità installazioni ridondante e distribuite per la memorizzazione dei cosiddetti Big Data.

6. Atenei a confronto

Partendo dalla nostra catena del valore della ricerca così come descritta nel Capitolo 4, Figura 9, abbiamo formulato un questionario intitolato “Survey sulla gestione dei dati prodotti dalla ricerca negli atenei italiani” che abbiamo utilizzato per intervistare i responsabili degli atenei coinvolti a vario titolo nella gestione dei dati della ricerca per un’indagine sulla situazione attuale.

Il metodo

La nostra intervista ha voluto analizzare le politiche degli atenei in merito all’open data, agli strumenti informatici utilizzati (o non utilizzati) per la conservazione dei dati e alle tecniche di produzione dei dati della ricerca.
Per la sezione la sezione Programmazione e indirizzo abbiamo fatto le seguenti domande:

Domande	**Possibili risposte**
Nel vostro ateneo sono previsti dei fondi per sostenere la conservazione dei dati digitali prodotti dalla ricerca?	Sì, fondi di dipartimento Sì, fondi di ateneo No Altro
Nel vostro ateneo sono previsti dei fondi per favorire le pubblicazioni open?	Sì, fondi di dipartimento Sì, fondi di ateneo No Altro

Abbiamo ritenuto importante suddividere il Sì per indagare sulla natura dei fondi, permettendo anche di specificare eventuali particolarità
Abbiamo suddiviso la sezione *Pianificazione e selezioni strumenti* in due sotto parti. Le domande per la prima parte sono le seguenti:

Domande	**Possibili Risposte**
È stato adottato uno strumento strutturato per la gestione dei dati prodotti dei progetti di ricerca (ES.: il Data Management Plan)?	Sì/No
Se è stato adottato uno strumento strutturato per la gestione dei dati prodotti dai progetti di ricerca, questo strumento viene utilizzato nelle fasi iniziali dei progetti di ricerca?	Sempre Solo in parte Solo per i Progetti Europei No
Se è stato adottato uno strumento strutturato per la gestione dei dati prodotti dai progetti di ricerca, è disponibile un template?	Risposta libera per specificare il template

Abbiamo ritenuto importante non focalizzarsi direttamente sul DMP per indagare se qualche ateneo avesse adottato eventuali altri strumenti. Nella seconda domanda abbiamo inserito esplicitamente come possibile risposta “Solo per i progetti europei” per via delle richieste del framework H2020. Le domande per la seconda parte sono le seguenti:

Domande	Possibili Risposte
Viene utilizzato un software applicativo dedicato alla gestione delle proposte dei progetti di ricerca?	Sì, modulo applicativo Cineca Si, modulo applicativo sviluppato da altro fornitore Sì, modulo applicativo sviluppato in proprio Sì, supporto non strutturato (Es.: Excel) Altro
Viene utilizzato un software applicativo dedicato alla gestione contabile dei progetti di ricerca?	Sì, modulo applicativo Cineca Si, modulo applicativo sviluppato da altro fornitore Sì, modulo applicativo sviluppato in proprio Sì, supporto non strutturato (Es.: Excel) Altro
Viene utilizzato un software applicativo dedicato alla gestione delle risorse umane dei progetti di ricerca?	Sì, modulo applicativo Cineca Si, modulo applicativo sviluppato da altro fornitore Sì, modulo applicativo sviluppato in proprio Sì, supporto non strutturato (Es.: Excel) Altro
Indicare quali altri applicativi in aggiunta ai precedenti vengono utilizzati all'interno del processo della ricerca.	Risposta libera

Abbiamo ritenuto utile distinguere tra modulo applicativo e supporto non strutturato per avere differenti casistiche nelle statistiche. Le domande coprono l'offerta ad oggi di moduli applicativi Cineca.
L'ultima domanda è stata lasciata di proposito generica per due motivi:

1. verificare se esiste qualche software che Cineca non considera nella sua offerta che invece viene utilizzato dagli atenei,
2. indagare quanti degli intervistati considerassero il repository istituzionale come parte del processo della ricerca.

Da un'indagine separata, confermata da Cineca, infatti sappiamo con certezza che tutti gli atenei intervistati hanno un'installazione di IRIS come repository per le pubblicazioni.
La sezione relativa a *Catalogazione e valutazioni risultati: disponibilità interna*, ha una sola domanda chiave.

Domande	Possibili risposte
Indicare quali applicativi software vengono utilizzati per permettere ai gruppi di ricerca di collaborare internamente mediante la condivisione dei dati risultati della ricerca. Precisare se si tratta di strumenti utilizzati a livello di gruppo di ricerca, di dipartimento o di ateneo	Risposta libera

L'obiettivo era quello di indagare se ci fossero eventualmente soluzioni strutturate per risolvere il problema. Le domande relative alla *Catalogazione e valutazioni risultati: disponibilità esterna*, sono suddivise in due parti.

Domande	**Possibili risposte**
A livello di ateneo ci sono delle policy per l'open data?	Sì/No
Le policy sono relative solo alle pubblicazioni o includono anche i dati digitali prodotti nella ricerca?	Dati digitali e Pubblicazioni Solo pubblicazioni
Esistono strumenti di conservazione centralizzata dei dati prodotti dalla ricerca? (es.: Dataverse, IRIS)	Sì/No

La seconda parte è completamente opzionale qualora non vengono usati a livello di ateneo dei data repository centralizzati.

Domande	**Possibili risposte**
Indicare quali data repository vengono utilizzati	Risposta libera
Sono state allocate delle risorse economiche per garantire la manutenzione sul lungo periodo dei repository?	Sì/No

Oltre a chiedere quale fosse l'eventuale data repository utilizzato è interessante sapere se l'ateneo stanzia dei fondi per mantenere i dati memorizzati nel lungo periodo.
Per quanto riguarda la sezione *Valorizzazione dei risultati: disponibilità interna* abbiamo indagato principalmente su come fossero organizzati i siti dedicati ai progetti di ricerca.

Domande	**Possibili Risposte**
I siti dei progetti vengono conservati in qualche modo (ad esempio staticizzati)?	Sì/No
Per quale motivo viene dismesso un sito di progetto di ricerca	Vincolo temporale (numero di anni) Vincolo tecnologico (obsolescenza) Vincolo economico (esaurimento risorse) Altro

In particolare oltre al meccanismo ovvio di staticizzare i siti web (vedasi ad esempio l'Internet Archive (40)), abbiamo chiesto quale fosse il motivo principale per cui un sito dedicato ad un progetto di ricerca viene dismesso.
Le ultime domande relative alla sezione *Valorizzazione dei risultati: disponibilità esterna* sono relative alla disponibilità nel lungo periodo dei dati digitali.

Domande	Possibili Risposte
Esistono delle policy per la conservazione dei dati digitali prodotti dai progetti chiusi?	Sì/No
Sono disponibili delle statistiche sulla raggiungibilità dei dati dopo 5 anni dal termine dei progetti di ricerca?	Sì/No

Il campione

Sono state inviate 24 richieste di compilazione del questionario ai diversi referenti degli atenei individuati, inizialmente, grazie a un'indagine preliminare partita da una richiesta del nostro tutor, Fabrizio Pedranzini, sfruttando il forum IT del CODAU che è letto dai dirigenti dei sistemi informativi degli atenei.
Sono state ricevute 19 risposte ai questionari così distribuite:

- 11 sono stati compilati da Università del Nord,
- 6 da Università del Centro
- 2 da Università del Sud.

Di seguito l'elenco completo delle Università che hanno risposto ai questionari.

1. Modena e Reggio Emilia
2. Politecnico di Milano
3. Politecnico di Torino
4. Sapienza
5. Scuola IMT Alti Studi Lucca
6. Scuola Normale Superiore di Pisa
7. Scuola Superiore Sant'Anna
8. Università Ca' Foscari Venezia
9. Università degli Studi di Trento
10. Università degli studi di Udine
11. Università del Salento
12. Università dell'Insubria
13. Università di Napoli Federico II
14. Università di Padova
15. Università di Siena
16. Università di Verona
17. Università Milano-Bicocca
18. Università Statale di Milano
19. Università degli studi di Roma Tor Vergata

Nonostante l'indagine preliminare abbia semplificato la ricerca, una delle problematiche che abbiamo affrontato è stata quella di trovare una figura che avesse visibilità globale della situazione del proprio ateneo e che quindi potesse rispondere al nostro questionario e, in alcuni casi, non siamo riusciti ad individuare qualcuno responsabile di raccogliere tutte le risposte.
Molti degli intervistati hanno dichiarato che si tratta di un problema caldo a cui l'organizzazione si sta dedicando. In un caso esisteva addirittura un delegato del rettore per l'Open Science e in altri casi l'attenzione al problema è così sentita che sono state create delle commissioni a livello amministrativo.

