JLIS.it 12, 3 (September 2021)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12703

JLIS.it

Linked Open Data native cataloguing and archival description

Marilena Daquino[(a)]

a) Università di Bologna, https://orcid.org/0000-0002-1113-7550

Contact: Marilena Daquino, marilena.daquino2@unibo.it
Received: 3 January 2021; **Accepted:** 23 June 2021; **First Published:** 15 September 2021

ABSTRACT

In the last years cultural heritage institutions have radically changed the way they publish their data. Publishing Linked Open Data (LOD) offers many advantages, in terms of innovation, visibility, and engagement with patrons. New data are served along with legacy services and data, via dedicated interfaces that allow developers and Digital Humanists to access specialised information. However, Linked data are living entities that change over time and require expensive curatorial activities, and should not be misaligned with respect to original data. To cope with this problem, several LOD-native cataloguing systems have been created. In this article an overview of current projects for LOD-native cataloguing is provided. Projects and systems are analysed with respect to related problems and benefits.

KEYWORDS

Linked Open Data; Library Linked data; Semantic Web; Crowdsourcing.

CITATION

Daquino, M. "Linked Open Data native cataloguing and archival description." *JLIS.it* 12, 3 (September 2021): 91–104. DOI: 10.4403/jlis.it-12703.

JLIS.it 12, 3 (September 2021)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12703

Introduzione

Nell'ultimo decennio numerose realtà disciplinari e professionali hanno rivisitato i metodi con cui pubblicare i propri dati al fine di supportare la ricerca e attrarre nuovi utenti. Queste trasformazioni, dettate da cambiamenti sociali e tecnologici, sono particolarmente incoraggiate – se non *de facto* imposte – dalla incalzante capacità di accesso e indicizzazione di enormi moli di dati da parte dei motori di ricerca. Questi ultimi influenzano i processi epistemologici degli utenti del web, sempre più governati da un accesso centralizzato a informazioni eterogenee e capaci di supportare indagini interdisciplinari sempre più precise e semanticamente rilevanti. Di conseguenza, punti di accesso a fonti di dati specialistici risultano penalizzati a vantaggio di aggregatori su ampia scala.

Biblioteche, musei e archivi hanno visto tali cambiamenti come una possibilità di rinnovamento e una chiamata alle armi. Da un lato, la possibilità di uscire dai silos della propria conoscenza e delle tradizioni professionali ha generato scenari di collaborazione interdisciplinare di lungo termine – etichettate sinteticamente con GLAM (tr. Galleries, Libraries, Archives and Museums). Dall'altro lato, l'urgenza di emergere dal *mare magnum* del Web e recuperare il secolare riconoscimento di intermediari tra l'utente e il patrimonio culturale, ha portato le istituzioni ad un momento di autocoscienza e di critica dei metodi con cui tale patrimonio informativo è gestito e riportato al cittadino sotto forma di dati. È noto come nel dominio librario e archivistico i dati siano spesso creati e conservati in sistemi gestionali proprietari che non consentono il riutilizzo dei dati nemmeno da parte degli stessi, espropriati, creatori e pertanto, questi non vengono indicizzati dai motori di ricerca e non risultano immediatamente accessibili agli utenti.

In tale contesto, la pubblicazione di Linked Open Data e l'utilizzo di concettualizzazioni (ontologie) comuni per organizzare la conoscenza sono divenuti tasselli indispensabili del rinnovamento. Per Linked Open Data si intende una modalità di pubblicazione di dati strutturati, elaborabili dalle macchine, inclusivi di associazioni tra concetti (astrazioni di entità reali) identificati univocamente e in maniera persistente tramite meccanismi e protocolli del Web e del *Semantic Web* (Bizer Heath e Berners-Lee 2011). Il valore aggiunto di queste tecnologie risiede nella possibilità di realizzare applicazioni in grado di fornire risposte semanticamente precise a domande complesse, sfruttando le relazioni tra dati provenienti da fonti di natura diversa – siano queste autorevoli, socialmente riconosciute come valide o servizi commerciali. Il livello di maturità di queste tecnologie è tale da renderle trainanti nello sviluppo dei motori di ricerca e pertanto i Linked Open Data sono fonti privilegiate per l'indicizzazione dei contenuti sul Web.

L'investigazione sulla fattibilità e l'applicazione di tecnologie per il Semantic Web nell'ambito GLAM ha dato vita ad una folta letteratura sull'argomento (Van Hooland e Verborgh 2014; Coyle 2013; Hyvönen 2012). Gli standard di metadati sono stati analizzati, rivisitati e trasformati in ontologie di dominio (Carriero *et al.* 2019; Mazzini e Ricci 2011; Hillmann *et al.* 2010; Doerr *et al.* 2008), così da non perdere il controllo sulla semantica del dominio stesso e fornire linee guida all'interno delle comunità professionali. Nuove comunità nazionali e internazionali, interdisciplinari,[1] vengono create da esperti di dominio e tecnici così da presiedere l'evoluzione e il passaggio alle nuove tecnologie.

[1] Si vedano: *American Library Association* http://www.ala.org/lita/about/igs/linked/lit-iglld, *Linked Open Data in Libraries Archives and Museums* (LODLAM) http://lodlam.net/, *Linked Data for Libraries* (LD4L) https://www.ld4l.org/, *Linked.art* http://linked.art/.

JLIS.it 12, 3 (September 2021)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12703

Grandi progetti di conversione vengono promossi dagli istituti che detengono i mezzi economici e tecnologici per dettare gli sviluppi.[2] Come risultato, ad oggi numerosi *dataset* sono disponibili online per la consultazione e il riutilizzo aperto, insieme a servizi specialistici per l'interrogazione e per il miglioramento dell'esperienza dell'utente.

