Parsifal: armonizzare la tradizione con la modernità. L'Authority file condiviso di URBE come nuovo terreno di collaborazione

Tiziana Possemato, Annalisa Di Sabato, Alessandra Moi

Abstract: L'Unione Romana Biblioteche Ecclesiastiche (URBE) è stata fondata nel 1991 ed è oggi composta da venti istituzioni accademiche romane, con lo scopo principale di creare una rete per la gestione informatica del patrimonio librario. Il progetto Parsifal nasce con l'idea di realizzare un catalogo unico, condiviso tra i partecipanti, in un portale che offra un unico punto di accesso al patrimonio bibliografico delle istituzioni, secondo i nuovi linguaggi del web per favorire la massima apertura e condivisione delle informazioni. Realizzata secondo il modello bibliografico BIBFRAME, la piattaforma tecnologica su cui poggia Parsifal è una componente dell'iniziativa Share Family. Il contributo evidenzia i punti di forza del progetto Parsifal, così come le criticità ancora da risolvere e chiude con un accenno agli scenari evolutivi dell'iniziativa, con particolare riferimento alla collaborazione con il mondo Wiki.

Parole chiave: Share Family; linked data; Parsifal; LOD Platform; riuso dati; Wikidata.

***Abstract**: The Unione Romana Biblioteche Ecclesiastiche (URBE) [Roman Union of Ecclesiastical Libraries] was founded in 1991, and is today composed of twenty academic institutions in Rome, with the primary purpose of creating a network for the computerised management of the library heritage. The Parsifal project was born of the idea of developing a single catalogue shared between the participants, in a portal which offers a single point of access to the bibliographic heritage of the institutions, using the new languages of the web to encourage maximum openness and sharing of information. Built according to the BIBFRAME bibliographical model, the technological platform on which the Parsifal project is based is a component of the Share Family initiative. The contribution emphasises the strengths of the Parsifal project, as well as the critical issues that still need to be resolved, and closes with a reference to the initiative's development prospects, with particular reference to collaboration with the Wiki world.*

***Keywords**: Share Family; linked data; Parsifal; LOD Platform; data re-use; Wikidata.*

Tiziana Possemato, @Cult, Italy, tiziana.possemato@atcult.it, 0000-0002-7184-4070
Annalisa Di Sabato, @Cult, Italy, annalisa.disabato@atcult.it, 0009-0003-8020-6059
Alessandra Moi, University of Milano-Bicocca, Italy, alessandra.moi@atcult.it, 0000-0003-0104-4999

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

1. Il progetto Parsifal dell'Unione Romana Biblioteche Ecclesiastiche (URBE)

L'Unione Romana Biblioteche Ecclesiastiche (URBE) è stata fondata nel 1991 ed è oggi composta da 20 istituzioni accademiche romane[1]. Scopo dell'Associazione URBE è stato quello di creare una rete per la gestione informatica del patrimonio librario delle biblioteche aderenti. Le biblioteche di URBE hanno due elementi unificanti:

- la natura delle istituzioni (istituzioni accademiche ecclesiastiche), con un patrimonio bibliotecario piuttosto unico per le tematiche trattate;
- la tradizione comune sulle norme di catalogazione (Anglo-American Cataloguing Rules vers. 2[2] prima e Resource Description and Access[3] poi), con relativi corsi di formazione destinati ai bibliotecari catalogatori, con l'obiettivo di standardizzare quanto più possibile le pratiche catalografiche in uso presso ciascuna biblioteca. Denominatore comune di questo progressivo piano di standardizzazione delle pratiche catalografiche è stata l'adozione nell'intera Rete URBE, dal 2017, della linea guida RDA (Original Toolkit) con organizzazione di un importante piano di formazione congiunto.

Ogni biblioteca produce e mantiene il proprio catalogo bibliografico e solo poche biblioteche gestiscono, ove possibile, un *authority file* locale. Il sogno di avere un catalogo unico su cui costruire una serie di servizi per un'utenza condivisa è rimasto irrealizzato per più di trent'anni. Il progetto *Parsifal*[4] nasce con l'idea di realizzare, con tecnologie e linguaggi nuovi, il disegno di un catalogo unico della rete, condiviso tra i partecipanti e pubblicato attraverso un portale fondato sulle tecnologie e i linguaggi del web semantico. Parsifal risponde all'esigenza individuata dai Rettori delle Pontificie Università Romane di dotare le biblioteche delle proprie istituzioni di un motore di ricerca altamente innovativo, che offra un unico punto di accesso al patrimonio bibliografico delle biblioteche. Il catalogo unico integrato conta a oggi circa 3 milioni di risorse bibliografiche e il patrimonio offerto (negli ambiti della teologia, filosofia, studi biblici, patristica, mariologia, diritto canonico, sociologia e altro) è spesso unico e introvabili altrove. Realizzata secondo il modello bibliografico BIBFRAME[5], con estensioni per garantire la compatibilità con l'IFLA LRM (Library Reference Model) (IFLA LRM 2020), la piattaforma tecnologica (LOD Platform)[6]

[1] <https://www.urbe.it>.

[2] Una panoramica sulle Anglo-American Cataloguing Rules è possibile averla visitando il sito *Librarianship Studies and Information Technology.* <https://www.librarianshipstudies.com/2018/12/anglo-american-cataloguing-rules-aacr.html>.

[3] <https://www.rdatoolkit.org>.

[4] <https://parsifal.urbe.it>.

[5] <https://www.loc.gov/bibframe>.

[6] La LOD Platform è la piattaforma tecnologica alla base di tutti i progetti della Share Family. Lo stack tecnologico della LOD Platform è composto da un insieme di strumenti e applicazioni:
- Authify, un modulo RESTFul che fornisce servizi di ricerca di record bibliografici e di autorità contenuti in dataset esterni, principalmente legati a fonti autorevoli pubblicate nel web

su cui poggia il progetto Parsifal è una componente dell'iniziativa *Share Family*[7] e nella medesima iniziativa internazionale si pone, dunque, il progetto Parsifal.

Il nome del portale, Parsifal, è un riferimento simbolico alla figura centrale del mito del Graal, narrato per la prima volta nel *Perceval* (poema cavalleresco di Chrétien de Troyes) e successivamente ampliato dal tedesco Wolfram von Eschenbach, nella sua opera *Parzival*. Questo cenno alla leggenda arturiana e alla ricerca del Graal da parte del giovane Parsifal riflette l'obiettivo primario della piattaforma di aiutare gli utenti a trovare, identificare, selezionare, ottenere e navigare informazioni su autori, opere e le loro relazioni.

Parsifal è pensato, così, come un'occasione per la Rete URBE per costruire un luogo comune e condiviso che esponga al mondo la ricchezza informativa del patrimonio posseduto e gestito dalle diverse istituzioni, che faciliti la reperibilità del patrimonio e che costituisca anche un'occasione di confronto e di scambio tra i catalogatori della rete. Tutto, senza dimenticare le radici, la tradizione e la specificità di ciascuna biblioteca partecipante.

(VIAF, Library of Congress Name Authority File, …) ma estensibili anche ad altre tipologie di dataset;
- Cluster Knowledge Base (CKB): distribuita su tre database diversi (PostgreSQL, Solr e un triple store per i dati in RDF) è il risultato dei processi di identificazione, arricchimento e clusterizzazione dei dati;
- RDFizer, il modulo RESTFul che automatizza l'intero processo di conversione dei dati in formato RDF;
- un layer di API per il consumo e l'aggiornamento/cura dei dati;
- il portale (definito anche "entity discovery portal") per la pubblicazione e la fruizione dei dati ottenuti dai diversi processi di identificazione, arricchimento, riconciliazione e conversione dei dati;
- un entity editor, lo strumento che, nella versione 3.0.0. della piattaforma, mette a disposizione dei catalogatori un potente strumento per la cura dei dati pubblicati sul portale.

A queste componenti principali si aggiungono altre componenti funzionali alla gestione dei dati e dei processi di aggiornamento degli stessi (gestione dei delta), tra cui un database - Cassandra - per la conservazione di tutti i record originali inviati dalle istituzioni partecipanti, e Chronos, una componente per l'armonizzazione e la sincronizzazione dei processi nei diversi database che compongono la piattaforma.

