From the Holocaust Victims Names to the Description of the Persecution of the European Jews in Nazi Years: the Linked Data Approch and a New Domain Ontology[1]

Laura Brazzo, CDEC – Fondazione Centro di Documentazione Ebraica Contemporanea Milano
Silvia Mazzini, Regesta.exe

Aim of this paper is to present the first case of application of LOD to the Names of the Victims of the Shoah and the related Shoah domain ontology. The authors will present the outcomes reached by the CDEC Foundation and Regesta.exe in the framework of a publication project of archival descriptions and digitized resources concerning the history of the Shoah in Italy.

The authors will give account of the key-phases of the project: reasons behind the adoption of Linked Data technologies; analysis of the starting available data; building up and architecture of the Shoah ontology. The paper will include the analysis of the carried out interlinking test and the critical situations revealed by the test; the planned development of the ontology in order to describe further aspects of the Nazi persecution of the Jews; examples of further fields of application of the Shoah ontology.

Musei e Memoriali della Shoah, così come celebrazioni e commemorazioni della Shoah, trovano nel ricordo dei nomi delle migliaia di vittime dello sterminio nazista, uno dei principali motivi della loro esistenza.

Nel corso degli ultimi decenni, a partire almeno dagli anni '80, parallelamente all'intensificarsi della ricerca storica e storiografica sulla Shoah, si sono moltiplicati anche i luoghi dedicati ai Nomi delle vittime: non più solo Monumenti *ad memoriam* collocati in luoghi geografici significativi per la storia e la memoria della Shoah, ma anche memorial books e, più di recente, memoriali virtuali.

La ricerca sui nomi delle vittime è ancora un campo fluido: grafie dei nomi così come dati anagrafici e biografici sono soggetti a revisioni dovute talvolta alla scoperta di nuove fonti documentarie, talvolta all'intervento diretto di discendenti delle vittime in grado di correggere le informazioni esistenti e più spesso di fornirne di nuove, prima mancanti. Da questo punto di vista il web appare dunque come il luogo più idoneo e congeniale alla pubblicazione dei nomi: sia per la sua intrinseca flessibilità, sia per l'alto numero di visitatori che è in grado raggiungere.

Il principale limite di tutte queste pubblicazioni, incluse quelle sul web, deriva dalla loro dispersione e impossibilità di connessione dei dati disponibili.

Il segnale forte di questo limite è arrivato nel 2010 con l'avvio del progetto europeo EHRI - European Holocaust Research Infrastructure [1, 2]che ha come obiettivo la creazione di una infrastruttura internazionale per gli archivi sulla Shoah, per la condivisione, tramite un unico punto di accesso, delle fonti e delle informazioni per la storia della Shoah.

[1] Questo documento è l'abstract del paper che sarà presentato alla Conferenza Internazionale Digital Humanities di Sydney (giugno 2015)

STATO DELLA RICERCA

Lo stato attuale delle pubblicazioni dei nomi delle vittime della Shoah si caratterizza per la duplicazione dei nomi e l'assenza di collegamenti fra di essi.

La vastità del fenomeno persecutorio nazista, la difficoltà di reperimento di fonti documentarie utili e al contempo la diversità delle fonti di volta in volta disponibili, hanno portato alla costruzione di numerosi elenchi, basati su criteri ogni volta diversi: la nazionalità delle vittime (come è nel caso della Germania o della Grecia o del Belgio o ancora dell'Austria), oppure il luogo di residenza al momento dell'arresto (come nel caso dell'Olanda), il paese da cui è avvenuta la deportazione (come nel caso della Francia o dell'Italia), il campo nazista di internamento (Auschwitz, Mauthausen, etc.).

Lo scopo principalmente commemorativo di queste liste ha fatto sì che per lungo tempo non sia stata presa in sufficiente considerazione da un lato la questione degli spostamenti, ovvero delle migrazioni e fughe di ebrei che hanno caratterizzato tutto il periodo che va dal 1933 al 1945, dall'altro il fenomeno della duplicazione dei dati relativi alle medesime persone. I nomi di ebrei, per esempio, nati in Germania, emigrati in Francia, fuggiti in Italia e deportati dall'Italia in Francia e infine ad Auschwitz, si ritrovano in almeno quattro banche dati diverse (quella tedesca, quella francese, italiana e quella di Yad Vashem) ciascuno associato ad informazioni che variano a seconda delle fonti documentarie utilizzate.

