Big Data in the Humanities

Quando la massa dei dati è troppo grande per un'analisi manuale e lo studio di queste fonti necessita di nuovi metodi e nuovi modelli interpretativi

Viral Texts:
nuove risposte dai Big Data culturali

- **Ryan Cordell dirige una ricerca che utilizzando metodi computazionali esamina - nell'ambito del progetto di "viral texts" - milioni di pagine digitalizzate di quotidiani USA del XIX secolo**

- **La ricerca grazie alla sua capacità di analizzare in poco tempo molti più dati di quanti ne possa studiare in una vita intera un singolo ricercatore, è in grado di far emergere nuovi modelli interpretativi con la conseguente scoperta di fenomeni precedentemente non percepibili in regime di scarsità di dati, come – ad esempio – le caratteristiche "virali" di testi appartenenti a vari generi popolari: poesie, racconti, notizie, appunti di viaggio, discorsi politici, vignette ecc.**

Network of "Viral Text" Sharing, 1836-1899

Il grafico illustra i collegamenti tra i giornali pre-guerra civile dall'archivio dei giornali Chronicling America del Library of Congress e l'archivio delle riviste di Making of America. I cerchi (nodi) del grafico rappresentano singoli quotidiani, mentre le linee tra loro (bordi) rappresentano testi condivisi scoperti dai ricercatori che lavorano sul progetto Viral Texts

Informatica e scienze umane

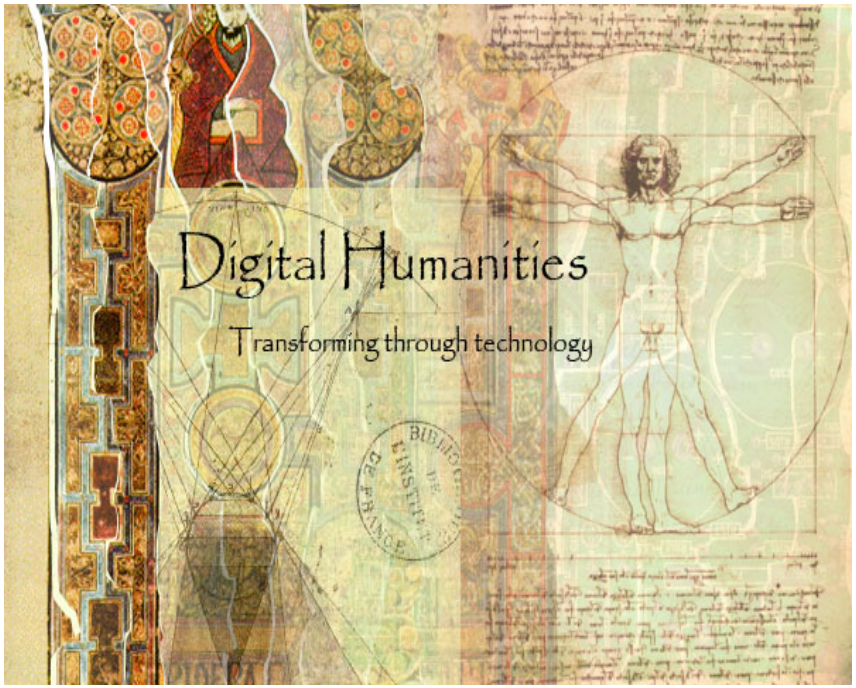
- **Rapporto molto antico basato su i concetti di linguaggio e comunicazione**

- **Inoltre, il legame è stato rinforzato dal fatto che il computer fa parte di quelle tecnologie definite “caratterizzanti” (come il libro e l’orologio), cioè**

    - **In grado di stabilire legami metaforici con la cultura del suo tempo**

Digital Humanities?

- **E’ un campo di studi, ricerca, insegnamento che nasce dall'unione di discipline umanistiche e informatiche.**
- **Comprende ricerca, analisi e divulgazione della conoscenza attraverso strumenti informatici.**
- **È attualmente oggetto di una profonda trasformazione che ha messo in moto – alla luce degli ultimi sviluppi delle tecnologie digitali - diverse riconsiderazione riguardo le sue metodologie e i suoi concetti fondamentali**



Digital Humanities?