La tabella sottostante riassume il numero dei passaggi, intesi come scambio di email successive, che abbiamo fatto per ottenere le risposte al questionario. Nella tabella inoltre sono riportate le aree amministrative del referente che ha risposto al questionario: in cinque casi abbiamo avuto direttamente un'intervista con il dirigente dell'area di riferimento.

AREA	**DIRIGENTE**	**NUMERO DI STEP**
PERFORMANCE/QUALITA'	SI	2
RICERCA	NO	2
ICT	NO	2
ICT	NO	1
ICT	SI	1
RICERCA	NO	2
BIBLIOTECHE	NO	2
RICERCA	NO	2
RICERCA	NO	2
RICERCA	SI	3
ICT	NO	1
RICERCA	SI	3
RICERCA	NO	2
RICERCA	NO	3
ICT	NO	2
ICT	NO	1
ICT	NO	1
BIBLIOTECHE	NO	3
RTD[5]	SI	2

[5] Responsabile per la transizione digitale https://www.agid.gov.it/it/agenzia/responsabile-transizione-digitale

Analisi dei dati

Programmazione e indirizzo

Siamo interessati a capire se gli atenei, tra le proprie strategie programmatiche e di indirizzo, sostengono e incentivano la conservazione dei dati prodotti dalla ricerca e le pubblicazioni open.
Dalla prima domanda risulta che la maggior parte degli atenei (il 57,9%) non prevede lo stanziamento, a nessun livello (ateneo, dipartimento o progetto), di fondi a sostegno della conservazione dei dati digitali prodotti dalla ricerca. Le risposte libere ci danno inoltre contezza del fatto che alcuni atenei non hanno chiara la questione della conservazione dei dati. Crediamo che ciò dipenda da una mancanza a livello organizzativo di una struttura trasversale che sia in grado di fornire supporto sia al personale tecnico che di ricerca per creare la "cultura del dato".

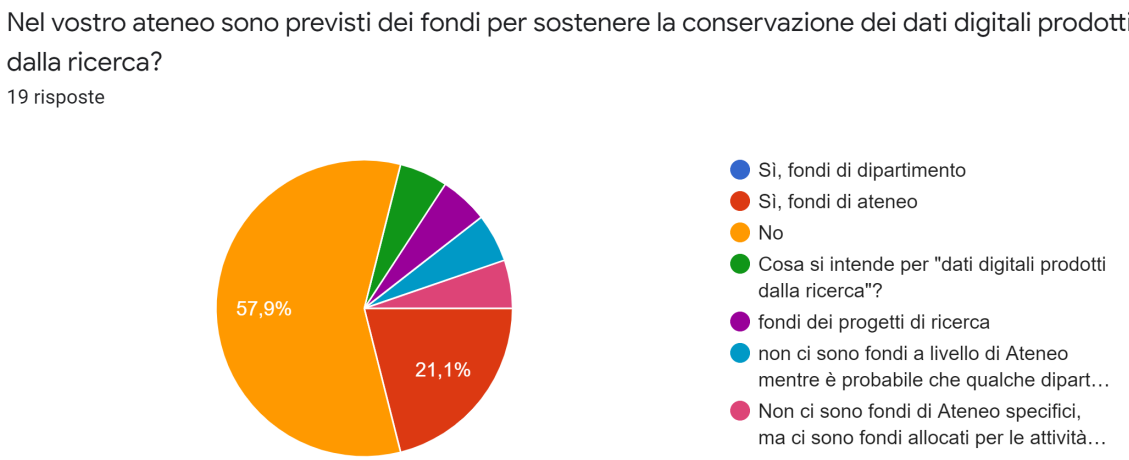

Figura 18 - Distribuzione delle risposte: fondi per la conservazione dei dati

Relativamente alle pubblicazioni open, il 42,1% degli atenei non prevede a livello di programmazione e indirizzo dei fondi per favorirle, mentre il restante 57,9% prevede fondi a vari livelli: ateneo, dipartimento, progetto di ricerca, ecc.

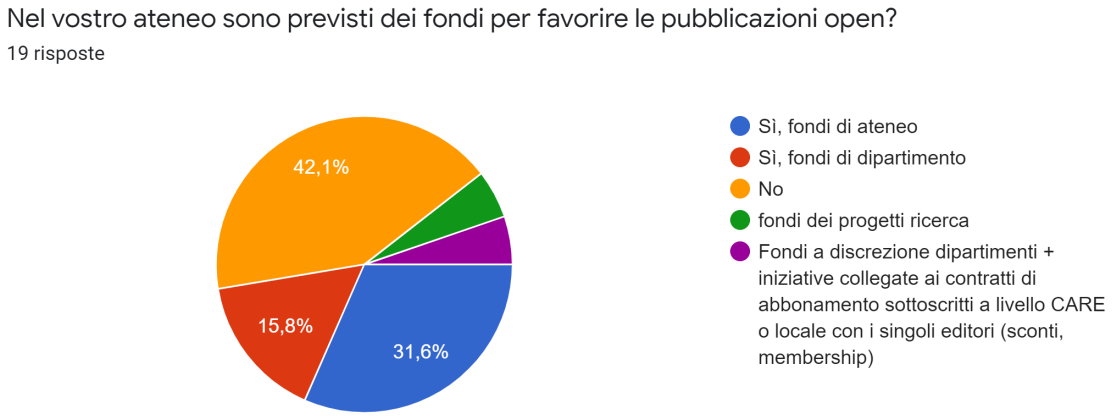

Figura 19 - Distribuzione delle risposte: fondi per le pubblicazioni

Pianificazione e selezione strumenti

A questo livello il nostro interesse si rivolge sia a valutare l'uso del DMP come strumento per la gestione dei dati, ma più in generale a considerare quando e in che misura i sistemi vengono utilizzati dagli atenei per gestire i dati prodotti dalla ricerca.

