Stefano Bargioni, Carlo Bianchini, Camillo Carlo Pellizzari di San Girolamo[1]

IRIS e Wikidata: un progetto per una migliore valorizzazione e fruizione degli archivi della ricerca scientifica italiana

1. Introduzione

IRIS (Institutional Research Information System) è un servizio di *repository* dei 'prodotti' della ricerca destinato alle università e agli istituti di ricerca italiani. È sviluppato e mantenuto dal consorzio Cineca[2]:

> IRIS consente l'archiviazione, la consultazione e la valorizzazione dei prodotti scaturiti dalle attività di ricerca. È un sistema unico e integrato con le altre soluzioni Cineca, nonché in grado di dialogare con i sistemi centrali nazionali ed internazionali per la gestione e la disseminazione delle pubblicazioni, conforme ai requisiti del MUR e della Commissione Europea per l'Open Access. Il repository è basato sulla piattaforma tecnologica internazionale DSpace. Il sistema è inoltre integrato con i più importanti provider di metadati editoriali e informazioni bibliometriche internazionali (Web Of Science, Scopus, CrossRef, PubMed…)[3].

Ogni IRIS permette il censimento e la gestione di tutti i 'prodotti' della ricerca, le pubblicazioni scientifiche prodotte come risultato delle ricerche svolte da un'istituzione; per esempio, libri, contributi su rivista, contributi in atti di convegno, curatele, tesi, brevetti ecc. In quanto strumento di controllo dei risultati della ricerca, IRIS risulta di primaria importanza per la VQR, ovvero la Valutazione della Qualità della Ricerca delle Università e degli Enti di ricerca svolta su base quinquennale dall'ANVUR (Agenzia Nazionale di Valutazione del sistema Universitario e della Ricerca)[4]. Tuttavia, in ogni deposito istituzionale esiste il rischio di pubblicazione di informazioni frammentate e di ridondanza e spesso la necessità di consultare più depositi per avere un quadro completo della situazione di un ricercatore[5].

[1] Il testo è stato scritto in totale collaborazione e condivisione: tuttavia vanno ascritti a Stefano Bargioni i paragrafi 1, 2.0, 2.1, 2.2, 3.1 e 3.2, a Carlo Bianchini i paragrafi 4 e 5 e a Camillo Carlo Pellizzari di San Girolamo i paragrafi 2.3 e 3.3.
[2] <https://www.cineca.it>. (Ultima consultazione di tutte le risorse online: 15 settembre 2023).
[3] Cineca, *IRIS: la gestione della Ricerca*», 9 agosto 2018, <https://tinyurl.com/45jknrjd>.
[4] Si veda il d.p.r. n. 76 del 2010.
[5] Miriam Baglioni [*et al.*], *(Semi)automated disambiguation of scholarly repositories*, «arXiv», (2023), <https://doi.org/10.48550/arXiv.2307.02647>.

IRIS è in uso dal 2015, ed è utilizzato da oltre 80 istituzioni accademiche italiane, prevalentemente universitarie. Le istanze di IRIS hanno in comune la componente software essenziale, vale a dire il database, i metodi di indicizzazione, il web server ecc.[6]. Differiscono per alcune soluzioni grafiche e per la scelta dell'organizzazione dei dati, soprattutto in coerenza con la natura e la dimensione dell'istituzione.

In quanto membri del GWMAB[7], un gruppo italiano che si occupa dei dati di musei, archivi e biblioteche in Wikidata, gli autori si sono proposti di trovare risposta, attraverso un'analisi dei dati di IRIS svolta tramite i dati di Wikidata, ai seguenti quesiti: in IRIS vengono gestite le entità 'persona', secondo il modello IFLA LRM, il più diffuso in ambito biblioteconomico e bibliografico[8]? Il sistema IRIS è in grado di fornire un quadro unitario e coerente della ricerca italiana, secondo quanto richiesto dalla legge[9]? Il contenuto di IRIS è interoperabile con il web semantico, e in particolare con alcuni sistemi rilevanti come Wikidata[10]?

2. Materiali e metodi

2.0 I dati nei depositi istituzionali IRIS

In ogni IRIS, i dati vengono alimentati tramite l'inserimento dei 'prodotti' da parte di uno degli autori appartenenti all'istituzione di pertinenza dell'IRIS. Durante l'inserimento, il software offre alcuni ausili, per esempio per evitare duplicazioni e per permettere il riconoscimento degli autori di quella istituzione già presenti nell'anagrafica[11]. Ogni autore, interno o esterno all'istituzione, riceve un identificatore univoco nell'ambito di ogni singolo IRIS: si tratta di un codice composto dalle lettere "rp" seguite da 4, 5 o 6 cifre (per esempio <https://iris.uniroma1.it/cris/rp/rp36632>). La funzione di identifica-

[6] Il server fisico è sempre lo stesso, dato che i nomi dei server negli indirizzi https sono quasi tutti alias di surplusfe-prod-01.cineca.it.

[7] *Gruppo Wikidata per Musei, Archivi e Biblioteche*. 24 luglio 2023, <https://tinyurl.com/7e5rvdua>.

[8] *IFLA Library Reference Model. A Conceptual Model for Bibliographic Information*, a cura di Pat Riva, Patrick Le Boeuf e Maja Zumer. Den Haag: IFLA, 2017.

[9] L'art. 3-bis della l. n. 1 del 1° gennaio 2009 ha modificato il precedente d.l. n. 180 del 10 novembre 2008, e prevede che sia costituita un'«Anagrafe nazionale nominativa dei professori ordinari e associati e dei ricercatori, contenente per ciascun soggetto l'elenco delle pubblicazioni scientifiche prodotte. L'Anagrafe è aggiornata con periodicità annuale».

