Linked data: un'opportunità per il riuso

«Dig*Italia*» 2-2023
DOI: 10.36181/digitalia-00081

Tiziana Possemato
@Cult, Casalini Libri

L'utilizzo delle fonti disponibili sul web costituisce un'enorme opportunità nei processi di conversione dei cataloghi bibliografici in linked open data: i processi di risoluzione delle entità, per identificare gli oggetti nel mondo reale, e i meccanismi di arricchimento dei dati, per potenziare l'esperienza di ricerca da parte dell'utente finale, rende indispensabile l'utilizzo di fonti diverse dai propri dati originari. Ma non tutto ciò che è disponibile sul web è realmente fruibile: fattori legati alla qualità dei dati e all'interoperabilità possono complicare o addirittura impedire lo scambio reciproco di dati informativi. Questa analisi presenta alcuni casi di utilizzo di fonti esterne per i processi sopra citati, evidenziando gli elementi di criticità che più facilmente si incontrano nelle fasi di riutilizzo dei dati e le possibili ipotesi di soluzione.

Il *riuso* nei progetti di conversione e pubblicazione di linked open data

Quando ho ricevuto l'invito a condividere esperienze e pensieri sul *riuso* dei dati, ho riflettuto sul titolo dell'incontro da cui questo contributo nasce: *Fare per non sprecare. Nei laboratori del riuso digitale*. Mi sono chiesta come parlare del riuso partendo dal mio laboratorio quotidiano, dall'esperienza di gestore e utilizzatore di dati e metadati nell'ambito dei progetti che sono parte del mio lavoro. Il riferimento è quello della iniziativa internazionale *Share Family*[1], che ha tra i suoi obiettivi:

– l'arricchimento dei dati originali con dati provenienti da fonti esterne;
– l'identificazione e riconciliazione delle entità attraverso meccanismi di entity resolution (risoluzione delle entità);
– la conversione dei cataloghi dai formati originali in linked data;
– la restituzione dei dati convertiti e arricchiti alle biblioteche partecipanti, per il riuso in sistemi differenti;
– la pubblicazione dei dati in linked data in una discovery platform, per la fruizione da parte di altre istituzioni e degli utenti finali;
– la creazione di un ambiente e di strumenti collaborativi per facilitare la condivisione di dati e servizi.

A ben vedere, in ognuno di questi obiettivi c'è una componente di *riuso* legata ai dati, alle esperienze e competenze condivise o ai servizi.

[1] Share-VDE è un'iniziativa che abbraccia istituzioni autorevoli in diversi ambiti e localizzate in diverse aree geografiche, tra il Nordamerica, il Canada e l'Europa. Obiettivo principale dell'iniziativa è quello di facilitare la catalogazione e l'esposizione dei record bibliografici in linked data, supportando così la transizione dall'ambiente catalografico tradizionale a modelli innovativi, attraverso l'applicazione del paradigma dei linked data. Per una più ampia panoramica su Share-VDE e la Share Family si rimanda alla pagina Wiki della iniziativa: <https://wiki.share-vde.org/wiki/ShareFamily:Main_Page>.

Quasi in ciascuna delle macro-fasi in cui un progetto di pubblicazione di linked open data (LOD) si articola, esiste un elemento di riuso che non è opzionale rispetto all'intero flusso di elaborazione dei dati. Quando si parla di conversione e pubblicazione di dati in linked data ci si riferisce, infatti, a un trattamento delle informazioni finalizzato a rendere i dati fruibili da macchine, senza dimenticare, ovviamente, la possibilità che i dati siano utilizzati dagli utenti finali per estendere le potenzialità di identificazione delle risorse di proprio interesse. Il contesto in cui tutto questo avviene è quello del web, nella vastità dei propri spazi e delle informazioni in esso pubblicate. È raro, dunque, che questi progetti non prevedano una forma di arricchimento del dato originale, estendendo le connessioni ad altri insiemi informativi e moltiplicando la rete semantica di cui ciascuna risorsa costituisce un nodo. Tim Berners-Lee, l'inventore del web e iniziatore del progetto linked data, suggerisce uno schema di implementazione a cinque stelle per i linked data. Il sistema è inteso come *cumulativo*: ogni stella aggiuntiva presuppone che i dati soddisfino i criteri dei passaggi precedenti, in un processo che rende i dati sempre più aperti e fruibili da altri. La quinta stella viene assegnata proprio quando venga soddisfatta questa condizione: creare link ad altri dataset pubblicati in linked open data (*interlinking*)[2]. Questo meccanismo è una delle fasi di *ar-*

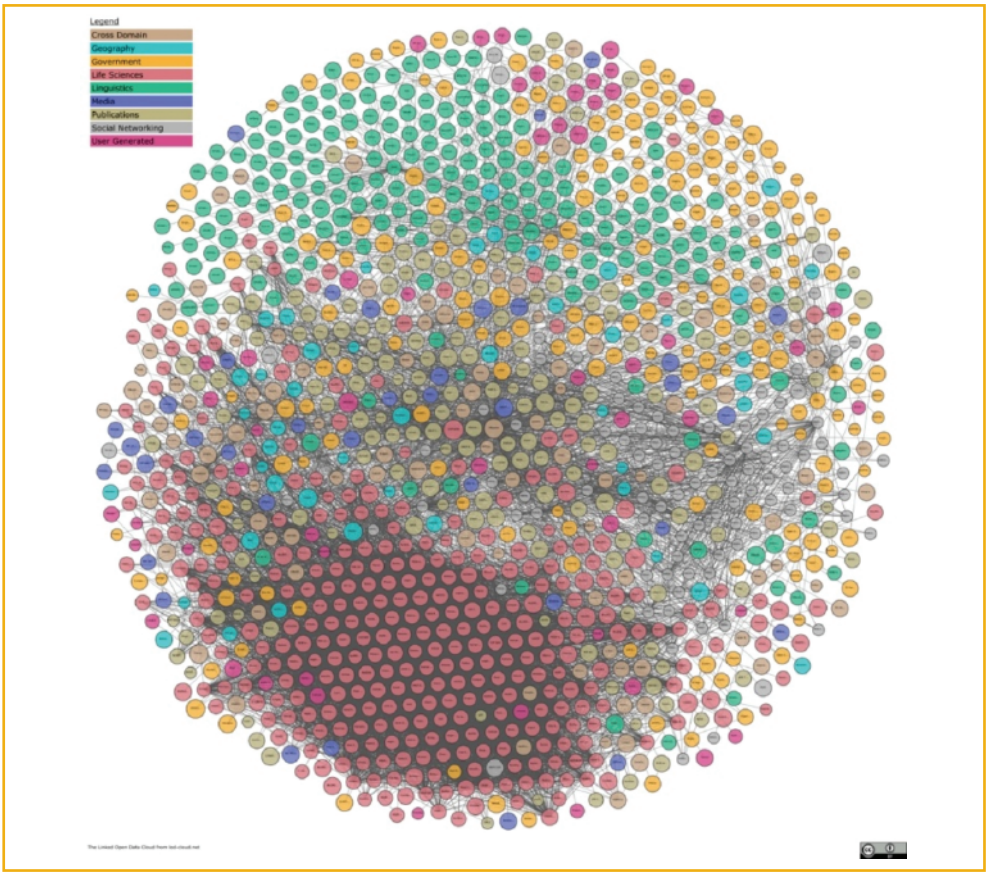

Figura 1. *The Linked Open Data Cloud: rappresentazione dei dataset pubblicati in linked open data e collegati l'uno all'altro. Il diagramma contiene 1.255 dataset con 16.174 link, alla data di maggio 2020*