Nonostante la lungimiranza di molte istituzioni e processi ormai quasi decennali di trasformazione dei dati, il connubio GLAM e Linked Open Data è ancora in una fase di maturità sociotecnica infantile. Nobili tentativi di mettere a disposizione dati per il riuso da parte di una non ben specificata super-competente audience di umanisti, bibliotecari e archivisti, si risolvono in momenti *una tantum* di generosa liberazione dei beni comuni, senza offrire una visione strategica per la gestione della conoscenza aperta. Ad esempio: come gestire la collaborazione tra istituzioni dei beni culturali che partecipano attivamente alla produzione di Linked Data? Come garantire l'aggiornamento, la manutenzione e la preservazione dei nuovi dati? Quale rapporto esiste tra questa nuova versione e i dati originari, ancora preservati e manutenuti negli originali sistemi proprietari? I dati sono oggetti vivi, soggetti a mutamenti (arricchimento, modifica, aggiornamento), e urge un'ampia collaborazione tra professionisti del patrimonio culturale e ricercatori per la definizione di strategie di lungo termine (Tomasi e Daquino 2015).

Il paradigma dei Linked Open Data ha richiesto di rivedere sistematicamente le informazioni riportate nelle descrizioni archivistiche e nelle schede catalografiche (destrutturando e ristrutturando tipologia, granularità e precisione), così da superare gli schemi documento-centrici della descrizione con approcci data-centrici, che valorizzano le relazioni con il contesto (storico, artistico, culturale, organizzativo, etc.) come *first-class citizen* oltre al solo oggetto documentato. Attualmente queste visioni convivono in due mondi separati. Da un lato troviamo i dati ancora gestiti in sistemi dinamici per la catalogazione, allorché chiusi e non modificabili, e dall'altro i dati organizzati secondo brillanti e sofisticati modelli concettuali, che risiedono altrove e subiscono altri processi, scollati dai primi.

Sistematizzare la relazione tra i Linked Open Data e i dati originari è un obiettivo di ricerca che pone non poche sfide, di natura tecnica e sociale, alle quali i soggetti interessati hanno risposto con soluzioni diverse. Molti hanno deciso di sfruttare i nuovi dati per integrare servizi esistenti, ad esempio fornendo percorsi di ricerca alternativi nei cataloghi online o in applicazioni dedicate all'esplorazione, oppure offrendo strumenti per la disambiguazione di entità significative, ad esempio con collegamenti a record d'autorità di Virtual International Authority File (VIAF)[3] o Online Computer Library Center (OCLC)[4]. Altri hanno invece proposto di risolvere tale divergenza a monte, creando sistemi di catalogazione nativi Linked Open Data, così da modificare radicalmente il paradigma della descrizione (Malmsten 2009).

Nel panorama GLAM, i pionieri di sistemi nativi Linked Open Data sono attualmente una minoranza. È invece significativa la popolazione di enti che ha avviato la sperimentazione proponendo progetti di arricchimento dei propri dati (questa volta Linked Open Data) tramite *crowdsourcing*. Ciò fa pensare che ci troviamo in una fase di sperimentazione, in cui gli istituti della cultura si fanno

[2] Per menzionare solo alcuni tra i progetti più noti a livello internazionale, si vedano: *WorldCat* http://www.oclc.org/data/data-sets-services.en.html il catalogo di OCLC, *Open Library* https://openlibrary.org/ il dataset bibliografico di Internet Archive e il subset Linked Data della *British National Bibliography* http://bnb.data.bl.uk/.

[3] http://viaf.org/.

[4] https://www.oclc.org/en/home.html.

JLIS.it 12, 3 (September 2021)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12703

promotori di iniziative di collaborazione per contribuire allo sviluppo di soluzioni native Linked Open Data stabili, riutilizzabili e sostenibili, valutandone benefici e limiti nella prospettiva di un cambiamento collettivo di paradigma. È in quest'ultima direzione che l'analisi proposta in questo articolo volge lo sguardo, esaminando le motivazioni, gli aspetti problematici, i limiti e i benefici che le tecnologie del Semantic Web possono offrire alla catalogazione e la descrizione archivistica.
Alcuni strumenti e progetti di rilievo nazionale e internazionale per la catalogazione e la descrizione archivistica collaborativa vengono descritti, così da chiarire il quadro in cui ci si muove. Non è obiettivo di questo contributo fornire una revisione esaustiva delle esperienze in atto o passate, per le quali si rimanda alla letteratura menzionata nella sezione dedicata. In secondo luogo, vengono esaminate le problematiche di natura socio-tecnica emerse dalla letteratura dei progetti collaborativi nativi Linked Open Data, esplorando requisiti, funzionalità, limiti, lezioni apprese e linee di ricerca future.

Progetti collaborativi nativi Linked Open Data nel dominio archivistico e librario

Con *crowdsourcing* ci si riferisce al contributo di un'audience non specificata, tramite invito aperto, in attività di varia natura proposte da un'istituzione, organizzazione, o azienda (Howe 2006). I casi più noti sono Wikipedia,[5] in cui gli utenti contribuiscono alla redazione di contenuti web di natura enciclopedica, e Amazon Mechanical Turks, la piattaforma americana per il reclutamento e lo svolgimento di mansioni a pagamento – come annotazione, trascrizione, correzione di testi – spesso utilizzata dalle istituzioni per sopperire alla mancanza di volontari specializzati.
Quando le campagne di crowdsourcing sono proposte da istituzioni culturali, diversi modelli di partecipazione pubblica sono messi in campo (Simon 2010; Carletti *et al.* 2015; Andro e Saleh 2017). Questi possono prevedere contributi limitati da parte dell'utente, in cui le attività richieste variano da *social tagging* (attraverso folksonomie o tassonomie definite dalle istituzioni) a trascrizione e correzione di testi e metadati. Possono essere progetti collaborativi nella realizzazione di nuove applicazioni o eventi supervisionati dall'istituzione, che per esempio chiede all'utente di fornire storie sugli oggetti culturali o di contribuire alla collezione con nuovi oggetti. Possono prevedere momenti di co-creazione e co-curatela, in cui l'intero design del progetto (e.g. un'esibizione virtuale) è condiviso dalle parti. Non ultimo, l'istituzione può ospitare progetti realizzati dagli utenti. A tali modelli corrispondono numerosissime iniziative proposte da biblioteche, musei, archivi e soggetti di natura diversa che ambiscono al coinvolgimento attivo di cittadini ed esperti di dominio nella creazione e curatela del patrimonio culturale.[6] A titolo di esempio, si pensi alla campagna del progetto Europeana 1914-1918,[7] dove gli utenti possono contribuire con memorabilia da digitalizzare e

[5] https://www.wikipedia.org/.
[6] Per brevità non si riporta qui la lista dei progetti di crowdsourcing promossi da musei, archivi, biblioteche, per la quale si rimanda il lettore alla letteratura: Oomen e Aroyo 2011; Ridge 2013; Ridge 2014; Blanke Kristel e Romary 2015; Carletti *et al.* 2015; Terras 2016; Andro e Saleh 2017; Koukopoulos *et al.* 2017; Bonacchi *et al.* 2019. Questi lavori offrono una panoramica di progetti, tassonomie di modelli di partecipazione pubblica, tipologie di attività e potenziale per la gestione del patrimonio culturale.
[7] https://www.europeana.eu/en/collections/topic/83-1914-1918.