[10] Hilary Thorsen, *Wikidata as a hub for identifiers*, (11 giugno 2020), <https://t.ly/677QP>.

[11] Il manuale di IRIS illustra il processo di registrazione dei 'prodotti' alle pagine <https://tinyurl.com/4kxztrtj> e <https://tinyurl.com/3sfc7khh>.

zione degli autori, tuttavia, generalmente non mostra né dati anagrafici (per esempio, l'anno di nascita) né identificatori internazionali, quali l'ORCID[12]. Se un ricercatore ha collegato il proprio profilo all'ORCID, infatti, il sistema lo include nei metadati del 'prodotto' (come spiegato in dettaglio più avanti), ma non tra i dati pubblici relativi all'autore.

L'associazione degli autori esterni – che non hanno un identificatore "rp" – è quindi affidata al solo cognome e al nome, spesso rappresentato dalla sola iniziale.

Le istituzioni accademiche che adottano IRIS sono state desunte dall'elenco mantenuto dal Cineca[13]. A quelle in elenco (80 a luglio 2021) è stata aggiunta <www.iris.unicampus.it>; è possibile che quest'ultima non sia presente nell'elenco suddetto per motivi di aggiornamento. L'elenco completo delle basi trattate in questo studio si trova nel file "operazioni di scrape.ods" su Zenodo[14].

2.1 Definizione delle proprietà in Wikidata

L'assegnazione di un identificatore IRIS a ogni ricercatore e la sua visibilità attraverso la piattaforma sono stati i presupposti di questa ricerca, che si basa sulla riconciliazione delle entità 'persona' presenti in ogni IRIS con i corrispondenti elementi Wikidata (e con qualsiasi altro database, presente o futuro, che assegni un identificatore univoco a un gruppo più o meno ampio di ricercatori italiani).

In Wikidata è stata creata una proprietà per gli autori presenti in ciascun IRIS. Sia per evitare di proporre 81 nuove proprietà in una sola volta, sia per l'impossibilità di procedere al prelevamento dei dati in poco tempo, le proprietà di questo progetto sono state proposte in un lasso di tempo di mesi, dal 12 luglio 2021 al 6 aprile 2023[15]. La prima proprietà Wikidata per autori IRIS, per l'IRIS di SNS, è stata creata il 27 luglio 2021. L'elenco delle proprietà è anch'esso incluso nel file "operazioni di scrape.ods" su Zenodo.

[12] <https://orcid.org/>.

[13] <https://tinyurl.com/5ftfvp4t>.

[14] Script, scrape, dati tabellari e query SPARQL sono stati caricati in Zenodo: Stefano Bargioni; Carlo Bianchini; Camillo Carlo Pellizzari di San Girolamo, *Dati e script di "IRIS, Wikidata, SBN: un progetto per una migliore valorizzazione e fruizione dei repository della ricerca scientifica italiana"*, «Zenodo», (2023), <https://doi.org/10.5281/zenodo.8345202>.

[15] Sulla procedura di creazione di una proprietà si veda Claudio Forziati; Valeria Lo Castro, *La connessione tra i dati delle biblioteche e il coinvolgimento della comunità: il progetto SHARE Catalogue-Wikidata*, «JLIS.it», 9, (2018), n. 3, p. 109-120, <https://doi.org/10/ggxj9n> e la linea guida: <https://www.wikidata.org/wiki/Wikidata:Property_creation>. Tutte le proposte sono elencate in <https://tinyurl.com/2utrbu3c>; tutte le proprietà approvate sono elencate in <https://tinyurl.com/2tpuzjc5> (lista che si aggiorna automaticamente sulla base di una query SPARQL).

2.2 Estrazione dei dati

Gli archivi IRIS non pubblicano un *dump* dei propri dati, o API che possano permettere di accedere ad essi in modo massivo; l'OAI PMH (disponibile all'indirizzo relativo "/oai") fornisce dati sui 'prodotti' della ricerca, ma tali dati sono incrementali e non includono gli identificatori degli autori. Si è quindi proceduto a uno *scrape* dei dati, estraendo le informazioni contenute nel codice html delle pagine web[16].

Nel caso degli IRIS, l'accesso all'elenco degli autori può avvenire tramite la pagina il cui indirizzo relativo è "/browse?type=author". L'elenco è paginato e ogni nome di autore è associato a un identificatore inglobato nell'indirizzo. L'identificatore NNNN univoco permette l'apertura della "Pagina ricercatore" con indirizzo relativo "/cris/rp/rpNNNN", dove sono presenti ulteriori, ma scarse, informazioni sull'autore, seguite dall'elenco, anch'esso paginato, dei suoi 'prodotti'. Nella "Pagina ricercatore" sono presenti il nome e cognome del ricercatore, nella forma indiretta "cognome, nome", e di solito l'afferenza, tipicamente un nome di dipartimento; non vengono mai visualizzati il giorno (o anno) o il luogo di nascita e di rado compaiono identificatori[17]. L'ORCID è spesso presente nel codice html tra i metadati dei singoli 'prodotti', ma non visualizzato dal browser (Figura 1)[18].