[2] *Five stars: all of the above and links to other Linked Open Data*, <https://www.w3.org/2011/gld/wiki/5_Star_Linked_Data>.

ricchimento dei dati e consente di fruire di informazioni che altri hanno strutturato e reso disponibili. La selezione delle fonti che si intenda utilizzare per questi collegamenti è un momento delicato nel flusso di lavorazione, e fortemente condizionato dal tipo di dato che si voglia arricchire (dati riferibili ad agenti, a opere, a luoghi geografici, a soggetti) e da altri criteri di selezione: l'*autorevolezza* è certamente l'elemento preponderante, insieme all'*esaustività* della fonte rispetto ad uno specifico dominio e alle caratteristiche di *aperura* e *accessibilità*. Anche l'elemento di *aggiornamento* influenza fortemente la selezione di una fonte: nei processi di arricchimento l'obsolescenza dell'informazione è un fattore che mette fortemente a rischio l'usabilità del dato. Tra le fonti autorevoli più largamente utilizzate nel dominio bibliografico annoveriamo certamente VIAF[3], ISNI[4], gli authority file della Library of Congress[5], FAST[6], GND[7], Wikidata[8], GeoNames[9], IDRef[10], il Nuovo Soggettario[11] della Biblioteca Nazionale Centrale di Firenze, più diversi altri vocabolari e liste di termini controllati. Ma il web contiene migliaia di dataset potenzialmente collegabili tra loro e la nota rappresentazione del cloud dei linked data (Fig. 1) è cresciuta così tanto da diventare quasi non rappresentabile nella sua interezza[12].
Un progetto di pubblicazione di un dataset in LOD, dunque, aggiungerebbe poco all'elemento informativo originale se non cercasse altrove quei dati che sono determinanti sia nella fruizione da parte dell'utenza finale che per l'identificazione delle entità nei processi di entity resolution[13].

Riusare i dati per arricchire l'informazione

La creazione di link a fonti esterne offre vantaggi sia a chi pubblichi (rende i dati più ricercabili, aumenta il valore informativo dei dati), che a chi consumi dati (per esempio, aumenta il potenziale di scoperta di informazioni, e consente di accedere direttamente alla struttura dei dati per poterli più ampiamente riutilizzare, eliminando le barriere solitamente create dai formati chiusi).
L'arricchimento prevede non solo l'associazione al dato di origine di URI assegnati alla stessa entità su database esterni, ma permette anche l'acquisizione, dall'esterno (e dunque il riuso) di informazioni non inizialmente presenti nei dati di origine.
In Fig. 2 un esempio di arricchimento di dati sul portale Share-VDE[14]. L'entità rappresentata è già il risultato di un'aggregazione "ragionata"[15] di dati bibliografici e di autorità provenienti dalle fonti che partecipano all'iniziativa. L'entità è poi arricchita con i link alle fonti esterne su cui la stessa entità sia stata rilevata e con dati informativi provenienti da Wikimedia e da Wikidata. La Fig. 3 mostra il risultato dei processi di arricchimento dei dati sul portale Parsifal[16]: anche in questo caso l'entità è arricchita con dati non inizialmente esistenti nei cataloghi di origine.

[3] https://viaf.org/.
[4] https://isni.org/.
[5] https://id.loc.gov/.
[6] https://www.oclc.org/research/areas/data-science/fast.html.
[7] https://www.dnb.de/EN/Professionell/Standardisierung/GND/gnd_node.html.
[8] https://www.wikidata.org/wiki/Wikidata:Main_Page.
[9] https://www.geonames.org/.
[10] https://www.idref.fr/.
[11] https://thes.bncf.firenze.sbn.it/.
[12] Il sito <http://cas.lod-cloud.net/> propone adesso la sezione *Subclouds by domain* proprio per consentire l'esplorazione in sottoinsiemi di dataset.
[13] Di entity modeling e dei meccanismi di risoluzione delle entità nel web (entity resolution) ho avuto modo di parlare in Tiziana Possemato, *Entity Modeling: Traces of an Evolving Path*, «JLIS.It» 13 (2022), n. 3, p. 12–28, <https://doi.org/10.36253/jlis.it-481>.
[14] https://wiki.share-vde.org/wiki/Main_Page.
[15] Per aggregazione "ragionata" intendo un'aggregazione frutto di regole complesse di entity resolution, che nascono dall'analisi dei dati e sono tradotte poi in algoritmi e processi di clusterizzazione.
[16] Parsifal è il progetto di pubblicazione del catalogo unico in linked data prodotto dall'integrazione dei cataloghi di 17

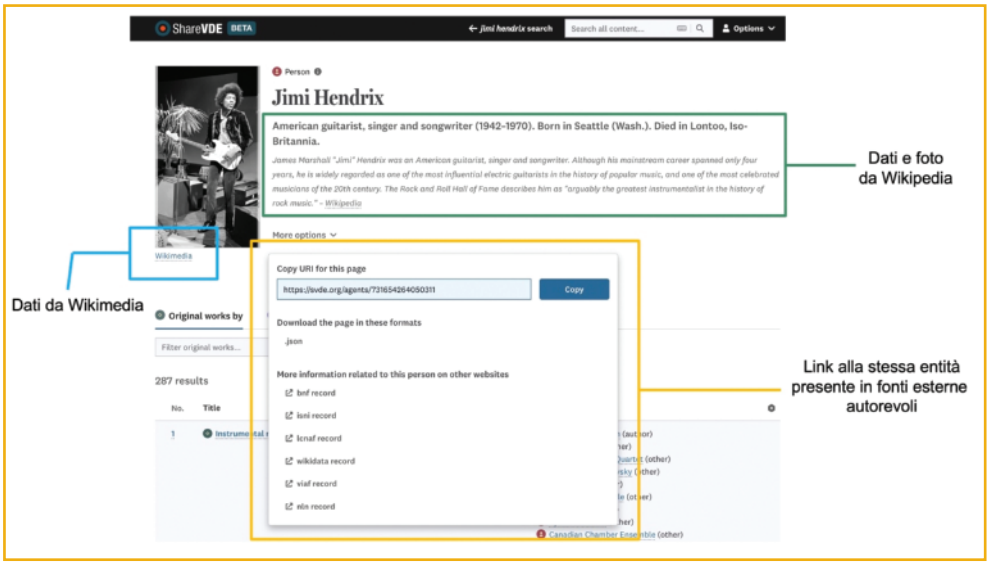

Figura 2. *Un esempio di arricchimento di dati sul portale Share-VDE per l'entità Jimi Hendrix*