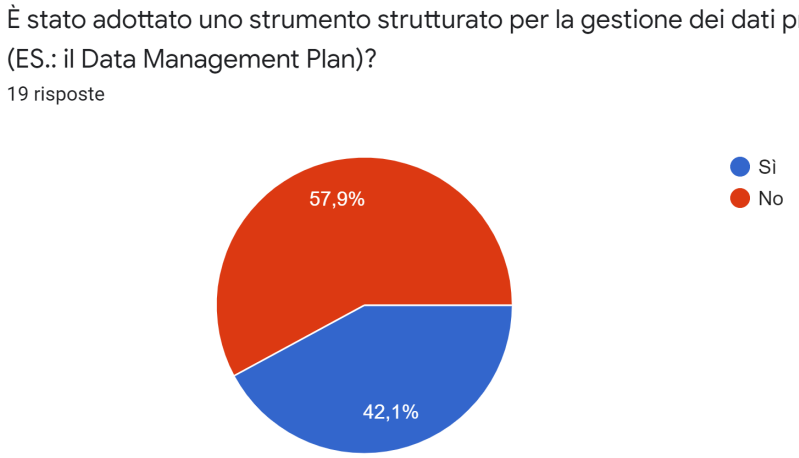

Figura 20 - Distribuzione delle risposte: DMP

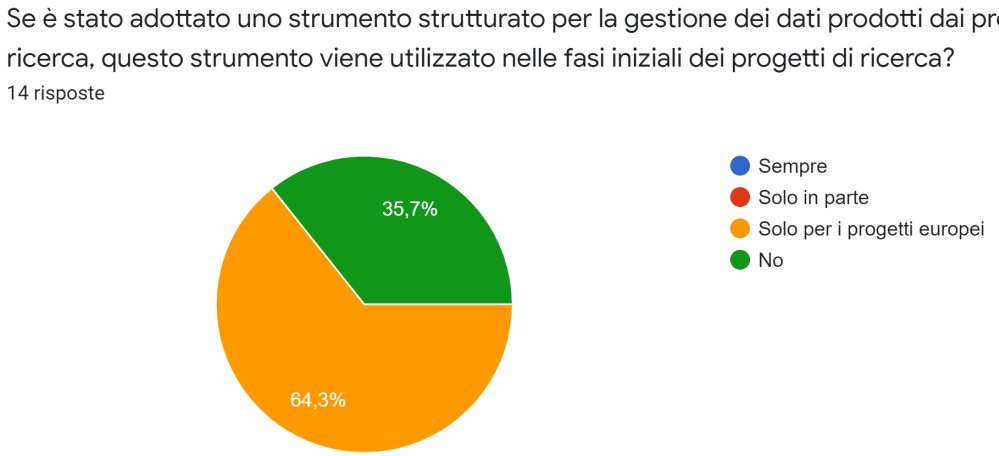

Figura 21 - Distribuzione delle risposte: DMP in dettaglio

Le risposte a questa domanda fanno supporre che, quando utilizzato, sia stato adottato uno strumento strutturato per la gestione dei dati esclusivamente laddove è imposto ossia solo per i progetti europei. È nostro parere invece che l'utilizzo di uno strumento come il DMP potrebbe essere estremamente utile per garantire la corretta gestione e conservazione dei dati prodotti da qualsiasi tipologia di progetto di ricerca (nazionale, interno, ecc.).
Relativamente alla domanda a risposta aperta "*Se è stato adottato uno strumento strutturato per la gestione dei dati prodotti dai progetti di ricerca, è disponibile un template?*":

- due atenei hanno atenei dichiarato di utilizzare un template,
- un ateneo ha specificato che "*Sono state elaborate linee guida e un template per la compilazione e viene data assistenza a chi (raramente) chiede aiuto*",
- due atenei hanno fornito l'indirizzo dei propri template (41) (42)
- un ateneo ha dichiarato di seguire le linee guida H2020.

Sull'utilizzo dei vari software da parte degli atenei sembra che la gestione delle proposte dei progetti non sia organizzata e che Cineca non sia riuscita a imporsi come fornitore.

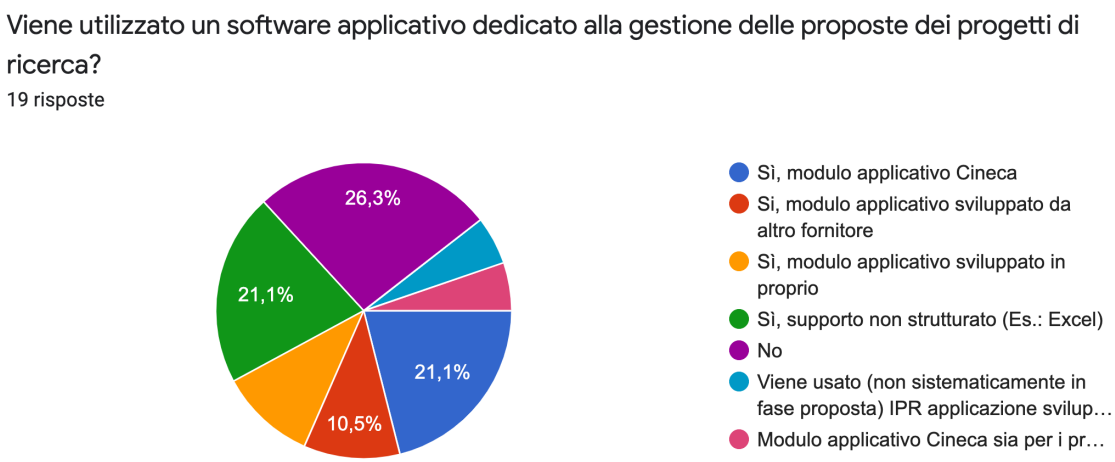

Figura 22 - Distribuzione delle risposte: software per le proposte di progetto

Per quanto detto nel Capitolo 4, è proprio in questa fase che dovrebbe nascere la prima versione del DMP e il fatto di avere un applicativo gestionale integrato con i sistemi informativi di ateneo aiuterebbe nella gestione del DMP stesso anche nelle fasi successive in cui verrebbe perfezionato.
La gestione contabile dei progetti di ricerca sembra invece passare per la maggior parte da software Cineca. L'idea di Cineca è stata di integrare U-GOV PJ (la parte della suite U-GOV dedicata ai progetti vedere Figura 8) con UGOV Contabilità tanto che è possibile mappare i piani dei conti, che hanno schemi differenti a seconda del tipo di progetto, direttamente con il piano dei conti in contabilità generale.
Nella figura riportiamo i risultati relativi al modulo applicativo per la gestione contabile.

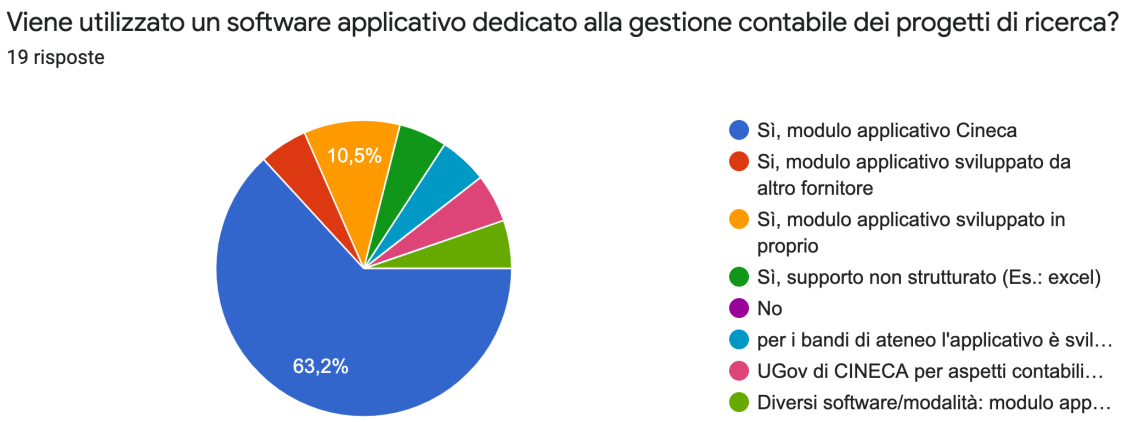

Figura 23 - Distribuzione delle risposte: software per la gestione contabile dei progetti

Per gestione delle risorse umane dei progetti sebbene ci sia una grossa intersezione con la parte di gestione contabile, Cineca è riuscita ad imporsi solo per il 36,8%. Il modulo software proposto da Cineca permette di gestire i *TimeSheet* delle ore di lavoro che le risorse umane spendono sui progetti. Solo in un caso viene utilizzato anche IRIS come nella visione iniziale di Cineca (Figura 8). Nella figura riportiamo i risultati relativi al modulo applicativo per la gestione delle risorse umane.

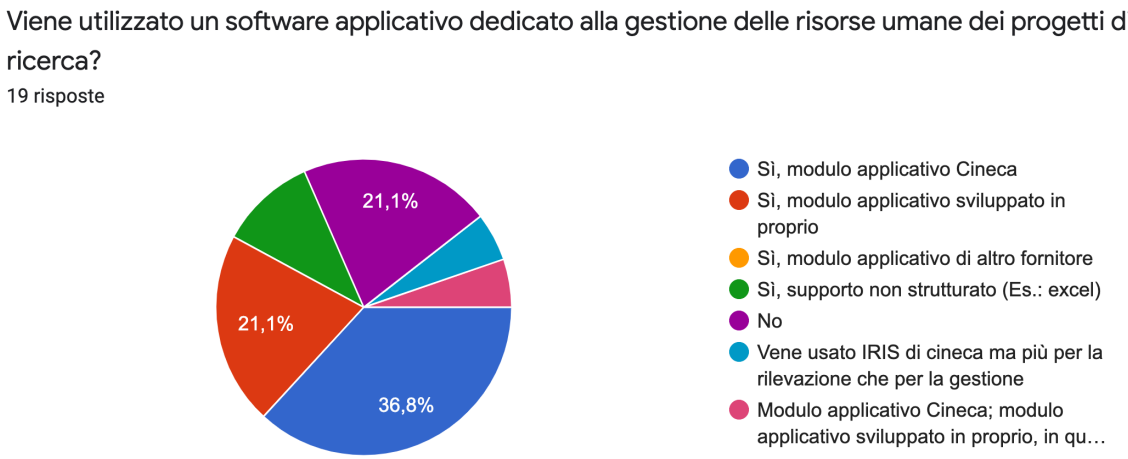

Figura 24 - Distribuzione delle risposte: software per la gestione delle risorse umane