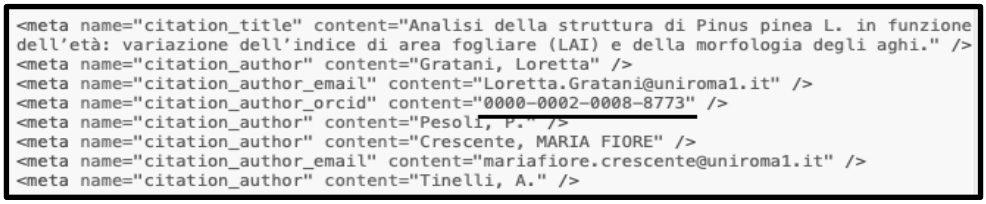

Figura 1 - Un esempio di ORCID nel codice html
(estratto da <https://iris.uniroma1.it/handle/11573/425694>)

[16] Fino al 2021, i sistemi IRIS erano basati su DSPACE CRIS. Questa versione è stata dismessa a gennaio 2022 a favore della versione basata su DSPACE6. Il cambio di versione non ha sostanzialmente avuto impatto sul presente studio, benché la nuova versione sia stata introdotta dopo il 27 luglio 2021, data di inizio della riconciliazione degli autori di un primo IRIS con Wikidata.

[17] Esempio: <https://arts.units.it/cris/rp/rp187922>.

[18] Si noti la difficoltà tecnica di associare con sicurezza l'ORCID al primo autore e non al secondo. Si intuisce che ci si debba affidare all'ordine sequenziale dei tag meta, pratica però sconsigliabile nell'ambito dei documenti strutturati.

2.3 Metodi di riconciliazione

Per procedere all'analisi dei dati presenti negli IRIS, si è provveduto a riconciliare gli identificatori degli autori in ciascun IRIS con gli elementi di Wikidata, ovvero a registrare nell'apposita sezione dell'elemento di Wikidata l'identificatore visibile nell'IRIS, procedendo con diversi metodi di riconciliazione: 1. tramite ORCID presente in IRIS; 2. tramite ORCID ricavato dal confronto dei titoli associati a un autore; 3. tramite il confronto tra gli ISBN delle monografie associate a un autore; 4. manualmente, tramite la creazione e l'uso di un catalogo Mix'n'Match (cfr. § 2.3.4).

2.3.1 Riconciliazione tramite ORCID presente in IRIS

La presenza dell'identificatore personale ORCID in IRIS[19] comporta la possibilità di farne uso ai fini della riconciliazione con Wikidata, tramite l'apposita proprietà P496[20].

Durante la fase di estrazione dell'elenco degli autori di ciascun IRIS, se in un 'prodotto' IRIS un autore ha associato un ORCID, e se il medesimo ORCID è presente in un elemento Wikidata (come valore della proprietà P496), viene generato un comando per il tool QuickStatements[21] con cui all'elemento Wikidata viene assegnato l'ID dell'autore nell'IRIS in esame.

2.3.2 Riconciliazione tramite ORCID ricavato per confronto titoli

Nel caso non sia disponibile un ORCID su IRIS, si procede a un confronto tra i titoli dei 'prodotti' dell'autore presenti nel repository e i titoli dei 'prodotti' presenti nelle schede di autori col medesimo nome sul sito di ORCID[22]. Se almeno due titoli presenti in un certo ORCID coincidono con quelli presenti in IRIS[23], si considera sicura la corrispondenza tra l'ID dell'autore nell'IRIS in esame e quello descritto in ORCID. A questo punto, si usa l'ORCID in questione per procedere alla riconciliazione come nel punto precedente.

Se comunque la riconciliazione non va a buon fine (perché non si riesce ad abbinare l'ID dell'autore nell'IRIS in esame a un ORCID, o perché l'ORCID abbinato ad esso risulta assente in Wikidata), come ultima possibilità viene generata una *entry* di Mix'n'Match (di seguito MnM), indicando se vi siano

[19] Per ORCID in IRIS si veda <https://tinyurl.com/yckw54zh>.
[20] <https://www.wikidata.org/wiki/Property:P496>.
[21] <https://quickstatements.toolforge.org/>.
[22] Per accedere ai titoli presenti in ORCID, si può fare uso delle API illustrate all'indirizzo <https://support.orcid.org>.
[23] La coincidenza di un solo titolo è stata ritenuta insufficiente, allo scopo di minimizzare casi di errata attribuzione.

parole coincidenti tra i titoli in IRIS e i titoli in orcid.org. Un'eventuale riconciliazione viene demandata al lavoro manuale futuro svolto su MnM da qualsiasi utente di Wikidata.

Il file "conteggi_qs_mnm.ods" su Zenodo riporta i conteggi e le percentuali al momento del caricamento di ogni MnM: questi conteggi danno una visione degli abbinamenti iniziali ottenuti dalle operazioni di *scrape*.