La tecnologia Linked Open Data, da questo punto di vista, sembra offrire importanti opportunità e strumenti per il superamento di tale situazione. Essa consente l'individuazione univoca dei nomi e la riconciliazione dei nomi stessi (ovvero informazioni e documenti relativi a quel nome/persona) provenienti da banche dati diverse presenti nel web, superando migrazioni, estrazione di dati e creazione di nuovi *repositories*.

I LOD PER LA DESCRIZIONE DELLA SHOAH

La Fondazione CDEC di Milano, principale istituto per la storia della Shoah in Italia, ha avviato a partire dalla fine del 2012 un progetto di lungo periodo per la creazione di una Linked Data Digital Library. Primo obiettivo del progetto era la pubblicazione delle risorse d'archivio per la storia e la memoria della Shoah in Italia. A questo scopo, la prima necessità emersa è stata quella di integrare in un unico *repository* descrizioni archivistiche e catalografiche, documenti digitalizzati e banche dati autonome create ciascuna per differenti scopi. La seconda esigenza è stata quella di individuare un elemento univoco di raccordo e quindi di accesso alle diverse fonti informative. Tale elemento è stato individuato nei nomi di persona - sui quali è stata imperniata l'intera struttura archivistica dell'istituto sin dalla sua nascita. Fondamentale in questo si è rivelato il database dei nomi delle vittime della Shoah in Italia, perno dell'intero processo di integrazione e pubblicazione dei dati.

I caratteri propri di questo database verranno illustrati nel corso dell'intervento al fine di presentare non solo uno dei momenti chiave dello sviluppo del progetto, ma anche e soprattutto per dimostrare come grazie alla rappresentazione concettuale delle informazioni disponibili, sia stato possibile ripensare l'approccio stesso ai nomi, considerandoli cioè non più a partire dalle diverse funzioni svolte o esperienze vissute dal soggetto (deportato, partigiano, autore, fotografo) ma dal loro essere "persona", ovvero entità reali ed univoche.

La costruzione di un ontologia sul dominio specifico “Shoah” ha costituito il naturale completamento del lavoro sui nomi: le classi e le proprietà che ne sono alla base hanno permesso infatti la realizzazione di una rete concettuale in grado di ricondurre ad una singola entità Persone, informazioni distribuite su più sistemi informativi.

L’ontologia definita, grazie alla sua naturale flessibilità, consentirà la progressiva modellazione nel tempo di altre specifiche porzioni di dominio non considerate nel progetto iniziale e al contempo lo sviluppo di una costruzione *bottom-up* decentralizzata, come suggerito dalle metodologie Linked Data.

Al momento l’ontologia comprende 9 *classes*, 27 *object properties* e 26 *datatype properties*; sono stati creati indici sui nomi dei campi di concentramento e sterminio nazisti, sui luoghi di raccolta e detenzione, sui convogli e gli eccidi avvenuti in Italia prima della deportazione. Attraverso la creazione di IRI univoche per tutte le risorse del dominio, è stato possibile ricondurre tutte le informazioni a “nodi” univocamente riconosciuti nel dominio senza la proliferazione di informazioni.

Nello sviluppo complessivo del progetto, un ruolo chiave è stato svolto dall’introduzione di una piattaforma di lavoro trasversale e comune ai vari settori dell’Istituto, in grado di soddisfare le esigenze specifiche di lavoro di ciascun settore. Tale piattaforma è stata realizzata seguendo le raccomandazioni del W3C sulle Linked Data Platform [3] ed è conforme all’ontologia creata. Attraverso questa piattaforma oggi tutti i dipartimenti del Centro - specializzati ciascuno in diverse aree di attività di conservazione e ricerca - possono lavorare a partire un unico elenco di Persone, con informazioni biografiche univoche e, attraverso le IRI univoche, collegare tali Persone a descrizioni archivistiche, catalografiche, documenti digitalizzati.

Il progetto ha cercato di sperimentare tutti i vantaggi legati alle tecnologie semantiche per la pubblicazione del patrimonio culturale (su questo tema si veda ad esempio [4] [5] [6]) e a tal fine è stato realizzato un test di interlinking tra i dati CDEC relativi ai deportati e i dati pubblicati dall’Archivio Centrale dello Stato secondo le medesime tecnologie (http://dati.acs.beniculturali.it)

nel Fondo archivistico “A4 Bis , internati stranieri e spionaggio 1939-1945”, della Direzione Generale di Pubblica Sicurezza del Ministero degli Interni. Assumendo come valide le sole connessioni che hanno dimostrato una totale sovrapposizione di nome, cognome, data e luogo di nascita, è stato possibile riconciliare circa un centinaio di nomi completando da una parte il profilo biografico su queste persone e aprendo, dall’altra, molte altre criticità legate alla presenza di varianti linguistiche dei nomi di persona e di luogo.

