DOI: 10.26314/GARR-Conf18-proceedings-08

Beni archivistici e big data. Il progetto GAWS: Garzoni, Apprenticeship, Work, Society

Andrea Erboso

Archivio di Stato di Venezia

Abstract. Gli archivi rappresentano uno dei più ricchi e importanti "database" di informazioni del passato la cui consultazione, tuttavia, è resa difficoltosa dalla varietà di supporti, tipologie e grafie che caratterizza la documentazione manoscritta. L'Archivio di Stato di Venezia ha collaborato con importanti istituti universitari europei per un progetto dedicato alla valorizzazione del proprio patrimonio archivistico. Il progetto GAWS ha consentito, dopo un lungo lavoro di digitalizzazione integrale del materiale, la realizzazione di un database di vastissime dimensioni, unico nel panorama degli studi storici. Sul database, sul processo di inserimento dei dati e sulle possibilità di interrogazione degli stessi sono stati testati nuovi approcci offerti dalle tecnologie informatiche.

Keywords. Archivi, Garzoni, Big data, Open data, Machine learning

Introduzione

Il dibattito sempre più attuale sull'uso di big data a supporto della ricerca sta interessando, ormai da diversi anni, il mondo degli archivi: non deve stupire che anche un settore principalmente e tradizionalmente rivolto agli studi umanistici si interessi dei risultati più recenti offerti dalle computer sciences e dalle tecnologie di intelligenza artificiale nello sviluppo di applicazioni per la gestione, l'organizzazione e la manipolazione di grandi quantità di dati. Gli archivi, infatti, sono costituiti da complessi documentali già di per sé ricchissimi di contenuti informativi, a loro volta ulteriormente ampliati dai legami e dalle interrelazioni che in maniera organica intercorrono tra le carte: nel loro complesso gli archivi rappresentano un potenziale di conoscenza senza eguali. La stessa disciplina archivistica, quasi anticipando certi sviluppi dell'attuale ricerca sui big data, ha da sempre cercato di affinare le metodologie che consentono di descrivere i complessi archivistici ad un livello di definizione tale da permettere ad un ricercatore di orientarsi in un patrimonio che per vastità esclude a priori la possibilità dell'approccio diretto e analitico. Le metodologie di ordinamento e descrizione degli archivi hanno proprio l'obiettivo di fornire gli elementi generali necessari a strutturare gerarchicamente l'informazione, consentendo con relativa rapidità di identificare le unità che possono contenere i dati ricercati con precisione.

L'esperienza degli archivi è fondamentale anche per il tema, strettamente collegato, degli open data: a condivisione della conoscenza acquisita dalle esperienze di valorizzazione degli archivi è uno degli obiettivi dell'amministrazione archivistica. I documenti conservati presso gli Archivi di Stato, infatti, sono beni culturali a tutti gli effetti e, in quanto tali, disciplinati dal Codice dei beni culturali (D.Lgs 42/2004), che ne sancisce la natura di bene

Conferenza GARR 2018 - Data (R)evolution - Selected Papers
Cagliari, 3-5 ottobre 2018 - ISBN: 978-88-905077-8-6

pubblico. La predisposizione di strumenti di ricerca innovativi e aggiornati, necessari alla fruizione e valorizzazione degli archivi stessi, è dunque uno dei principali compiti degli Archivi di Stato che ne devono garantire la completa pubblicità e condivisibilità.

1. Il progetto GAWS

Un esempio di applicazione di tecnologie tipiche dei big data al contesto archivistico è il progetto GAWS, Garzoni Apprenticeship Work and Society, finanziato dall'Agence Nationale de la Recherche ANR e dal Fonds National Suisse FNS e frutto della collaborazione tra l'Università di Lille 3, l'Università di Rouen, l'EPFL di Losanna e l'Archivio di Stato di Venezia. Il progetto è in via di conclusione e ha già prodotto un complesso database che sarà pubblicato nell'autunno del 2019 (https://garzoni.hypotheses.org/).