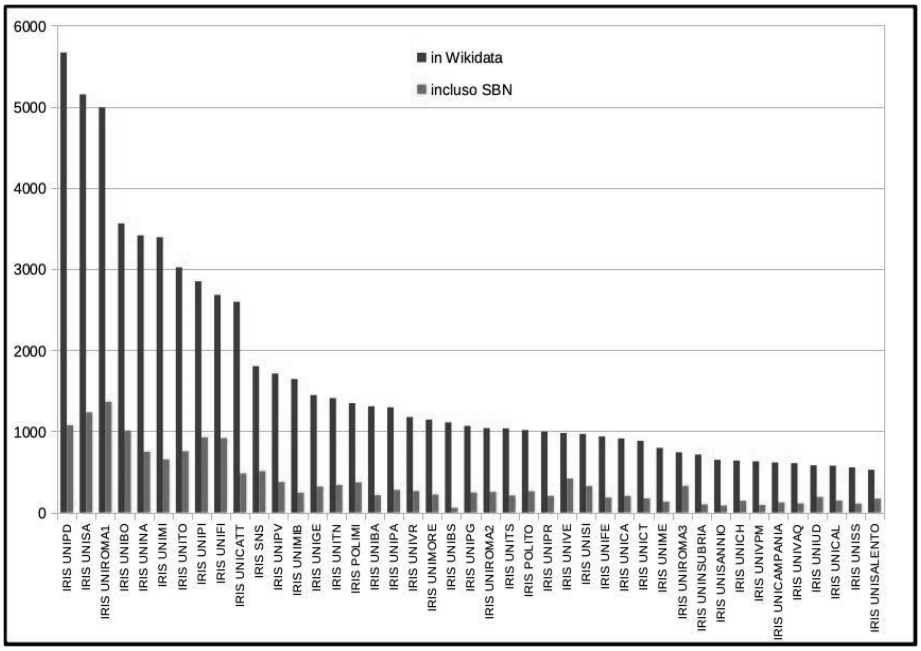

Figura 2 – Distribuzione degli abbinamenti tra Wikidata e IRIS con almeno 500 identificatori in Wikidata[24]

2.3.3 Riconciliazione tramite confronto di ISBN con SBN

Ogni IRIS raggruppa i 'prodotti' in determinate classi di propria scelta, ma tutti ne hanno una relativa alle monografie[25]. Con buona probabilità, i 'prodotti' di questa tipologia includono l'ISBN tra i metadati. Può quindi essere condotto uno scorrimento delle monografie per cercare una coincidenza tra gli ISBN presenti in IRIS e quelli presenti nell'OPAC SBN: nel caso in cui si trovi una

[24] Si vedano, per dati aggiornati, le seguenti query: numero di singole persone, per IRIS, abbinate a un elemento Wikidata, <https://w.wiki/73Li>; numero di singole persone, per IRIS, abbinate a un elemento Wikidata contenente anche un ID SBN, <https://w.wiki/73Ld>.
[25] Esempi: <https://tinyurl.com/4kmumdtk>.

corrispondenza per una monografia di un autore nell'OPAC SBN, si estrae dall'OPAC SBN il VID (cioè l'identificatore SBN) dell'autore; nel caso in cui un elemento Wikidata contenga quel VID (tramite P396[26]), è possibile inserire in quell'elemento l'ID dell'autore nell'IRIS in esame.

Per mancanza di spazio, qui viene omessa una descrizione più precisa di queste operazioni complesse. Si rimanda al codice Perl usato per questo progetto e pubblicato nel repository Zenodo nel file "Perl scripts.zip" (due eseguibili e cinque librerie). I conteggi di questa operazione sono raccolti nel file "conteggi match ottenuti tramite ISBN.ods" su Zenodo.

Il file "conteggi_iris_sbn.ods" su Zenodo presenta i conteggi di elementi Wikidata con almeno un IRIS, e quelli con SBN, e relative percentuali e un grafico.

2.3.4 Creazione dei cataloghi Mix'n'Match e riconciliazione

A partire dallo *scrape* di tutti i nomi degli autori, è stato possibile costruire un file .tsv adatto ad alimentare il riconciliatore di Wikidata denominato Mix'n'Match (abbreviato MnM)[27]. In base alla sintassi di MnM[28], la struttura tabellare del file dev'essere costituita almeno da due informazioni: identificatore dell'entità ("id") e nome dell'entità ("name"). Ove presente, è stata aggiunta una colonna ("P496") con l'ORCID e la colonna della descrizione ("desc"), composta in un modo particolare, come spiegato di seguito.

Per quanto riguarda la riconciliazione, ogni catalogo MnM raggruppa le *entry* in quattro gruppi:

1) quelle già abbinate in modo sicuro a Wikidata (grazie ai meccanismi di cui ai punti precedenti, o manualmente);
2) quelle abbinate automaticamente a uno o più elementi di Wikidata, sulla base della somiglianza tra la stringa del nome della *entry* e gli elementi testuali (etichette, descrizioni, alias, proprietà con tipo di dato stringa) presenti negli elementi di Wikidata; ciascun utente di MnM può convalidare o rifiutare questi abbinamenti automatici;
3) quelle non abbinabili a Wikidata (un utente può marcare come non abbinabile a Wikidata una *entry* che confla persone diverse, oppure non più esistente, oppure impossibile da identificare con sicurezza);
4) quelle non abbinate in alcun modo con elementi di Wikidata (perché non è stato possibile effettuare un abbinamento automatico, o perché tale abbinamento automatico è stato rifiutato da un utente).

[26] <https://www.wikidata.org/wiki/Property:P396>.
[27] <https://mix-n-match.toolforge.org/>.
[28] <https://meta.wikimedia.org/wiki/Mix'n'match/Import>.