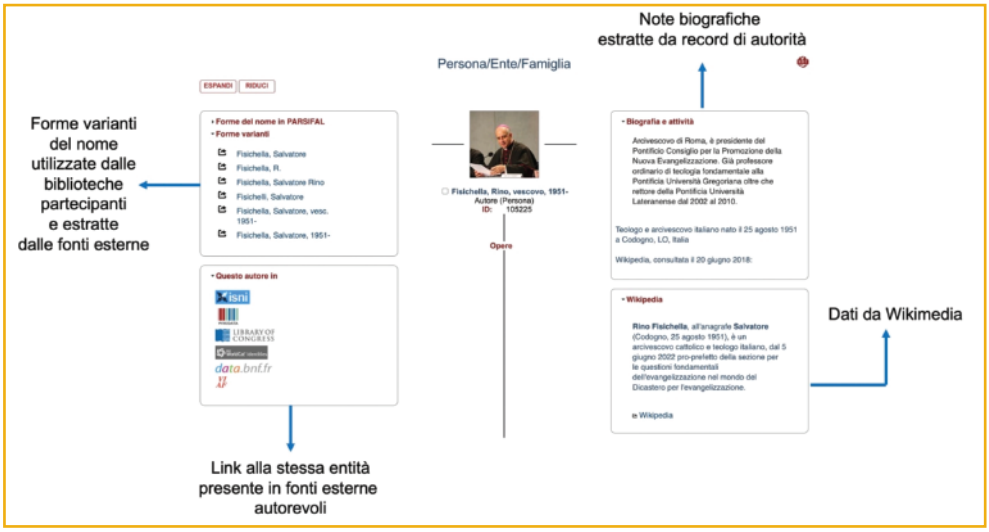

Figura 3. *L'entità Rino Fisichella sul portale Parsifal, arricchita con dati provenienti da diverse fonti*

Il valore dell'arricchimento attraverso il riuso di informazioni già selezionate e strutturate in altre fonti non richiede una presentazione estesa: gli esempi riportati sono due tra centinaia di progetti e siti che utilizzano questa enorme opportunità di selezionare le fonti di interesse e nello stesso tempo ottimizzare i tempi di lavorazione dei dati, per un'informazione più efficace. Ma soprattutto, il riuso di dati per l'arricchimento rende all'oggetto descritto un *contesto*[17] che estende incredibilmente il potere informativo indirizzato ai ricercatori.

biblioteche di Università e Istituzioni pontificie romane della rete URBE – Unione Romana Biblioteche Ecclesiastiche. Il progetto è parte della Share family ed è consultabile al link <https://parsifal.urbe.it/parsifal/home>.

[17] Si veda quanto dice Alex Stinson nell'articolo *Wikidata in collections: building a universal language for connecting GLAM catalogs:* «This kind of "context beyond the collection" is particularly important for materials that necessar-

La riflessione critica va tuttavia spostata sugli investimenti necessari per mantenere aggiornate le informazioni di arricchimento e di quanto il riuso debba essere collaborativo se lo si vuol rendere efficace. E di questo parleremo successivamente.

Il riuso dei dati nei processi di entity resolution

Ciascuna cosa che sia parte del nostro mondo, sia essa concreta (una persona, un libro, un'opera d'arte, un oggetto del nostro quotidiano) sia essa astratta (un concetto, un'idea) può essere rappresentata in modi differenti all'interno della stessa base di dati (come nel caso di una persona descritta con il nome anagrafico o con lo pseudonimo), o in modo ancor più evidente, in basi dati differenti (per esempio, nel caso di un titolo di un libro trascritto in scrittura latina o in scrittura araba). Questa difformità ha radici diverse, a partire dalla nostra abitudine a guardare le cose da una certa prospettiva. Un'entità, nella sua realtà, è sempre molto più complessa di come riusciamo a rappresentarla e pensare di poterne dare un quadro esaustivo è dichiaratamente un'illusione. Ciascuna entità viene, dunque, rappresentata con uno o più *profili* in relazione a diversi fattori correlati: fattori culturali, linguistici, dipendenti dall'applicazione di differenti regole descrittive, ma soprattutto *contestuali.* Il contesto entro il quale un'entità si esprime è uno degli elementi che più influenzano la modalità di presentazione al mondo dell'entità stessa. Nell'ambito del web, possono accumularsi centinaia di profili diversi della stessa entità, e i processi di entity resolution devono riconoscerli come riferibili ad uno stesso oggetto per poi utilizzarli per estendere la possibilità di identificazione. Più i profili sono ricchi e riferibili a fonti autorevoli, più i processi di identificazione possono essere estesi e qualitativamente certificabili. Esempi di riuso dei dati con questa finalità di disambiguazione sono quelli che utilizzano alcuni elementi descrittivi per riconoscere la medesima entità nascosta dietro pseudonimi, eteronimi, alias e altre identità alternative, così come i processi, in alcuni casi più complessi, che disambiguano entità apparentemente coincidenti (omonimie). Perché due profili diversi siano riconducibili, dalle macchine e dunque da processi automatizzati, ad una medesima cosa, bisogna trovare un *gancio* tra i profili diversi, quindi un primo *anello di congiunzione* in quella che possiamo definire una *catena identificativa*. I database prodotti dagli istituti culturali e in particolare dalle biblioteche sono particolarmente ricchi di informazioni che possono assolvere questa funzione. I record bibliografici così come quelli di autorità prodotti dalle biblioteche registrano spesso le informazioni di collegamento tra un'identità anagrafica (per esempio Alberto Pincherle) e l'identità pseudonimica (Alberto Moravia). In un record di autorità in formato Marc 21 o Unimarc, questo anello è costituito dal blocco di campi *4XX - See Reference Tracing Block*[18] che collega, a favore delle macchine, la stringa di testo *Moravia, Alberto* (forma pseudonimica) con la forma *Pincherle, Alberto* (definito, secondo le regole REICAT applicate nell'esempio estratto dall'OPAC SBN e qui utilizzato, come *nome reale*).

ily require interpretation, such as art or historical objects. Take for example at the Museum of Modern Art in New York (MOMA): the Museum has integrated Wikidata and associated Wikipedia articles into "artist" pages in their online catalogue. […] The Museum decided to supplement the artist profiles with first Wikipedia articles, and then with the Wikidata ids: by doing so, both the public and reusers of the MOMA data can add even more context from Wikidata or the other vocabularies connected to the collection». Alex Stinson, *Wikidata in Collections: Building a Universal Language for Connecting GLAM Catalogs,* «*Down the Rabbit Hole*» (blog), 9 aprile 2018, <https://medium.com/freely-sharing-the-sum-of-all-knowledge/wikidata-in-collections-building-a-universal-language-for-connecting-glam-catalogs-59b14aa3214c>.

