**La grande bellezza dei grandi dati**

**Roma, 26 novembre 2015**

Da tempo l'Archivio di Stato di Venezia, il Politecnico di Losanna e l'Università Ca' Foscari hanno promosso una comune iniziativa di ricerca per le nuove frontiere della digitalizzazione della documentazione archivistica, valendosi dell'esperienza acquisita dall'Archivio di Stato stesso in lunghi anni di creazione e predisposizione di contenuti digitali all'utenza.

Gli archivi in realtà sono un granaio di fatti fondamentale, forse unico, per la conoscenza del nostro passato ed il suo riverbero sul presente, solo in piccola parte consultabile rispetto ad una richiesta potenzialmente enorme.

Fin dalle epoche mesopotamiche e mediorientali, passando per il mondo greco romano, se pur meno ricco di testimonianze documentarie a noi pervenute, e giungendo al medioevo ed all'età moderna, nei documenti archivistici si riflette la vita quotidiana di intere comunità , di Istituzioni e di persone.

Gli archivi son un vero patrimonio di Big-data.

Il ventunesimo secolo si è aperto nel segno dei Big-data. Lo sviluppo delle reti informatiche, e dei contenuti atomizzati riorganizzati dalle stesse macchine ha permesso di disporre di una quantità di dati del tutto incomparabile con le epoche passate, ma bisognosa di gestione, senza tornare a strade analogiche e predeterminate per la loro interpretazione, ma sfuggendo al rumore senza significato di nuvole informatiche inevitabilmente costituitesi.

Si sta passando dall'Internet delle connessioni ad essere umani all'Internet delle cose.

Non si è più solo in presenza di connessioni fra i calcolatori, fra dispositivi mobili ed esseri umani, ma le connessioni riguardano anche, e principalmente, gli oggetti.

Si pensi semplicemente al settore delle automobili, l'invenzione della fine del diciannovesimo secolo forse più diffusa nel ventesimo. Se fino a qualche anno fa la componente meccanica delle automobili si aggirava intorno al 70%, mentre oggi appare una componente minoritaria, anche nei costi, e lo sviluppo dell'elettronica condivisa in rete consente prestazioni impensabili. Tutti conosciamo le automobili a guida elettronica in corso di sperimentazione. Ma anche il guidatore può essere connesso mentre guida con dispositivi applicati alla sua stessa maglietta che ne misurino, ad esempio, il tasso alcolico, le pulsazioni ed altri innumerevoli parametri considerati utili da monitorare.

Pensiamo in campo medico come il monitoraggio della nostra salute possa avvenire con informazioni provenienti da diversi punti di origine, in casa, sotto sforzo etc. I sensori applicati sull'organismo sono in grado di fornire in tempo reale un numero di informazioni enorme e dettagliato, sincroniche e diacroniche al tempo stesso, che qualcuno, o qualcosa, dovrà valutare.

Si pensi all'industria, al settore della grande distribuzione, all'agricoltura che può essere monitorata senza soluzione di continuità, all'ambiente, con applicazioni che potrebbero ridurre,ad esempio, gli enormi rischi connessi al degrado idrogeologico.

Del resto i Big-data si stanno accrescendo in modo esponenziale. Basti pensare che il 90% dei dati scambiati sul web è stato prodotto negli ultimi due anni, e la tendenza continua ad aumentare.

Ed è questo però un problema di fondo che limita l'applicazione dei Big-data ai più diversi settori; il loro stesso, strabocchevole numero, che ne rende difficoltoso un utilizzo coerente.

Immersi in una marea di dati, che possono essere eterogenei, distribuiti da ogni punto possibile, ma poi da raccogliere, memorizzare in luoghi deputati, piattaforme, data center, che tengano conto dei fondamentali problemi della privacy che diventano in tale complessità sempre più stringenti, possiamo smarrirci.

Questo diventa il vero, non eludibile, problema dei Big-data. E' impensabile che a valle della produzione e trasmissione dei dati possano esservi semplicemente operatori umani che li decrittino e li interpretino. Nessun sistema industriale, o di welfare, potrebbe sostenere un costo simile.

Tornando all'esempio precedente dell'utilizzo in campo medico è evidente che la decrittazione e l'interpretazione dei dati a disposizione deve essere compiuta da medici, ma nessun sistema sanitario nazionale potrebbe permettersi di disporre di operatori sanitari in tale numero da poter assicurare il servizio. Non parliamo di quelli privati.

Risulta abbastanza evidente che soltanto utilizzando altre macchine, predisposte per svolgere funzioni di estrazione ed interpretazione dei dati, si potrà evitare di disperdere la grande potenzialità dei Big-data, facendone veicolo per il miglioramento della qualità della vita.

In questo sta il cuore del dibattito che oggi è in corso sul tema, e delle sperimentazioni che si compiono nei laboratori di ogni parte del mondo. Occorre trovare degli algoritmi, delle metodologie che siano in grado di estrarre le informazioni che ci interessano distinguendole dal rumore delle informazioni inutili ai nostri scopi.