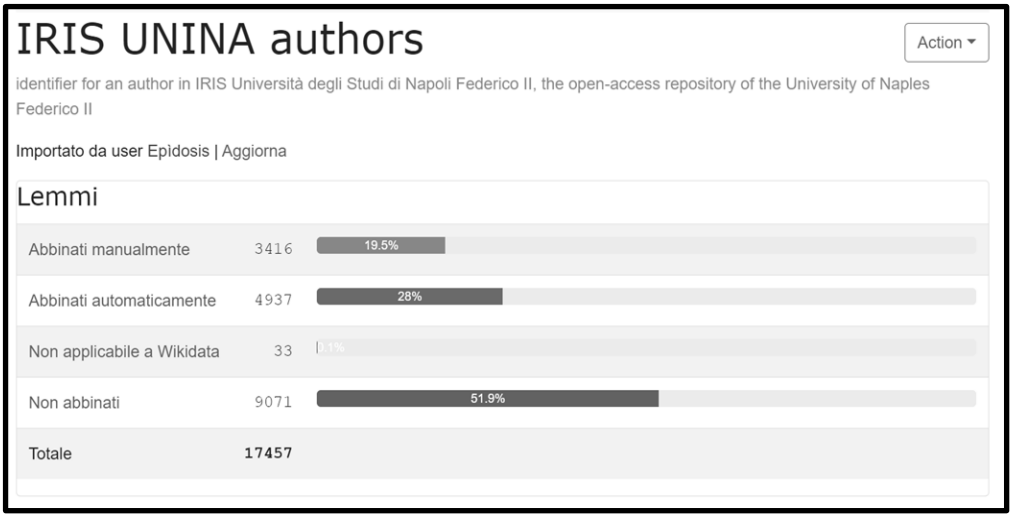

Figura 3 – Immagine da <https://mix-n-match.toolforge.org/#/catalog/4670> (10/07/2023)

Il gruppo 2) può essere riconciliato manualmente, anche avvalendosi di IOCT[29], uno strumento parte di questo progetto e che consiste in un'estensione del proprio browser[30]. IOCT permette di confrontare i titoli presenti in ORCID e nel repository IRIS in modo molto simile al confronto usato al § 2.3.2, con la differenza sostanziale che il confronto, con il passare del tempo, può risultare migliore, specialmente se l'autore carica propri titoli in ORCID. Anche il gruppo 4) può essere riconciliato manualmente, soprattutto creando gli elementi mancanti.

Per quanto riguarda l'esito della riconciliazione manuale, nel file "progress.ods" su Zenodo sono inclusi tabella e grafici della crescita globale degli abbinamenti in MnM degli 81 archivi, in base a conteggi settimanali prelevati automaticamente dal 1 agosto 2022 al 30 giugno 2023; ci sono due grafici, uno per tutti i singoli IRIS e uno per il totale, riportato in Figura 4.

[29] **I**RIS **O**RCID **C**onfronto **T**itoli. Il codice di questa estensione è disponibile all'indirizzo <https://www.wikidata.org/wiki/User:Bargioni/MnM_gadgets/ioct.js>.
[30] Occorre aver previamente installato l'estensione Code Injector, disponibile per i principali browser. Per Firefox si trova all'indirizzo <https://tinyurl.com/3wwf9d27>.

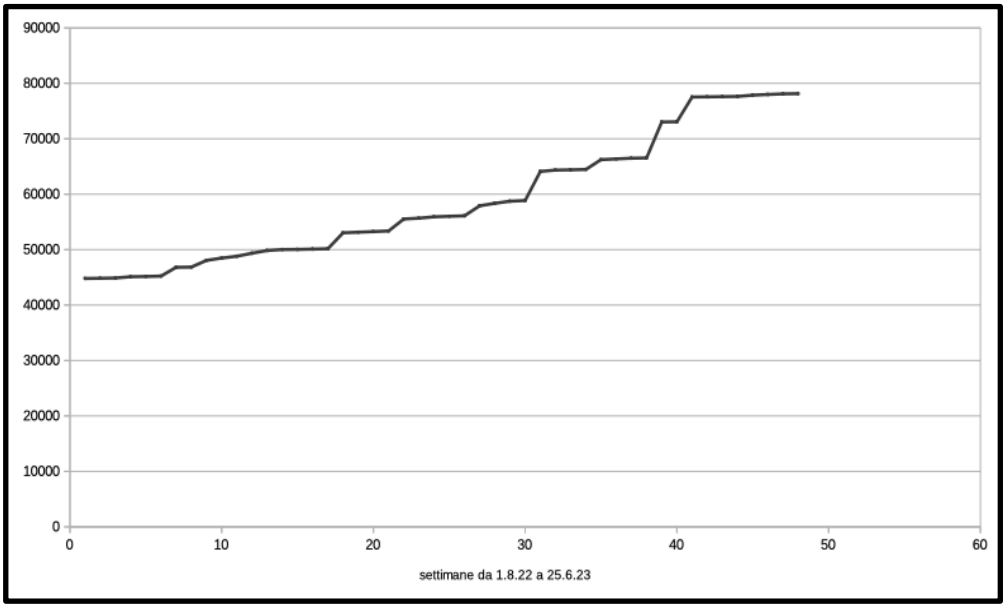

Figura 4 – Andamento globale nel tempo delle riconciliazioni degli IRIS in Wikidata

3. Analisi dei dati

3.1 Una panoramica della ricerca italiana attraverso gli IRIS

Da un'analisi complessiva (al 15 luglio 2023), gli IRIS studiati censiscono complessivamente 489.268 ricercatori e 6.475.415 'prodotti' (di cui 3.537.068 articoli di riviste e 212.193 monografie – o *libri* –[31], mentre per le altre tipologie di 'prodotto' non è possibile fornire dati attendibili in quanto non sono pienamente omogenee tra i diversi IRIS). I dati forniti riguardano il totale degli oggetti censiti in tutti gli IRIS, ma gli oggetti (ricercatori, articoli, monografie ecc.) possono essere censiti separatamente in più di un IRIS[32].

Al fine dell'identificazione degli autori, è sicuramente utile la presenza del nome separato dal cognome. Tra i 'prodotti' della ricerca censiti in IRIS sono presenti tipologie di pubblicazioni censite anche altrove (per esempio, le mo-

[31] Gli script per i conteggi e i risultati sono conservati nel file "conteggi_autori_titoli_iris.zip" su Zenodo.