- **L'evoluzione di questo campo di studi ha visto una fase iniziale definita, Humanities compuntig**
  - **Progetto della prima fase l'Index Thomisticus di padre Roberto Busa**

- **Comunque, ambedue le fasi partono dalla stessa domanda: qual è il contributo che il computer e l'informatica posso dare all'avanzamento delle scienze umane?**

- **La differenza tra Humanities Computing e Digital Humanities è dovuta essenzialmente al progresso dell'innovazione tecnologica**
  - **Dall'avvento di Internet , alla nascita del Web e dell'ipertesto (Ted Nelson "Xanadu)**

Le Digital Humanities davanti ai "Massive cultural digital objects"

- **includono corpus di grandi dimensioni come:**
  - **I milioni di libri digitalizzati da Google Books**
  - **milioni e milioni di immagini, micro-messaggi ecc. condivisi su servizi di social network**
  - **sistemi di informazione geografica come Google Earth ecc.**

- **In questi casi il tradizionale rapporto 1:1 del singolo studioso di fronte a un documento non è più proponibile**

Studiare
i “Massive cultural digital objects”

- **Cosa può essere estratto da questi enormi data set?**
- **quali interpretazioni si possono ricavare da queste estrazioni?**
- **È possibile estrarre più conoscenza interrogando un data set relativo a 25 milioni di libri digitalizzati oppure può essere più proficuo leggerne soltanto 5 molto attentamente?**

Le fonti di ricerca si «smaterializzano»

- **La digitalizzazione ha aggiunto una nuova dimensione rispetto alla materialità delle fonti tradizionali**
- **Questo ha comportato ricerche su scale più ampie, che comprendono molti altri tipi di fonti**
- **Nel corso dei prossimi anni, i ricercatori e le istituzioni del patrimonio (archivi, biblioteche, audio e video, centri di conoscenza) saranno di fronte a una sfida comune, quella i Big Data culturali, ovvero le nuove fonti: digitalizzazioni, 'digital born' , blog, pagine web, social media ecc.**
- **Nuovi metodi di ricerca automatica devono essere in grado di porre domande e ricavare risposte significative da la nuova gamma di fonti digitali**

“Strutturare” il rapporto tra Big Data e Digital Humanities

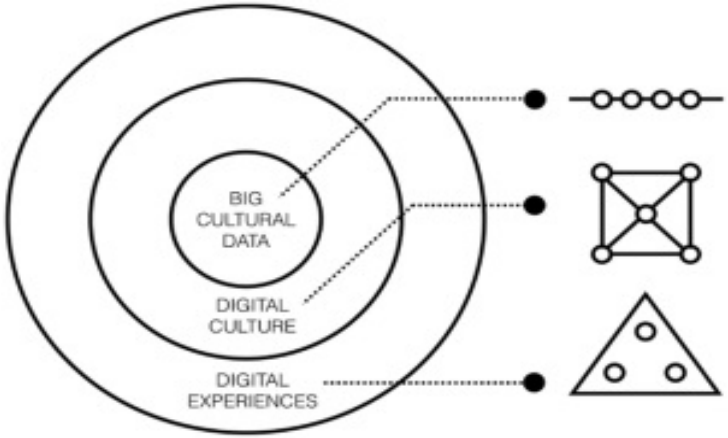
Per anni, i digital humanities sono stati organizzati sulla base di approcci computazionali e sono stati soprattutto oggetto di ricerca e riflessioni critiche riguardanti gli effetti delle tecnologie digitali su cultura e conoscenza

Secondo uno studio del Digital Humanities Laboratory (DHLAB), de Lausanne, Lausanne, Switzerland:

- **l'elaborazione e l'interpretazione dei dati deve avvenire in un contesto più ampio, che può essere definito “cultura digitale”**
- **L’approccio “cultura digitale” può essere segmentato in sottodomini corrispondenti a gruppi di relazioni tra le varie entità**

“Strutturare” il rapporto tra Big Data e Digital Humanities

- **I tre cerchi illustrano tre livelli di contestualizzazione e realizzazione di Big Datai culturali.**
- **Il primo cerchio contiene le ricerche sui grandi database culturali e sul nuovo tipo di comprensione che questi database consentono.**
- **Il secondo cerchio corrisponde alla ricerca sull'interdipendenza tra il discorso collettivo, le comunità su larga scala, il software di mediazione e gli attori IT globali che si può verificare nel contesto di ciò che può essere chiamato in larga misura "Cultura Digitale".**
- **L'ultimo cerchio contiene ricerche sulle nuove esperienze digitali nell’ambito di grandi set di dati culturali nel mondo fisico.**