L'obiettivo principale della LOD Platform è:
- creare un ecosistema di dati collegati in cui le entità BIBFRAME beneficiano il più possibile della ricchezza di dati ereditata dai cataloghi originali delle biblioteche e dalle fonti esterne;
- fornire una fonte di dati autorevole e viva attraverso la Cluster Knowledge Base;
- riconciliare i dati di diverse biblioteche in un catalogo unico e integrato e arricchirli con informazioni provenienti da fonti esterne (ad esempio aggiunta di URI a entità da VIAF, ISNI, Wikidata ecc.);
- consentire la pubblicazione dei dati in una modalità che sia comprensibile e usabile agli utenti finali e ricercatori con esperienze e competenze diverse;
- esporre i dati in diversi formati, affinché possano essere consumati da umani e da macchine attraverso approcci e modalità diverse.

Per maggiori informazioni sulla LOD Platform: <tinyurl.com/2a5bs4ca>.

[7] <https://wiki.share-vde.org/wiki/ShareFamily:Main_Page>.

2. La felice coesistenza della tradizione e delle nuove tendenze

La tradizione catalografica e informatica della Rete URBE è piuttosto lunga, anche per via di uno degli scopi fondativi della rete stessa (la gestione informatica del patrimonio librario delle biblioteche aderenti). Il progetto di informatizzazione della rete, partito ai primi degli anni '90 con la scelta di un unico applicativo catalografico per tutte le biblioteche partecipanti, si è poi nel tempo trasformato lasciando a ciascuna biblioteca la libertà di scegliere il proprio applicativo, mantenendo saldo il principio di un'adesione omogenea alle regole di catalogazione e al MARC 21 come formato di strutturazione dei dati. Il punto di partenza per la creazione di un ecosistema comune nel progetto Parsifal è quanto consolidato in più di trent'anni di scelte teoriche condivise ma di forte operatività locale e autonoma: ciascuna biblioteca ha prodotto il proprio catalogo bibliografico disponibile in formato MARC 21, applicando, ove più rigidamente ove meno, le regole catalografiche RDA; alcune biblioteche hanno anche prodotto authority file locali, disponibili anche questi in formato MARC 21. Questo il punto di partenza di Parsifal, radicato nella tradizione catalografica della Rete URBE.

A partire dagli export dei singoli cataloghi, bibliografici e di authority, i dati entrano nel flusso di elaborazione della piattaforma tecnologica generando nuove descrizioni, non prima esistenti, raccolte in una knowledge base comune chiamata Cluster (o Entity) Knowledge Base (CKB): ciascun nuovo "record" qui conservato si riferisce a un'entità di tipo persona, famiglia o ente, oppure a un'opera, ed è il risultato di un processo di analisi e identificazione delle entità che usa algoritmi di comparazione degli attributi descrittivi presenti nei record di origine e presenti in fonti autorevoli esterne. L'utilizzo di dati provenienti da altri sistemi nazionali e internazionali è funzionale a disambiguare, ove possibile, l'entità analizzata e arricchire la descrizione di informazioni aggiuntive, non originariamente presenti nei cataloghi locali. Il processo di clusterizzazione utilizza passaggi diversi ma complementari per realizzare l'obiettivo principale, che è l'identificazione corretta dell'entità a partire da descrizioni diverse (per fattori culturali, linguistici, di tradizione catalografica):

- l'analisi e la gestione dei metadati di authority relativi agli agenti (persone, famiglie, enti) per un primo processo di pre-clusterizzazione; in questa fase, una prima forma di cluster agente è creata, in attesa poi di essere alimentata da altri attributi ereditati da altre fonti;
- il trattamento dei record bibliografici provenienti dalle diverse biblioteche, con la creazione o l'alimentazione dei cluster agente già prima abbozzati;
- la creazione/alimentazione di un cluster di tipo opera a partire da un titolo preferito (quello che un tempo era identificato come titolo uniforme) oppure da un punto di accesso autorizzato;
- in assenza di titoli preferiti e punti di accesso autorizzati, la creazione/alimentazione di un cluster di tipo opera a partire da un titolo proprio bibliografico (tag 245 per i record in MARC 21);
- l'arricchimento dei dati in MARC 21 (con URI assegnati nei processi precedenti);

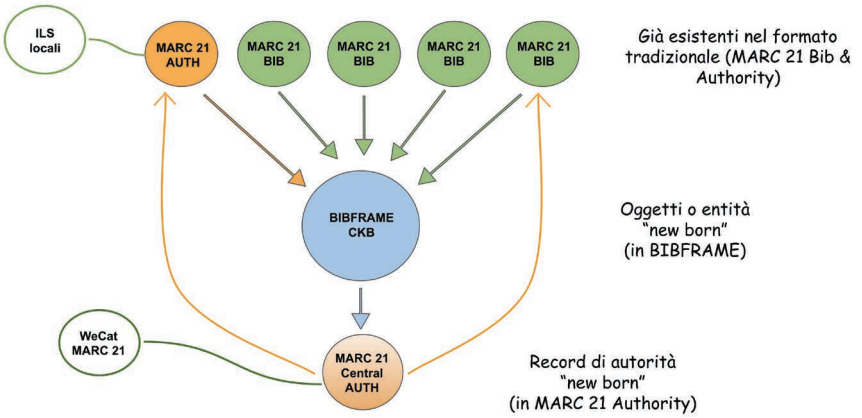

Figura 1. La felice coesistenza dei dati tradizionali in MARC 21 con i "nuovi nati" in BIBFRAME.

- una particolare procedura di raffronto tra i cluster di tipo agente simili e le opere a essi associate rafforza gli elementi di identificazione e riduce il numero di agenti duplicati.

Di seguito si riporta, in forma semplificata, un esempio di processo reale di clusterizzazione, estratto dalla CKB del progetto Parsifal, attraverso la comparazione di stringhe provenienti da tre record di Authority prodotti da istituzioni diverse e relativi allo stesso agente (Ibn Batuta, 1304-1377), più un record di authority proveniente da una fonte esterna (VIAF). Di ciascun record di authority si estrapolano solo alcuni tag, quelli utili a illustrare un processo di clusterizzazione semplificato.

Record di authority #1

100 1 $aIbn Batuta,$d1304-1377
400 1 $wnax@$a*Ibn Baṭūṭaẗ, Muḥammad ibn ʿAbd Allâh*
400 1 $wnax@$aابن بطوطة، محمد بن عبد الله
400 1 $wnax@$aIbn Battutah, Muhammad ibn Ábd Allah
400 1 $wnax@$a*Ibn Baṭṭūṭah, Muḥammad ibn ʿAbd Allāh*
400 1 $wnax@$aIbn-Baṭṭūṭa, Muḥammad Ibn-ʿAbdallāh
400 1 $wnax@$aBattuta, Ibn
400 1 $wnax@$aBattutah, Ibn
400 1 $wnax@$aBatuta, Ibn
400 1 $wnax@$aEbn-Baṭuṭa, Muḥammad Ibn-ʿAbdallāh Ibn-Muḥammad
400 1 $wnax@$aIbn Batouta, …

Record di authority #2

100 1 $aIbn Battuta, Abu Abdallah,$d1304-1377
400 1 $wnax@$a*Ibn Batuta,1304-1377*

400 1 $wnax@$aIbn Battuta, Abu-Abdallah Muhammad Ibn-Abdallah
400 1 $wnax@$aIbn Battuta, Sams al-Din Abu Abd Allah Muhammad b. Abd. Allah b.
400 1 $wnax@$a*Ibn Baṭṭūṭaʾ, Muḥammad ibn ʿAbd Allâh*
400 1 $wnax@$aIbn Battuta
400 0 $wnax@$aIbn Battutah
400 0 $wnax@$aIbn Batuta, Mohammed
400 0 $wnax@$aIbn Batutah
400 1 $wnax@$aIbn Batūtah, Muhammad ibn ʿAbd Allāh
400 1 $wnax@$aIbn Baṭouṭah al- Maghrabī$bʿAbd Allāh ibn Moḥammad ibn Ibrāhīm al-Lawātī