[18] Il blocco di tag 4xx contiene forme varianti del nome dell'entità analizzata, da cui è necessario fare riferimento per arrivare all'accesso principale scelto per quella entità. Si veda *UNIMARC Manual. Authorities Format*, a cura di IFLA UBCIM Programme, 2nd rev. and enl., UBCIM Publications, new ser., v. 22. München: K.G. Saur. 2001.

=LDR 00701nx a2200217 45
=001 IT\ICCU\CFIV\000648
=005 20230103121517.4
=010 \\$a0000000122760711
=100 \\$a20140716aitaa50 ba0
=101 \\$aita
=102 \\$aIT
=152 \\$aREICAT
=200 \1$aMoravia$b, Alberto
=300 \\$a1907-1990 // Pseudonimo di Alberto Pincherle, narratore, critico letterario e cinematografico, cofondatore della rivista Nuovi argomenti, deputato. Nato e morto a Roma.
=**400 \1**$3IT\ICCU\CFIV\000649$9**Nome reale**$a**Pincherle**$b**, Alberto**

Il primo anello della catena identificativa è stato così definito, e il ruolo di individuare dietro due profili diversi la stessa entità è stato assolto, sia che il processo di risoluzione dell'entità funzioni nell'ambito dello stesso database (nel database dell'OPAC SBN, in questo caso), sia che il processo di risoluzione colleghi database diversi. In questo ultimo scenario, un database esterno a SBN potrebbe interrogare i dati dell'OPAC per assumere l'informazione sulla relazione di pseudonimia qui risolta.
Abbiamo, però, menzionato quanto alcuni processi di entity resolution possano essere ben più complessi in altre situazioni, come quelle di *omonimia*. Rimanendo all'esempio di Alberto Pincherle, quanto registrato nel tag 400 del record di autorità chiarisce in modo inequivocabile la relazione tra i due nomi[19] riferibili al nome anagrafico e allo pseudonimo. Che possa esistere un Alberto Pincherle diverso da quello che utilizza come pseudonimo il cognome Moravia può essere suggerito dalla presenza di un record di autorità che espliciti altri elementi qualificanti come gli estremi cronologici. Ma senza avere la certezza che questa sia una norma rigida e diffusamente rispettata. La tradizione catalografica pre-RDA[20] non chiedeva di aggiungere elementi che potessero disambiguare due stringhe identiche, se non in casi eclatanti di omonimia[21]. L'orizzonte di riferimento è quello del catalogo locale, dove diventa semplice, a ogni operazione catalografica, verificare che non esistano, nell'ambito del catalogo stesso, degli omonimi. Ma quando i dati escono dal catalogo locale per essere presentati e riutilizzati nel web, la problematica di centinaia di stringhe identiche riferite a cose diverse diventa macroscopica.
Per arricchire la catena identificativa i processi di entity resolution devono aggiungere altri anelli, come, appunto, quello delle date di nascita e morte di una persona che, se non presenti nel proprio catalogo, devono essere recuperate altrove, nelle fonti informative disponibili sul web. Nel caso dell'OPAC SBN, l'entità Alberto Pincherle storico italiano del cristianesimo è ben identificata con le date di nascita e morte (*Pincherle, Alberto <1894-1979>*) e dunque questa fonte, se fosse disponibile secondo le tecno-

[19] Qui parlo volutamente di relazione tra *nomi* e non tra *identità diverse* perché l'ambito di esistenza dell'esempio utilizzato è quello della catalogazione tradizionale, ancora non orientata all'entity modeling e dunque all'idea di identificare, in questo caso, un agente con le sue diverse identità.

[20] RDA – Resource Description and Access, già nella prima versione (Original RDA), che sostituiva le regole AACR2 – Anglo American Cataloging Rules, 2nd edition – prevede, nell'istruzione 9.3 che le date associate alla persona, (data di nascita e data di morte) siano *elementi essenziali*. Alle date di nascita e morte RDA chiede di aggiungere il periodo di attività della persona come elemento essenziale, solo se necessario per distinguere una persona da un'altra con lo stesso nome.

[21] REICAT indica, al paragrafo 15.3.1 A, quanto segue, a proposito delle qualificazioni cronologiche: «Per distinguere persone con lo stesso nome si indicano, se possibile, l'anno della nascita e, per i defunti, l'anno della morte. Se le date non sono note con certezza si possono usare indicazioni approssimative dell'epoca o del periodo di vita o di attività della persona». *Regole italiane di catalogazione – REICAT*, a cura della Commissione permanente per la revisione delle regole italiane di catalogazione, Roma: ICCU, 2009.

logie del web e quindi utilizzabile nei processi di entity resolution operati dalle macchine, sarebbe certamente utile per la disambiguazione delle due entità. Il problema, però, rimane quello di una forma *Pincherle, Alberto* senza date di nascita e morte utilizzata in alcuni contesti come forma preferita del nome o, indifferentemente, come forma variante e dunque associata sia all'una (lo storico del cristianesimo) che all'altra entità (lo scrittore e giornalista italiano). Quell'anello che, nel caso di pseudonimia, aveva risolto l'ambiguità tra il nome anagrafico e lo pseudonimo, trasferito in un contesto più ampio, rischia di diventare addirittura una criticità e un elemento di ambiguità. Su questo elemento torneremo tra poco. Per ora, proviamo a seguire il percorso di costruzione di una catena identificativa nell'ambito di processi automatizzati.

La selezione di fonti autorevoli accessibili nel web, dove certe tematiche di disambiguazione siano state poste e in alcuni casi anche risolte, diventa dunque obbligatoria. Nel caso che stiamo utilizzando come esemplificativo di un processo, la fonte Wikidata, disponibile nel web con tecnologie che nascono per il riuso, potrebbe risolvere diversi casi di disambiguazione, essendoci dietro la creazione di ciascun item una forte componente di controllo umano, distribuito tra la vasta comunità di partecipanti. In questa fonte le due entità Alberto Pincherle[22] e Alberto Moravia[23] sono chiaramente identificate, e la proprietà *different from*[24] sancisce e registra questa distinzione. La fonte Wikidata suggerisce anche l'uso di altre proprietà che potrebbero essere utilizzate nei processi di entity resolution laddove le date di nascita e morte delle due entità, nei dati di partenza, non fossero presenti o anche dove fosse necessario aggiungere altri anelli alla catena identificativa: la proprietà *occupation*[25], per esempio, sup-