Alla domanda a risposta aperta “Indicare quali altri applicativi in aggiunta ai precedenti vengono utilizzati all'interno del processo della ricerca“ abbiamo ricevuto diverse risposte interessanti che riportiamo singolarmente.
In un caso è stato segnalato l'utilizzo del software smartsheet (43) che è una piattaforma in cloud che permette il lavoro in collaborazione sui progetti.
Un’altra risposta ricevuta è stata la seguente: *“è stato sviluppato un gestionale per coprire un buco nel supporto alla gestione di progetti, contratti, prove e servizi. Si integra con altre applicazioni (UGOV per la parte di contabilità, Timesheet) e si integrerà con altri ambiti (IRIS per i progetti e i prodotti ricerca, Titulus per il protocollo, Applicazione di gestione Sicurezza e DVR Attività, Applicazione a supporto gestione GDPR) in un progetto pluriennale*”. Molto interessante, a nostro parere, il fatto che questo applicativo funga da collante tra i diversi applicativi Cineca e non.
Una terza risposta ricevuta ci dice che esiste un applicativo sviluppato da un fornitore esterno solo per la gestione dei progetti di ateneo. In un caso invece è stato segnalato che è in valutazione Research Professional di ExLibris (44) sul supporto delle domande-offerte relative a filoni di ricerca e finanziamenti.
In un'altra risposta è stato segnalato l’utilizzo di “*applicativi di selezione e risorse umane (es. candidatura assegno di ricerca); applicativo per gestire i finanziamenti interni; piattaforma di generazione cedolini e attestazione di costo annuo (collegata a CSA e a timesheet)*”.
Infine in due hanno segnalato in questa sezione l’utilizzo di IRIS.
Dalle risposte in questa sezione si deduce che Cineca è riuscita a imporsi e a fornire soluzioni software standard per gli atenei in maniera effettiva. Sembra quasi che Cineca abbia sfruttato la potenzialità di U-GOV Contabilità che è usato in tutte le Università (dall’introduzione dell’obbligo della contabilità economico-patrimoniale) per proporre moduli software interoperabili con la parte contabile. Poiché questa scelta sembra essere stata vincente, allo stesso modo, crediamo che Cineca potrebbe utilizzare il fatto che IRIS sia utilizzato da tutte le Università per migliorare la gestione dei processi relativi alla ricerca.

Catalogazione e valutazione dei risultati: disponibilità interna

In questa sezione ci siamo focalizzati sulla catalogazione per uso interno all'ente da parte dei gruppi di ricerca. Abbiamo quindi domandato quali applicativi vengono utilizzati per permettere ai gruppi di ricerca di collaborare internamente mediante la condivisione dei dati risultati della ricerca.
Dalle risposte si evince che gli strumenti di condivisione, quando utilizzati, vengono usati limitatamente al singolo gruppo di ricerca e “*i gruppi di ricerca sono molto autonomi e molto spesso non abbiamo una visibilità su strumenti usati e consuetudini di gestione e collaborazione*” o ancora “*Non c’è una gestione uniforme, esclusiva o veicolata a livello di Ateneo. Ciascun gruppo di ricerca adotta sistemi, software o piattaforme per la gestione e la condivisione dei dati relativi all’implementazione del progetto per il proprio gruppo*”[6].
A livello di gruppo di ricerca, i sistemi di condivisione citati sono molto vari: smartsheet, Google Drive, Office 365 e Teams, storage su cloud, DropBox, Excel, Word, email/attachment, data repository disciplinari.

[6] Estratto di risposte al survey

Un ateneo sta attivando una “piattaforma per la condivisione/disseminazione dei dati della ricerca”, mentre due atenei utilizzano un repository istituzionale dei dati[7].

Catalogazione e valutazione dei risultati: disponibilità esterna

Scopo di questa sezione è comprendere in che misura gli atenei ritengono che i dati prodotti dalle ricerche rappresentino un mezzo per dimostrare all’esterno la qualità dei propri studi, infatti le domande si sono concentrate sulle politiche degli atenei relative alle modalità di divulgazione e fruizione dei dati prodotti verso chiunque, all'esterno dell'ente.
Da questa rilevazione emerge che il 68,4% degli atenei possiede delle policy per l’open data e, di questi, solo 23,5% include, nelle proprie policy, anche i dati prodotti dalle ricerche.
Da notare che la percentuale degli atenei che possiedono policy per l’open data (68,4%) è coerente con la percentuale di atenei che prevede lo stanziamento di fondi appositi per assicurare la conservazione dei dati digitali della ricerca (60%).
Al contrario, il fatto che solo il 23,5% degli atenei includa nelle proprie policy anche i dati prodotti dalla ricerca fa pensare che le attività necessarie alla gestione e conservazione dei dati non siano in generale frutto di strategie di ateneo ma che siano lasciate alle iniziative personali dei ricercatori. Una ragione per cui le policy riguardano principalmente le pubblicazioni si potrebbe ricercare nell’attuale normativa sulla valutazione dei risultati della ricerca a livello nazionale che utilizza come indicatore principale proprio le pubblicazioni prodotte creando un sistema dove ciò che conta è l’articolo finale e non il processo che lo ha generato. Noi ci troviamo invece d’accordo con la frase “quello che conta è il viaggio e non l'arrivo”.
I due grafici delle figure sottostanti riassumono quanto detto.

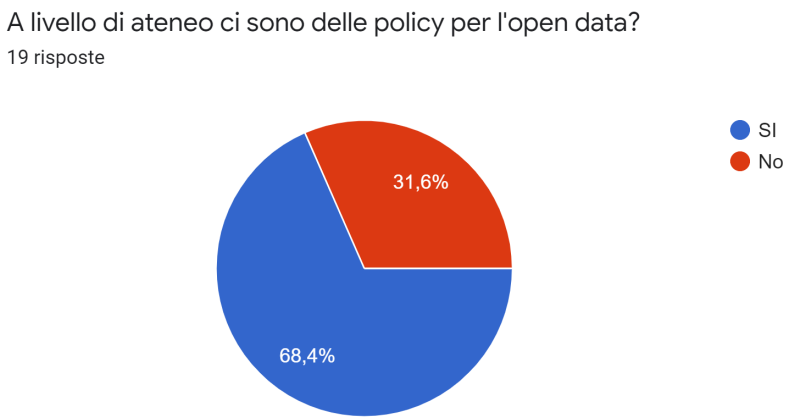

Figura 25 - Distribuzione delle risposte: policy per l’open data

[7] L’Università Federico II di Napoli utilizza fedOA (http://www.fedoa.unina.it). L’Università di Milano utilizza Dataverse, organizzato per dipartimento, laboratorio, gruppo di ricerca. Il Politecnico di Milano sta sperimentando da tre anni MendeleyData, uno strumento Elsevier

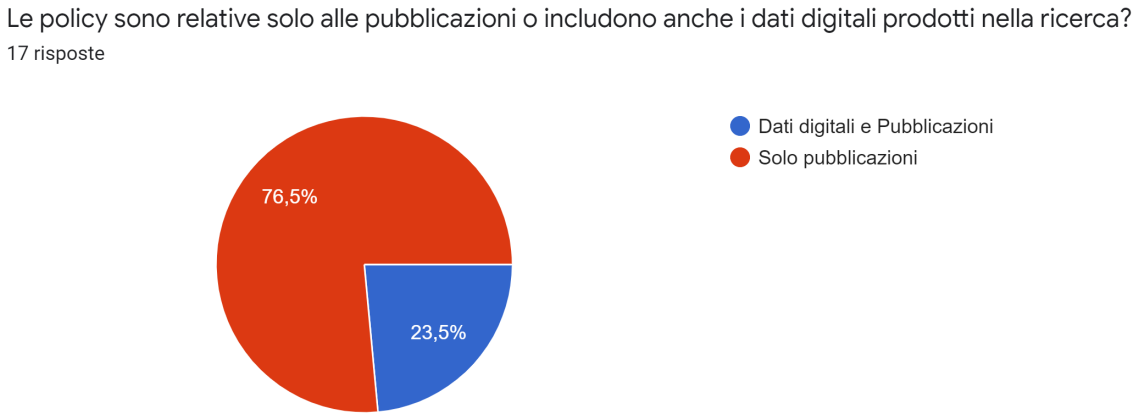

Figura 26 - Distribuzione delle risposte: policy per i dati digitali

Il 73,7% degli intervistati dichiara di possedere un sistema di conservazione centralizzato dei dati prodotti dalla ricerca, ma alla domanda "Indicare quali data repository vengono utilizzati" le 14 risposte che ci sono pervenute si possono riassumere così:

Figura 27 - Distribuzione delle risposte: strumenti per la conservazione dei dati

- quattro atenei dichiarano di utilizzare IRIS (esclusivamente per le pubblicazioni),
- un ateneo utilizza Dataverse,
- un ateneo Eprints insieme a altri repository disciplinari accreditati dalla comunità scientifica di riferimento,
- tre atenei utilizzano Zenodo,
- un ateneo utilizza Eprints,
- quattro atenei non sanno o non ne hanno nessuno.