Record di authority #3

100 1 $a*Ibn Baṭṭūṭah, Muḥammad ibn ʿAbd Allāh, $d1304-1369?*
400 1 $wnax@$aIbn Baṭṭūṭa, Abū ʿAbdallāh
400 1 $wnax@$aIbn Baṭṭūṭa, Šamš al-Dīn Abū ʿAbd Allāh Muḥammad b. ʿAbd Allāh b. Muḥammad b. Ibrāhīm b. Muḥammad b. Ibrahīm b. Yūsuf al-Lawatī al-Tanǧī
400 1 $wnax@$aIbn Baṭṭūṭa$bMuhammad ibn Abddallah
400 1 $wnax@$aIbn Baṭṭūṭa$bMuḥammad ibn ʿAbd Allāh
400 1 $wnax@$aIbn-Battuta, Abu-Abdallah Muhammad
400 1 $wnax@$aIbn-Baṭṭūṭa, Abū-ʿAbdallāh Muḥammad Ibn-Abdallāh
400 1 $wnax@$aIbn-Baṭṭūṭa, Abū-ʿAbdallāh Muḥammad
400 1 $wnax@$aIbn-ʿAbdallāh, Muḥammad Ibn-Baṭṭūṭa
400 1 $wnax@$aMuḥammad Ibn ʿAbdallāh aṭ-Ṭanǧī Ibn Baṭṭūṭa, Abū ʿAbdallāh
400 1 $wnax@$aابن بطوطة، محمد بن عبد الله،

Record di authority #4 (preso da fonte esterna)

100 1 $aابن بطوطة، محمد بن عبد الله،
400 0 $wnax@$aMuḥammad Ibn-ʿAbdallāh Ibn-Baṭṭuṭa
400 0 $wnax@$aMuḥammad ibn ʿAbd Allāh ibn Baṭūṭah
400 0 $wnax@$aMuḥammad ibn ʿAbd Allāh ibn Baṭūṭa
400 0 $wnax@$aMuḥammad ibn ʿAbd Allāh
400 0 $wnax@$aal-Tangi
400 0 $wnax@$aابن بتته،
400 0 $wnax@$aابن بطوطة

In corsivo sono riportate alcune delle stringhe *comuni* ai vari record che costituiscono una sorta di catena identificativa tra le diverse fonti. Il sistema applica un insieme di algoritmi logici per cercare, analizzare, identificare e creare un cluster a partire dal record MARC, passando attraverso un processo di analisi delle stringhe e la creazione di forme di stringa normalizzate (sort-form), necessarie per l'applicazione di processi di matching (punteggio di similarità), swapping,

confronto dei dati, filtro e raffinamento dei risultati con esclusione delle forme non riconducibili al cluster, assegnazione di pesi alle stringhe più utilizzate e altre logiche di analisi del testo. Alla fine del processo viene creata una classifica (ranking) assegnando a ciascun cluster un peso, che sarà poi utilizzato per scegliere, a fronte di cluster riconducibili alla medesima entità, quello maggiormente rappresentativo, e facendo confluire in esso gli altri cluster (le forme non già presenti nel cluster scelto).

I dati risultanti da questi processi di analisi e comparazione sono salvati in un database relazionale locale, in cui saranno presenti:

- le stringhe corrispondenti alle forme preferite provenienti dai tag 1xx dei record di authority;
- le forme varianti provenienti dai tag 4xx dei record di authority;
- le forme provenienti dai record bibliografici;
- le forme provenienti dalle fonti esterne.

Alla fine di questo processo sarà creato il cluster per l'entità (nell'esempio, l'entità agente Ibn Batuta, 1304-1377, cluster ID 17581). Il cluster, seppur creato attraverso record di authority, non viene mostrato sul portale agli utenti finali, finché non sia a esso associato un dato bibliografico (che esprime la presenza di una risorsa bibliografica riferibile a quell'agente).

Il sistema dei pesi assegnati anche a ciascuna stringa (una stringa proveniente da un record di authority ha certamente un peso maggiore rispetto a una stringa assegnata a un dato proveniente da un record bibliografico) orienta poi le logiche di presentazione sul portale di consultazione della forma preferita a fronte delle molte forme varianti presenti per quella entità. Nel caso di più forme preferite provenienti da diversi record di authority, altre logiche di selezione, definite dalla Commissione di Catalogazione, sono applicate, per scegliere tra tutte la forma più significativa rispetto all'utenza tipo del catalogo unico integrato.

Il modello utilizzato per costruire i nuovi dati è, appunto, BIBFRAME con le estensioni previste e implementate nell'ambito della Share Family. La knowledge base così prodotta è un punto di arrivo dei processi di trasformazione dei dati e offre uno strumento potente agli utenti della rete. Ma costituisce anche un punto di partenza per un nuovo livello di servizio condiviso tra i catalogatori della rete: a partire dalla knowledge base di URBE viene prodotto in MARC 21 un Authority File Centralizzato (AFC). Un formato tradizionale e ben noto a tutte le biblioteche per uno strumento del tutto nuovo, creato con il contributo di ciascuna biblioteca e a cui ciascun catalogatore della rete è chiamato a operare per migliorare, con la propria competenza, la qualità di quanto prodotto automaticamente dalla macchina. Lo stesso authority file viene esportato quotidianamente per essere utilizzato localmente (nel caso in cui la biblioteca necessiti di un authority file locale) e per migliorare ulteriormente i cataloghi bibliografici di ciascuna biblioteca, i quali poi, esportati nuovamente con procedure notturne, contribuiscono a loro volta ad aumentare la qualità della CKB, in un flusso virtuoso e continuo. Lo strumento messo a disposizione delle biblioteche della rete per operare in modo collaborativo su questo nuovo Authority File Centralizzato

è WeCat, il modulo di catalogazione del sistema OLISuite: ogni biblioteca della Rete URBE che partecipi al progetto Parsifal ha un account per operare sull'AFC e migliorare, a beneficio di tutti, la qualità dei dati catalografici.

Questa architettura ibrida, apparentemente così semplice, offre diverse opportunità alle biblioteche, ma anche alcune sfide da gestire e superare. Nel seguito di questo saggio, focalizzeremo l'attenzione sui punti di forza e di debolezza del progetto Parsifal.

3. Opportunità e punti di forza del progetto Parsifal

Nonostante la lunga tradizione di scelte catalografiche comuni, le differenze in termini di forma delle stringhe descrittive per lo stesso "oggetto", da biblioteca a biblioteca, sono in alcuni casi molto rilevanti. Non dimentichiamo che nella pratica catalografica italiana è molto diffusa la catalogazione originale piuttosto che quella derivata (copy cataloguing), per tutta una serie di ragioni storiche, politiche e anche tecnologiche. Mai in passato, nonostante molti tentativi fatti nell'ambito della Rete URBE, che comunque è espressione della tradizione italiana, è stato possibile definire accordi solidi su come strutturare i metadati, con il risultato che la stessa entità è presente in cataloghi diversi in forme dissimili.

Il progetto Parsifal ha in qualche modo incoraggiato le biblioteche al dialogo, attraverso la creazione di una Commissione di Catalogazione, costituita da membri provenienti da diverse biblioteche della rete, che ha analizzato e orientato i processi di identificazione delle entità e ha guidato le decisioni su come presentare i dati sul portale (lasciando in molti casi invariate le scelte locali e consentendo, così, alle biblioteche di adottare, ove necessario, politiche locali diverse rispetto a quelle centralizzate). Le scelte catalografiche maturate dalla Commissione sono diventate regole logiche, costruite per informare e orientare le scelte delle macchine. Così, per esempio, nella definizione della regola per scegliere la *forma preferita* del nome di un autore, in caso di forme eterogenee provenienti dai diversi cataloghi, è stato assegnato un "peso" a ciascun *tipo* di campo nome presente sui record di authority o bibliografici (regola dell'autorevolezza del tipo di fonte): tra la forma *Francesco d'Assisi* presente in alcuni dei record bibliografici (nel tag che il MARC 21 Bibliographic riserva per la registrazione della stringa del nome di un autore) e la forma *Franciscus Assisiensis, santo, 1182-1226* presente invece nei record di authority (nel tag che il MARC 21 Authority riserva per la registrazione della stringa del nome di un autore), la Commissione ha deciso di assegnare un peso maggiore alla forma proveniente da un record di authority rispetto al peso assegnato alla forma proveniente da un record bibliografico. Questa logica, tradotta in regola comprensibile alla macchina, ha orientato la scelta della forma preferita del nome di un autore (quindi di un'entità di tipo Agente) in tutti i casi in cui applicabile (quindi, ove presenti forme derivanti dai record bibliografici e forme derivanti dai record di authority). Un altro esempio di regola di discrimine sulla forma da scegliere tra le varie esistenti è quella che in ambiente anglosassone viene definita del *most common used*: nel caso in cui la regola precedente (della autorevolezza del tipo di fonte, bibliografica o di autho-