Consapevoli delle criticità legate all’interlinking automatico da cui possono scaturire “sameAs” scorretti (si vedano ad esempio [7] e [8]), il testbed di interconnessione con un dataset esterno è stato molto efficace poiché ha attribuito una maggiore certezza al dato pubblicato e parallelamente ha ampliato il grafo di conoscenza di partenza. Questa interconnessione tra risorse è possibile e semplice attraverso le tecnologie LOD che sono oggi a disposizione, ma necessita senz’altro di un accurato controllo manuale.

Un ulteriore vantaggio legato alla pubblicazione LOD riguarda il *reasoning* dei dati esposti attraverso il quale è possibile inferire nuova conoscenza dedotta dalla struttura dei dati utilizzata per la rappresentazione; il *reasoning* applicato sulle relazioni familiari, ad esempio, è in grado di far emergere inconsistenze ed errori di cui altrimenti non si avrebbe conoscenza e allo stesso tempo può aiutare nella ricostruzione virtuale degli alberi genealogici.

CONCLUSIONI E SVILUPPI FUTURI

Gli sviluppi futuri legati a questo progetto sono diversi e molto ambiziosi: se da una parte l'obiettivo principale di creazione di una piattaforma comune per il lavoro interno del Centro è stato realizzato con successo, dall'altra si sono aperte numerose strade che si vorrebbe investigare per capire quanto questo modello sia riutilizzabile su una più ampia scala. I numerosi istituti e centri di ricerca dedicati alla storia della Shoah sparsi in Europa, negli Stati Uniti, in Israele, potrebbero essere il principale banco di prova e sperimentazione del modello messo a punto, nonché soggetti attivi di una collaborazione decentralizzata. I vantaggi sarebbero molteplici, a cominciare dalla possibilità di ottenere dati più precisi sul numero effettivo delle vittime attraverso la riconciliazione di nomi duplicati o triplicati. La rintracciabilità in un unico punto di accesso, di informazioni provenienti da più fonti, spesso sconosciute, consentirebbe inoltre l'arricchimento della conoscenza non solo dell'identità anagrafica delle persone ma anche delle loro personali biografie. Laddove vi sia la disponibilità dei dati, la tecnologia linked data consentirebbe inoltre la ricostruzione virtuale dei nuclei famigliari distrutti o dispersi.

Il carattere proprio dell'ontologia, da un lato, e la natura del processo descritto dall'altro, lasciano intravedere inoltre possibilità di applicazione dell'ontologia stessa per la descrizione di fenomeni simili.

References

[1] Ehri Project http://www.ehri-project.eu/project-description

[2] Speck, R., Blanke, T., Kristel, C. et alii (2014) "The Past and the Future of Holocaust Research: From Disparate Sources to an Integrated European Holocaust Research Infrastructure". Retrieved from http://arxiv.org/ftp/arxiv/papers/1405/1405.2407.pdf

[3] http://www.w3.org/TR/ldp/

[4] Coyle, K. (2013) "Library linked data: an evolution". Retrieved from http://leo.cineca.it/index.php/jlis/article/view/5443/7889

[5] Edelstein, J. et alii (2013) "Linked Open Data for Cultural Heritage: Evolution of an Information Technology" Retrieved from

[6] Summers, E., Salo, D. (2013) "Linking Things on the Web: A Pragmatic Examination of Linked Data for Libraries, Museums and Archives." Library of Congress. Retrieved from http://arxiv.org/ftp/arxiv/papers/1302/1302.4591.pdf

[7] Halpine, H. et alii (2009) "When owl:sameAs isn't the Same: An Analysis of Identity in Linked Data"

[8] Glaser, H., Halpine, H. (2012). "The Linked Data Strategy for Global Identity"

Retrieved from http://eprints.soton.ac.uk/333924/3.hasCoversheetVersion/IC-16-02-Lnkd.pdf

http://www.whysel.com/papers/LIS670-Linked-Open-Data-for-Cultural-Heritage.pdf