DOI: 10.26314/GARR-Conf16-proceedings-06

I corpora digitali: dall'obsolescenza tecnologica, alla salvaguardia e alla condivisione

Eva Sassolini, Sebastiana Cucurullo, Alessandra Cinini

Istituto di Linguistica Computazionale "Antonio Zampolli", ILC-CNR

Abstract. Studio e implementazione di un protocollo di recupero, conservazione e valorizzazione di testi e corpora digitali interessati da problemi di obsolescenza tecnologica. Le strategie di salvaguardia adottate si spingono oltre il salvataggio dei testi e la conservazione in un formato di rappresentazione in linea con gli standard internazionali (XML TEI), si pongono come obiettivo la valorizzazione di questo patrimonio attraverso nuove modalità di fruizione dei contenuti. Lo scopo è affiancare le funzionalità classiche di analisi testuale, che da sempre caratterizzano le nostre attività di ricerca, a nuove modalità grafiche e visuali di fruizione dei dati e, in alcuni casi, migrare verso dispositivi mobili e tecnologie App. In questo articolo, oltre al protocollo di recupero, presentiamo due sperimentazioni di valorizzazione di contenuti testuali. Nel primo caso proponiamo tecniche di visual analytics applicate ad un corpus testuale semi strutturato riguardante corrispondenza redatta in lingua italiana del 1600. Nel secondo caso abbiamo realizzato un'applicazione per sistema Android finalizzata all'interrogazione di dati testuali relativi ad un progetto di censimento di architetture moderne della regione Liguria.

Keywords. Testi digitali, Analisi testuale, Preservazione dei dati, Standardizzazione, Diffusione dei risultati

Introduzione

L'Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC) rappresenta da sempre un punto di riferimento per la comunità scientifica nazionale ed internazionale per lo studio e la realizzazione di procedure per l'analisi automatica di testi e di materiale lessicale (Picchi 2003). Queste attività hanno prodotto una grande quantità di materiali testuali, spesso arricchiti da un variegato e prezioso apparato di annotazioni. Oggi, conservati in vari formati e tracciati record, rappresentano un patrimonio culturale di inestimabile valore da salvaguardare e valorizzare: migliaia di testi e corpora d'autore o di riferimento per aspetti linguistici, storico-culturali e giuridici. Come noto però ogni oggetto digitale è destinato prima o poi ad avere problemi di obsolescenza tecnologica. Nel dibattito attuale la questione della sostenibilità e conservazione dei dati digitali costituisce forse il principale nodo da risolvere. Sono molte le voci autorevoli che hanno sollevato il problema, come è accaduto nella Conferenza UNESCO del 2012 alla quale fu dato il significativo titolo: The Memory of the World in the Digital Age: Digitization and Preservation (Nota 1). Così come le iniziative internazionali finalizzate alla preservazione e conservazione a lungo termine dei materiali digitali. Il progetto Digital Preservation Europe (DPE) è un esempio di iniziativa internazionale per la costruzione e formalizzazione di "buone pratiche" (Nota 2), tra i cui obiettivi dichiarati ci sono sia le tecniche e processi di gestione di "memorie digitali" che le azioni congiunte a livello internazionale: "Digital preservation combines policies, strategies and actions to ensure access to reformatted and born digital content regardless of the challenges of media failure and technological change. The goal of digital preservation is the accurate rendering of authenticated content over time" (ALA 2007).

Anche in Italia è ormai forte la consapevolezza del problema, si pensi a quanto fatto dalla Presidenza del Consiglio dei Ministri con l'Agenzia per l'Italia Digitale (AgID), in cui sono state predisposte linee guida che illustrano le procedure e gli strumenti per la conservazione dei documenti informatici da parte delle PA

(Nota 3). Rimangono purtroppo indietro ambiti specifici come il recupero e conservazione di testi digitali che hanno subito un'elaborazione software a fini linguistici o filologici, per i quali rischiamo ancora la perdita per la mancanza di opportune iniziative di recupero.

1. Il progetto di recupero

Il progetto di recupero ILC è nato pochi anni fa come iniziativa interna (Sassolini et al. 2014) e le attività che lo caratterizzano proseguono oggi con la collaborazione di molte istituzioni pubbliche e private, impegnate sullo stesso fronte, che condividono con ILC le finalità di preservazione e valorizzazione delle proprie risorse digitali. Esistono criteri di priorità per la scelta dei testi da recuperare e tengono conto di:

- caratteristiche linguistiche, storiche e culturali dei testi;
- recupero di casi di studio significativi;
- importanza dei materiali, spesso legati alla realizzazione di autorevoli progetti nazionali e internazionali (Cinini et al. 2013).