rity) non possa essere applicata, uno specifico *contatore* viene attivato in fase di analisi dei dati (da parte della macchina) per identificare la forma del nome più comunemente utilizzato nei cataloghi (siano essi bibliografici o di authority) per la specifica entità: tra la forma *Benedictus a sancto Philadelpho O.F.M. santo, 1526-1589* e la forma *Benedictus a Sancto Philadelpho, 1526c.-1589*, la prima forma è stata selezionata automaticamente come quella preferita per la visualizzazione, perché più diffusamente utilizzata nei cataloghi. Tra le funzioni della Commissione di Catalogazione anche quella di definire delle norme di standardizzazione nella presentazione di alcuni particolari campi che arricchiscono le descrizioni degli autori. La finalità, ancora una volta, è quella di suggerire pratiche descrittive comuni e uniformi, ove possibile, mettendo nel contempo in atto strategie di trasformazione e presentazione del dato per rendere più chiara e leggibile all'utente finale l'informazione ricercata. Così, sono state create regole per la conversione di numeri romani in numeri arabi, normalizzati con un punto finale (per cui, per esempio, *Gandulphus Bononiensis, sec. XII* diventa *Gandulphus Bononiensis, secolo 12.)* per la conversione delle abbreviazioni registrate nel campo data in forma estesa (per cui, per esempio, *Abbo Floriacensis, s., ca.945-1004* diventa *Abbo Floriacensis, santo, circa 945-1004),* per la conversione in forma estesa delle abbreviazioni di titoli e termini associati al nome. In presenza di più qualificatori nella stessa occorrenza di sottocampo relativa a questi titoli, la selezione di uno solo dei valori previsti è indicata in un'apposita tabella definita dalla Commissione di Catalogazione (in questo modo, per esempio, *Augustinus, s., vesc., 354-430* diventa *Augustinus, santo, 354-430*). Questi sono alcuni esempi di regole finalizzate a orientare poi i processi di standardizzazione e di presentazione dei dati catalografici ai fini di rendere più coerente e comprensibile all'utenza finale le informazioni pubblicate.

Tuttavia, le logiche di correzione e standardizzazione concordate con la Commissione di Catalogazione, così come quelle ereditate dai confronti con la più ampia comunità internazionale che ruota intorno all'iniziativa Share Family, non sono sufficienti a risolvere tutte le casistiche di anomalie e difformità riscontrate mettendo insieme i dati provenienti dai diversi cataloghi. La regola del *most common used*, per esempio, non può funzionare nel caso in cui sia presente lo stesso numero di occorrenze per ciascuna forma del nome, derivanti da record bibliografici, senza alcun record di authority a definire la forma preferita da mostrare. Ecco, questo è il caso in cui i catalogatori sono chiamati a un altro tipo di intervento, attraverso uno strumento collaborativo che è l'Authority File Centralizzato: le descrizioni degli autori presenti nella Cluster Knowledge Base costruita attraverso i processi di identificazione delle entità e di conversione secondo il modello BIBFRAME, costituiscono il punto di partenza per la creazione di un catalogo di record di autorità, nuovo patrimonio comune di tutte le biblioteche della rete. Il set di informazioni disponibili nella Cluster Knowledge Base sono stati esportati e convertiti nel formato MARC 21, molto familiare ai catalogatori della rete, e caricati nel modulo di catalogazione WeCat. Ogni biblioteca accede allo stesso strumento (WeCat) e allo stesso authority file per migliorare, attraverso le proprie competenze, la qualità del catalogo comune.

Figura 2. Un esempio di autore presente sia come entità BIBFRAME che come record di autorità in MARC 21 (visibile in entrambi i formati sul portale Parsifal).

Particolari meccanismi di tracciamento delle relazioni tra la descrizione riferibile a un autore presente nella CKB e il relativo record in MARC 21 confluito nell'Authority File Centralizzato, garantiscono che il flusso di aggiornamento e allineamento tra le diverse basi dati non si interrompa.

Per perseguire l'obiettivo di una più attenta e profonda cura della qualità dei dati e per facilitare e ottimizzare le lavorazioni sono state pensate e sviluppate nel modulo WeCat nuove funzionalità: la funzione di *merge*, per fondere record di autorità riferiti alla stessa entità (e dunque non correttamente identificati e clusterizzati dai processi macchina nella CKB) è un esempio di queste funzionalità avanzate sviluppate *ad hoc* per migliorare l'esperienza del lavoro cooperativo e condiviso. Il lavoro di pulizia, correzione, arricchimento dei record di authority, fatto congiuntamente dai catalogatori, rientra nei flussi di aggiornamento delle piattaforme e confluisce nuovamente sul catalogo unico della rete a beneficio sia degli utenti finali che degli stessi catalogatori, e della più vasta comunità scientifica e del web.

L'incredibile risultato di questo approccio, con flussi di dati che da MARC 21 sono trasformati in linked data secondo il modello BIBFRAME e poi di nuovo in MARC 21, per ricominciare in un flusso ciclico e virtuoso, dà la dimensione di quelli che sono i risultati forse più rilevanti di questo progetto:

- avere costruito un'occasione e gli strumenti necessari per lavorare in una modalità condivisa, su obiettivi comuni all'intera Rete URBE;
- la produzione di dati autorevoli, in alcuni casi introvabili in altre fonti, disponibili per la comunità del web, soprattutto in considerazione della specificità e unicità delle risorse che la rete mette a disposizione. Il set di dati disponibile in BIBFRAME rappresenta, quindi, un valore aggiunto molto importante per la Rete URBE: fornire dati in BIBFRAME è un modo per rendere visibile al

mondo il patrimonio in alcuni casi unico delle biblioteche, uscendo dal contesto strettamente locale.

Ma il risultato migliore è, forse, quello dei catalogatori, che hanno accettato la sfida di entrare nei nuovi meccanismi catalografici e organizzativi, dando prova della volontà, abbastanza condivisa nel gruppo, di adattare la propria mentalità alla nuova era della pratica catalografica, passando da un approccio alla metadatazione centrato sul record a una visione fondata sulla *gestione e modellizzazione delle entità*, secondo il modello che i linguaggi e le tecnologie del web propongono. Questo è un importante risultato che i catalogatori della rete stanno cercando di perseguire in una modalità pratica e concreta, seguendo il processo di conversione dei dati, partecipando all'evoluzione che, passo dopo passo, sta portando a raffinare le logiche di identificazione delle entità, definendo una modalità di intervento sui dati che sposta necessariamente il panorama operativo dal contesto tradizionale della singola biblioteca al nuovo ambiente condiviso.

Ma il cammino per realizzare tutti gli obiettivi è ancora in corso e alcuni ostacoli devono essere ancora superati.

4. Criticità ancora aperte

Nei processi di identificazione di un'entità riveste un peso importante la qualità dell'informazione di origine, che, nel caso degli *agenti* (autori, collaboratori, traduttori ecc.) viene utilizzata per comparare le stringhe (*string matching*) e capire se le entità nascoste dietro un insieme di attributi diversi siano la stessa entità oppure entità diverse. Le descrizioni presenti nei cataloghi sono spesso significative per un agente umano, ma non per una macchina. Casi come questo:

- Abbrescia, Domenico, O.P.
- Abbrescia, Domenico M., O.P., 1922-1996
- Abbrescia, Domenico Maria, O.P.
- Abbrescia, Domenico Maria

non possono essere facilmente registrati e intesi da una macchina come sicuramente riferentesi alla stessa persona. Casi di omonimia o di nomi diversi dietro lettere puntate sono frequentissimi nei nostri cataloghi e nel mondo reale. In molti di questi casi anche il riutilizzo di metadati provenienti da altre fonti autorevoli non aiuta, come vedremo successivamente. Rimanendo in questo esempio, il primo risultato dei processi di identificazione delle entità ha prodotto quattro diversi *cluster*, quindi quattro differenti *entità* riconosciute dietro le quattro differenti stringhe del nome: la macchina non riesce a ragionare sull'ipotesi che *Abbrescia, Domenico Maria* e *Abbrescia, Domenico M., O.P., 1922-1996* possano riferirsi alla stessa persona e, nel dubbio, crea "entità" differenti. Questo meccanismo, replicato *n* volte, può creare moltissimo rumore nel catalogo unico. L'analisi per risolvere questa casistica ha suggerito di appoggiarsi a un criterio logico ampiamente riconosciuto in ambito bibliografico: *ogni opera è identificabile attraverso il suo autore; ogni autore è identificabile attraverso la sua opera*. Applicando questo principio, sono