Frédéric Kaplan

Digital Humanities Laboratory (DHLAB), École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

Venice Time Machine: esempio di un progetto «Digital Humanities»

- **I ricercatori del Digital Humanities Laboratory (DHLAB) hanno sviluppato le tecnologie e l'infrastruttura**
- **Per convertire in formato digitale l'enorme quantità di manoscritti amministrativi dell' Archivio di Stato di Venezia**
- **Realizzando una ricerca storica che è anche un nuovo tipo di sistema informativo**

Big Data in the Humanities: the Need for Big Questions

- **Barry C Smith direttore dell'Istituto di filosofia di Londra, dal suo Blog si è posto alcune domande importanti:**
  - **Le scienze umane, per loro natura, sono focalizzate sul particolare, su specifici oggetti, sulla piccola scala corrispondente a persone, luoghi e cose, considerate quasi sempre rispetto al loro peculiare contesto storico**
  - **L'approccio attuale dei Big Data in the Humanities si limiterebbe spesso al tentativo di cercare un senso nella complessità di dati non strutturati attraverso tools di visualizzazione**
  - **Inoltre la ricerca umanistica finora si è servita per lo più di dati provenienti da fonti cartacee digitalizzate, ma cosa cambierà quando gli oggetti di studio diventeranno sempre più "born digital" ?**
  - **Quando gli storici si troveranno immersi nella "Datafication"? Quando, insomma, eventi, persone, luoghi e cose saranno sempre di più trasformati in un coacervo di "dati informatici"? A quel punto sarà inevitabile il salto di paradigma conflittuale rispetto ai metodi tradizionali di ricerca?**
  - **Le scienze umane potrebbero – a breve – dover fare i conti con una difficile scelta tra studiare le cause o le correlazioni tra dati**
  - **Sarà davvero possibile trasformare l'esame automatico di grandi datasets in conoscenza, in una nuova visione del mondo altrettanto valida rispetto a quella costruita nei secoli sulla "causalità"?**

Gestire e strutturare i Big Data della cultura: i progetti europei

- **Aggregatori dei “big data della cultura”:**
  - **Europeana**
  - **European Library**
- **e-Infrastructure per la gestione dei “big data della cultura”**

Europeana

- **E’ il più grande aggregatore di dati culturali provenienti da diverse istituzioni dei 28 paesi membri dell'Unione europea in 30 lingue**

- **Ha un data set di oltre 50 milioni di metadati (2017) che descrivono e rimandano a libri, audiovisivi, immagini, manoscritti, documenti sonori, applicazioni in 3D ecc.**

- **E’ un portale multilingue dietro a quale opera un sistema di aggregazione e indicizzazione di dati afferenti a risorse digitali provenienti da oltre 1500 istituzioni – per la maggior parte musei, biblioteche, archivi, centri culturali, ecc. - sparse sul territorio europeo**

Europeana:
“armonizzare” i dati e gestione diritti

- **Modello dati EDM**
  - **uno standard intermedio studiato appositamente per rendere più semplice il processo di trasformazione dei metadati in entrata e nello stesso tempo di garantire la persistenza dei record anche in caso di future evoluzioni degli standard di riferimento**
  - **Utilizzato come modello unico, rappresenta un approccio mirato all’armonizzazione dei dati in modo non monolitico attraverso l’applicazione dei principi del web semantico, con l’obiettivo di integrare – in un ambiente aperto – i vari modelli utilizzati all’intero dei datasets del patrimonio culturale europeo**
- **La gestione dei diritti**
  - **licenza CCO 1.0 Public Domain che rende possibile di utilizzare i metadati pubblicati per qualunque scopo, compresi scopi commerciali**
  - **Unica condizione richiesta ai fornitori è la sottoscrizione dell’accordo DEA che comunque concede libertà alle istituzioni che rilasciano i propri dati di decidere quanti e quali (limitando ad esempio i permessi solo ad alcuni set di metadati) rendere interoperabili attraverso Europeana.**
  - **il riuso aperto dei dati. Una politica con ricadute importanti per tutti i soggetti pubblici e privati aggregati al progetto: maggiore visibilità e traffico verso i rispettivi siti, sviluppo di servizi innovati e stimolo per l’industria creativa, maggiori opportunità di generare entrate mediante distribuzione di risorse digitali ecc**