JLIS.it 12, 3 (September 2021)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12703

popolare l'archivio digitale, o il progetto americano StoryCorps,[8] dove i cittadini statunitensi possono condividere interviste e storie per essere archiviate ad uso delle future generazioni.
Dalla letteratura si evince come il carattere sporadico di molte iniziative sia legato ad un coinvolgimento del pubblico mirato alla sensibilizzazione e disseminazione del patrimonio culturale, avente più una valenza comunicativa che non un obiettivo di effettiva co-curatela del patrimonio culturale. Dalla letteratura emerge anche la difficoltà di integrare i dati risultanti delle campagne all'interno dei sistemi gestionali esistenti e distinguere nettamente il contributo dell'utente da quello del catalogatore o esperto di dominio. I dati ottenuti vengono spesso "sigillati" in applicazioni dedicate all'esplorazione, come esibizioni virtuali in cui i contributi del pubblico fungono da supporto alla navigazione, nella speranza che questi contribuiscano ad una migliore esperienza dell'utente. Ambienti dedicati al contributo degli utenti e l'esplorazione, sebbene separati dai sistemi gestionali catalografici e archivistici, sono stati sviluppati da istituzioni per mantenere attivo il bacino di volontari e restituire un risultato immediato ai loro sforzi. Tra queste meritano menzione *Citizen Archivist Dashboard*,[9] un ambiente collaborativo per il tagging, la trascrizione e la scansione di documenti archivistici promosso da National Archives and Records Administration (NARA) e le campagne promosse da Smithsonian Institution Archives,[10] in cui volontari partecipano alla trascrizione (peer-reviewed) di fonti testuali. I risultati delle trascrizioni alimentano servizi di esplorazione e ricerca full-text nell'ambiente online dedicato alla navigazione delle fonti documentali.
Tra le iniziative promosse, solo un numero esiguo prevede la creazione di Linked Open Data. Ancora meno sono i progetti che prevedono l'intera gestione del processo di creazione, curatela e preservazione dei dati integralmente in Linked Open Data. Tra questi, merita menzione il software Accurator (Dijkshoorn *et al.* 2019), realizzato dal Rijksmuseum di Amsterdam per supportare l'annotazione di immagini di opere d'arte con aspetti iconografici e specialistici (e.g. l'annotazione della specie di animali ritratti nell'opera). Rivolto ad un pubblico di esperti di dominio (e di nicchia), l'inserimento di termini è supervisionato dall'utilizzo di termini provenienti da vocabolari controllati e tassonomie, come Iconclass[11] e il Getty Art & Architecture thesaurus (Baca e Gill 2015), nella loro versione Linked Open Data. Le annotazioni vengono integrate alla base di conoscenza in Linked Open Data del museo, a sua volta generata a partire dai dati catalografici custoditi in un sistema gestionale non nativo Linked Open Data.
Un progetto innovativo, anche questo dedicato alla collaborazione tra esperti di dominio, è *Listening Experience Database (LED)* (Adamou *et al.* 2019). Il progetto prevede la collaborazione di ricercatori di ambiti disciplinari affini alla storia della musica e la musicologia per la realizzazione di un database di fonti bibliografiche che riportano esperienze di ascolto musicale. L'annotazione degli estratti testuali avviene attraverso un ambiente online interamente basato su tecnologie del Semantic Web, che consente l'inserimento dati, la gestione editoriale dei contenuti (peer-reviewed) e l'esplorazione. Durante l'inserimento dei dati l'utente ottiene dei suggerimenti per la disambiguazione a partire da

[8] https://storycorps.org/.
[9] http://archives.gov/citizen-archivist/.
[10] https://transcription.si.edu/.
[11] http://iconclass.org/help/lod.

JLIS.it 12, 3 (September 2021)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12703

record della British National Bibliography[12] e DBpedia.[13] Ogni nuovo record viene gestito come un oggetto (un grafo) a sé stante che nel corso del processo editoriale viene annotato con la provenienza delle informazioni (e.g. il catalogatore) e lo stadio di peer-review, prima di poter essere pubblicato e andare a popolare l'ambiente online per l'esplorazione.

Simile al precedente è il progetto ARTchives,[14] un recente progetto collaborativo internazionale di censimento di fondi archivistici creati da storici dell'arte. Il sistema di catalogazione consente l'inserimento dati, la gestione del processo editoriale e la visualizzazione del catalogo di fondi archivistici per finalità di ricerca storiografica. Il sistema di inserimento dati e l'applicazione per l'esplorazione dei dati sono integrati e sono nativamente basati su tecnologie del Semantic Web. L'inserimento dati è supportato nella descrizione tramite una serie di funzionalità di (1) deduplicazione, autocompletamento e riconciliazione automatica a Wikidata, (2) estrazione di conoscenza e Named Entity Recognition in lunghi testi descrittivi e (3) utilizzo di vocabolari controllati provenienti da Wikidata – a loro volta riconciliati con i più importanti thesauri di dominio (e.g. Getty AAT per descrizione degli aspetti artistici). Come in LED, il processo editoriale è rappresentato attraverso grafi, annotati con la *provenance* e lo stadio del record nel processo editoriale. Un editorial board revisiona i record prima di pubblicarli. Questi aggiornano immediatamente il catalogo online, esplorabile ai fini della ricerca storiografica. I dati sono allineati con Wikidata a livello di schema e a livello di entità, in previsione di un export in Wikidata, così da garantirne visibilità, riuso e preservazione anche dopo la fine del progetto.