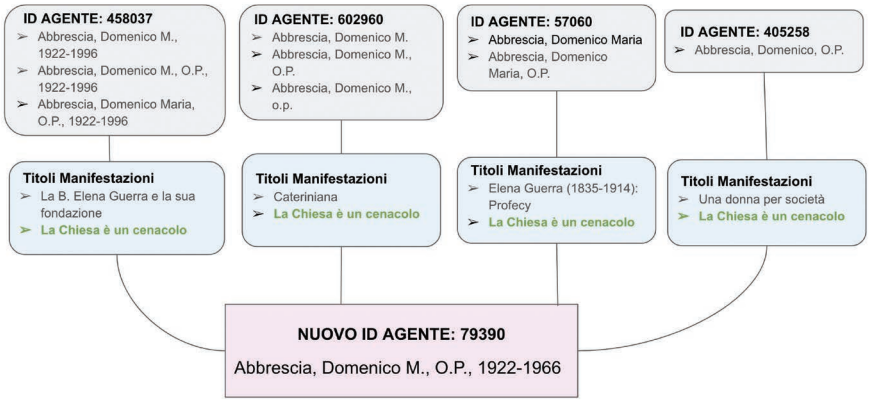

Figura 3. Il passaggio per le opere associate a ciascun "cluster" per creare *catene identificative* e ridurre il rumore delle mancate identificazioni

state scritte regole e procedure di confronto delle opere associate a ognuna entità marcata come "potenzialmente a rischio di duplicazione". Come illustrato in figura 3, le opere associate nel database a ciascuna entità sono state messe automaticamente a confronto per individuare un possibile *anello* comune, che crea quella che possiamo immaginare come una *catena identificativa*, in cui ogni anello raffina il processo di matching, fino alla riconciliazione di tutti i cluster degli agenti in uno.

Questo meccanismo del passaggio attraverso le *opere* per scorgere lo stesso *autore* dietro descrizioni diverse può funzionare solo nel caso in cui in fase di prima elaborazione dei dati siano prodotti cluster diversi marchiati come *simili.* Esiste nella base dati di URBE un'altra casistica di errore, prodotta quasi paradossalmente proprio dal meccanismo che, per meglio identificare un'entità, riutilizza i dati di fonti esterne autorevoli (come VIAF, ISNI, i database delle grandi biblioteche nazionali ecc.). Dico 'paradossalmente' perché il meccanismo di *hyperlinking* e cioè di collegamento di una entità presente nel proprio dataset con entità che consideriamo *medesime* in altre fonti, è ciò che il sistema di valutazione previsto da Tim Berners-Lee per i Linked Open Data premia con la *quinta stella*[8]: riconoscere due o più entità, descritte in fonti differenti, come le medesime aumenta il grado di disambiguazione e identificazione di ciascuna di esse, e consente anche il riuso dei dati, e quindi l'arricchimento della descrizione della entità con informazioni non già presenti nel sistema di origine. Ma tutto questo funziona molto bene quando i dati sono corretti. In caso di errore o di lacuna presente sulla fonte autorevole che interrogo per disambiguare e arricchire il mio set informativo, la probabilità di ereditare anche l'errore è molto alta. E in

[8] «In your RDF, have the identifiers be links (URLs) to useful data sources» il che significa che il proprio dataset pubblicato nel web acquista maggior valore e credibilità quanto più i dati siano collegati a quelli di dataset esterni. <https://dvcs.w3.org/hg/gld/raw-file/default/glossary/index.html#x5-star-linked-open-data>.

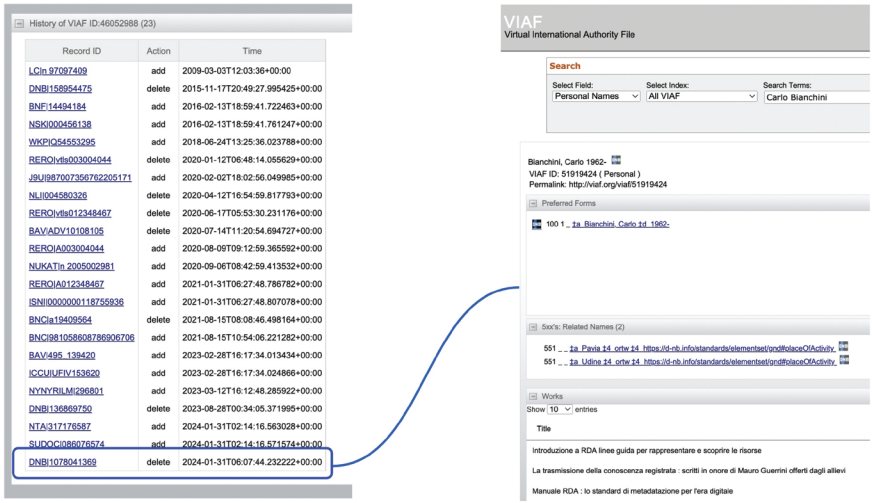

Figura 4. La storia tracciata sulla voce VIAF per *Carlo Bianchini, 1967-* da cui si evince che la forma errata *Carlo Bianchini, 1962-* è stata cancellata il 31-01-2024.

questi casi la definizione di algoritmi automatici per correggere gli errori è poco efficace. Un esempio di questa errata identificazione è quello dell'autore *Carlo Bianchini, 1967-* che in VIAF presentava, al momento dello scarico dei dati ai fini del riutilizzo in Parsifal per gli arricchimenti, un'errata associazione di una stringa relativa all'entità *Carlo Bianchini, 1962-* (più, ancora oggi visibile, un'errata associazione di una stringa relativa all'entità *Carlo Bianchini, 1932-*). Il riutilizzo dei dati del VIAF in fase di disambiguazione delle entità ha prodotto in Parsifal un accorpamento delle prime due entità sopra citate (*Carlo Bianchini, 1967-* e *Carlo Bianchini, 1962-*) che avrebbero dovuto, invece, essere riconosciute come entità differenti. Come risolvere questa casistica di errore, che a partire da una fonte rischia di estendersi a molti altri sistemi, è ancora in fase di indagine. Ma il caso è interessante e apre a una riflessione sul *riuso collaborativo*[9] dei dati e di come non possa definirsi sufficiente il riutilizzo se le fonti non riescono a dialogare: il dato errato su VIAF, relativo a Carlo Bianchini con la doppia data di nascita (1967 e 1962) è stato corretto il giorno 31-01-2024 (come evidente nella figura 4); ma VIAF non mette a disposizione sistemi o procedure che consentano agli utilizzatori dei suoi dati di essere informati automaticamente rispetto ad aggiornamenti, sì da poter riallineare le proprie basi di dati. Per risolvere questo problema di incomunicabilità tra le fonti, l'iniziativa Share Family, di cui Parsifal è parte, ha elaborato lo strumento dell'Entity Registry, un registro che, come il VIAF, tracci

[9] Ho avuto l'occasione di parlare di riuso e delle opportunità e criticità di questa pratica, durante il Convegno di studi *Fare per non sprecare. Nei laboratori del riuso digitale*. Si veda Possemato 2023, 134-46.

tutte le modifiche storiche apportate a un'entità ma che sia, poi, pronto a esporre queste informazioni nelle stesse modalità, tecnologie e linguaggi dei linked data, dunque in modo fruibile da terze parti. Ma su come affrontare il problema di informazioni erroneamente prese da fonti autorevoli esterne, torneremo brevemente al termine di questo lavoro.

