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

E. Sassolini, M. Sassi, S. Cucurullo, A. Cinini

Istituto di Linguistica Computazionale "Antonio Zampolli"

eva.sassolini|manuela.sassi|nella.cucurullo|alessandra.cinini{@ilc.cnr.it}

introduzione

L'Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC) nella sua storia cinquantennale ha accumulato una grande quantità di materiali testuali che oggi sono conservati in vari formati e tracciati record. Tali risorse, spesso arricchite con un variegato e prezioso apparato di annotazioni, 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. Per molti di questi rischiamo ancora la perdita per la mancanza di opportune iniziative di recupero. L'urgenza di definire una procedura che metta al sicuro le risorse dall'inevitabile processo di obsolescenza tecnologica è comunque limitata dalla necessità di impostare un'operazione ragionata e quanto più possibile rispettosa dei dati, che conduca a formati di rappresentazione standardizzati senza perdita di informazioni. La comunità internazionale considera le risorse digitali una parte centrale dei beni culturali e molte istituzioni sono coinvolte in iniziative internazionali finalizzate alla preservazione e conservazione a lungo termine dei materiali digitali. Il progetto Digital Preservation Europe[1] (DPE) è un esempio iniziativa internazionale per la costruzione e formalizzazione di "buone pratiche". Tra gli obiettivi dichiarati troviamo sia le tecniche e processi di gestione di "memorie digitali" che le azioni congiunte a livello internazionale.

Il progetto di il recupero

ILC da sempre rappresenta un punto di riferimento per la comunità scientifica nazionale ed internazionale per lo studio e la realizzazione di procedure per l'analisi automatica dei testi e di materiale lessicale. Sin dai primi tentativi di utilizzare il calcolatore per analizzare dati linguistici in istituto si è sviluppato un aggregato ricco di conoscenze, strumenti e materiali, che può avvalersi del supporto e della collaborazione di studiosi di varie discipline (linguisti, lessicologi, lessicografi, filologi, letterati, ecc.): un background di competenze, standard di elaborazione e di codifica, procedure di elaborazione ed infine un grande archivio di materiale testuale. Il nostro gruppo di ricerca, da sempre impegnato nello sviluppo, gestione e adattamento di sistemi di analisi testuale, ha quindi potuto attingere a queste esperienze per formulare specifici metodi e tecniche per il recupero di materiali testuali digitali.
Il progetto di recupero è nato pochi anni fa come iniziativa fortemente voluta da ILC e prosegue oggi con la collaborazione di molte istituzioni pubbliche e private, impegnate sullo stesso fronte. Il nostro approccio al recupero è stato necessariamente cauto: inizialmente abbiamo lavorato su testi che erano stati prodotti per essere indicizzati con le prime procedure di analisi testuale presenti all'ILC sin dalla fine degli anni '70 del secolo scorso. I criteri adottati per la scelta dei testi hanno tenuto conto sia della necessità di trovare casi di studio significativi, sia dell'importanza dei materiali, spesso legati alla realizzazione di autorevoli progetti nazionali e internazionali.

[1] "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 2007:2)

Il lavoro di recupero

Si è trattato di misurarsi con 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 più livelli. Il più ostico riguarda il software con il quale, per esempio, sono state redatte alcune edizioni critiche o i complessi schemi di annotazione linguistica. Il problema invece più strutturale riguarda 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, stabilisce le codifiche ammesse e le applicazioni software capaci di decodificare file simili e di 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 quindi un "protocollo" costituito da una serie d fasi più o meno articolate di *transizione*, ossia una serie di formati intermedi tramite i quali un testo conservato in un formato obsoleto viene ricondotto ad uno standard.
Le fasi di recupero possono essere diverse perché date dalla composizione dei passi necessari a ricondurre una codifica dei caratteri, quindi di tutte le annotazioni, al formato Unicode, con i passi finalizzati al mapping del formato del file allo standard XML-TEI, senza perdita di informazioni.