Il progetto ha preso avvio da una fonte documentaria conservata presso l'Archivio di Stato: la serie degli Accordi dei garzoni del fondo della Giustizia Vecchia, una raccolta sistematica di 53.890 contratti di apprendistato tra garzone e maestro, relativi a oltre 1.300 professioni e raccolti in 32 registri (per un totale di 14.236 carte manoscritte).

La particolarità che rende unica al mondo questa serie è che, a differenza di quanto accade in tutta l'Europa dell'età moderna dove il contratto di apprendistato è considerato di natura privatistica e quindi sottoscritto dalle parti di fronte al notaio, a Venezia questa pratica era espressamente vietata. Nella Serenissima era lo Stato - la magistratura della Giustizia Vecchia appunto - ad occuparsi della registrazione di questo tipo di accordi (Bellavitis et al., 2017). Questa particolare disposizione ha garantito alla documentazione una omogeneità sorprendente con la registrazione costante della stessa tipologia di informazioni per tutto l'arco di tempo per cui si conserva la fonte (1575-1772 con lacune).

L'obiettivo del progetto è stato, dunque, sperimentare un metodo per strutturare le informazioni contenute nella fonte e renderle esplorabili in modalità altrimenti impossibili.

2. Descrizione e inserimento dei dati

In una prima fase si è proceduto alla digitalizzazione sistematica di tutto il materiale che è stato poi caricato in un'interfaccia di navigazione e lavoro compatibile con gli standard IIIF (Ehrmann et al., 2016).

Contemporaneamente un team di storici e archivisti si è occupato di effettuare uno studio preliminare sulla fonte ed estrarre da essa il modello delle informazioni che un gruppo di informatici ha adoperato per sviluppare l'ontologia e l'architettura informatica. La tipologia di contenuti presenti nei singoli accordi e inseriti nel database sono:

- il nome del garzone con alcuni elementi identificativi (patronimico, provenienza);
- la professione che il garzone vuole imparare;
- il nome del maestro che insegnerà il mestiere, anch'esso con alcuni elementi identificativi (localizzazione della bottega);
- il nome di una persona che funge da garante e che molto spesso è un parente del garzone stesso;
- le condizioni materiali del contratto;

- la durata dell'apprendistato in anni (regolamentata da ogni singola corporazione);
- le condizioni salariali.

L'inserimento dei dati è avvenuto direttamente a partire dalla riproduzione digitale del

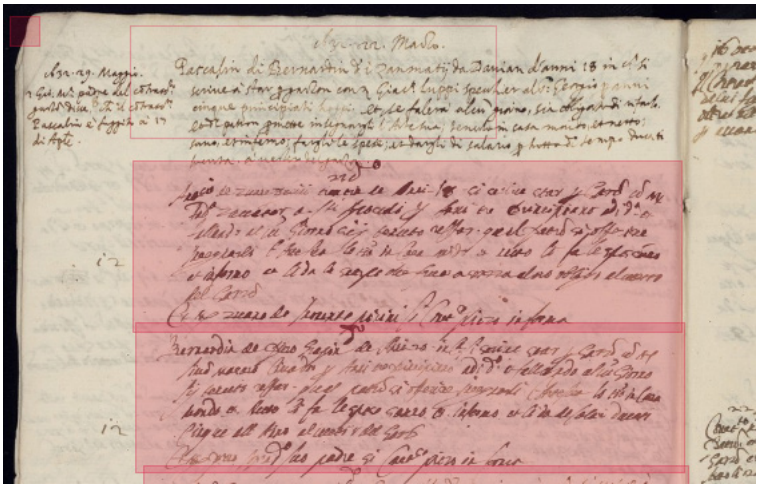

Fig. 1
Esempi di segmentazione della riproduzione digitale del documento

documento che è stata segmentata in riquadri, uno per contratto.
Direttamente dal segmento di immagine è possibile aprire la maschera di inserimento dei dati per la trascrizione della informazioni. Questa attività, condotta da un team di specialisti, ha prodotto più di 470.000 trascrizioni, alle quali sono state associate più di 420.000 tag semantiche individuate nello studio preliminare sulla fonte. Ogni trascrizione con le relative tag è stata annotata direttamente sull'immagine come fosse un suo metadato: in tutto il database, dunque, ogni informazione rimane costantemente associata alla riproduzione del documento stesso.