Nel nostro percorso progettuale abbiamo incontrato una grande varietà di formati dei file, che ha reso il lavoro di recupero estremamente complesso.

L'obsolescenza tecnologica è infatti un problema che va affrontato a vari livelli (Figura 1). Il più ostico riguarda il software con il quale, per esempio, sono state redatte alcune edizioni critiche o complessi schemi di annotazione linguistica. Il più diffuso riguarda invece testi che presentano un formato ormai superato e spesso mancante di una specifica di corredo per la corretta comprensione. Una specifica di formato fornisce infatti i dettagli necessari per costruire un file da un testo e viceversa, stabilisce le codifiche ammesse e le applicazioni software capaci di decodificarne il formato e restituirne il contenuto. Mancando questo tassello la ricostruzione è ardua e non sempre si riesce a ottenere una riproduzione esatta della risorsa. Il progetto di recupero è diventato oggi un "protocollo" costituito da una serie di fasi più o meno articolate di decodifica, ossia una serie di passi tramite i quali un testo conservato in un formato obsoleto viene ricondotto ad uno standard.

2. Dal recupero alla salvaguardia

Una volta assolte tutte le fasi di recupero il testo è pronto per essere messo a disposizione della comunità scientifica. La salvaguardia è prioritaria ma non basta, perché i singoli archivi possano trasformarsi in una rete di conoscenza condivisa e distribuita a livello nazionale e internazionale, serve un'integrazione all'interno di infrastrutture di ricerca che supportino la creazione, la fruizione, la distribuzione e la valorizzazione delle risorse. La recente partecipazione dell'Italia alla rete europea CLARIN-ERIC (Common Language Resources and Technology Infrastructure) è apparsa come un'occasione importante per approdare alla condivisione non solo dei risultati del lavoro di recupero e conservazione ma anche dello stesso protocollo.

La creazione del consorzio CLARIN-IT, del

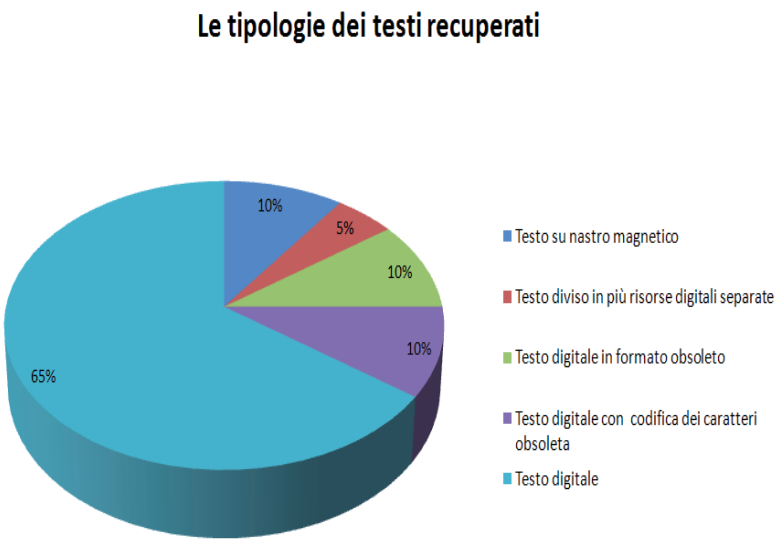

Fig. 1 Sintesi delle tipologie di testi trattati nella fase di recupero

quale ILC costituisce uno dei pilastri infrastrutturali, consentirà alle comunità di ricerca nel settore delle scienze umane e sociali di federare le infrastrutture locali esistenti e le loro risorse. Questa iniziativa auspichiamo possa trasformarsi in un catalizzatore per lo sviluppo di una rete di eccellenza italiana ed europea per la ricerca nel settore delle Digital Humanities. L'obiettivo di valorizzazione e condivisione si inserisce nel più ampio panorama degli studi digitali nelle scienze umane e sociali, che si impegnano a preservare, documentare e rendere accessibili i dati, con l'utilizzo di standard di metadati e di annotazione condivisi internazionalmente, e a indicizzare i dati stessi in piattaforme comuni.