Progetti come LED e ARTchives presentano alcuni limiti immediatamente percepibili. La tempistica di sviluppo dell'ambiente è legata alle esigenze di un progetto, la generalizzazione del sistema per la descrizione delle risorse eterogenee non è sempre data, così come l'immediatezza di utilizzo del sistema di catalogazione in nuovi progetti non è immediata e la manutenzione del servizio dipende unicamente dall'istituzione promotrice. Allo stesso tempo i vantaggi delle soluzioni native Linked Open Data sono significativi. Gli strumenti a supporto del catalogatore basati su tali tecnologie forniscono strumenti sempre più adeguati ad assicurare qualità dei dati, omogeneità nella descrizione e diversità nelle forme di espressione del catalogatore (in linguaggio naturale) senza rinunciare a dati strutturati.

Alcuni dei limiti sopra menzionati sono stati oggetto dello sforzo lungimirante di piattaforme per la catalogazione che emergono in seno alla comunità internazionale di Digital Humanities. Tra questi troviamo i prodotti nati dall'iniziativa aziendale di regesta.exe per la descrizione archivistica, che hanno contribuito alla realizzazione dei modelli ontologici che riproducono gli standard archivistici e i corrispondenti schemi XML (Aprea 2018). Tramite soluzioni come XDams,[15] l'archivista può inserire i dati sotto forma di descrizioni testuali, ma il sistema non fornisce servizi per la disambiguazione e la riconciliazione tra le risorse menzionate, né per la trasformazione in Linked Open Data. Un successivo progetto della stessa azienda, Bygle[16], è stato proposto per implementare

[12] https://bnb.data.bl.uk/.
[13] https://wiki.dbpedia.org/.
[14] http://artchives.fondazionezeri.unibo.it/.
[15] https://www.xdams.org/.
[16] https://www.regesta.com/2015/01/28/nasce-bygle/.

JLIS.it 12, 3 (September 2021)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12703

il paradigma dei Linked Data sin dalla creazione e gestione dei dati. Al momento il software non è più manutenuto.

Rilevante per la comunità GLAM è OmekaS,[17] un *content management system* per la catalogazione di oggetti eterogenei e la creazione di mostre virtuali. L'ambiente consente di organizzare la descrizione delle risorse tramite ontologie (esistenti o definite dall'utente) e prescrivere l'utilizzo di vocabolari controllati (come i succitati Getty vocabularies, VIAF, Library of Congress Subject Headings[18]). I dati, gestiti in un sistema non nativo Linked Open Data, sono tuttavia interrogabili come tali tramite dedicate *Application Programming Interfaces* (API) che ne forniscono una vista "semantizzata" (grazie al formato JSON-LD, che consente a sviluppatori del semantic web di sfruttare l'espressività di RDF e a sviluppatori tradizionali di riferirsi a dati nel formato JSON). Utenti con ruoli diversi possono contribuire al processo di catalogazione, sebbene l'ambiente non offra strumenti a supporto del processo editoriale (come avviene invece in LED o ARTchives).

Soluzioni nate al di fuori del dominio archivistico e librario includono interessanti approcci, come Semantic MediaWiki, software aperto attualmente alla base di Wikipedia, o Wikibase, l'ambiente alla base di Wikidata.[19] Gli aspetti collaborativi e dedicati al processo editoriale sono qui preponderanti, mentre il ciclo di vita dei dati è gestito in ambienti non nativi Linked Open Data. Non di meno, i dati possono essere interrogati utilizzando i linguaggi di interrogazione standard (SPARQL) ed essere esportati in RDF. Queste soluzioni richiedono una minore curva di apprendimento da parte dei catalogatori, essendo sviluppati per utenti con diverse competenze. Nonostante ciò, di difficile applicabilità qui sono proprio i principi alla base dei Linked Open Data, ovvero la possibilità di integrare dati provenienti da fonti esterne (o interne, creando collegamenti tra record creati all'interno dello stesso progetto) durante la catalogazione stessa. Sebbene sia possibile riferirsi a tali fonti esterne (mediante URI identificativi) e interne (mediante ID locali attribuiti ai record), strumenti per il supporto e la velocizzazione della catalogazione sono assenti e richiedono al catalogatore di reperire autonomamente i codici identificativi. Si pensi ad esempio alla modifica di una entità (*item*) in Wikidata. L'utente dovrà innanzitutto scoprire (letteralmente) nella documentazione del progetto Wikidata quale identificativo è associato ad un campo di inserimento dati (*proprietà*), previo scoprire quali proprietà esistono e quali possono applicarsi al tipo di entità in questione. In secondo luogo, dovrà scansionare manualmente la base di conoscenza per reperire l'identificativo dell'item che vuole associare come valore della proprietà. Non esiste infatti un meccanismo di propedeuticità (basato su template come avviene in Omeka) e di suggerimenti (come in LEDo ARTchives) che permetta all'utente di contribuire in maniera intuitiva e facilitata.

Problemi sociotecnici nei progetti collaborativi nativi Linked Open Data

La breve panoramica di progetti e tecnologie utilizzate da istituti culturali evidenzia come non esistano ad oggi soluzioni in grado di soddisfare ogni problematica, siano queste di natura sociale o tecnologica. In particolare, l'obiettivo delle campagne di crowdsourcing influenza e determina

[17] https://omeka.org/s/.
[18] https://id.loc.gov/authorities.
[19] Si rimanda il lettore a questo blog post per una dettagliata descrizione delle differenze tra le due piattaforme https://professional.wiki/en/articles/managing-data-in-mediawiki#.