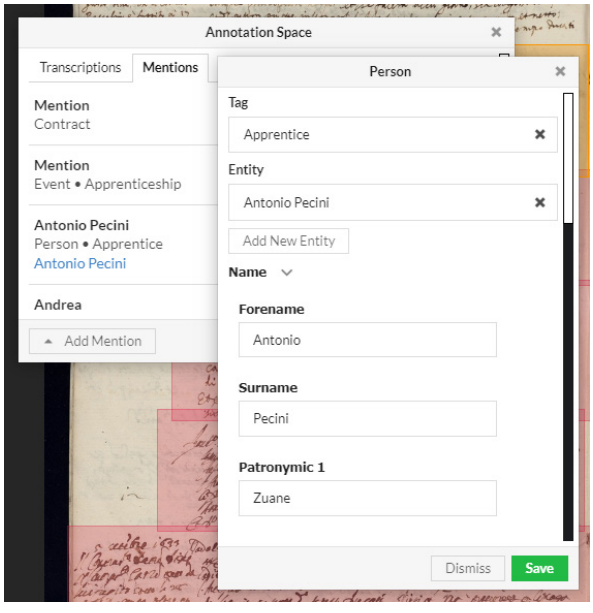

Fig. 2
Esempio della maschera di trascrizione e inserimento dati

2.1 Tecniche di machine learning per il data entry

L'aspetto più critico della compilazione è stato quello riguardante i nomi di persona: tra i 187.266 nominativi, i casi di omonimia e similarità sono stati frequentissimi. Le modalità di compilazione hanno previsto la trascrizione letterale del nome come presentato nella fonte, la cosiddetta person mention, e l'attribuzione di una person id normalizzata.
I casi problematici emersi dalla compilazione sono riconducibili a due fattispecie:

- diverse mention riferite alla stessa persona (stesso individuo citato più volte);
- mention identiche riferite però a persone diverse (omonimia).

La possibilità di disambiguare tra i possibili omonimi e i casi di variante grafica dello stesso nome è stata offerta dalle tecniche di machine learning messe a punto appositamente per il progetto. All'atto di attribuzione del person id normalizzata infatti, il software di compilazione proponeva al compilatore i nomi simili già inseriti nel database, offrendo in aggiunta una sintesi dei dati correlati al nominativo: in questo modo al compilatore venivano forniti tutti gli elementi disponibili (e il link diretto alla fonte per la verifica puntuale) per scegliere criticamente se assegnare un nuovo id (omonimia) o attribuendone uno già esistente (variante grafica). Man mano che il database si è popolato di informazioni, maggiore è stata la precisione del tool nell'indicare al compilatore la rosa di casi tra i quali operare la disambiguazione, eliminando sempre più il rumore causato dai risultati non pertinenti.

