POLITECNICO
MILANO 1863

Scuola del Design
Design della Comunicazione

Visualizzare *Internet Archive* attraverso la visualizzazione diretta

Relatore
Gabriele Colombo

Tesi di Laurea Magistrale di
Lucia Ferrari
matr. 916789

Anno Accademico 2020–2021

Indice

Indice delle figure

Cap. 2 Visualizzare *Internet Archive*

Cap. 3 Visualizzazione delle informazioni e visualizzazione diretta

Abstract

La nostra società diventa ogni giorno sempre più digitalizzata: in un panorama ormai dominato dalle nuove tecnologie, la nostra cultura produce sempre più artefatti in forma digitale, in una continua interrelazione tra vecchi e nuovi media.

In questo contesto, *Internet Archive*, una biblioteca digitale non profit che offre uno spazio digitale permanente per l'archiviazione di diversi tipi di risorse, si è concentrato sulla garanzia della disponibilità e accessibilità universale della conoscenza umana, archiviando ogni giorno contenuti da tutto il mondo, visualizzabili in qualsiasi momento da chiunque.

La quantità di materiale che viene così reso disponibile è sempre più consistente. In presenza di una disponibilità esorbitante di artefatti digitali e digitalizzati, l'accesso e l'analisi di questo patrimonio può essere di difficile gestione. Uno dei modi più efficaci per aiutare le persone ad affrontare il sovraccarico di informazioni è quello di provare a visualizzarle: mappando i dati visivamente è più facile individuare le informazioni importanti e capire le correlazioni più significative che diversamente sarebbero difficili da scoprire.

In questa tesi di ricerca, dopo aver illustrato le caratteristiche e i servizi offerti da *Internet Archive*, vengono analizzati alcuni casi studio che utilizzano la visualizzazione delle informazioni per rappresentare i materiali digitali presenti nell'Archivio.

L'analisi dei casi studio fa emergere alcune criticità nella visualizzazione di *Internet Archive*: la navigazione non è semplice o intuitiva; inoltre, non sempre l'interfaccia grafica utilizzata e le visualizzazioni che da essa derivano sono le più adeguate a rappresentare il risultato della ricerca, essendo in molti casi troppo specifiche e per questo non mutuabili per un uso più generalizzato su qualunque tipo di materiale presente in *Internet Archive*.

Alla luce di quanto analizzato e delle criticità emerse, verrà infine proposto un progetto di implementazione dell'attuale interfaccia di *Internet Archive*. Con l'obiettivo di superare i limiti individuati e migliorare la ricerca e la visualizzazione dei risultati, verranno presentate nuove modalità di visualizzazione dei materiali digitali archiviati, che consentano all'utente di analizzare l'elevata quantità di informazioni disponibili e di individuare le informazioni chiave, rappresentandole adeguatamente.

L'approccio scelto per la progettazione è quello della della visualizzazione diretta — un metodo che ha avuto un certo sviluppo a partire dai primi anni 2000 — che nella visualizzazione dei risultati di ricerca non riduce gli artefatti dell'archivio in elementi grafici astratti, ma crea invece nuove rappresentazioni visive in cui viene mostrata la forma visiva originale degli artefatti stessi.

Introduzione

In un mondo dove le nuove tecnologie aprono scenari in costante evoluzione e offrono grandi opportunità, assistiamo ad una proliferazione senza precedenti degli artefatti digitali, accompagnata da una vasta attività di digitalizzazione degli artefatti analogici, che rende centrale il tema della conservazione delle risorse culturali.

Lo scopo per cui *Internet Archive* è stato creato 25 anni fa è quello della archiviazione in modo permanente di contenuti digitali e digitalizzati, puntando inoltre sulla possibilità di renderli disponibili per qualunque utente in qualunque momento.

L'accesso ad una disponibilità così ampia di materiali può essere complicato e in qualche caso non permette di sfruttare le potenzialità offerte da *Internet Archive*.

Questa tesi di ricerca affronta diversi argomenti con l'obiettivo di descrivere lo stato dell'arte di *Internet Archive*, analizzare i diversi modi in cui sono stati valorizzati i materiali dell'Archivio, individuare nuove opportunità di implementazione dell'interfaccia grafica attualmente usata per visualizzare le collezioni culturali digitali, e realizzare infine un prototipo basato su tali opportunità.