Il diagramma rappresentato in figura 1 è una estrema semplificazione del reale flusso dei dati, la cui complessità è invece rappresentata nella figura 5, in un circuito di informazioni che, attraverso processi automatizzati (e in parte manuali) collega i poli, costituiti dalle singole biblioteche, il cuore del sistema costituito dalla Cluster Knowledge Base delle entità e l'Authority File Centralizzato. Ciascuna azione prodotta sui dati, in qualsiasi di questi nodi, produce una modifica allo stato delle informazioni, che può o deve essere comunicata agli altri nodi in relazione a come il flusso di dati cammini. Questa complessità, che è costruita per garantire l'allineamento tra le basi dati e il riverberare degli effetti positivi degli interventi fino ai cataloghi locali, richiede alle biblioteche un nuovo sforzo collaborativo: le biblioteche stanno ora discutendo le modalità operative per migliorare il flusso e sfruttare al massimo l'opportunità che la nuova organizzazione offre loro. Nell'ambito del progetto Parsifal sono state delineate le linee guida per ottimizzare questi flussi, pur facendo salvo il principio di legittimità di ciascuna biblioteca di operare in modo autonomo, in relazione alle proprie procedure interne e alle possibilità di modificare le proprie abitudini. Molti scenari possibili sono ora allo studio, e richiedono un nuovo sforzo di armonizzazione e allineamento tra le parti, per mettere a frutto tutto quanto fin qui realizzato. In questa revisione del flusso di lavoro, ove le operazioni catalografiche superano necessariamente i confini di ciascuna biblioteca, molta parte ce l'hanno gli strumenti di controllo e allineamento dei dati che sono stati realizzati nell'ambito del progetto, come per esempio la ricca reportistica giornaliera prodotta a termine dei processi notturni e inviata alle singole biblioteche per informare dei cambiamenti avvenuti[10]; oppure come l'estensione dei processi di riallineamento tra la Cluster Knowledge Base e l'Authority File Centralizzato e viceversa. Molto potrà essere realizzato in futuro di concerto con i gestori degli ILS locali, per sviluppare protocolli e strumenti di dialogo diretto tra i nodi centrali e le periferie. Ma moltissimo sarà determinato proprio dalle capacità che le biblioteche avranno di guardarsi come *rete*, rinunciando a parte delle tradizioni e delle abitudini locali per raccogliere lo stimolo a operare in modo collaborativo, come poli diversi (seppur sempre autonomi e indipendenti) di un medesimo organismo.

[10] Un esempio di report giornalieri inviati alle biblioteche:
- Nuovi cluster per fonte (nuovi cluster prodotti dai processi delta dei record bibliografici locali e dal delta dell'Authority File Centralizzato).
- Cluster aggiornati per fonte (cluster aggiornati dai processi delta dei record bibliografici locali e dal delta dell'Authority File Centralizzato).
- Cluster inattivi per fonte (biblioteca).
- Record scartati per fonte (biblioteca).
- Record eliminati nell'Authority File Centralizzato

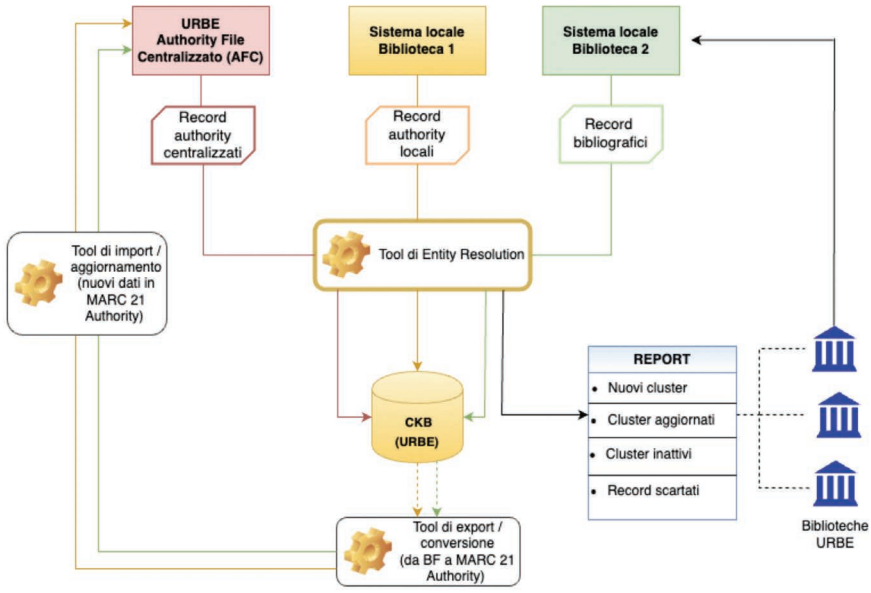

Figura 5. Una schematizzazione del flusso di dati, tra i poli (le biblioteche), la Cluster Knowledge Base delle entità (CKB) e l'Authority File Centralizzato (AFC).

Intanto, un'altra prova importante di collaborazione nell'ambito del progetto c'è stata con la costituzione della *Commissione per l'interfaccia del portale* che, in considerazione della tipologia di utenza cui i cataloghi delle biblioteche della Rete URBE sono principalmente destinati tenendo anche conto del nuovo contesto operativo che l'occasione del progetto sta costruendo, ha definito e condiviso con lo staff di analisti e tecnici che ha seguito lo sviluppo della piattaforma, le linee guida per la costruzione dell'interfaccia di ricerca, ora disponibile all'indirizzo parsifal.urbe.it.

5. Futuri scenari evolutivi di Parsifal

Molti sono gli stimoli che Parsifal sta offrendo ai bibliotecari e a tutti coloro che stanno cooperando per la riuscita del progetto. Certamente l'estensione del controllo di autorità dall'entità *Agente* (con i suoi autori, traduttori, collaboratori ecc.) all'entità *Opera* è uno dei prossimi obiettivi del gruppo di lavoro, per garantire nel tempo a questo tipo di entità lo stesso livello qualitativo oggi in fase di costruzione per gli agenti.

Strettamente legata a questa attività, e considerando la tipologia di risorse oggi presenti nel catalogo unico integrato, con molte opere tradotte o presenti in molteplici edizioni, l'applicazione dell'*Hub* prevista da BIBFRAME potrebbe essere un'opportunità per ottimizzare la presentazione delle informazioni. L'Hub in BIBFRAME è descritto come un'entità astratta che funge da ponte tra due

Figura 6. L'opera *Schloss* di Franz Kafka come presentata oggi sul portale *Parsifal*. Nella sezione a sinistra sono evidenti i raggruppamenti per versioni linguistiche.

opere[11]. Definito con una specifica classe, si configura nell'applicazione pratica più come una sorta di contenitore vuoto (quindi, privo di una sua propria specificità) in cui far confluire, per esempio, più espressioni di una medesima opera, secondo diversi criteri di raggruppamento: tutte le traduzioni francesi dell'opera *Metamorphoseon libri XV* di Ovidio; tutte le traduzioni italiane dell'*Odissea* di Omero, oppure tutti gli spartiti per canto e pianoforte del *Macbeth* di Giuseppe Verdi, e così via[12]. In un catalogo in cui le notizie relative a una medesima opera possono essere molto numerose, rendendo faticosa la navigazione dei risultati da parte di un ricercatore, l'applicazione dell'entità Hub consentirebbe di migliorare la fruizione dei dati. Parzialmente questo risultato è già ottenuto sul portale con un meccanismo che non tocca però i dati (come mostrato in figura 6), ma lavora sul solo livello della presentazione. L'applicazione dell'Hub di BIBFRAME permetterebbe di modellare in modo più profondo i dati, consentendo anche alle

[11] «An abstract resource that functions as a bridge between two Works», <https://id.loc.gov/ontologies/bibframe.html#c_Hub>.

[12] Qui abbiamo fatto riferimento all'*Espressione* per meglio chiarire la finalità dell'*Hub* in BIBFRAME, ricordando che nel modello BIBFRAME il concetto di *Opera* e di *Espressione* coincidono e sono espressi come *Work*. Ricordiamo anche che nel modello ontologico FRBR poi evoluto in IFLA LRM, ciascuna singola traduzione di un'opera (per citare il caso delle traduzioni) rappresenta una *Espressione*. L'*Hub*, dunque, potrebbe essere applicato per creare quel livello di raccordo tra il *Work* inteso come *Opera* e i singoli *Work* intesi come ciascuna traduzione di essa.

macchine di trarre vantaggio da queste forme di raggruppamento di opere con una lunga storia editoriale.