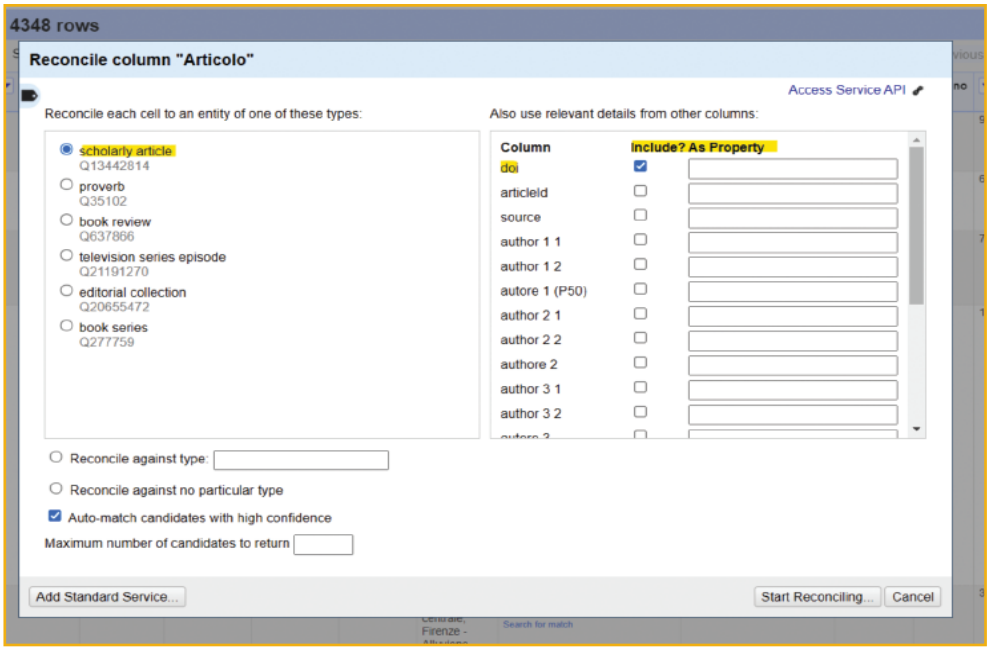

Figura 4. *La funzione di aggiunta di parametri da utilizzare per riconciliare le entità nel tool OpenRefine*

[22] https://www.wikidata.org/wiki/Q28357631.
[23] https://www.wikidata.org/wiki/Q161933.
[24] <https://www.wikidata.org/wiki/Property:P1889>: «item that is different from another item, with which it may be confused».
[25] <https://www.wikidata.org/wiki/Property:P106>: «occupation of a person».

porterebbe i processi di identificazione, ove incrociati con i *soggetti* associati alle opere ascrivibili all'uno e all'altro autore. Le macchine possono certamente essere istruite per *capire* che il soggetto "Cristianesimo – Storia" o "Storia del Cristianesimo" è più facilmente riferibile all'occupazione "storico" (e dunque ad Alberto Pincherle) che non all'occupazione "giornalista", "poeta", "scrittore".
L'estensione dell'orizzonte degli elementi registrati su fonti diverse e utilizzati per migliorare i processi di identificazione laddove le tradizionali stringhe del nome (cognome, nome con date di nascita e morte) si rivelassero insufficienti può essere riassunta nel passaggio dallo *string matching* all'*entity matching*: in questa diversa modalità di riutilizzo dei dati per i processi di entity resolution, come brevemente tracciato attraverso l'esempio di Alberto Pincherle/Alberto Moravia, l'identificazione avviene attraverso l'aggiunta, in passi differenti, di elementi riferibili all'entità e non alla strutturazione di una stringa di testo. La funzione di arricchimento di elementi identificativi nei casi di mancata disambiguazione delle entità che il tool OpenRefine[26] mette a disposizione (Fig. 4), è esemplificativa di un metodo che, proposto per processi semi manuali, può essere adottato in larga scala nei processi completamente automatizzati.

L'utilizzo delle proprietà registrate nei diversi profili riferibili a una medesima entità, intese come *attributi* e *relazioni*, amplifica notevolmente la qualità dei risultati dei processi di identificazione. Ma le relazioni significative che aiutano questo processo dipendono molto dal contesto in cui i dati siano espressi e al quale si riferiscono: in ambito bibliografico la relazione più significativa in termini di identificazione delle entità è quella tra l'autore (o creatore) e la sua opera, e viceversa. Non si tratta, però, di una relazione diffusamente espressa nei dati tradizionali: i cataloghi bibliografici solo in casi di autori classici o molto prolifici suggerivano la creazione di campi del tipo Nome/Titolo, quindi quei campi che consentono un raggruppamento dell'autore con le sue opere; pochissimi i record di autorità dedicati alla descrizione di opere e pochissimi i campi riferibili a titoli di opere nei record bibliografici[27]. Questa significativa lacuna nei cataloghi bibliografici, insieme ad altri limiti legati alla strutturazione dei dati nelle registrazioni tradizionali, amplifica una problematica importante nel riuso delle fonti nei processi di identificazione. Di seguito, una estrema esemplificazione di una logica di clusterizzazione[28] e dei rischi che processi massivi, completamente automatizzati, comportano.

Il rischio e i limiti del riuso

Gli esempi di entità riferibili ad *Alberto Moravia* e *Alberto Pincherle* sono rappresentati in Fig. 5 con le etichette scelte in molte fonti di dati come preferite e alcune delle diverse forme del nome definite come varianti, con o senza elementi cronologici che le qualifichino. Il diagramma registra anche alcune delle relazioni con le opere ascrivibili a ciascuna delle due entità.

[26] OpenRefine come strumento di riconciliazione e clusterizzazione delle entità è usato in moltissimi progetti di conversione e pubblicazione dei dati in RDF <https://openrefine.org/>. Offre diverse funzioni di gestione dei dati anche se richiede comunque un consistente intervento manuale e non può, dunque, essere pensato come sufficiente in progetti che trattano quantità di dati rilevanti.

[27] Il riferimento qui è ai titoli di opere e non di manifestazioni e quindi a quello che nella tradizione catalografica pre-RDA veniva identificato come titolo uniforme e che può differire anche sensibilmente dal titolo riportato nella specifica pubblicazione.

[28] Per processi di clusterizzazione si intendono qui meccanismi di identificazione delle entità descritte attraverso set di metadati e la generazione di un *cluster* che rappresenta l'entità stessa con i suoi attributi e le sue relazioni con altre entità. Un cluster è, dunque, un oggetto costruito dalla macchina nei processi di entity resolution cui è associato un identificatore, solitamente un URI nel caso di pubblicazione in termini di RDF.

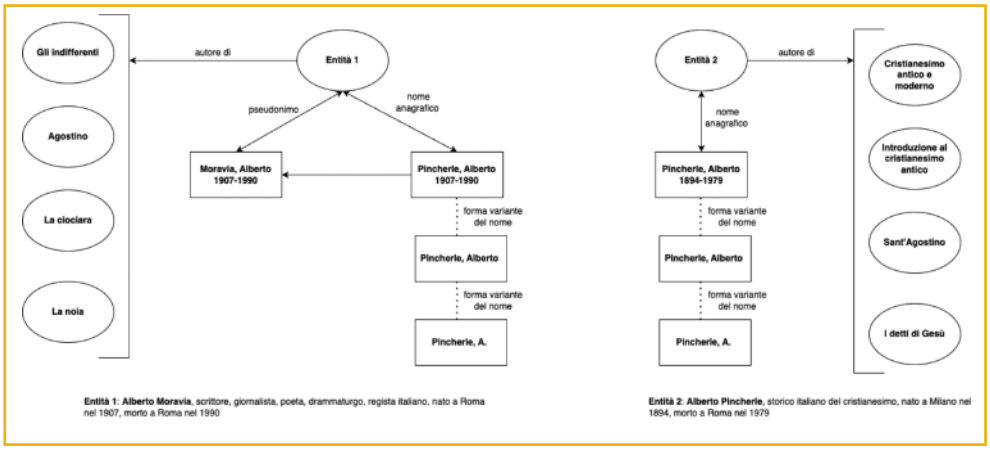

Figura 5. *Le entità Alberto Moravia e Alberto Pincherle con una rappresentazione schematizzata delle forme varianti del nome e delle relazioni con le rispettive opere*