Le risposte aperte sono contrastanti con la risposta precedente in cui la maggior parte degli intervistati ha dichiarato di possedere un sistema centralizzato per la conservazione dei dati, questo perché probabilmente la maggior parte degli intervistati aveva in mente le pubblicazioni e non i dati quando ha risposto.

Le 14 risposte sulla disponibilità di risorse economiche per garantire la manutenzione sul lungo periodo del repository sono così suddivise:

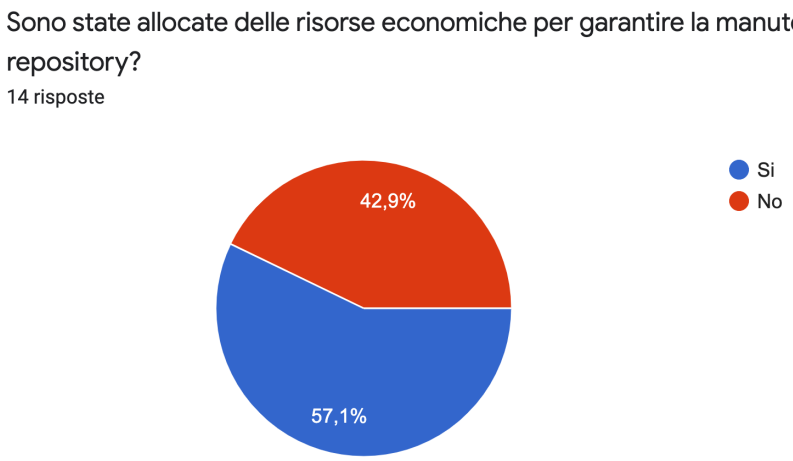

Figura 28 - Distribuzione delle risposte: risorse economiche per la manutenzione

In più della metà dei casi sono state stanziate delle risorse economiche per la manutenzione del repository. Questo è particolarmente vero ad esempio per il software Cineca che spesso prevede contratti triennali per l'hosting e le licenze.

Valorizzazione dei risultati: disponibilità interna

Scopo di questa sezione è comprendere in che misura gli atenei ritengono che i dati prodotti dalle ricerche possano essere in qualche modo riutilizzabili anche in futuro da parte degli altri ricercatori all'interno dell'ente.
Abbiamo ritenuto utile investigare, oltre che all'ovvio utilizzo di un data repository per indicizzare, sul ciclo di vita dei siti web (conservazione a lungo termine e loro dismissione) che spesso vengono prodotti per i progetti di ricerca e che costituiscono lo storico digitale delle ricerche. Nello specifico su 19 intervistati solo 3 hanno dichiarato di conservare i siti staticizzandoli.

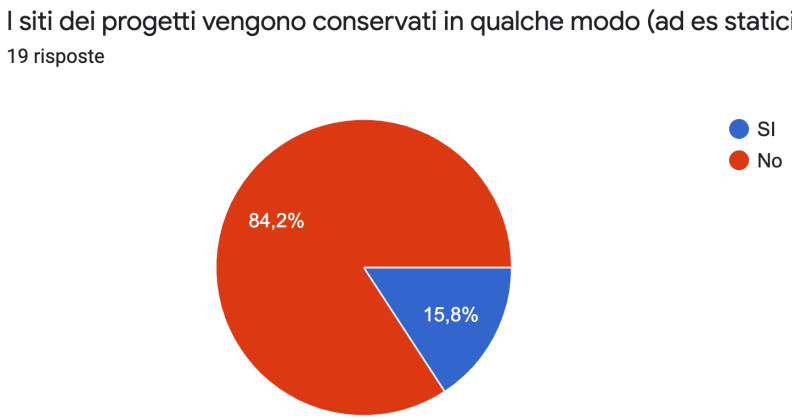

Figura 29 - Distribuzione delle risposte: usanza di staticizzare i siti web

Alla domanda *"Per quale motivo viene dismesso un sito di un progetto di ricerca"* abbiamo ricevuto molte risposte diverse che riteniamo utile riportare per esteso:

Risposta	**Num. risposte**
Vincolo temporale (numero di anni)	3
Vincolo tecnologico (obsolescenza)	1
Vincolo economico (esaurimento risorse)	2
Dietro consenso del responsabile scientifico	1
A seguito di una valutazione scientifica	1
Manca una politica istituzionale di conservazione volta a regolare tale pratica	1
A discrezione dei dipartimenti	1
A discrezione del gruppo di ricerca	1
A volte per l'esaurimento dei fondi e a volte per il disinteresse degli stessi ricercatori	1
Informazione non nota a livello centrale	3
In caso di mancata manutenzione	1
I siti che sviluppiamo internamente vengono staticizzati dopo la fine del progetto; non tutti i siti però sono sviluppati internamente	1

L'intervistato dell'ultima risposta ha detto che nel suo ateneo esistono delle politiche per definire i siti dei progetti in maniera standard con un framework specifico. Riteniamo che questa sia sicuramente una politica molto utile che dovrebbe essere utilizzata come buona pratica da parte di tutti.

Valorizzazione dei risultati: disponibilità esterna

Scopo di questa sezione è comprendere in che misura gli atenei ritengono che i dati prodotti dalle ricerche possano essere in qualche modo consultabili anche in futuro da parte degli altri ricercatori di altri enti.
Qui abbiamo chiesto sia relativamente alle policy di conservazione dei dati digitali dei progetti chiusi, che potrebbero ad esempio essere richiesti da altri, sia se in qualche modo vengono conservate delle statistiche di accesso a lungo termine sui dati digitali che eventualmente vengono conservati anche in un data repository.
Riportiamo il riepilogo delle risposte nelle due figure sottostanti.

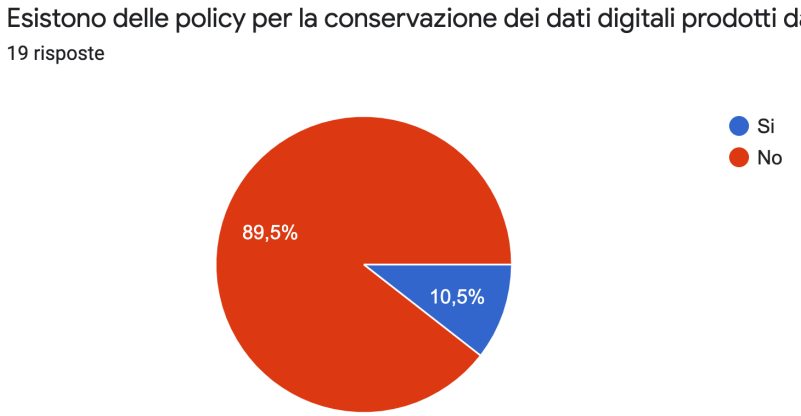

Figura 30 - Distribuzione delle risposte: policy di conservazione dei dati prodotti da progetti di ricerca chiusi

Figura 31 - Distribuzione delle risposte: statistiche di raggiungibilità dei siti dei progetti dopo 5 anni

Purtroppo le risposte che abbiamo ricevuto sono state quasi tutte negative e questo può significare che non esiste ancora, da parte degli atenei, un interesse nella conservazione a lungo termine forse perché, essendo una questione relativamente nuova, non si è ancora creata una "cultura della valorizzazione del dato prodotto", e quindi, una volta terminata la ricerca, il dato digitale viene abbandonato con le conseguenti problematiche tecnologiche e di sicurezza di cui abbiamo già parlato (cfr. paragrafo Com'è cominciato tutto: il caso SNS) anziché diventare elemento centrale da valorizzare.