[32] La deduplicazione dei 'prodotti' può essere controllata al momento della *submission* di ogni 'prodotto', <https://tinyurl.com/ycyzt6sj>, oppure da un utente amministratore dell'Archivio <https://tinyurl.com/yjarapcp>.

nografie sono descritte anche nell'OPAC SBN e gli articoli scientifici – soprattutto se dotati di DOI – si trovano anche in banche dati quali Scopus o Google Scholar); tuttavia negli IRIS sono presenti tipologie di prodotti che sarebbero pressoché impossibili da reperire altrimenti (es. recensioni, capitoli di miscellanee, voci di enciclopedie ecc.), ma che sono fondamentali per dare un'immagine completa della produzione scientifica di un ricercatore. Infine talvolta negli IRIS, ove la normativa sul diritto d'autore lo consenta (in preprint, in postprint o in versione editoriale) e l'autore abbia caricato un pdf, è anche disponibile il full-text[33].

3.2 I dati di IRIS a confronto con Wikidata

Il file "iris_non_applicabili.ods" su Zenodo, riprodotto in parte in Figura 5, riporta i conteggi di autori che per qualche motivo non possono essere considerati validi (nomi che raccolgono omonimi, autori senza 'prodotti', ecc.). Gli autori 'non applicabili', in totale 479 al 25 giugno 2023, sono stati individuati con il lavoro effettuato finora sui MnM (cfr. § 2.3.4, punto 3). Si tratta pertanto di un dato incompleto, ma già utile per gli amministratori degli IRIS al fine di individuare errori catalografici da correggere.

Le due università per le quali il numero di 'non applicabili' è attualmente più alto, come si evince dalla figura seguente, sono l'Università di Salerno e la Scuola Normale Superiore, poiché per entrambe sono stati attuati dei progetti sistematici di abbinamento con Wikidata, che hanno tra l'altro permesso di individuare questi errori presenti nei rispettivi IRIS[34].

Invece gli abbinamenti in Wikidata riportano 335 casi di due ID per una stessa persona e 12 di tre ID per una stessa persona nel medesimo IRIS[35].

[33] Il full text di un prodotto scientifico potrebbe essere stato pubblicato da un ricercatore anche senza rispettare pienamente il diritto d'autore, su siti come ResearchGate e Academia.edu. Si veda in merito Rudj Gorian, *Autori, bibliotecari, open access: osservazioni empiriche e riflessioni su pratiche, comportamenti e ruoli nella piattaforma IRIS dell'Università di Trento*. Trento: Università degli studi di Trento, Dipartimento di lettere e filosofia, 2021, <https://tinyurl.com/yw5tuzvz>.

[34] Si vedano <https://tinyurl.com/3xps8zym> e <https://tinyurl.com/drkb6haj>.

[35] La query SPARQL è stata effettuata il 15 luglio 2023 e si trova a <https://w.wiki/73ny>.

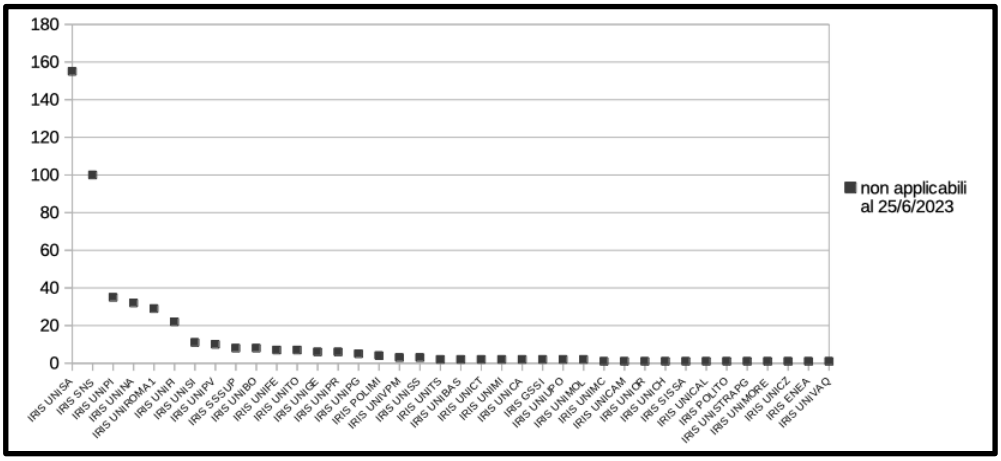

Figura 5 – Numero di autori non validi (non include gli IRIS a zero casi)

3.3 Difficoltà nell'identificazione e nell'abbinamento

Considerando le caratteristiche tipiche di una voce di autorità, le schede IRIS dei ricercatori sono carenti ai fini dell'identificazione e per l'abbinamento con Wikidata sotto diversi punti di vista, qui di seguito elencati:

- forma del nome: la forma del nome del ricercatore presentata da IRIS è tipicamente il nome anagrafico, probabilmente tratto da un'anagrafica dell'ateneo; tuttavia, questo nome anagrafico è registrato con difformità nell'uso delle maiuscole[36] e generalmente in codifica ASCII, quindi senza l'uso di diacritici[37], talvolta suppliti in modi non standard[38]; inoltre, anche quando il nome anagrafico non è quello più usato dall'autore nelle sue pubblicazioni[39], IRIS non presenta alcuna forma alternativa per i nomi dei ricercatori;

[36] In alcuni casi sono interamente in maiuscolo sia il cognome sia il nome, in altri casi solo il cognome, in altri casi nessuno dei due (cfr. ad esempio <https://tinyurl.com/jbwf6jpm> e <https://tinyurl.com/ycy4tr8h>).