3. Le iniziative di valorizzazione del testo

Come ulteriore passo in questo processo di valorizzazione, abbiamo sperimentato tecniche di visual analytics per realizzare viste e sintesi grafiche dei contenuti da integrare con le applicazioni di analisi testuale di cui ILC dispone (Rydberg-Cox 2011). I dati utilizzati per la sperimentazione provengono da un corpus testuale risalente alla prima metà del 1600, il cui contenuto è costituito da lettere, redatte in un linguaggio prevalentemente informale. Il materiale appartiene ai 20 volumi dell'edizione Favaro della biblioteca di Galileo Galilei, conservata a Firenze presso il Museo Galileo. Degli otto volumi contenenti il "Carteggio" abbiamo scelto il volume XV, relativo all'anno 1633, perché arco temporale denso di rilevanti eventi storici per Galileo, come il processo e condanna da parte dell'Inquisizione nel giugno del 1633. Ogni lettera è strutturata in "campi" e questo ne facilita l'organizzazione in dati matriciali. Per ragioni di omogeneità abbiamo eliminato quei documenti che non presentavano almeno i campi principali quali: titolo, mittente e destinatario. Le rappresentazioni visuali realizzate utilizzano diverse modalità grafiche (Moretti 2005), ognuna corredata della possibilità di interagire con il motore di analisi testuale attraverso le funzionalità più classiche, cambia solo la modalità di formulazione della query, che viene legata all'interazione con l'oggetto grafico utilizzato:

1. Diagrammi a barre per produrre la sintesi dei soggetti che ricevono e spediscono missive:
   a. Quali e quanti sono i personaggi che scrivono a Galileo;
   b. A chi lo scienziato scrive con più assiduità;
2. Diagrammi temporali per mettere in relazione missive ed eventi storici, per:
   a. Valutare come lo scambio di messaggi sia strettamente connesso al diffondersi per esempio della notizia della condanna di Galileo;
   b. Mettere in evidenza triangolazioni/gruppi di corrispondenze legate da una stretta temporalità, individuando così l'esistenza di possibili "temi di discussione";
3. Strutture grafiche ad albero per l'analisi della lingua adottata per le missive. Per esempio emersione di arcaismi evidenti nelle occorrenze delle parole (forme):
   a. In una prima rappresentazione a cluster è mostrato l'insieme delle forme articolato in sottoinsiemi di parole, dimensionati secondo le frequenze di attestazione, organizzati in categorie grammaticali (sostantivi, verbi aggettivi, ecc.) o strutturali (formule di apertura, di cortesia, di saluto);
   b. In una seconda modalità, gli stessi dati matriciali, sono utilizzati per la costruzione di un albero espandibile interattivamente. Criteri di razionalizzazione dei dati hanno imposto un taglio delle forme con frequenze basse. In questa seconda modalità è maggiore l'interazione con il motore di analisi testuale.

4. La migrazione verso tecnologie mobile

Sempre con l'obiettivo di migliorare la capacità di fruizione dei dati testuali abbiamo fatto un secondo esperimento nell'ambito dei beni culturali sfruttando esperienze consolidate in questo settore (Sassolini et al. 2013). L'obiettivo era quello di rendere fruibili da mobile i dati raccolti nell'ambito del progetto di ricerca "Censimento e schedatura di complessi di architettura moderna e contemporanea in Liguria" (Nota 4). Le intenzioni del progetto erano duplici da un lato divulgare contenuti nuovi o difficilmente reperibili, organizzarli, standardizzarli e metterli

a disposizione della comunità, dall'altro offrire una modalità di divulgazione intuitiva e diffusa.

Le modalità di interazione in ambito mobile hanno oggi canoni standardizzati dove la consultazione dei contenuti testuali si intreccia con i dati geografici: attivando sistemi di notifiche, sono suggerite dinamicamente all'utente informazioni correlate al luogo in cui si trova o ad oggetti che sta osservando. Attualmente è stata sviluppata una applicazione per sistema operativo Android, che affianca la consultazione delle architetture censite con visualizzazione su mappa, a quella dei contenuti descrittivi delle architetture di maggior rilievo (Figura 2).

L'utente può navigare nei testi, con le funzioni base di Information Retrieval (IR) e visualizzare le opere rispondenti ai criteri di ricerca. Sulla mappa si evidenziano ad esempio le aree in cui un determinato progettista ha operato, attribuibili ad uno stile architettonico o legate alla presenza di specifici interventi (ricostruzione, riqualificazione, etc.). Cluster basati sulla prossimità geografica possono suggerire all'utente itinerari di visita. La sperimentazione è stata fatta con un campione ancora esiguo di documenti ma, nella prospettiva in cui saranno presenti una quantità significativa di materiali testuali, è prevista una maggior interazione tra le due modalità di consultazione: testuale e su mappa.