Le tradizioni e le scelte catalografiche espresse nei record bibliografici e di autorità abbiamo già detto quanto possano essere difformi e non sempre attente, per naturale distanza cronologica dal concetto di entity modeling, ai problemi di identificazione delle entità rispetto a quelli relativi alla costruzione di stringhe descrittive uniformi. Tra le forme varianti registrate per l'una e l'altra entità si trovano stringhe perfettamente coincidenti (*Pincherle, Alberto* e *Pincherle, A.*) entrambe prive di notazione cronologica. Quell'anello che costituisce un valore nel riconoscimento della stessa entità dietro forme del nome differenti, può diventare un elemento di criticità nei processi massivi, e come tali poco controllabili, di riutilizzo di fonti esterne. Nella Fig. 6 è schematizzato un possibile processo di entity resolution che aggancia però le medesime forme varianti del nome presenti in entrambe le entità nella fonte VIAF[29] (le forme rappresentate in rosso nel diagramma di Fig. 6).

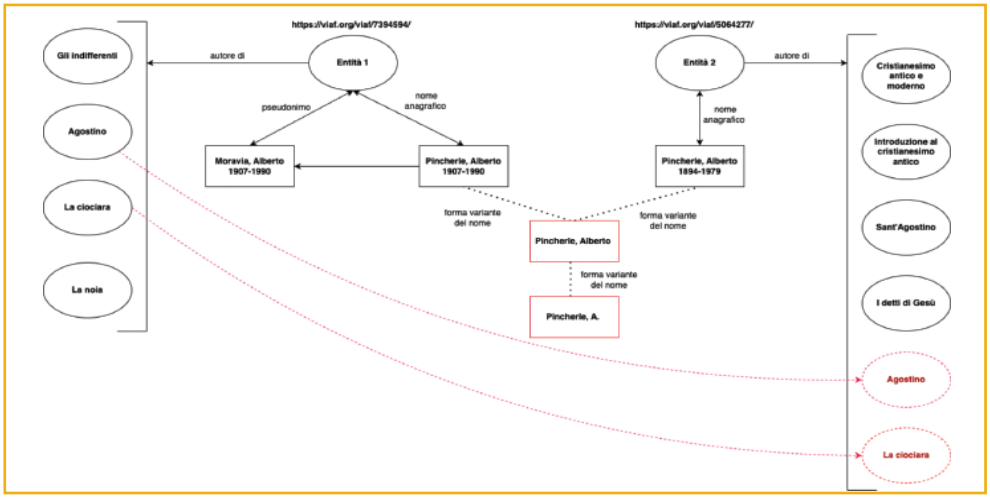

Figura 6. *Una schematizzazione dell'utilizzo delle forme varianti del nome, associate alla forma preferita di ciascuna entità, nei processi di entity resolution e le criticità prodotte dal meccanismo automatizzato in VIAF*

[29] La pagina di VIAF dedicata all'entità *Pincherle, Alberto*, ID 5064277, è stata visitata l'ultima volta il 14 maggio 2023.

Il risultato di questa sovrapposizione di forme varianti del nome è che la sezione dei Titoli collegata ad Alberto Pincherle (lo storico del cristianesimo) presenta gravi problemi di attribuzione delle opere, proprio per via di quell'anello in comune tra le due entità, evidentemente utilizzato in diversi record che hanno contribuito alla costruzione del cluster con identificativo VIAF 5064277 (si veda la Fig. 7 dove è riprodotta una porzione della pagina dedicata ad Alberto Pincherle su VIAF).

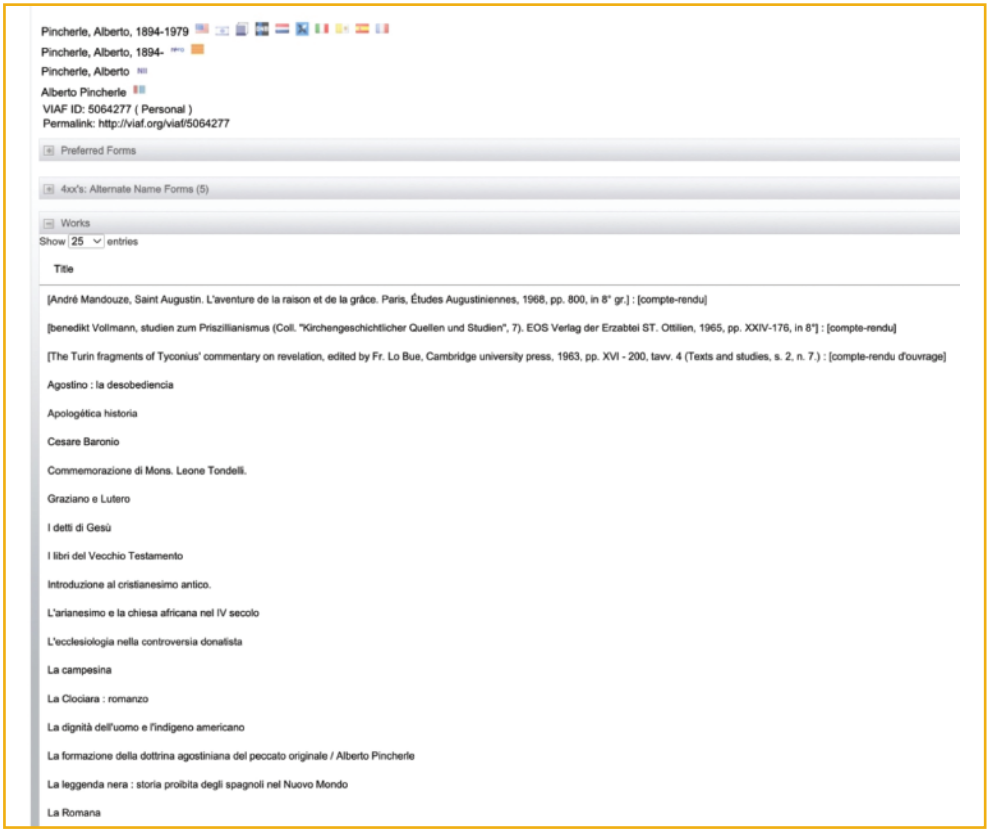

Figura 7. *L'entità Pincherle, Alberto, 1894-1979 con associate diverse opere di Alberto Moravia (si noti, tra le altre, Agostino, La campesina, La ciociara, La romana)*