[37] Esempio: Ana Acebrón Muñoz in <https://tinyurl.com/4pdmeksv>.

[38] In particolare è frequente la resa della vocale finale accentata con un apostrofo (ad esempio Niccolò Abriani in <https://tinyurl.com/mv74ss9m>).

[39] Accade tipicamente per primi o secondi nomi correntemente omessi, o presentati sotto forma di iniziale, e per secondi cognomi omessi. Es. Maria Cecilia Ceccarelli <https://tinyurl.com/vbat82mu>, che compare come Cecilia Ceccarelli in ORCID <https://orcid.org/0000-0001-9664-6292> e nelle sue pubblicazioni.

- descrizione: le informazioni fornite sul ricercatore generalmente sono o del tutto assenti o limitate al dipartimento di afferenza[40], talvolta accompagnato dagli estremi cronologici di attività[41], solo molto di rado da ulteriori informazioni quali settore scientifico-disciplinare (SSD) e ruolo[42]; le date di nascita, principale strumento di disambiguazione tra omonimi negli archivi di autorità[43], non vengono mai mostrate, nonostante l'ampio numero di omonimie (si vedano i primi 500 nomi in comune tra gli 81 IRIS[44] che vanno da 22 a 6 cataloghi distinti, con una media di 8,45 potenziali omonimi per catalogo);
- connessione con altri database: in generale le schede IRIS non mostrano alcun identificatore esterno, salvo rari casi in cui (tuttavia, come detto in 2.3.1, gli identificatori ORCID sono talvolta ricavabili dai metadati delle schede dei 'prodotti') mostrano l'ORCID del ricercatore, talvolta corredato da ulteriori identificatori[45] (ciò, comunque, dimostra che è già tecnicamente possibile mostrare pubblicamente un ampio novero di identificatori); tra gli ORCID, da un'analisi a campione si nota che una parte di essi non contiene alcuna informazione (né cenni biografici, né altri identificatori, né pubblicazioni, ma solo nome e cognome)[46] e risulta quindi inutile ai fini dell'identificazione.

Sulla base di questi punti, nella gran parte dei casi l'unica informazione utile per identificare un ricercatore è costituita dai 'prodotti' a lui attribuiti; tuttavia, un'analisi a campione dà come risultato alcune centinaia di casi (cfr. § 3.2) in cui si reperiscono 'prodotti' con più autori in cui uno o più autori sono stati abbinati erroneamente (tipicamente alcuni dei coautori non affiliati a una data università sono stati confusi con persone omonime o parzialmente omonime – stesso cognome e stessa iniziale del nome – affiliate a quell'università)[47].

Considerando non i singoli IRIS, ma l'insieme degli IRIS, si può aggiungere che la presenza dei medesimi 'prodotti' e dei medesimi ricercatori in un ampio numero di IRIS (senza alcuna forma di link reciproco), costituisce duplicazioni

[40] Esempio: <https://tinyurl.com/24eubfwp>.
[41] Esempio: <https://tinyurl.com/yckpmds5>.
[42] Esempio: <https://tinyurl.com/2jad9f4r>.
[43] Cfr. per esempio le *Norme per il trattamento di informazioni e dati comuni a tutte le tipologie di materiale* in SBN riguardo alle qualificazioni, <https://tinyurl.com/245dxduv>.
[44] I dati si possono vedere a: <https://tinyurl.com/4hhh9e43>.
[45] Esempio: <https://tinyurl.com/2uw93kkd>.
[46] Esempio: <https://tinyurl.com/4dffu9b4> ricavato da <https://tinyurl.com/epvnr4st>.
[47] Esempio: <https://tinyurl.com/4eavvyvd> ha vari autori, tra cui Giuliana Fiorentino, ma in realtà si tratta di Giuseppe Fiorentino; <https://tinyurl.com/29vs9vnp> ha vari autori, tra cui Vincenzo Bianco, ma in realtà si tratta di Vittorio Bianco.

di autori o di prodotti che non facilitano l'identificazione univoca di queste entità.

Il lavoro fin qui descritto potrà essere esteso in futuro anche agli autori presenti nell'archivio delle pubblicazioni[48] del Consiglio Nazionale delle Ricerche (CNR) tramite la proprietà Wikidata P11886[49] (creata il 20 luglio 2023) e il relativo catalogo MnM[50].

4. Discussione

Dall'analisi dei dati disponibili e visibili emerge che l'identificazione dei singoli ricercatori e dei loro prodotti è problematica, soprattutto per una impostazione di base del sistema degli IRIS a livello nazionale, ovvero la mancanza di un archivio di autorità nazionale che consenta l'identificazione univoca di ogni ricercatore, la registrazione delle relative forme varianti del nome, di dati descrittivi (per esempio, il passaggio da un'università all'altra) e di identificatori nazionali e internazionali. Il problema è aggravato dal fatto che un medesimo ricercatore riceve due identificativi distinti nell'IRIS e nell'UNIFIND[51] della stessa università, pur essendo i due prodotti realizzati entrambi dal Cineca[52].

Oltre a questo vizio di impostazione, anche la creazione dei dati presenta punti deboli: i dati dei 'prodotti' vengono caricati dagli autori e solo successivamente, con modalità che variano sensibilmente da IRIS a IRIS, vengono potenzialmente rivisti da persone terze[53], non necessariamente provviste di adeguata formazione in ambito catalografico; questa è una delle motivazioni che hanno consentito la proliferazione di errori, tra cui il succitato problema dei coautori identificati erroneamente. Inoltre, a fronte di tale mole di errori, manca un sistema omogeneo per la segnalazione dei medesimi: non esiste un indirizzo standard per ciascun IRIS che possa raccogliere queste segnalazioni[54].