JLIS.it 12, 3 (September 2021)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12703

requisiti e problemi. A titolo di esempio, l'avvicinamento del pubblico ai beni culturali tramite nuove forme di esplorazione e coinvolgimento (per esempio attraverso forme di *gamification*) richiede un intervento tecnologico differente rispetto a quanto necessario per un sistematico contributo dell'utente nell'arricchimento e l'aggiunta di nuove risorse culturali – il quale può variare dalla richiesta di minime competenze tecniche da parte dell'utente per il caricamento di un'immagine con pochi metadati associati, ad una conoscenza più capillare degli schemi e standard di catalogazione per un contributo sostanziale nella co-curatela. Quest'ultimo scenario è sicuramente quello che pone più sfide ed è quello indispensabile al sostentamento di un cambiamento radicale nel paradigma nella catalogazione e descrizione archivistica.

In questa sezione vengono esaminati alcuni aspetti socio-tecnici emersi dall'analisi delle soluzioni tecnologiche descritte nella sezione precedente, così da fornire una chiave di lettura e di confronto a supporto della valutazione di futuri progetti nativi Linked Data in ambito archivistico, biblioteconomico e museale. I punti emersi sottolineano il punto di vista del catalogatore e dello sviluppatore del sistema di catalogazione, mentre il punto di vista dell'utente finale dei dati non è preso in considerazione se non in senso lato, poiché è solo parzialmente affetto da considerazioni di carattere strutturale e gestionale. Le problematiche sociotecniche che sottintendono sistemi collaborativi di catalogazione e descrizione archivistica nativi Linked Open Data possono essere sommariamente raggruppati come segue.

La condivisione della concettualizzazione

L'organizzazione della conoscenza è l'aspetto principale che vede coinvolti catalogatori e sviluppatori nel design del sistema. Se da un lato il catalogatore l'organizzazione dei dati deve rispondere ad esigenze professionali e comunitarie (come l'adesione a standard di dominio, l'utilizzo di vocabolari controllati o metodi per l'identificazione delle risorse), dal lato dello sviluppatore, l'organizzazione deve essere funzionale alla gestione e manutenzione dei dati nel tempo. Problemi di nomenclatura, norme redazionali, procedimenti qualitativi per il trattamento dei dati rientrano in quei requisiti non funzionali (ovvero a cui il sistema non deve saper ovviare dal punto di vista tecnico) che dipendono di volta in volta dal progetto in essere e dalle pratiche della comunità. Problemi relativi ad aspetti di carattere ontologico devono essere risolti a monte dal sistema, come ad esempio restrizioni sul tipo di dati da inserire (e.g. date, descrizioni testuali libere, identificativi di risorse, restrizioni sulla classe di risorse), adesione a modelli esistenti (e.g. tramite l'importazione di modelli ontologici esistenti) e possibilità di esportazione in formati e modelli di dati alternativi (e.g. in oggetti JSON, serializzazioni RDF, tabelle CSV).

Il sistema dovrebbe essere preferibilmente agnostico nel merito dei requisiti di dominio (i.e. gli specifici standard catalografici), così da garantire un utilizzo flessibile di modelli alternativi e la possibilità di condividere tali modelli assieme ai dati che vengono realizzati. Tra i servizi sopra descritti, OmekaS offre diverse funzionalità importanti in questo senso. Consente infatti di esportare separatamente i dati in JSON-LD, le ontologie e gli specifici template utilizzati per la catalogazione - contenenti il sottoinsieme di proprietà/campi (e.g. autore, anno, editore), tipi di valori (e.g. un termine da vocabolario controllato, una stringa, una risorsa identificata da un URI) utilizzati per descrivere una certa risorsa (e.g. un libro). Al momento però OmekaS non consente di restringere la classe a cui

JLIS.it 12, 3 (September 2021)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12703

le informazioni inserite devono appartenere (e.g. specificare che il valore del campo *autore* deve essere un URI che identifica una risorsa appartenente alla classe *Persona*). Tali aspetti di alto livello (a livello di schema) sono propedeutici per garantire la qualità dei contenuti (a livello di dato).

La qualità e l'autorialità dei contenuti

È questo l'aspetto che forse più preoccupa il catalogatore, essendo un progetto collaborativo, per sua natura, un luogo popolato da contributi di utenti con competenze diverse. Metodi di costrizione e di omogeneizzazione della catalogazione hanno il duplice effetto di restringere la variantistica di casi eccezionali (con soluzioni ad hoc che non sempre sono sostenibili nel tempo) e di proporre una semplificazione della realtà che tende a creare a sua volta nuove eccezioni, dovute ad errori di interpretazione da parte dell'utente.

Si pensi ad esempio alla varietà di soggetti che possono essere attribuiti ad un'opera d'arte o una fotografia. Questi potranno concernere il soggetto primario dell'opera – quello riconoscibile immediatamente dall'occhio umano sulla base della propria esperienza, come una persona, un oggetto inanimato, un luogo familiare –, un soggetto secondario – ad esempio un soggetto iconografico come "l'Ultima cena", che richiede esperienza nel campo della storia dei tipi –, oppure idee e concetti tipici del tempo e della cultura a cui l'opera appartiene (un movimento artistico, una tendenza stilistica). Questi aspetti, classificabili nitidamente da un esperto di dominio, possono essere solo intuitivamente percepibili da un osservatore meno esperto, che può invece voler contribuire con la propria esperienza personale (osservazioni sul colore, emozioni, memorie e storie). Le limitazioni imposte dai vocabolari controllati, per quanto esaustivi in un dato dominio, non sempre sono complementari con altri schemi (e.g. come combinare termini provenienti dai vocabolari del Getty AAT con termini meno specialistici provenienti ad esempio da Wikidata?) e spesso prevengono l'utente dalla possibilità di esprimere appieno la propria esperienza.

L'utilizzo di folksonomie è stato proposto nel tempo per ovviare a tale ragione, creando però l'effetto collaterale di separare inesorabilmente il lessico degli esperti (basato su tassonomie iper-specialistiche e controllate) dal lessico dell'osservatore (libero, non strutturato e inconciliabile con quello dell'esperto). In tal senso, un sistema di catalogazione collaborativo dovrebbe supportare il catalogatore esperto e meno esperto nella descrizione utilizzando un lessico comune ad entrambi, senza rinunciare ad una strutturazione sofisticata del sapere sottesa a tale lessico. A tal fine, l'utilizzo di Linked Open Data e di basi di conoscenza curate sia da utenti generici, sia da esperti di dominio per garantire l'allineamento semantico tra i lessici, diviene un aspetto fondamentale per la realizzazione di progetti collaborativi. Un esempio utile è senza dubbio Wikidata, utilizzato dal progetto ARTchives per definire un ampio vocabolario multilingue di termini allineati alla terminologia di dominio archivistico, librario e museale (e.g. con link agli identificativi in VIAF, LOC, Getty).