A questo riguardo, e il Politecnico di Losanna ce ne fornisce un esempio paradigmatico, sono sorti studi e sperimentazioni sul tema dell'intelligenza artificiale, seppur da decenni esso è in campo, a partire dalle indicazioni dello stesso Von Newman, il padre dei calcolatori, che tentò un'imitazione dei meccanismi del cervello umano, di cui allora, del resto, si sapeva pochissimo.

I computer tradizionali compiono le operazioni di estrazione in modo sequenziale, isolando specifici aspetti dei dati e trattandoli in successione. Non così lavora il cervello umano, dove prevalgono modalità che potremmo chiamare di tipo parallelo, in grado di comparare i dati isolando solo quelli interessanti allo scopo che ci si prefigge.

Si fa spesso l'esempio del riconoscimento facciale. Il computer esamina una dopo l'altra le caratteristiche del volto e formula alla fine il suo giudizio, mentre il nostro cervello è in grado di dedurre da aspetti combinati la riconoscibilità di una persona.

Insomma c'è tanto da fare,e tanto si sta facendo, nei laboratori delle più grandi Università, per arrivare ad una gestione dei Big-data utile all'umanità, servendosi anche del grande patrimonio della cultura umanistica, che sola può essere in grado di uscire dalle strettoie dello specialismo.

Il grande progetto delle Digital Humanities promosso dal Politecnico di Losanna, ed al suo interno la Venice time machine, in collaborazione con l'Archivio di Stato di Venezia e l'Università Ca' Foscari intende esplorare queste nuove frontiere.

La rivoluzione dei Big-data ha solo lambito in questi decenni il mondo degli archivi. In primo luogo esistono applicazioni, non molto diffuse, per la gestione informatica del servizio di sala studio agli studiosi, che di per sé immettono su nuove strade non solo la celerità dello stesso servizio, ma anche la sua efficacia in termini di conservazione dei documenti, dal momento che ogni unità archivistica è tracciata, il suo work-flow conosciuto e ridotte considerevolmente situazioni di smarrimento, o peggio, di furti.

L'Archivio di Stato di Venezia ha in questi ultimi anni messo in opera un sistema di gestione informatizzata del servizio di sala di studio, e dei depositi, molto apprezzato dagli studiosi, il cui cuore pulsante è la banca dati nel sistema, che descrive i fondi archivistici e tutte le loro interne partizioni, in modo che lo studioso non debba indicare di sua iniziativa la segnatura, il che porterebbe fatalmente a richieste del tutto scompensate dei fondi esistenti, ma la debba semplicemente catturare dal sistema, dove il computer fornisce automaticamente tutti i dati necessari. Il grande lavoro non è stata quindi la creazione del software ma la realizzazione di una grande banca dati archivistica, che però è nelle possibilità di ogni Archivio di Stato in Italia, partendo che dagli strumenti che a partire dagli anni sessanta si sono elaborati, quali la Guida generale degli Archivi di Stato. Occorre anche dire che, prevedendo il sistema di Venezia un preciso topografico dei depositi, inserito nella banca dati, il computer informa immediatamente l'addetto del luogo dove è conservato il pezzo richiesto. Questo ha portato ad un dimezzamento delle ore uomo necessarie per il prelievo e la ricollocazione.

Ma l'Archivio di Stato di Venezia si propone nei prossimi anni di fare di più. Dotando, gradualmente, ogni unità archivistica di sensori elettronici sarà possibile debellare per sempre il problema dello smarrimento della documentazione, perché riarchiviata male, che in un grande Archivio, ed in una grande biblioteca, è cosa purtroppo di tutti i giorni.

Veniamo quindi all'affascinante tema dei fondi archivistici quali depositari di infiniti dati, e quindi grandi depositi di Big-data, che nel corso della storia gli archivisti hanno affrontato con metodologie sempre più complesse e raffinate, e che hanno prodotto grandi risultati.

Come tutti sappiamo la via analogica per la ricerca archivistica è stata quella di partire dai soggetti produttori, dalle Istituzioni, dalle famiglie, dalle persone, Tali soggetti hanno creato naturalmente i propri archivi, li hanno organizzati secondo criteri utili alla propria attività ed al reperimento celere delle informazioni, con l'attenzione rivolta sempre al rapporto costi benefici di ogni intervento.

Gli archivisti hanno rispettato questi originari ordinamenti, introducendovi degli elementi di sistematizzazione e descrizione specialistici che non hanno mai inteso sovvertire il criterio analogico di navigazione attraverso i fondi, sorretti dalla conoscenza della storia delle Istituzioni e dalla conoscenza, altrettanto importante ma troppo spesso trascurata, delle forme che gli

archivi stessi hanno assunto nel tempo e delle caratteristiche degli strumenti di corredo di cui si sono dotati.

Operazioni massive di riproduzione digitalizzata di documentazione archivistica, e creazioni di ampie banche dati relative a specifici oggetti documentari sono state condotte nel mondo degli Archivi, ed hanno fatto evidenziare positività e criticità.