Tornando all'importanza delle relazioni tra l'autore e la sua opera nella identificazione delle entità (la capacità di riconoscere un autore attraverso la sua opera e di identificare l'opera attraverso l'attribuzione al suo autore) si capisce come possa essere delicato l'utilizzo di fonti autorevoli generate da processi macchina massivi per sciogliere certe ambiguità e come possa essere elevato il rischio di amplificare e replicare infinite volte il medesimo errore. La soluzione a questa criticità non è semplice, e dovrebbe coinvolgere una quantità di dataset sempre più estesa. Tra le strade praticabili per ridurre le criticità, il controllo dei risultati dei processi di identificazione attraverso strumenti che prevedano un coinvolgimento diretto di esperti dei diversi settori. Il modello Wikidata, in questo senso, è significativo perché chiama una comunità diffusa a collaborare in attività di analisi dei dati e di cura della qualità degli stessi. Solo per citare uno dei tanti esempi: nell'ambito del Gruppo Wikidata per Musei, Archivi e Biblioteche[30] si è

[30] https://www.wikidata.org/wiki/Wikidata:Gruppo_Wikidata_per_Musei,_Archivi_e_Biblioteche.

creato un più ristretto gruppo di lavoro dedicato alla registrazione delle Riviste di biblioteconomia[31]. La pagina della discussione sul progetto registra alcune fragilità da tenere presenti durante la gestione dei dati, tra cui un elenco di omonimie non risolte da OpenRefine nei processi di clusterizzazione e gestite manualmente dal gruppo di partecipanti[32]. La strada della collaborazione è qui certamente tracciata, ma nasce e si esaurisce nell'ambito dello stesso progetto. Per ottenere risultati rilevanti su ampia scala è necessario invece esportare questi modelli al di fuori delle singole esperienze, per sfruttare al massimo il potenziale che le tecniche di riuso mettono a disposizione ma che diventano insostenibili (anche da un punto di vista economico) se non largamente condivise.

Il riuso efficace se collaborativo

L'esposizione dei dataset sul web con metodologie che ne consentano il riutilizzo favorisce l'impianto di indagini sulla qualità dei dati attraverso tool di comparazione e di analisi statistica dei risultati di elaborazione. Il servizio Wikidata Query Services[33], per esempio, consente di creare in modo guidato delle query significative sui dati indicizzati per poi analizzarne i risultati. Si veda per esempio, in Fig. 8, il grafo prodotto dalla esecuzione di una query[34] che rileva gli autori collegati all'entità *Alberto Moravia* con la relazione ad alcune opere rilevanti, elementi che possono costituire un punto di partenza per analisi di maggior dettaglio dei dati che hanno prodotto quel risultato[35].

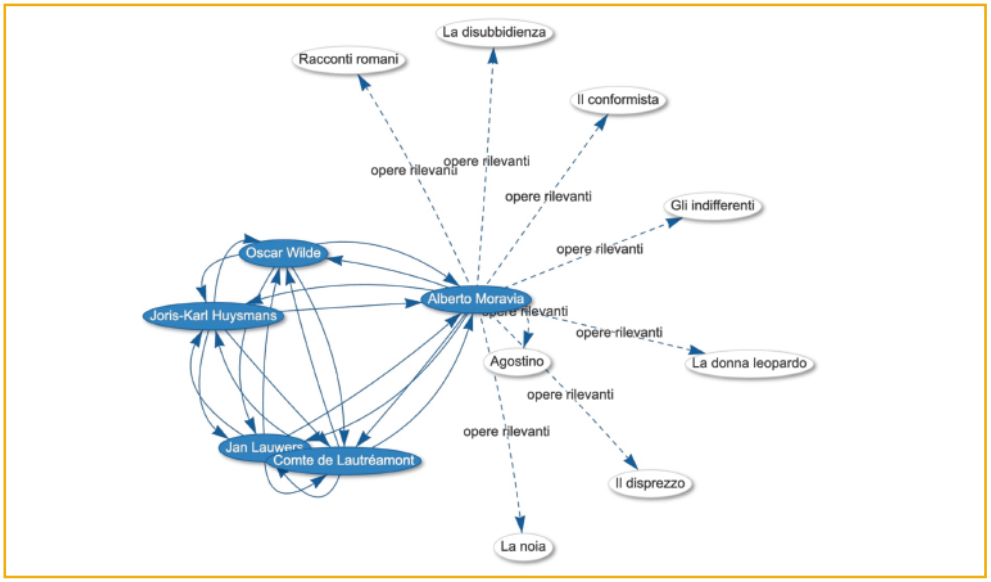

Figura 8. *La rappresentazione su Wikidata dell'entità Alberto Moravia con gli autori correlati e le opere correlate*

[31] Il progetto ha lo scopo di utilizzare Wikidata come dataset per l'analisi, la visualizzazione e la scoperta di informazioni relative alle riviste italiane di biblioteconomia: <https://www.wikidata.org/wiki/Wikidata:Gruppo_Wikidata_per_Musei,_Archivi_e_Biblioteche/Riviste_di_biblioteconomia>.

[32] Si veda la sezione *Promemoria conflazioni* che registra, a mo' di memo, tutti i casi di omonimia ambigui già risolti.

[33] https://query.wikidata.org/.

[34] https://w.wiki/6hvh.

[35] Di riuso di dati e collaborazione fruttuoso tra progetti diversi ho parlato con Claudio Forziati in Claudio Forziati — Tiziana Possemato, *Riuso, interoperabilità, influenza: la cooperazione virtuosa tra i progetti SHARE e Wikidata*, Milano: Editrice Bibliografica, 2019, <http://eprints.rclis.org/34350/>.

Nella Fig. 9 è riportato un altro esempio di rilevazione di anomalie attraverso la registrazione di una violazione di valore singolo in VIAF per l'entità *Euripides*[36]: ben otto diversi identificativi sono stati rilevati su VIAF per la medesima entità, denunciando così un grave problema di riconciliazione tra i diversi profili riferibili a Euripides[37].

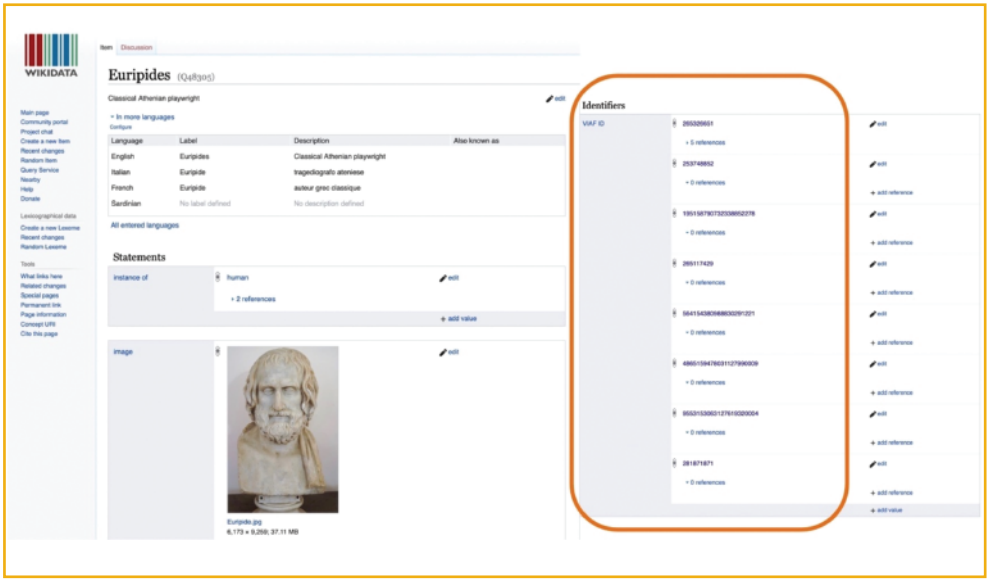

Figura 9. *Wikidata registra una violazione di valore singolo in VIAF, ma VIAF non riceve la registrazione.*