European Library

- **E’ un progetto finalizzato all’integrazione dell’universo bibliotecario europeo**
- **Offre l'accesso attraverso Internet alle risorse di 48 biblioteche nazionali europee.**
- **E’ possibile la navigazione tra oltre 28 milioni di oggetti digitali e circa 175 milioni di records bibliografici**
- **Le risorse, sia digitali che tradizionali, includono libri, riviste, giornali e altro materiale. È possibile ricercare e scaricare materiali digitali: alcuni gratuiti, altri a pagamento.**

European Library: l'integrazione dei dati bibliografici

- **Ad esempio, il caso del multilinguismo**
  - **le classificazioni a soggetto delle biblioteche usano sistemi dipendenti dalle varie lingue**
    - **Progetti per allineare i vocabolari**
      - **Tra i metodi utilizzati, c'è l'accesso multilingue MACS che copre inglese, francese e tedesco**

“Europeana 1914-1918”: altro esempio di integrazione dati a livello europeo

- **Un importante progetto di digitalizzazione e pubblicazione di fonti storiche primarie e secondarie della Prima Guerra Mondiale**

- **L’European Library ha messo a disposizione metadati sulla Prima Guerra Mondiale provenienti da 11 istituti d’Europa.**

- **Per rendere possibile l’accesso multilingue, si è partiti dalle intestazioni a soggetto della Library of Congress espandendole poi mediante le traduzioni nelle lingue dei vari fornitori di contenuti.**

Infrastruttura per la gestione dei “big data della cultura”: esempio CLARIAH

- **E’ un progetto sostenuto da un consorzio di 40 istituzioni della conoscenza e del patrimonio, enti pubblici e le aziende dell‘organizzazione olandese per la ricerca scientifica (NWO)**

- **Lo scopo è una infrastruttura di ricerca per le arti e le scienze umane**

- **CLARIAH fornisce ai ricercatori:**
  - **Accesso a grandi raccolte di dati digitali**
  - **Accesso ad applicazioni innovative e di facile utilizzo per il trattamento di tali dati.**
  - **Data curation: dati e applicazioni vengono gestiti in maniera sostenibile in modo che possano essere utili anche in futuro per i ricercatori**

La gestione dei dati in CLARIAH

- **La crescente disponibilità di enormi quantità di dati digitali è una delle ragioni principali della progettazione di una infrastruttura come CLARIAH**

- **Le enormi quantità di dati rendono impossibile ricercarli in modo tradizionale. Il ricercatore deve utilizzare software di ricerca per trovare parti potenzialmente rilevanti e ignorare quelle irrilevanti, e anche per eseguire analisi dei dati**

- **I dati sono disponibili in molti tipi. I tipi principali sono i testi della lingua naturale, i dati audio-visivi ei dati strutturati (database). Tutti e tre i tipi sono rappresentati in CLARIAH**

Big data della cultura: progetto Library of Congress

Il 16 maggio la Library of Congress ha annunciato che ha reso disponibili 25 milioni di records del suo catalogo

25 Million Free Records of Bibliographic Metadata

LIBRARY OF CONGRESS

Big data della cultura: progetto Library of Congress

- **Il dataset messo a disposizione copre 45 anni di attività della LOC: dal 1968 al 2014**
- **Ogni record fornisce informazioni standardizzate tra cui titolo, autore, anno pubblicazione, soggetto, note ecc.**
- **I dati riguardano una vasta gamma di documenti della biblioteca: libri, file computer, manoscritti, mappe, musica e materiali visivi.**
- **Il rilascio del dataset prevede due modalità:**
  - **Accessibilità gratuita mediante il sito data.gov**
  - **Distribuzione a pagamento in formato MARC per grandi clienti commerciali e biblioteche di tutto il mondo**

Riuso dei dati e innovazione: progetto Library of Congress

Beacher Wiggins, direttore della LOC per le acquisizioni e l'accesso bibliografico:

«Oltre al loro tradizionale valore, i dati bibliografici rilasciati potranno essere utilizzati per una vasta gamma di ricerche culturali, storiche e letterarie… da una più efficiente condivisione delle informazioni alla visualizzazioni e e altre possibilità che possiamo cominciare a prevedere…speriamo che questi dati vengano analizzati da scienziati sociali, analisti di dati, sviluppatori, statistici e tutti coloro che possono fare un lavoro innovativo con grandi set di dati per migliorare l'apprendimento e la formazione di nuove conoscenze "