5. Conclusioni

Il protocollo di recupero è periodicamente aggiornato ma le modifiche sono limitate ai singoli casi che non rientrano in nessuna casistica trattata. Per quanto riguarda la salvaguardia e la condivisione internazionale, ciò che offre il consorzio CLARIN-IT è una sicura risposta al processo di recupero. Aprirsi ad una platea così vasta pone però nuovi interrogativi, nell'infrastruttura CLARIN esiste infatti la possibilità di formulare ricerche "federate", che permettono di proiettare una singola ricerca sull'intera rete. In questa prospettiva applicazioni classiche di accesso ai contenuti sono poste davanti a nuove sfide, che noi vogliamo raccogliere rivolgendoci anche alle nuove tecnologie di rappresentazione sintetica e grafica dei dati. Pensiamo ad un allargamento della platea dei fruitori, non solo gli "addetti ai lavori" o gli studiosi, ma anche utenti comuni, tipicamente più orientati all'utilizzo di dispositivi mobili, che hanno familiarità con rappresentazioni grafiche delle informazioni. Il nostro intento è porre l'esigenza di una maggiore diffusione di una cultura digitale che non esaurisca il suo compito all'interno delle comunità scientifiche, ma che sia in grado di adeguarsi all'evoluzione delle tecnologie e delle modalità di fruizione dei contenuti.

Fig. 2 Scheda descrittiva dell'architettura selezionata

Riferimenti bibliografici

Sassolini, Eva. Cucurullo, Sebastiana. Sassi, Manuela. (2014), Methods of textual archive preservation. In: CLiC-it 2014 – First Italian Conference on Computational Linguistics, (Università di Pisa e CNR, Pisa, Italia, 9-10 dicembre 2014), Proceedings, vol. I, ISBN 978-886741-472-7, pp. 334 – 338, Pisa University Press, Pisa.

Cinini, Alessandra. Cucurullo, Sebastiana. Picchi, Paolo. Sassi, Manuela. Sassolini, Eva. Sbrulli, Stefano. (2013), I testi antichi: un patrimonio culturale da conservare e riutilizzare, In: 27a DIDAMATICA 2013, Tecnologie e Metodi per la Didattica del Futuro, (Pisa, 7-8-9 maggio 2013), Proceedings, pp. 867-870, AICA, Pisa.

Rydberg-Cox, Jeff. (2011), "Social networks and the language of greek tragedy", Journal of the Chicago Colloquium on Digital Humanities and Computer Science, Vol. 1, No. 3.

ALA (American Library Association). (2007), Definitions of digital preservation, Chicago: American Library Association, Available at: http://www.ala.org/ala/mgrps/divs/alcts/resources/preserv/defdigpres0408.pdf

Moretti, Franco. (2005), Graphs, Maps, Trees: Abstract Models for a Literary History, London; New York: Verso.

Picchi, Eugenio. (2003), PiSystem: sistemi integrati per l'analisi testuale, In A. Zampolli, N. Calzolari, L. Cignoni, (eds.), Computational Linguistics in Pisa - Linguistica Computazionale a Pisa, Linguistica Computazionale, Special Issue, XVIII-XIX, (2003), Pisa-Roma, IEPI, Tomo II, pp 597-627.

Note

Nota 1. http://www.unesco.org/new/en/communication-and-information/events/calendar-of-events/events-websites/the-memory-of-the-world-in-the-digital-age-digitization-and-preservation/

Nota 2. Digital Preservation Europe (DPE) was a collaborative European digital preservation project that ran from 2006 to 2009, aimed at creating collaborations and synergies among many existing national initiatives across the European Research Area.

Nota 3. http://www.agid.gov.it/sites/default/files/linee_guida/la_conservazione_dei_documenti_informatici_rev_def_.pdf.

Nota 4. MIBACT per la Liguria, Regione Liguria e Dipartimento DSA di Scienze per l'Architettura dell'Università degli Studi di Genova).

Eva Sassolini
eva.sassolini@ilc.cnr.it

Nella propria attività di ricerca in campo informatico ha maturato esperienza nello sviluppo e adattamento di sistemi di analisi testuale e nella realizzazione di strumenti di acquisizione e gestione di corpora testuali e nella loro conversione in standard internazionali di rappresentazione per la conservazione a lungo termine e la valorizzazione

Sebastiana Cucurullo
nella.cucurullo@ilc.cnr.it

Esperienza nello sviluppo e adattamento di sistemi software per l'analisi testua-le. Attività di ricerca: acquisizione e trattamento di testi e corpora testuali, standardizzazione dei formati (XML-TEI) e gestione di database

Alessandra Cinini
alessandra.cinini@ilc.cnr.it

Da anni svolge attività di ricerca come informatico presso ILC ed ha esperienza nell'acquisizione di materiali digitali dal web, nell'annotazione semantica dei testi, e nella costruzione di risorse linguistiche di dominio nell'ambito dei Beni Culturali.