Una riflessione necessaria, dunque, riguarda la capacità delle comunità di aggiornare i propri dati e di comunicare e condividere gli aggiornamenti in forme efficaci[38]. Parte della comunità del web sta interrogandosi su questo aspetto così rilevante per la qualità dell'informazione offerta e la sostenibilità dei progetti. Ai metodi di aggiornamento tradizionali[39] si affiancano adesso protocolli più puntuali, nati nel contesto del web semantico e per questo già naturalmente attenti a come comunicare aggiornamenti non su insiemi di record bibliografici ma sulle singole triple costituenti un dataset in RDF[40]. Interessante in questa direzione è l'iniziativa internazionale del gruppo di lavoro costituito all'interno della comunità LD4P[41] focalizzato specificatamente sulle pratiche di scambio di dati di autorità in ambito LOD. Il gruppo

[36] https://www.wikidata.org/wiki/Q48305.

[37] Questo dato è stato rilevato il 14 maggio 2023. Lo stesso test, fatto nel giugno 2022, aveva rilevato cinque violazioni di valore singolo a fronte delle otto rilevate quasi un anno dopo.

[38] «Unfortunately, there is no automatic reciprocity between VIAF and Wikidata: when a Wikidata item gets a link to a VIAF cluster, VIAF does not have an automated way to add a reciprocal link to the Wikidata item. Likewise, when a VIAF cluster gets a link to a Wikidata item, Wikidata has no automatic way to add a reciprocal link to the VIAF cluster». Carlo Bianchini — Stefano Bargioni — Camillo Pellizzari di San Girolamo, *Beyond VIAF. Wikidata as a Complementary Tool for Authority Control in Libraries*, «Information Technology and Libraries» 40 (giugno 2021), p. 1–31.

[39] Il riferimento è alla pubblicazione di dump periodici di dataset o a protocolli quali l'OAI-PMH per la ricezione degli aggiornamenti.

[40] Il riferimento è, per esempio, a protocolli quali ATOM Feeds: <https://datatracker.ietf.org/doc/html/rfc4287> oppure alle raccomandazioni del W3C Activity Streams <https://www.w3.org/TR/activitystreams-core/>.

[41] *Linked Data for Production* è il risultato della collaborazione tra sei istituzioni (Columbia, Cornell, Harvard, Library of Congress, Princeton e Stanford) per supportare la transizione dei flussi di lavoro di produzione dei servizi tecnici delle biblioteche dai formati di dati tradizionali (MARC) a quelli basati su LOD.

di lavoro[42] si è concentrato sulla definizione di buone pratiche per le agenzie bibliografiche e gli istituti culturali che intendano condividere i propri dati, con particolare attenzione alle diverse tipologie di aggiornamenti, alle modalità e agli strumenti per garantire la comunicazione tra i diversi attori[43]. Uno dei risultati più significativi del gruppo è stata la pubblicazione di raccomandazioni (ancora in corso di revisione) per definire API che forniscano un tracciamento adeguato delle modifiche ai metadati delle entità durante l'intero ciclo di vita delle stesse. Le raccomandazioni sono basate proprio sulle specifiche definite dall'Activity Streams 2.0 e sono destinate a produttori e consumatori di linked data, con lo scopo di facilitare lo scambio e garantire l'interoperabilità tra istituzioni[44].

Conclusioni

Una comunità collaborativa, se ben organizzata, può produrre un'onda gravitazionale che si espande e incontra altre comunità, con altre energie da condividere. Tutto questo, purché si definiscano alcuni elementi importanti relativi all'interoperabilità e al riuso: l'elemento dell'*apertura* delle fonti è un punto di partenza, insieme all'attenzione alla *qualità* del dato e all'utilizzo di *standard e protocolli* che facilitino la condivisione. In questa ottica, diventa fondamentale favorire e incoraggiare riflessioni su come ottimizzare il dialogo tra poli che condividono dati e servizi, per evitare che quell'onda gravitazionale propaghi la ricchezza ma anche le criticità e le anomalie delle fonti utilizzate. Il web, considerato una vetrina per pubblicare i propri dati e renderli disponibili, deve diventare un terreno di confronto attivo e di reale interscambio di dati ed esperienze tra istituzioni. I progetti di pubblicazione di LOD non possono costruire nuovi silos informativi, ma creare occasioni per confrontarsi su temi quali la manutenzione costante, la sostenibilità a lungo termine dei progetti, l'aggiornamento continuo e la comunicazione tra istituzioni. L'amplificazione che l'esposizione nel web offre, anche in termini di problemi e criticità irrisolte, non deve spaventare. Deve, invece, diventare occasione di confronto e di continuo miglioramento e di apertura delle procedure di gestione dei dati e dei servizi a essi collegati. L'alternativa, pericolosa quando pensata nella vastità del web, è l'obsolescenza e la inutilizzabilità delle informazioni prodotte.

The usage of the sources available on the web constitutes an enormous opportunity in the conversion processes of bibliographic catalogs in linked open data: the entity resolution processes, to identify objects in the real world, and the data enrichment mechanisms, to enhance the end user's search experience, makes it essential to use other sources other than their own data. But not everything available on the web is really usable: factors related to data quality and interoperability can complicate or even prevent the reciprocal exchange of information data. This analysis presents some cases of usage of external sources for the processes mentioned above, highlighting the critical elements that are more easily encountered in the phases of data reuse and the possible hypotheses of solution.

[42] https://wiki.lyrasis.org/display/LD4P3/WP2%3A+Production+LD+Authority+Support.

[43] «This working group focuses on best practices in support of authorities that want to share their metadata as linked data. In this charter, the group will explore and make recommendations for change management of authoritative data. We will define the common types of changes, determine the format for conveying changes to downstream consumers, and specify tooling options for publishing changes that allow for cached data to be synced and updated. This work is specifically looking to provide recommendations that allow for the creation of a common toolset that works across authorities supported by different institutions».
<https://wiki.lyrasis.org/display/LD4P3/Charter+2+-+Best+Practices+for+Authoritative+Data+Working+Group>.

[44] https://ld4.github.io/entity_metadata_management/api/0.1/recommendations.html#objectives-and-scope.

L'ultima consultazione dei siti web è avvenuta nel mese di dicembre 2023