Testo sorgente	**Perc.**	**Fasi di transizione richieste (FT)**	**Meta dati**
Testo su nastro magnetico	10%	Molte FT	Ricerche su materiali cartacei storici di ILC
Testo diviso in più risorse digitali separate	5%	FT>3	Recuperati da schede cartacee di progetto
Testo digitale in formato obsoleto	10%	FT>2	Recuperati da schede cartacee di progetto
Testo digitale con codifica dei caratteri obsoleta	10%	2<FT<3	Recuperati da: - Schede cartacee - Documentazione digitale
Testo digitale	65%	1 FT	Recuperati da documentazione digitale

Le azioni di salvaguardia e condivisione

Una volta assolte tutte le fasi necessarie al recupero del testo e delle annotazioni ivi contenute, il file che viene prodotto è un documento XML-TEI P5 con codifica utf8 pronto per essere messo a disposizione della comunità scientifica. A queste esigenze di doverosa salvaguardia, se ne affianca un'altra, non meno importante di condivisione dei risultati: perché i singoli archivi possano trasformarsi in una rete di conoscenza condivisa e distribuita a livello nazionale e internazionale, è auspicabile la loro 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[2] (Common

[2] Infrastruttura europea per creare, coordinare e rendere le risorse linguistiche e le tecnologie disponibili e prontamente utilizzabili: www.clarin.eu

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 di CLARIN-IT, del quale ILC costituisce uno dei pilastri infrastrutturali, consentirà alle comunità di ricerca nel settore delle scienze umane e sociali di trasformare la vasta collezione di risorse e infrastrutture locali esistenti, attualmente scollegate, in un'unica infrastruttura di ricerca nazionale, integrandola al contempo con la rete esistente e in corso di sviluppo a livello europeo. Questa iniziativa potrebbe trasformarsi così in un catalizzatore per lo sviluppo di una rete di eccellenza italiana ed europea per la ricerca nei settori del trattamento automatico del linguaggio e del testo nel contesto più ampio delle Digital Humanities. Abbiamo già iniziato a popolare il nodo italiano con i testi che mano a mano vengono recuperati, l'intento è iniziare un percorso di valorizzazione e permettere una condivisione più ampia di quanto rendiamo disponibile. Questa prospettiva si apre al 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.

Le prospettive future

La condivisione e salvaguardia internazionale offerte dall'infrastruttura europea sono una sicura risposta al processo di recupero, ci siamo domandati però se questo esaurisse la funzione di stimolo che la ricerca ha l'obbligo di svolgere. Inoltre affidare il patrimonio testuale ad una platea così vasta pone nuovi interrogativi, infatti l'integrazione in CLARIN prevede tra l'altro regole e formati standard per la formulazione di ricerche "federate" che offrono la possibilità di proiettare una singola ricerca sull'intera rete dell'infrastruttura. In questa prospettiva applicazioni classiche di accesso ai contenuti sono poste davanti a nuove sfide.
Valutando cosa è già stato fatto da altri per esempio da "Labex Obvil", laboratorio di eccellenza di Parigi che unisce competenze interdisciplinari provenienti da diverse discipline, da specialisti di letteratura e scienze della cognizione ad informatici, abbiamo intrapreso uno studio delle recenti tecniche di visual anaytics per la produzione di elaborazioni grafiche dei contenuti degli archivi digitali e le abbiamo applicate, al momento, ad alcuni casi di studio. L'iniziativa francese infatti, che si pone come un osservatorio della vita letteraria[3], utilizza risorse offerte dalle tecnologie informatiche sviluppate per il trattamento della lingua e dei testi, per esaminare sotto vari aspetti la letteratura francese classica e contemporanea. Sostanzialmente affianca le modalità classiche di fruizione dei testi a quelle più recenti di rappresentazione grafica e visuale dei contenuti. In questa prospettiva noi pensiamo anche 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. Naturalmente questo impone un ripensamento delle strategie di valorizzazione di questo patrimonio recuperato, proponendo nuovi scenari applicativi indirizzati ad una platea sempre più giovane e tecnologicamente avanzata. 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.

[3] http://obvil.paris-sorbonne.fr/