Non sempre però la natura didascalica della descrizione archivistica (fatta di lunghi testi introduttivi, note storiche, contestualizzanti o specializzanti) si concilia con la necessità di avere dati strutturati per il loro riuso e interrogazione puntuale. Riprendendo il caso della soggettazione delle opere (siano queste fotografie, documenti archivistici, artefatti museali), nei sistemi di annotazione collaborativa viene chiesto all'utente di compiere un enorme sforzo sintetico, riducendo la propria esperienza e la

JLIS.it 12, 3 (September 2021)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12703

propria interpretazione a categorie considerate accettabili – ed in qualche modo riconducibili ad altre opere – definibili in poche, selezionate, parole. Allo stesso tempo, al catalogatore o archivista viene chiesto invece di fornire lunghe descrizioni testuali, utili a contestualizzare l'oggetto e fornire chiavi di lettura, le quali solo raramente vengono a loro volta destrutturate in categorie sintetiche e concetti. Quando questo processo di destrutturazione avviene, esso avviene richiedendo al catalogatore di duplicare gli sforzi, ergo di esprimere prima liberamente il messaggio in lunghe descrizioni e di inserire poi solo alcuni dei termini chiave in un secondo campo, quest'ultimo elaborabile dalla macchina in ricerche terminologiche e aggregazioni. Compito dei sistemi collaborativi di catalogazione e descrizione archivistica deve essere quello di supportare diversi metodi di espressione, garantendo il massimo *comfort* nella descrizione, per esempio associando automaticamente termini e categorie condivisibili a lunghi testi in linguaggio naturale. Lo stato dell'arte nei software per Named Entity Recognition e Natural Language Processing consente infatti di supportare il catalogare/utente nella riconciliazione di entità con vocabolari e basi di conoscenza Linked Open Data (come visto nel caso sopra menzionato di LED e ARTchives), sia di estrapolare da lunghe descrizioni testuali le entità rilevanti che caratterizzano il discorso, evitando così da un lato la duplicazione degli sforzi (chiedendo di fornire sia descrizioni testuali sia *keyword*) e consentendo dall'altro la massima espressione della propria conoscenza mediante linguaggio naturale – senza dover rinunciare alla strutturazione dei dati. I Linked Open Data sono lontani dall'essere una panacea universale ai problemi di qualità e non prevengono dal necessario processo di controllo e revisione da parte di esperti sui contenuti generati da utenti di provenienza diversa. La natura dei progetti collaborativi si intende basata su un rapporto di fiducia tra collaboratori al fine di produrre un bene comune. Pertanto un sistema deve saper rispondere ad esigenze di "democratizzazione" del sapere, concedendo a tutti i contraenti uguali diritti nel poter esprimere la propria conoscenza.
Non di meno, un annoso problema nei sistemi di catalogazione collaborativa è la definizione dei ruoli, dei privilegi e della possibilità di intervenire sui contenuti creati da altri, eventualmente con la possibilità di censurare l'accesso a determinate risorse da parte di classi di utenti ritenuti occasionali o non sufficientemente preparati per modificare le descrizioni. In altri termini, mentre la comunità più ampia invoca la necessità di sistemi sofisticati per la catalogazione, così da poter esprimere le più complesse esigenze descrittive, spesso i contraenti non ritengono i collaboratori in grado di comprenderne le potenzialità di un tale strumento senza un controllo centralizzato. Pertanto, la necessità di un processo di curatela - che richiede un controllo editoriale da parte di membri scelti della comunità - rimane un problema irrisolvibile tramite la sola tecnologia. La definizione delle *policy* per la qualità dei dati, l'accesso e il controllo sui dati realizzati rimane tra gli aspetti più dispendiosi sottintesi alla creazione di sistemi collaborativi.

La manutenzione del sistema e l'aggiornamento dei dati

La liberazione della conoscenza tramite le tecnologie del Semantic Web non previene dalla necessità di sviluppare meccanismi maturi per l'aggiornamento e l'integrazione dei dati originali. Nel contesto di progetti collaborativi, questi ultimi restano custoditi nel gestionale in cui i catalogatori operano quotidianamente, mentre quelli creati collaborativamente, serviti come Linked Open Data, si trovano spesso in servizi dedicati, scollati dai precedenti. Le inevitabili difficoltà nel mantenere non uno, ma

JLIS.it 12, 3 (September 2021)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12703

due, distinti e non sempre dialoganti, sistemi informativi è forse il fattore principale nella scelta di persistere nell'utilizzo o di abbandonare le tecnologie del Semantic Web.
Essendo i dati soggetti a mutamenti nel tempo - siano questi modifiche, revisioni, preparazione per la pubblicazione finale – anche il processo editoriale in senso stretto deve essere rappresentabile mediante gli stessi modelli e protocolli del Semantic Web. Come anticipato nella descrizione di LED e ARTchives, il processo editoriale può a sua volta essere rappresentato sotto forma di grafo, o grafi, i quali vengono iterativamente annotati con informazioni aggiornate sugli agenti che intervengono nell'editing (autori, collaboratori, revisori, editori), sugli aspetti diacronici (versioning delle annotazioni nel tempo), e sullo stadio di revisione (modifica, revisione, pubblicazione). La rappresentazione formale del processo editoriale è a discrezione dello sviluppatore, che può utilizzare vocabolari standard come PROV (Lebo *et al.* 2013) o diversi contenitori (*named graphs*) (Carrol *et al.* 2005). PROV è l'ontologia promossa dal W3C per rappresentare processi e azioni svolte da agenti che hanno un effetto (creazione, modifica, eliminazione) su entità di varia natura. Tramite l'utilizzo di grafi multipli è poi possibile distinguere i singoli contributi autoriali (il record creato da un utente, lo stesso record dopo la modifica di un altro utente, il record finale rivisto dagli editori pronto per la pubblicazione) e annotare il loro stadio nel processo editoriale. In questo senso, specifici modelli per relazionare i contenitori, come il data model Nanopublication (Groth Gibson e Velterop 2010), consentono di esprimere formalmente le relazioni tra asserzioni, il contesto di provenienza e il contesto di pubblicazione.