Come anticipato, un altro importante filone evolutivo del progetto sarà quello focalizzato sul miglioramento dei processi di identificazione delle entità, condizionati dalla qualità del dato di origine (mancanza di elementi qualificanti), ma anche dalla qualità delle fonti esterne interrogate per la disambiguazione e l'arricchimento. Abbiamo già citato prima il caso di *Carlo Bianchini* nel database di VIAF. Ma numerose sono le anomalie che si registrano nelle fonti disponibili nel web che, come detto, diventano degli amplificatori del problema laddove manchino meccanismi reciproci di informazione e aggiornamento tra le varie basi di dati. Emblematico in Parsifal è il caso di *Carlo Mazzone* (Parsifal ID: 526759). Il Catalogo unico di URBE contiene diverse opere associate all'autore *Carlo Mazzone, sacerdote*. VIAF pubblica sia l'entità *Carlo Mazzone, sacerdote* (VIAF ID: 121846540) sia l'entità *Carlo Mazzone, allenatore di calcio* (VIAF ID: 564170671811416650006). Gli algoritmi di identificazione di Parsifal, che comparano i dati prodotti dalle elaborazioni interne con quelli delle fonti esterne, hanno intercettato correttamente l'ID VIAF relativo al sacerdote. Ma, tra gli identificatori esterni che lo stesso VIAF utilizza per arricchire la voce relativa al sacerdote, ci sono due ID che "sviano" i processi di identificazione: quello ISNI relativo, però, a *Carlo Mazzone, allenatore di calcio* (ISNI: 0000 0000 8041 3059) (figura 7) e quello di Wikidata, anche questo relativo però a *Carlo Mazzone, allenatore di calcio* (ID Q1042327), poi cancellato dalla voce relativa al sacerdote, il 31-01-2024 (figura 8). Come il caso di Carlo Bianchini, anche questo caso, come tanti altri, mette seriamente in dubbio l'efficacia del riutilizzo delle fonti esterne.

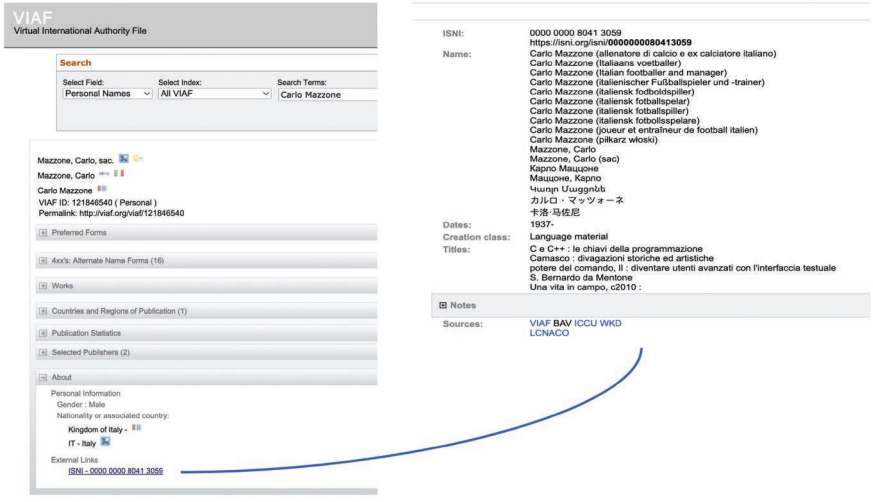

Figura 7. L'entità *Carlo Mazzone, sacerdote* (VIAF ID: 121846540) in VIAF, associata anche all'identificatore ISNI relativo, però, a *Carlo Mazzone, allenatore di calcio* (ISNI: 0000 0000 8041 3059).

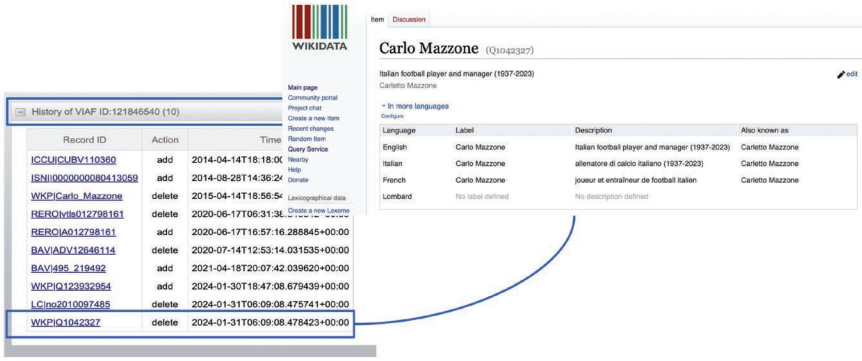

Figura 8. La pagina relativa all'entità *Carlo Mazzone, sacerdote* (VIAF ID: 121846540) che registra, nella sezione *History*, la cancellazione (avvenuta il 31-01-2024) del collegamento con l'ID Wikidata relativo a *Carlo Mazzone, allenatore di calcio* (ID Q1042327).

Ma il dubbio non può coinvolgere l'intera fonte che, come il VIAF o come ISNI, rimane una fonte "autorevole". Per risolvere questi gravi problemi di disambiguazione, i bibliotecari di Parsifal, con il supporto dello staff di analisti e tecnici della Share Family, sta studiando l'introduzione di un nuovo predicato da aggiungere ai dati originali di Parsifal (in modo manuale, nei flussi di correzione delle voci di authority tramite WeCat, e tramite logiche e procedure automatiche), che indichi alle macchine la diversità di una entità rispetto a una apparentemente equivalente presente in altre fonti. La proprietà "diverso da" offrirà ai bibliotecari e alle macchine un nuovo strumento per distinguere e identificare come *diverse* due entità apparentemente *medesime*. Questa evoluzione del sistema WeCat/Authority e dei meccanismi di identificazione e clusterizzazione delle entità di Parsifal è necessariamente demandata a una fase evolutiva del progetto, insieme ad altre strategie di incremento della qualità dei processi che saranno il frutto di una relazione di collaborazione e condivisione dei bibliotecari della Rete URBE.

6. Oltre le biblioteche: Parsifal e la sua integrazione in Wikidata

Il dialogo con le fonti esterne non si esaurisce, tuttavia, nell'interrogazione del VIAF e nelle nuove modalità che si stanno delineando per una sua più efficace integrazione. Ormai da alcuni anni la fonte Wikidata, frutto dell'attività aperta e collaborativa di utenti di tutto il mondo, si presenta come dataset sì numericamente più contenuto di VIAF o ISNI ma qualitativamente più affidabile, in quanto liberamente editabile anche tramite azioni manuali di correzione e arricchimento. Ben distante dalla visione selettiva del VIAF, in cui possono confluire dati provenienti esclusivamente da specifiche e ben selezionate biblioteche e agenzie bibliografiche nazionali, Wikidata viene sempre più spesso scelto come interlocutore privilegiato per l'avvio di importanti progetti Linked Open Data,

nei termini sia di riuso dei dati esistenti, sia viceversa come piattaforma in cui far confluire i propri dati così da garantirne una migliore visibilità. Non a caso, fin dalle prime fasi di sviluppo di Parsifal sono state prospettate diverse ipotesi di integrazione e interrogazione della fonte Wikidata da parte dei bibliotecari della Rete URBE, tra cui un possibile affiancamento di Wikidata alla fonte VIAF nei processi di riconciliazione. Tuttavia, se l'uso di Wikidata nei processi di riconciliazione appare come ipotesi ancora in fase di valutazione e analisi per la sua complessità, un ulteriore e altrettanto interessante filone evolutivo di Parsifal riguarda il riversamento in Wikidata di specifici set di dati, particolarmente significativi della ricchezza informativa e delle peculiari risorse delle biblioteche aderenti. Si parla dunque di insiemi omogenei di dati, facilmente individuabili sulla base di un preciso oggetto di interesse e che siano particolarmente rappresentativi della storia e tradizione culturale della Rete URBE. Tra questi, è stato proposto, tramite un progetto ancora in bozza, di avviare il dialogo con Wikidata a partire dalla produzione scientifica che ruota attorno alle case editrici degli enti ecclesiastici facenti parte del progetto Parsifal, con la finalità sia di inquadrare e delineare la loro evoluzione storica, sia di evidenziare il ruolo centrale che ancora oggi esse rivestono nella vita culturale e accademica degli enti di cui fanno parte. La prima fase di questo progetto prevede l'individuazione di tutte le case editrici coinvolte tramite un'attività di analisi a partire dai cluster presenti nella CKB, attività questa non banale a causa dei frequenti cambi e variazioni che interessano i nomi degli editori. A partire da questa prima analisi, si procederà allo scarico di tutti i titoli delle pubblicazioni legate alle case editrici in questione, a prescindere dalla loro tipologia (monografie, riviste, singoli articoli, ecc.), che verranno mano a mano caricate su Wikidata tramite export massivi di tipo automatico attraverso appositi strumenti (es. *OpenRefine*[13]). Ovviamente tali caricamenti massivi dovranno essere sottoposti anche ad azioni di verifica manuale, così da garantire un'alta qualità dei dati e da poter sfruttare al meglio tutti gli strumenti di riuso ma anche di valutazione, analisi e visualizzazione messi a disposizione di Wikidata[14]. L'attività svolta, infine, non sarà limitata al solo ambito di Wikidata ma implicherà lo sviluppo nel portale Parsifal di un'apposita sezione dedicata agli editori della Rete URBE e alla loro produzione scientifica, così da offrire agli utenti finali un sistema di facile fruizione, integrato in un unico portale bibliografico e senza la necessità del passaggio a un sistema esterno come quello di Wikidata.