[48] <https://publications.cnr.it/authors>.

[49] <https://www.wikidata.org/wiki/Property:P11886>.

[50] <https://mix-n-match.toolforge.org/#/catalog/6002>.

[51] UNIFIND è una piattaforma, sviluppata dal CINECA, che consente al personale accademico di un'università di raccogliere e presentare i propri dati, comprese le pubblicazioni caricate in IRIS (cfr. <https://wiki.u-gov.it/confluence/pages/releaseview.action?pageId=327778509>).

[52] Un esempio: si confronti <https://tinyurl.com/2p9dmh49> con <https://iris.unife.it/cris/rp/rp80967> per lo stesso docente Enrico Albertini dell'Università di Ferrara.

[53] Ci sono due livelli sopra il livello personale di *submission*: dipartimentale <https://tinyurl.com/2w7cb2v6> e di ateneo, o di utente amministratore <https://tinyurl.com/yck2t2fp>.

[54] Tramite le pagine OAI di ogni IRIS /oai/request?verb=Identify si nota che (al 15 luglio 2023) sono impostati 22 indirizzi noreply@cineca.it evidentemente inutilizzabili, mentre altri 59 non seguono alcun accordo tra loro, e alcuni sono indirizzi personali invece che di ruolo. Sarebbe

La mancanza di un sistema unico nazionale di identificazione di autori e 'prodotti' è un impedimento fondamentale per produrre un quadro d'insieme dei prodotti della ricerca in Italia.

5. Conclusioni

L'analisi fin qui svolta ha mostrato sia i diversi punti di forza del sistema IRIS, tra i quali va evidenziato il coinvolgimento diretto degli autori nel caricamento dei propri 'prodotti', sia alcuni aspetti non presi sufficientemente in considerazione, ma che – se introdotti – potrebbero conferire maggiore unitarietà agli archivi e renderli maggiormente interoperabili con il web semantico, garantendo tutti i relativi vantaggi.

In IRIS manca un archivio di autorità unico nazionale e quindi, a livello nazionale, i singoli ricercatori non vengono gestiti come previsto per le entità 'persona' del modello IFLA LRM. Inoltre, i dati analizzati mostrano che l'insieme degli IRIS universitari non è in grado di fornire un quadro unitario e coerente della ricerca italiana, secondo quanto previsto a termini di legge. Il contenuto di IRIS è potenzialmente interoperabile con il web semantico, e in particolare con alcuni sistemi rilevanti come Wikidata, ma, dato che al momento i dati più significativi a tale scopo sono presenti ma non pubblici, l'interoperabilità non è possibile.

La riconciliazione tramite Wikidata avviata con questo progetto e senza il coinvolgimento delle istituzioni responsabili dei singoli IRIS è una dimostrazione della potenzialità dei linked open data in termini di diffusione e visualizzazione dei prodotti e dei ricercatori e di controllo della coerenza e della qualità dei dati.

Diversi accorgimenti potrebbero aiutare a migliorare l'efficacia del sistema nazionale degli IRIS per fornire un quadro unitario e coerente della ricerca italiana, secondo quanto richiesto dalla legge:

- una migliore identificazione degli autori, oggi carente a livello locale e, di conseguenza, a livello globale, potrebbe essere ottenuta con la pubblicazione aperta di almeno un identificatore internazionale quale l'ORCID e di alcuni dati anagrafici quale data e luogo di nascita. Si potrebbero formare connessioni e percorsi di navigazione tra cataloghi diversi, specialmente con Wikidata, l'OPAC SBN e le rispettive biblioteche degli atenei;

per esempio auspicabile uniformare tutti all'indirizzo segnalazioni_iris@dominio.it e pubblicarlo in homepage. Va in ogni caso notato che alcuni IRIS offrono in effetti in homepage un indirizzo per contatti, ma diverso da quello della pagina *identify* del server OAI.

- un contributo significativo potrebbe venire da parte dei ricercatori mediante la cura del proprio account ORCID, con l'aggiunta di dati personali e di pubblicazioni;
- sarebbe auspicabile che venisse adottato un sistema unico, a livello di software IRIS o degli IRIS nazionali, per la segnalazione degli errori catalografici e in particolare dei casi di erronea identificazione di alcuni coautori presenti nelle pubblicazioni;
- per una migliore visione di insieme degli archivi IRIS, sarebbe auspicabile uniformare i nomi dei server (per esempio alla forma iris.dominio.it), eventualmente come sinonimi dei nomi attuali;
- nello sviluppo e nell'estensione dei servizi offerti dal Cineca agli atenei, vanno anche tenute presenti le necessità della corretta identificazione delle entità che è alla base dei linked open data e vanno evitate duplicazioni di ID per il medesimo ricercatore come avviene oggi per IRIS e UNIFIND prodotti dal Cineca. Sarebbe auspicabile invece che ogni UNIFIND usasse i medesimi ID del relativo IRIS o che almeno l'uno rimandi all'altro tramite un link (cosa che attualmente non avviene).

Sono infine ipotizzabili sviluppi futuri ampiamente automatizzabili, come per esempio usare il servizio OAI-PMH per seguire le creazioni negli archivi di nuovi nomi e di nuovi 'prodotti'.

L'obiettivo di una più precisa identificazione dei ricercatori e dei prodotti della ricerca a loro associati è non solo auspicabile ma necessaria per la correttezza della VQR.