7. La nostra proposta

Il percorso e le analisi fatte ci hanno permesso di comprendere diversi aspetti della gestione dei dati digitali della ricerca. Ripercorrendo a ritroso il percorso fatto possiamo dire che il questionario ci ha chiarito l'organizzazione degli atenei relativamente alla gestione dei dati digitali; l'analisi della filiera della ricerca ci ha permesso di individuare le fasi coinvolte nella gestione e conservazione dei dati digitali e gli approfondimenti sulle raccomandazioni della comunità europea ci hanno fornito gli strumenti operativi (DMP e repository di dati).
Possiamo quindi dire che la gestione dei dati digitali della ricerca è un problema impegnativo e necessita di un coinvolgimento di tutti i livelli organizzativi dell'ente, ma non solo.
Si potrebbe pensare di cercare una soluzione inserendola a livello strategico dell'ateneo (piano strategico) per poi dare vita a diverse azioni a tutti i livelli che implementano la strategia.
Non ci dobbiamo però dimenticare le considerazioni che sono emerse dal questionario che ci hanno portato a pensare che ci sia anche una questione di livello nazionale (relativo alla concezione della valutazione della ricerca) o al fatto che effettivamente le linee guida della comunità europea funzionano almeno per i progetti di ricerca europei e quindi la soluzione che vorremmo proporre potrebbe iniziare proprio a questo livello e a cascata coinvolgere tutti gli enti che producono dati di ricerca. Prendendo infatti spunto dal *Data Curation Centre* (45) inglese potremmo pensare che un organismo di natura nazionale possa fornire le linee guida comuni.
La struttura della soluzione che vorremmo proporre è quindi multilivello e può essere schematizzata come in figura.

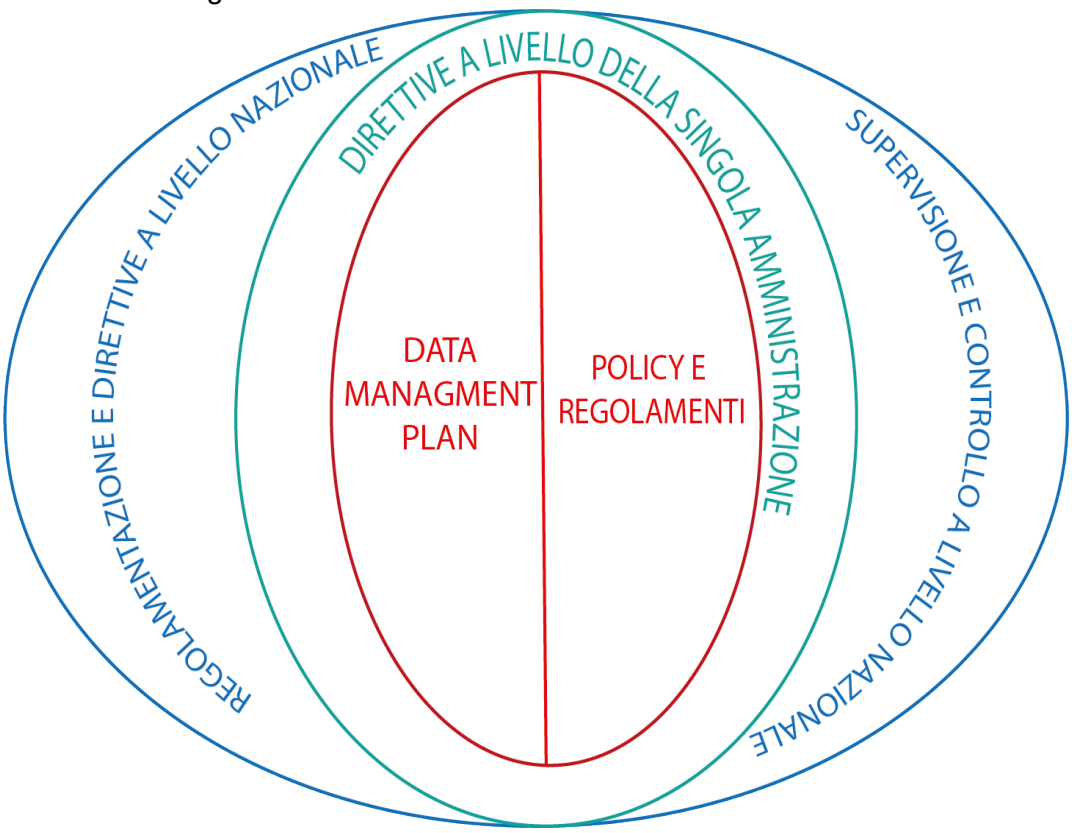

Figura 32 - Schematizzazione del framework proposto

A livello nazionale sarebbe appunto opportuna **l'istituzione di un organismo che fornisca le linee guida** compatibilmente con i prossimi framework di progetto della Comunità Europea, e supervisioni la produzione dei dati dei centri di ricerca.

A livello amministrativo gli enti potrebbero agire su diversi fronti:

- tecnico
- riorganizzazione amministrativa.

Sul fronte tecnico si potrebbe ipotizzare di **integrare i DMP direttamente nel software di gestione per la presentazione delle domande di progetto**. Ci immaginiamo infatti che sia molto utile integrare i sistemi informativi degli atenei direttamente con gli applicativi gestionali per tenere traccia delle domande di progetto presentate. Come risulta dalle risposte al survey gli atenei non hanno adottato un software simile sebbene Cineca lo renda disponibile. Le domande di progetto infatti vengono presentate direttamente sui portali degli enti finanziatori. In questo modo, non solo si perde a livello centrale traccia del lavoro svolto, ma non rimane alcun riferimento, ad esempio, dei progetti che sono stati finanziati internamente dall'ateneo. Considerato il successo di IRIS dovuto soprattutto all'interoperabilità con loginmiur (35) e alla produzione automatiche di metriche compatibili, Cineca stessa potrebbe pensare di integrare il suo software con i portali degli enti finanziatori. Il modello di software interoperabile pare riscuotere un grande successo presso gli atenei sebbene richieda da parte di Cineca un continuo lavoro di integrazione. Un'utile caratteristica che il software di memorizzazione delle proposte e dei DMP dovrebbe avere è quella di permettere il *versioning* specialmente per la sezione relativa ai DMP che come abbiamo visto sono un documento vivo.

Il passo successivo sul fronte tecnico è **l'installazione di un repository per i dati** come base per eventualmente fornire supporto a strumenti di collaborazione tra gruppi più evoluti.
Il repository di dati più completo free e open source attualmente esistente è Dataverse, auspichiamo che le Università abbiano una versione di Dataverse gestita centralmente. Come spiegato precedentemente Dataverse ha delle caratteristiche uniche che nessun altro data repository attualmente pare possedere. A livello nazionale per esempio si potrebbe collaborare allo sviluppo e alla personalizzazione di un Dataverse nazionale anche mediante estensione e plugin specializzati.
Dal punto di vista della scalabilità Dataverse è progettato per crescere e sebbene l'indicazione a livello centrale sia di spostare le sale server nel cloud, attualmente l'ostacolo principale per installare Dataverse in cloud è proprio il costo elevato dello spazio disco in hosting. Pensiamo che anche per Cineca mantenere in cloud un applicativo come Dataverse non sia sostenibile specialmente per le grosse quantità di dati che andrebbe a memorizzare (specialmente per i progetti che coinvolgono i Big Data). A livello nazionale si potrebbero produrre delle linee guida su come effettuare localmente delle installazioni di Dataverse che siano anche corrette dal punto di vista del *disaster recovery*. Tutto sommato Dataverse potrebbe essere considerato come un applicativo per la ricerca e, per questa tipologia di applicativi, non esiste ancora l'obbligo di migrare il servizio in cloud (46).

Sul fronte della ristrutturazione organizzativa di ateneo si potrebbe **definire un consiglio (board) trasversale che coinvolga diverse figure professionali**. Nelle nostre interviste qualcuno ha chiamato questo consiglio *Data Monitoring Board* (42). Tra le figure professionali coinvolte oltre che ovviamente ricercatori che operano su grandi quantità di dati, anche informatici, archivisti (che curano di dati), legali ed esperti di comunicazione. Per le figure a

livello legale queste devono essere specializzate per esempio in regolamentazione relativa alla proprietà intellettuale e alla privacy. Per quanto riguarda invece gli esperti di comunicazione, immaginiamo che tra i loro compiti ci sia quello di erogare corsi di formazione sul data management sia ai ricercatori che al personale tecnico amministrativo di supporto e la divulgazione dei dati prodotti anche all'esterno dell'ateneo.
È molto importante che a questo livello vengano evidenziati tutti i vantaggi del DMP ed è necessario fare in modo che i ricercatori lo percepiscano come uno strumento utile alla ricerca anziché l'ennesimo onere/aggravio amministrativo, per questo motivo la formazione del personale risulta essenziale.
Queste modifiche amministrative potrebbero anche avere un impatto sui processi amministrativi relativi alla ricerca.