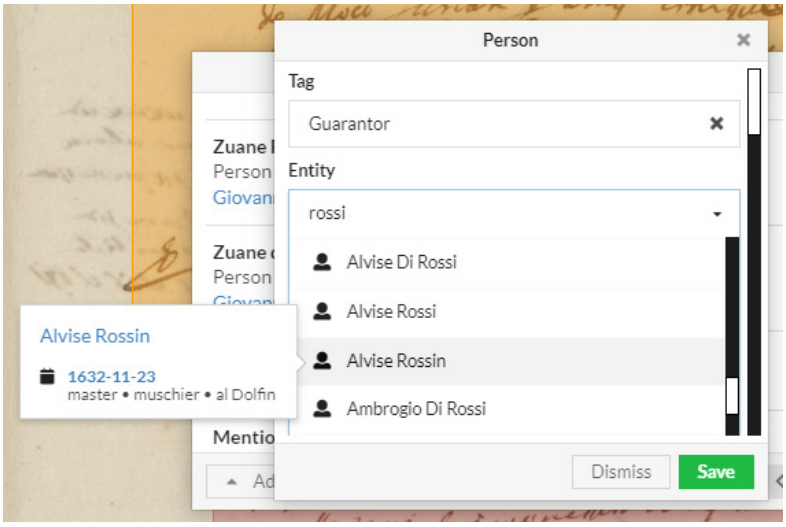

Fig. 3
Esempio di funzionamento del tool di suggerimento per la disambiguazione

3. Modalità di interrogazione

L'interfaccia di esplorazione dei dati, attualmente in fase di test, consente di interrogare il database con ricerche a testo libero, filtri e query sparql oltre che, naturalmente, per segnatura archivistica, navigando tra le immagini. Soprattutto il sistema di filtri si sta rivelando particolarmente efficace e consente di estrarre liste di contratti sempre più precise

che lo studioso potrà consultare online, scegliendo quali informazioni visualizzare tra le molte contenute nei contratti, oppure potrà scaricare nei formati xlsx, ods e json per elaborarli con software più complessi, per esempio per analisi statistiche su lunghi periodi.

In tutti i casi, sia nella consultazione online, che nei dati scaricati, sarà presente il collegamento con l'immagine del documento, evidenziata nella parte che contiene le informazioni. Questa possibilità è frutto di una precisa scelta metodologica, relativamente nuova negli studi umanistici, che consentirà agli studiosi di verificare agevolmente la fonte citata: la trascrizione, con i suoi inevitabili errori, diventa pertanto un punto di accesso all'informazione, senza sostituirla.

4. Conclusioni

L'esperienza del progetto GAWS è stata estremamente fruttuosa per sperimentare nuovi metodi di valorizzazione del patrimonio archivistico: da un lato, la digitalizzazione del materiale ha consentito di rendere disponibili agli studiosi i registri degli Accordi dei garzoni, preservando per il futuro gli originali dall'usura dovuta alla consultazione, dall'altro la banca dati offre opportunità di ricerca prima impensabili. La possibilità di interrogare il database a partire da un nome o da una professione, per fare un esempio, consente di far emergere aspetti che la consultazione tradizionale, limitata alla sola lettura degli accordi registrati in successione cronologica, non permetterebbe, come le reti di relazioni sociali tra maestri, garzoni e garanti o i flussi migratori legati all'apprendistato.

Il progetto ha poi offerto un terreno di prova per la creazione di un team multidisciplinare che ha visto collaborare fianco a fianco archivisti, storici, storici dell'arte, informatici, scienziati del linguaggio, in un clima collaborativo e di reciproca contaminazione, costituendo un incoraggiante precedente per la progettazione di nuovi interventi di valorizzazione del patrimonio archivistico.

Riferimenti bibliografici

Bellavitis A., Frank M., Sapienza V., a cura di (2017), Garzoni Apprendistato e formazione tra Venezia e l'Europa in età moderna, Mantova, Universitas Studiorum.

Ehrmann M., Colavizza G., Topalov O., Cella R., Drago D., Erboso A., Zugno F., Bellavitis A., Sapienza V., Kaplan F. (2016), From Documents to Structured Data: First Milestones of the Garzoni Project, DH Commons Journal, (2), pp. 1-18.

Autore

Andrea Erboso - andrea.erboso@beniculturali.it

Laureato in Storia dell'Arte Moderna presso l'Università Ca' Foscari di Venezia, ha conseguito il diploma in Archivistica, Paleografia e Diplomatica presso l'Archivio di Stato di Venezia. Dal 2018 è funzionario archivista di Stato presso l'Archivio di Stato di Venezia. Si è occupato principalmente di arte veneziana del secondo Cinquecento, di botteghe e apprendistato nelle arti. Ha collaborato a progetti di digitalizzazione documentaria e descrizione archivistica con importanti istituzioni internazionali.