La continuità e sostenibilità nel tempo

Campagne estemporanee in progetti collaborativi rischiano di essere dismesse data l'indisponibilità di un impegno costante da parte dei collaboratori e dalla difficoltà di reclutamento di nuovi collaboratori, nonché a cause della mancata sussistenza economica del progetto. Mentre alcuni modelli di partecipazione pubblica sembrano dimostrare il contrario (si veda il caso esemplare di Wikipedia), i progetti collaborativi specialistici faticano a trovare meccanismi di *engagement* e di premiazione in grado di eguagliare casi esemplari. Alcuni meccanismi virtuosi possono contribuire alla sostenibilità dei progetti. Generalmente, più utenti collaborano ad un progetto, più il progetto è visibile a nuovi utenti, i quali considereranno più facilmente di contribuire ad un progetto di successo. Nei progetti specialistici (si veda il caso degli Smithsonian Institution Archives) l'esplorazione immediata dei risultati del lavoro individuale può fungere da espediente motivazionale, concedendo la percezione di un immediato ritorno dei propri sforzi in un sistema vivo, attento e ricettivo, possibilmente senza innecessari impedimenti editoriali. Tali sistemi accettano un'ammissione di imperfezione, favorendo la costante modifica e il miglioramento incrementale dei contributi autoriali, invece di osteggiare la pubblicazione e la fruizione immediata con un controllo centralizzato sul contenuto.
Al momento i progetti collaborativi di ambito culturale sono spesso etichettati come sperimentazioni e campagne di sensibilizzazione. Poiché tali progetti raramente prevedono l'integrazione dei dati creati collaborativamente nei cataloghi delle istituzioni promotrici per la loro preservazione, un rischio concreto è che assieme alla progettualità vengano accantonati anche i dati e le tecnologie con cui i dati sono stati raccolti, gestiti e pubblicati. Oltre alla sostenibilità dei sistemi e dei progetti collaborativi,

JLIS.it 12, 3 (September 2021)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12703

la sostenibilità del dato è un aspetto critico. La duplicazione e il deposito dei dati in più contenitori digitali può aumentare, da un lato, la probabilità che questi non vengano eliminati definitivamente, e dall'altro, la visibilità dei dati stessi da parte dei motori di ricerca, innescando così processi virtuosi di collaborazione, continuità e preservazione. Non di meno, permane lo scollamento tra le dinamiche che i dati creati collaborativamente (Linked Open Data) e i dati catalografici e archivistici subiscono, al quale le soluzioni esistenti non hanno ancora saputo dare una risposta soddisfacente.

Conclusioni

Dalla letteratura sui progetti collaborativi in ambito archivistico, museale e librario emergono gli sforzi della comunità interdisciplinare (GLAM e Digital Humanities) verso la creazione di soluzioni sostenibili native Linked Open Data per la catalogazione e la descrizione archivistica. In particolare, progetti collaborativi di curatela e campagne di crowdsourcing risultano essere propulsori di innovazione, favorendo il riuso (non impositivo) di vocabolari standard, di metodi per la rappresentazione formale di diverse forme di autorialità e la previsione di forme di disseminazione dei risultati immediate.

La creazione di strumenti a supporto della catalogazione ambisce a garantire espressività e diversità senza rinunciare alla consistenza e alla qualità dei dati. Lo sviluppo, la cura e la disseminazione del patrimonio culturale risultano essere un importante fattore di resilienza dei sistemi tecnologici in quanto promotori di nuove domande di ricerca e di avanzamento (Daga *et al.* 2021). Ne sono esempio gli sviluppi in ambiti di *Conversational Artificial Intelligence* e *Language understanding*, auspicabili bacini di strumenti capaci di supportare il coinvolgimento e la condivisione del sapere.

In ultimo, dalla panoramica dei progetti di piccola-media dimensione nati in seno alla comunità archivistica e delle Digital Humanities, si evince la necessità di rafforzare le esistenti strategie per la sostenibilità dei dati e, indirettamente, di prevedere una maggiore assunzione di responsabilità verso i cittadini, che finanziano queste sperimentazioni e che contribuiscono con la loro conoscenza in progetti collaborativi. La preservazione dei dati – ancor più della preservazione dei sistemi in sé – rimane un fattore socio-tecnico irrisolto nella valutazione delle tecnologie del Semantic Web, dovuto per lo più alla natura problematica dei progetti collaborativi, che raramente prevedono l'integrazione di nuovi dati all'interno dei cataloghi esistenti. L'analisi di alternative non disgiuntive per la conservazione, come il riversamento in depositi (e.g. Zenodo, Figshare) ed in bacini vivi (e.g. Wikidata), può aumentare le chance di preservazione e incoraggiarne l'adozione, il miglioramento e la continuità.

Una maggiore maturità nello sviluppo di tecnologie a supporto della catalogazione prevedrà non solo l'apertura dei dati tramite Linked Open Data – spesso in maniera svilente considerata la conclusione, se non erroneamente l'obiettivo, di progetti digitali in ambito culturale – ma anche la gestione sistematica di meccanismi di riversamento e, se necessaria, di duplicazione dei dati per l'assunzione di responsabilità condivisa nella curatela e nella preservazione.

JLIS.it 12, 3 (September 2021)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12703

Riferimenti bibliografici

Aprea, Giovanni. 2018. "Uno sguardo sugli strumenti digitali applicati agli archivi." *Bibliothecae.it* 7:287–319. DOI: 10.6092/issn.2283-9364/8450.

Adamou, Alessandro, Simon Brown, Helen Barlow, Carlo Allocca, and Mathieu d'Aquin. 2019. "Crowdsourcing Linked Data on listening experiences through reuse and enhancement of library data." *International Journal on Digital Libraries* 20:61–79. DOI: 10.1007/s00799-018-0235-0.