Sullo stesso livello abbiamo posto DMP e Policy e Regolamenti. Per quanto riguarda il DMP vengono definite le informazioni che debbono essere richieste per ogni progetto aggiornando una checklist come quella descritta nel paragrafo relativo al DMP (Capitolo 3). Una checklist iniziale potrebbe essere definita a livello nazione (dall'organismo dedicato) e ogni singolo ateneo potrebbe specializzarla in base alle proprie particolarità. Nella nostra visione il DMP si specializza durante tutto il periodo di vita del progetto. Nella proposta iniziale di progetto non è necessario che questo sia compilato approfonditamente, ma in ogni caso avere un'idea dei dati coinvolti è utile a scrivere meglio una proposta di progetto vicina alla realtà. Se il progetto viene poi accettato, per esempio prima di ogni deliverable si potrebbe specializzare il DMP con lo stato dell'arte attuale. Il software di gestione delle proposte di progetto, descritto precedentemente, dovrebbe aiutare a tenere traccia delle varie revisioni del DMP.
Una cosa molto importante che si potrebbe fare anche a livello di template è la distinzione tra disponibilità interna e disponibilità esterna dei dati. Per esempio restringere alcuni dati per motivi relativi alla privacy nella sezione relativa alla disponibilità esterna, o specificare un certo tempo di embargo. Allo stesso tempo invece permettere nella sezione relativa alla disponibilità interna, accesso aperto solo ad una parte ai gruppi di ricerca dell'ateneo.

L'altro livello del nostro framework alla pari del DMP è quello delle policy e dei regolamenti. Qui si potrebbe produrre una vera e propria check list per guidare i ricercatori.
Ci immaginiamo a questo livello delle policy che forniscano delle scelte standard per:

- l'implementazione delle tecnologie da usare per i siti web dei progetti incentivando all'uso di strumenti open e standard;
- i contratti di appalto a consulenti software con delle clausole di manutenibilità a lungo termine e/o la possibilità di avere delle copie statiche dei siti alla fine del contratto;
- i contratti con gli editori in modo che questi siano in linea con i principi di accesso aperto alla ricerca garantendo al contempo il controllo della comunità scientifica sulla propria produzione intellettuale e un prezzo equo per i servizi forniti dagli editori (es. MIT framework for publisher contracts (47));
- la conservazione dei dati da parte dei ricercatori (utilizzo del repository di ateneo).

Con il doppio livello del DMP e delle policy e regolamenti si chiude la descrizione del nostro framework. Nelle conclusioni (Capitolo 8) abbiamo riportato delle idee per sviluppi futuri. Riteniamo infatti che la struttura del nostro framework sia abbastanza generica e flessibile e ben si adatterebbe a possibili rifiniture se applicato a contesti più specifici.

8. Conclusioni e possibili sviluppi futuri

In questo capitolo finale vorremmo riassumere le conclusioni a cui siamo giunti nei singoli capitoli e dare alcune linee per possibili sviluppi futuri del lavoro.
Lo studio della conservazione dei dati digitali prodotti dalla ricerca non è un problema nuovo. Già nel 2013 gli autori di (49) avevano pubblicato uno studio comprensivo riguardante i dati prodotti all'interno degli istituti CNR italiani procedendo ad una intervista puntuale di tutti i ricercatori: l'approccio è stato quello di un report svolto dal punto di vista di chi fa ricerca.
Nel nostro lavoro invece abbiamo analizzato il problema da un punto di vista gestionale-amministrativo, cercando di analizzare l'organizzazione attuale degli atenei e proporre un diverso assetto nella gestione operativa. Il tema della gestione dei dati digitali anche da un punto di vista gestionale è un tema caldo, è infatti di questi giorni la notizia che il MIT ha cessato il contratto con Elsevier (48) perché quest'ultima non propone un contratto in linea con il MIT Framework for Publisher Contracts (47).
Nel Capitolo 1 abbiamo descritto il problema pratico ancora prima di iniziare a documentarci sulla letteratura disponibile. Va detto che il fatto che i dati dei vecchi siti dei progetti fossero richiestissimi è stata la nostra fortuna: le statistiche ci hanno infatti confermato che effettivamente questi vecchi dati hanno una valenza ancora oggi e che vengono acceduti costantemente dai ricercatori dell'ente ma anche esternamente. Siccome l'infrastruttura che memorizza i dati è tecnologicamente obsoleta ci siamo chiesti dove fosse stato il problema e se si poteva fare qualcosa per anticiparlo.
Nel Capitolo 2 abbiamo ricercato una definizione puntuale di quello che rappresentano per noi i dati digitali della ricerca di cui avevamo intenzione di occuparci per circoscrivere il problema e abbiamo analizzato due articoli molto interessanti (6) (7): il primo in particolare mediante un metodo empirico dimostra come la probabilità che i dati siano ancora esistenti, probabilità condizionata dal fatto di aver ottenuto una risposta utile al messaggio di richiesta dei dati, decresce del 17% ogni anno da quanto è stata pubblicata la ricerca. Nel secondo articolo invece viene proposto uno studio relativo alle motivazioni che portano gli studiosi alla condivisione dei dati ed in particolare, motivazioni personali come il fatto di percepire un beneficio per la carriera, oppure un rischio e dalle motivazioni personali alla condivisione. Un altro aspetto fondamentale sembra essere l'influenza amministrativa. Il risultato di questi due studi è molto interessante: il primo ci dice che i dati della ricerca subiscono un declino col tempo, il secondo che la disponibilità a condividere dati è principalmente un fatto di natura personale e può essere migliorato se imposto da una regolamentazione a livello amministrativo.
Nel Capitolo 3 invece abbiamo guardato alle specificità dei progetti Data Pilot (ORD) definiti all'interno del progetto H2020. Un utile strumento richiesto per questi tipi di progetti è il Data Management Plan (DMP). Ci siamo resi conto che diversi enti finanziatori richiedono di strutturare il DMP con informazioni differenti e che effettivamente Italian Open Science Support Group aveva stilato una check list utile di informazioni basilari che ogni DMP dovrebbe avere (20).
Nel Capitolo 4 abbiamo spostato la nostra attenzione sugli applicativi utilizzati all'interno dell'Università andando a intervistare proprio il consorzio Cineca che sviluppa applicativi standard per le Università. In particolare abbiamo analizzato una white paper di Cineca (22) che illustra sia la filiera della ricerca come catena del valore dei processi coinvolti che vanno dalla proposta di un progetto fino alla relativa rendicontazione. Insieme alla descrizione dei