A partire da questo primo progetto sugli editori, l'ambizioso obiettivo è quello di individuare altre aree e ambiti specialistici, che diano conto dell'unicità e della ricchezza informativa della Rete URBE, così da replicare il processo sopra deline-

[13] *OpenRefine* è un tool open access utilizzato e diffuso a livello internazionale per la manipolazione e la pulizia dei dati. Il tool è inoltre provvisto di un sistema di integrazione API che gli permette di interrogare e, nel caso di Wikidata, di interagire con le fonti esterne. <https://openrefine.org>.

[14] Un esempio tra questi tool è certamente Scholia che, interrogando Wikidata in tempo reale, presenta una serie di tabelle e grafici sulla produzione di specifici autori, enti di ricerca, editori, o su una precisa tematica. <https://scholia.toolforge.org>.

ato ad altri set di dati potenzialmente di interesse per la comunità internazionale.

Oltre a questo progetto futuro, tuttavia, l'integrazione di Parsifal con Wikidata è già una realtà. Primo e indispensabile passo a qualunque forma di riversamento dati è infatti rappresentato dall'associazione tra gli identificatori dei cluster di Parsifal e gli item di Wikidata.

Questo meccanismo, che si dettaglierà a breve, ripercorre una strada già intrapresa da analoghi progetti della Share Family, tra cui *SHARE Catalogue*[15], il cui uso di Wikidata come "ponte" tra i dati bibliografici e il web ha costituito un indispensabile antecedente.

Ma come garantire concretamente questo collegamento? Trattandosi di una fonte autorevole, dotata di URI persistenti, è stata avviata una Property proposal[16] affinché gli identificatori di Parsifal vengano considerati in Wikidata come una specifica proprietà, allo stesso modo degli identificatori di VIAF, ISNI e altri. La Property proposal per gli identificatori di Parsifal è stata avviata dall'utente Epidosis (alias di Carlo Camillo Pellizzari di San Girolamo) sotto richiesta di Stefano Bargioni ed è stata approvata in breve tempo a seguito del parere positivo di diversi utenti della comunità Wikidata. La proprietà, si sottolinea, interessa per il momento gli ID dei soli cluster agente di tipo persona, in quanto entità sottoposte a un maggiore controllo di autorità.

Successivo step ha invece riguardato il caricamento di un set di informazioni minime, ma significative, sugli agenti di tipo persona di Parsifal, con l'obiettivo di garantire un'associazione automatizzata tra gli item di Wikidata e i corrispettivi cluster di Parsifal. Tale associazione è stata resa possibile tramite lo strumento Mix'n'match[17], che permette il caricamento di appositi "cataloghi", strutturati secondo formati ben precisi (solitamente file con estensione .csv o .tsv), funzionali a garantire un meccanismo di matching automatico tra gli item Wikidata e i dati del catalogo in questione. Ovviamente, tali meccanismi di match vanno poi ulteriormente raffinati, e dunque confermati o meno, tramite un lavoro manuale che può essere svolto da qualunque utente.

Il catalogo di Parsifal, caricato anche in questo caso da un'azione coordinata di Epidosis e di Stefano Bargioni[18], conta attualmente ben 236.767 associazioni automatiche con Wikidata e 162.187 associazione confermate manualmente[19]: un numero significativo se si pensa che il lavoro di importazione su Mix'n'match è stato fatto di recente.

Come si diceva inizialmente, un primo passo dunque, ma assolutamente necessario per garantire una prima forma di scambio con il vasto mondo wiki.

[15] Il progetto è ampiamente documentato in Forziati e Lo Castro 2018.

[16] <https://www.wikidata.org/wiki/Wikidata:Property_proposal/Parsifal_cluster_ID>

[17] <https://mix-n-match.toolforge.org>

[18] Cogliamo l'occasione di questa seconda citazione per ringraziare Stefano Bargioni e Camillo Carlo Pellizzari di San Girolamo del loro prezioso contributo, frutto di una sempre viva passione per il mondo wiki.

[19] <https://mix-n-match.toolforge.org/#/catalog/6216>

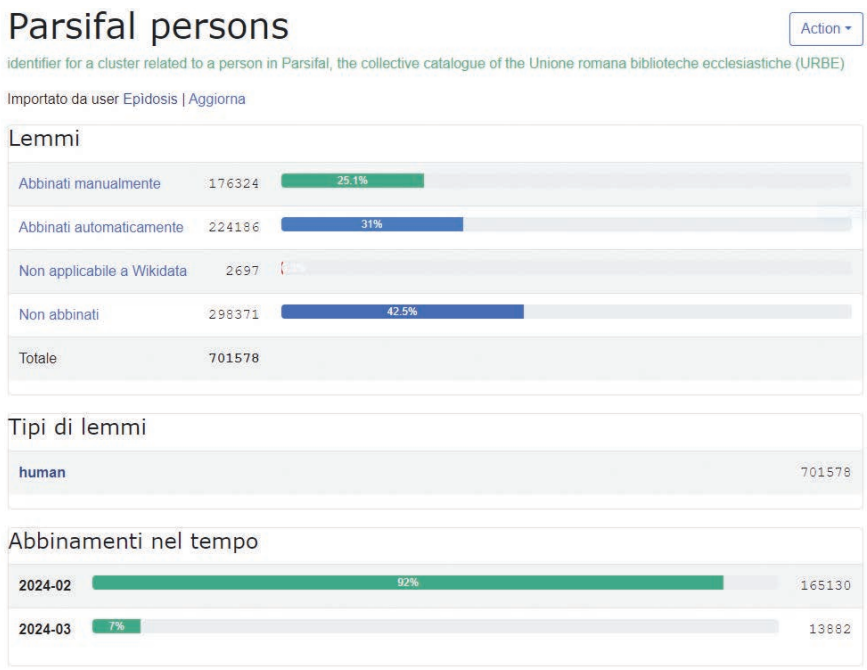

Figura 9. La pagina relativa al catalogo di Parsifal importato dentro Mix'n'match.

La Rete URBE, con il suo prezioso e in alcuni casi unico patrimonio, ha dato così il via, con l'iniziativa Parsifal e con tutte le estensioni di questo progetto, a un cammino che ha come fondamento la pratica della collaborazione concreta tra Istituzioni e l'apertura a forme di condivisione nazionale e internazionale, sia nell'ambito della Share Family che nel più ampio panorama del web.

Riferimenti bibliografici

Forziati, Claudio, e Lo Castro Valeria. 2018. "La connessione tra i dati delle biblioteche e il coinvolgimento della comunità: il progetto SHARE Catalogue-Wikidata." *JLIS.it*, IX, 3.

IFLA. 2020. *IFLA Library Reference Model. Un modello concettuale per le informazioni bibliografiche*, a cura di Pat Riva, Patrick Le Boeuf, Maja Žumer, Edizione italiana. Roma: ICCU.

Possemato, Tiziana. 2023. "Linked data: un'opportunità per il riuso". In "Atti del Convegno di studi *Fare per non sprecare. Nei laboratori del riuso digitale*." *DigItalia*, XVIII, 2. <https://doi.org/10.36181/digitalia-00081>.