Andro, Mathieu, and Imad Saleh. 2017. "Digital Libraries and Crowdsourcing: A Review." In *Collective Intelligence and Digital Archives: Towards Knowledge Ecosystem*, edited by Szoniecky, Samuel and Nasreddine Bouhaï, 135–162. Wiley.

Baca, Murtha, and Melissa Gill. 2015. "Encoding multilingual knowledge systems in the digital age: the getty vocabularies." *NASKO* 5:41–63. DOI: 10.7152/nasko.v5i1.15179.

Bizer, Christian, Tom Heath, and Tim Berners-Lee. 2011. "Linked data: The story so far." In *Semantic services, interoperability and web applications: emerging concepts*, 205–227. IGI Global.

Blanke, Tobias, Conny Kristel, and Laurent Romary. 2017. "Crowds for clouds: recent trends in humanities research infrastructures." In *Cultural Heritage Infrastructures in Digital Humanities*, edited by Agiatis, Benardou, Erik Champion, Costis Dallas, Lorna M. Hughes, 15. London: Routledge.

Bonacchi, Chiara, Andrew Bevan, Adi Keinan-Schoonbaert, Daniel Pett, and Jennifer Wexler. 2019. "Participation in heritage crowd-sourcing." *Museum Management and Curatorship* 34:166–182. DOI: 10.1080/09647775.2018.1559080.

Carletti, Laura, Gabriella Giannachi, Dominic Price, Derek McAuley, and Steve Benford. 2013. "Digital humanities and crowdsourcing: An exploration." *In Proceedings of the 2013 Museums and the Web Conference, Portland, OR, USA*, 17–20.

Carriero, Valentina *et al.* 2019. ArCo. The Italian cultural heritage knowledge graph. In *International Semantic Web Conference*, Cham: Springer, 36–52.

Carroll, Jeremy J., Christian Bizer, Pat Hayes, and Patrick Stickler. 2005. "Named graphs, provenance and trust." In *Proceedings of the 14th international conference on World Wide Web*, 2005, 613–622.

Coyle, Karen. 2013. "Linked Data: an evolution." *JLIS.it* 4:53. DOI: 10.4403/jlis.it-5443.

Daga, Enrico *et al.* 2021. "Integrating citizen experiences in cultural heritage archives: requirements, state of the art, and challenges." *JOCCH* 14:3 [in pubblicazione].

Doerr, Martin, Chryssoula Bekiari, Patrick LeBoeuf, and Bibliothèque nationale de France. 2008. "FRBRoo, a conceptual model for performing arts." In *2008 Annual Conference of CIDOC, Athens, 2008*, 15–18.

Groth, Paul, Andrew Gibson, and Jan Velterop. 2010. "The anatomy of a nanopublication." *Information Services & Use* 30:51–56. DOI: 0.3233/ISU-2010-0613.

Hillmann, Diane, Karen Coyle, Jon Phipps, and Gordon Dunsire. 2010. "RDA vocabularies: process, outcome, use." *D-Lib magazine* 16:6.

JLIS.it 12, 3 (September 2021)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12703

Howe, Jeff. 2006. "The rise of crowdsourcing." *Wired Magazine* 6:5.

Hyvönen, Eero. 2012. "Publishing and using cultural heritage linked data on the semantic web." *Synthesis Lectures on the Semantic Web: Theory and Technology* 2:1–159.

Lebo, Timothy, Satya Sahoo, Deborah McGuinness, Khalid Belhajjame, James Cheney, David Corsar, Daniel Garijo, Stian Soiland-Reyes, Stephan Zednik, and Jun Zhao. 2013. "Prov-o: The prov ontology." *W3C recommendation*. https://www.w3.org/TR/prov-o/.

Dijkshoorn, Chris, Victor de Boer, Lora Aroyo and Guus Schreiber. 2019. "Accurator: nichesourcing for cultural heritage." In *Human Computation Journal* 6:12–41. DOI: 10.15346/hc.v6i1.91.

Koukopoulos, Zois, Dimitrios Koukopoulos, and Jason J Jung. 2017. "A trustworthy multimedia participatory platform for cultural heritage management in smart city environments." *Multimedia Tools and Applications* 76:25943–25981. DOI: 10.1007/s11042-017-4785-8.

Malmsten, Martin. 2009. "Exposing library data as linked data." *IFLA satellite preconference sponsored by the Information Technology Section Emerging trends in (2009)*.

Mazzini, Silvia, and Francesca Ricci. 2011. "EAC-CPF Ontology and Linked Archival Data." In *Proceedings of the 1st International Workshop on Semantic Digital Archives (SDA 2011)*, 72–81. http://ceur-ws.org/Vol-801/paper6.pdf.

Oomen, Johan, and Lora Aroyo. 2011. "Crowdsourcing in the cultural heritage domain: opportunities and challenges." In *Proceedings of the 5th International Conference on Communities and Technologies*, 138–149. DOI: 10.1145/2103354.2103373.

Ridge, Mia, (ed.). 2014. *Crowdsourcing our cultural heritage*. Ashgate Publishing.

Ridge, Mia. 2013. "From tagging to theorizing: deepening engagement with cultural heritage through crowdsourcing." *Curator: TheMuseum Journal* 56:435–450. DOI: 10.1111/cura.12046.

Simon, Nina. 2010. *The participatory museum*. California: Museum 2.0.

Terras, Melissa. 2016. "Crowdsourcing in the digital humanities." In *A New Companion to Digital Humanities*, edited by Schreibman, S., Siemens, R., and Unsworth, J., 420–439. Wiley-Blackwell.

Tomasi, Francesca and Marilena Daquino. 2015. "The archival domain in a disciplinary-integrated ontological perspective." *JLIS.it* 6:13–38.

Van Hooland, Seth, and Ruben Verborgh. 2014. *Linked Data for Libraries, Archives and Museums: How to clean, link and publish your metadata.* London: Facet Publishing.

Ultimo accesso a tutti le URL: 31 maggio 2021.