processi è evidenziato il relativo applicativo gestionale Cineca coinvolto. Abbiamo proposto quindi una modifica alla catena del valore che tenga traccia anche della possibilità di pubblicare dati digitali della ricerca sia all'interno che all'esterno dell'ente che abbiamo chiamato semplicemente disponibilità interna e disponibilità esterna (Figura 9). Nella visione originale di Cineca il processo si conclude con il deposito dell'articolo nell'Institutional Repository IRIS, senza considerare i dati digitali prodotti.
La disponibilità interna nella nostra visione è relativa ai dati digitali pubblicati ad uso interno dell'ente e dei gruppi di ricerca, mentre la disponibilità esterna è relativa ai dati digitali che l'ente vorrebbe pubblicare e rendere utilizzabili da parte di tutti. Abbiamo infine notato che l'introduzione della disponibilità interna e della disponibilità esterna si ripercuote anche sui processi e sui software applicativi e per questo abbiamo proposto la relativa modifica (Figura 11).
Nel Capitolo 5 abbiamo studiato i data repository fornendo un'analisi quantitativa sull'utilizzo di Zenodo (data repository aperto e ospitato dal CERN (23)) da parte delle Università italiane. La tendenza all'utilizzo di Zenodo sembra in crescita sebbene abbia avuto un lieve arresto nel 2017, anno in cui è stato reso disponibile IRIS (Institutional Repository) (34) da parte di Cineca. Abbiamo anche provato a fare un'analisi di cosa è effettivamente memorizzato su Zenodo utilizzando un semplice algoritmo di machine learning sulle keyword degli articoli (Figura 15). La somiglianza delle keyword sembra suggerire che Zenodo sia utilizzato soprattutto in ambito scientifico. La scelta di Zenodo tuttavia dovrebbe essere solo in mancanza di una valida alternativa di un data repository istituzionale o specifico per un particolare dominio applicativo.
Relativamente alle scelte che gli atenei avrebbero a disposizione per l'installazione di un data repository in locale, la scelta dovrebbe ricadere sicuramente su Dataverse (24). Come ampiamente illustrato Dataverse rappresenta lo stato dell'arte dei data repository open source per le sue particolarità che lo rendono unico. Altre scelte potrebbero ricadere su Eprints o dSPACE ma perdendo tantissimo in flessibilità.
Nel Capitolo 6 abbiamo formulato un survey da inviare alle Università italiane suddividendo le domande in sezioni sulla base della catena del valore introdotta nel Capitolo 4. Riteniamo che formulare preventivamente una suddivisione in sezioni delle domande ci abbia aiutato tantissimo nella formulazione delle domande permettendoci di non tralasciare nessun aspetto. Il questionario è stato compilato in totale da 19 Università situate per la maggior parte al Nord Italia. I risultati sono ampiamente descritti nel capitolo, ma sembra che quello della gestione dei dati digitali sia un problema caldo che tutti gli atenei si stanno ponendo e stanno cercando di risolvere.
Nel Capitolo 8 abbiamo descritto, sulla base dell'esperienza maturata, la nostra proposta per risolvere il problema. Il nostro framework a quattro blocchi (Figura 32) inizia a livello più ampio a livello nazionale per concludersi a livello più stretto con Data Management Plan insieme a policy e regolamenti per creare il risultato gestionale sperato.
Per un lavoro futuro si potrebbe pensare di estendere il questionario a tutti gli atenei e magari effettuare uno studio anche per singolo dipartimento. Il survey è un metodo veloce di analisi ma ha diverse problematiche, prima tra tutte la difficoltà nel reperire qualcuno che abbia tutte le risposte e poi il fatto che le domande possono essere non capite. Delle interviste puntuali in luogo cercando di capire come effettivamente funziona sarebbero sicuramente la soluzione da preferire. Prendendo spunto da queste interviste si potrebbe stilare una checklist che estende il nostro framework e creare una vera e propria matrice puntuale di regole da seguire

a livello di policy e regolamenti. Se è vero che per ora non è possibile agire a livello nazionale sicuramente è possibile migliorare la gestione fino al livello amministrativo.

Bibliografia

1. *Consorzio OpenAIRE.* [Online] https://www.openaire.eu/.
2. OpenAIRE Research Data Management Briefing paper. *Consorsizio OpenAIRE.* [Online] https://www.openaire.eu/briefpaper-rdm-infonoads.
3. Queensland University of Technology. *Guidelines for the Management of Research Data at QUT.* [Online] https://www.library.qut.edu.au/research/data/documents/GDL_GuidelinesResearchData.pdf.
4. Occioni, Marisol e Vignocchi, Marialaura. Presentazione. *How to get started with the research data management plan.*
5. Ganguly, Raman. Data from Research Processes: from raw data to open access published data. *Figura da sito.* [Online] http://phaidra.univie.ac.at/o:387241.
6. *The Availability of Research Data Declines Rapidly with Article Age.* Timothy H.Vines, Arianne Y.K.Albert, Rose L.Andrew, Florence Débarre, Dan G.Bock, Michelle T.Franklin, Kimberly J.Gilbert, Jean-Sébastien Moore, Sébastien Renaut, Diana J.Rennison. 1, Vol. Current Biology, 24.
7. *Social scientists' data sharing behaviors: Investigating the roles of individual motivations, institutional pressures, and data repositories.* Adler, Youngseek Kim e Melissa. 4, Vol. International Journal of Information Management, 35.
8. *What is the EC Open Research Data Pilot?* [Online] https://www.openaire.eu/what-is-the-open-research-data-pilot.
9. European Open Science Cloud (EOSC). [Online] https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud.
10. Metadata Standards Directory Working Group. [Online] http://rd-alliance.github.io/metadata-directory/.
11. Dublic Core Metadata Initiative. [Online] https://dublincore.org/.
12. DataCite Metadata Schema. [Online] https://schema.datacite.org/.
13. The Creative Commons License Options. [Online] https://creativecommons.org/about/cclicenses/.
14. *How to complete an outputs management plan.* [Online] https://wellcome.ac.uk/grant-funding/guidance/how-complete-outputs-management-plan.
15. FAIRsharing.org. *A curated, informative and educational resource on data and metadata standards, inter-related to databases and data policies.* [Online] https://fairsharing.org/ .
16. Erc Data Management Plan Template. [Online] https://erc.europa.eu/content/erc-data-management-plan-template.
17. Digital Object Identifier. [Online] https://www.doi.org/.
18. DMPONLINE. [Online] https://dmponline.dcc.ac.uk/.
19. Example DMPs and guidance. [Online] https://www.dcc.ac.uk/resources/data-management-plans/guidance-examples.
20. Italian Open Science Support Group . *Materiali.* [Online] https://sites.google.com/view/iossg/materialimaterials.
21. Chi Siamo. *Consorzio Cineca.* [Online] https://www.cineca.it/it/content/chi-siamo.
22. U-GOV Ricerca - Il sistema per la gestione della ricerca in ateneo. *White paper.*
23. ZENODO. [Online] https://zenodo.org/.
24. The Dataverse Project. [Online] https://dataverse.org/.
25. UniProt.org. [Online] https://www.uniprot.org/.
26. The Image Data Resource (IDR). [Online] https://idr.openmicroscopy.org/.

27. 4SCIENCE. “The Dataverse Project” Open source research data repository software. *Presentazione.*
28. re3data.org. *REgistry of REsearch data REpositories.* [Online] http://re3data.org/.
29. CoreTrustSeal Trustworthy Data Repositories Requirements 2020–2022. [Online] https://zenodo.org/record/3638211.
30. ZENODO | Developers. [Online] https://developers.zenodo.org/.
31. ZENODO | oai2d. [Online] https://zenodo.org/oai2d.
32. Elasticsearch a distributed, RESTful search and analytics engine. [Online] https://www.elastic.co/elasticsearch/.
33. word2vec. *Tool for computing continuous distributed representations of words.* [Online] https://code.google.com/archive/p/word2vec/.
34. IRIS day registrazione evento. [Online] https://streaming.cineca.it/DefaultPlayer/div.php?evento=irisday.
35. Loginmiur unico punto di ingresso per i docenti ed i ricercatori. [Online] https://loginmiur.cineca.it/front.php/login.html.
36. OpenAIRE | Explore. [Online] https://explore.openaire.eu/.
37. Open Archives Initiative Protocol for Metadata Harvesting. [Online] https://www.openarchives.org/pmh/.
38. Eprints. [Online] https://www.eprints.org/uk/.
39. dSPACE. [Online] https://www.dspace.com/en/pub/home.cfm.
40. Internet Archive. [Online] https://web.archive.org/.
41. Data Management Plan per i ricercatori del politecnico di Torino. [Online] http://www.biblio.polito.it/open_access/horizon_2020/il_data_management_plan_dmp/data_management_plan_per_i_ricercatori_del_politecnico.
42. Data Management Plan per i ricercatori della Ca' Foscari. [Online] https://www.unive.it/pag/19979/.
43. smartsheet. [Online] https://it.smartsheet.com/.
44. *Research Professional.* [Online] https://knowledge.exlibrisgroup.com/Research_Professional.
45. Digital Curation Centre. [Online] https://www.dcc.ac.uk/.
46. AGID Processo di razionalizzazione dei data center pubblici e formazione dei PSN. [Online] https://www.agid.gov.it/it/argomenti/data-center.
47. MIT Framework for Publisher Contracts. [Online] https://libraries.mit.edu/scholarly/publishing/framework/.
48. MIT, guided by open access principles, ends Elsevier negotiations. [Online] http://news.mit.edu/2020/guided-by-open-access-principles-mit-ends-elsevier-negotiations-0611.
49. Luzi, Daniela, Caruso, Maria Girolama e Crescimbene, Cristiana. Indagine sui dati della ricerca nel settore ambientale: percezioni e pratiche dei ricercatori Cnr. *IRPPS Working Papers.* 2014.

Indice delle figure