Se il dato archivistico ha pieno significato ove inserito in un contesto è evidente che le operazioni di riproduzione documentaria non possono prescindere da una progettazione unitaria dell'operazione complessiva, a partire da un necessario ordinamento dei fondi, da una corretta descrizione archivistica a partire dagli standard internazionali, da una conoscenza delle modalità di organizzazione originarie della documentazione, che va rispettata, e degli strumenti che in quella sede si erano venuti formando.

Però il più vasto, in gran parte potenziale, pubblico interessato ai dati della storia difficilmente intende sottoporsi a lunghi studi istituzionali ed a percorrenze defatiganti. Ed ecco che le banche dati e le digitalizzazioni si sono concentrate su aggregati documentari nati ordinariamente come omogenei, ma che non sono in senso stretto né fondi né serie archivistiche.

Così l'anagrafico, con informazioni tratte da registri dei nati e dei morti, dagli stati delle anime, dalle liste di leva, ma potenzialmente allargabili ad un numero indefinito di fonti.

Così le censuazioni fiscali e catastali, che hanno interessato per il periodo moderno le dichiarazioni dei proprietari, essendo anagrafica la base delle rilevazioni. Nella collaborazione fra L'Archivio di Stato di Venezia ed il Politecnico di Losanna si sta portando avanti in questi anni la creazione di una banca dati e immagini sul catasto veneziano, che parte già dalla messa in rete per un presedente lavoro dell'Archivio di Stato, della redecima del 1514, e vuole portarsi fino al periodo napoleonico, utilizzando sistemi automatici di lettura e collegamento fra le informazioni.

A partire dal periodo napoleonico, e anche prima nella Lombardia austriaca, la disponibilità di catasti geometrico-particellari ha portato a privilegiare le mappe, ed è sorta un'ampia letteratura sulle modalità di ripresa e la loro aderenza agli originali per gli aspetti censuari.

Le riproduzioni di aggregati archivisti di altri fondi, quali i registri delle deliberazioni dei *Consilia* che stiamo portando avanti a Venezia, si sono avvalse degli strumenti originari di ricerca per orientarsi nel mare magnum delle riproduzioni.

Il limite più difficilmente aggirabile delle esperienze predette è il costo delle stesse operazioni.

Estrarre dati da documenti archivistici significa disporre di personale altamente qualificato, che sia in grado di effettuare controlli analitici sui nomi, sui luoghi, sulle cose da censire senza incorrere in errori sempre in agguato e possa compiere contestualmente gli opportuni collegamenti con il complesso archivistico.

Quando i dati devono essere estratti da diversi fondi, e contesti completamente diversi, i costi si moltiplicano.

Molti Laboratori nel mondo si sono posti l'obiettivo di sperimentare sistemi di lettura automatica dei documenti manoscritti e di gestione dei dati all'interno dei software dedicati.

Anche il Politecnico di Losanna ha dato il via ad un'iniziativa analoga, ed è prossima la firma di un protocollo d'intesa con l'Archivio di Stato di Venezia per sperimentare su fondi archivistici significativi le modalità di intervento in esame. Ce ne parlerà il prof. Frederick Kaplan in questo convegno.

Molto dispendiose anche le riproduzioni in senso proprio, ed il costo maggiore è sempre quello del personale che deve materialmente operare su scanner, aprendo i documenti ed i registri e sistemandoli in modo che non si danneggino.

Analogamente fra Losanna e Venezia è in corso una sperimentazione per la riproduzione di filze manoscritte o registri attraverso la Tomografia assiale computerizzata, che permetta di leggere e riprodurre pienamente ogni singola carta, senza aprire o manipolare il documento. Sono state già condotte sperimentazioni su nostra documentazione notarile e la prof.ssa Fauzia Albertin ce ne darà preziosi ragguagli.

C'è poi il capitolo della conservazione a lungo termine di tutto il complesso dati-immagini di cui ci parleranno illustri relatori in questo convegno, che ascolteremo con grande interesse.

Si pone, inevitabilmente, a fronte dei costi presumibili, il problema della partecipazione della comunità al costo complessivo, attraverso il pagamento dei servizi, o di parte di essi, che non possono essere considerati semplicemente attinenti al welfare universale, defalcando il diritto allo studio che mai deve essere messo in discussione.

Concludo queste mie brevi osservazioni ,che non vogliono rubare tempo al dibattito, con una considerazione: esiste nella società civile una richiesta enorme, amplissima, spesso inespressa, di informazioni sulla propria storia, della propria città, del proprio paese, e dell'intera comunità mondiale nell'epoca della globalizzazione.

Attrezzarsi non solo per creare Big-data ma anche per diffonderli è un compito che il mondo archivistico deve porsi, ma per farlo deve interfacciarsi con la ricerca scientifica e tecnologica, studiando meccanismi che, in analogia con altri settori di cui si accennava, vedano la presenza di macchine in grado di produrre ed utilizzare gli stessi Big-data.

L'esperienza di collaborazione fra l'Archivio di Stato di Venezia, il Politecnico di Losanna e Ca' Foscari può essere forse un momento di inizio significativo in tal senso.

Raffaele Santoro

Direttore dell'Archivio di Stato di Venezia