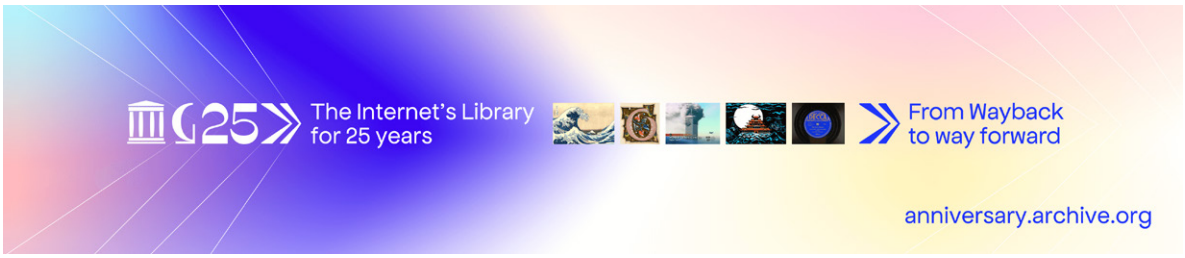
Per quanto riguarda la struttura dei contenuti trattati, nel primo capitolo viene definito l'ambito di lavoro in cui si colloca il tema affrontato in questa tesi, spiegando cos'è *Internet Archive*, descrivendo i vari tipi di risorse culturali in esso archiviate e presentando i diversi progetti sviluppati dall'*Archive Team*.

Nel secondo capitolo vengono analizzati alcuni casi studio che offrono una panoramica su diversi progetti di visualizzazione già realizzati e mostrano come alcune tipologie di materiali sono state valorizzate. Vengono anche evidenziate le criticità di questi progetti, che costituiscono possibili aree di miglioramento nell'utilizzo dei contenuti conservati in *Internet Archive*.

Il terzo capitolo ha l'obiettivo di contestualizzare l'ipotesi di implementazione oggetto di questa tesi di ricerca. Dopo aver richiamato il concetto di visualizzazione delle informazioni, viene introdotto l'ampio studio di Lev Manovich *What is visualisation?* (2011), che spiega la pratica della visualizzazione diretta, nella quale gli artefatti digitali vengono visualizzati nella loro forma visiva originaria. Vengono poi presentati tre suoi importanti progetti che utilizzano

questo approccio per la visualizzazione di differenti materiali digitali e forniscono importanti punti di riferimento per individuare possibili aree di miglioramento nelle visualizzazioni.

Dopo aver descritto il contesto generale in cui si colloca questa tesi e aver analizzato alcuni importanti riferimenti culturali, nel quarto capitolo viene descritto il progetto di implementazione dell'attuale interfaccia grafica di *Internet Archive*, che va nella direzione di migliorare l'accesso e l'esplorazione delle collezioni culturali digitali. Le scelte di design alla base del progetto vengono spiegate e motivate e viene inoltre descritto il processo di realizzazione del *mock up*, presentando nuove modalità che consentono di visualizzare in modo più efficace i risultati delle ricerche avviate per analizzare le collezioni di immagini archiviate in *Internet Archive*, e suggerendo la possibilità di estendere le nuove modalità di visualizzazione a tutti i tipi di materiali digitali presenti nell'archivio.



1. *Internet Archive*

In un panorama ormai dominato dalle nuove tecnologie, la nostra cultura produce sempre più artefatti in forma digitale, in una continua interrelazione tra vecchi e nuovi media.

La maggior parte delle società umane attribuisce grande importanza alla conservazione degli artefatti della propria cultura. Senza tali artefatti, la società non ha memoria e perde la possibilità di imparare dai propri successi ed errori per evolvere e progredire.

L'obiettivo di *Internet Archive* è di aiutare a preservare questi artefatti e di creare una biblioteca Internet per ricercatori, storici e studiosi, ma anche per il grande pubblico.

Per questo *Internet Archive* è stato definito *The archive of them all*, un'"enciclopedia di ogni cosa", e rappresenta una Biblioteca di Alessandria dell'era moderna.

Fin dalla sua nascita, *Internet Archive* si è concentrato sulla garanzia della disponibilità e accessibilità universale della conoscenza umana, creando una biblioteca digitale per archiviare permanentemente contenuti digitali da tutto il mondo, visualizzabili da chiunque in qualsiasi momento.

In aggiunta alla sua funzione primaria di archiviazione, *Internet Archive* è un'organizzazione che si batte per un Internet libero e aperto.

1.1. Cos'è *Internet Archive*

Internet Archive è una biblioteca digitale non profit che ha lo scopo dichiarato di consentire un accesso universale alla conoscenza. Essa offre uno spazio digitale permanente per l'accesso a diversi tipi di risorse: siti web, libri, audio, immagini e immagini in movimento, software, musica e altri artefatti culturali in forma digitale.

Internet Archive fu fondata da Brewster Kahle nel 1996 e fa parte della International Internet Preservation Consortium (IIPC), un'organizzazione internazionale di biblioteche e altre istituzioni che coordina gli sforzi di conservazione dei contenuti Internet.

Gli uffici amministrativi hanno sede a San Francisco, mentre i centri elaborazione dati sono collocati a San Francisco, a Redwood City e a Mountain View, in California. La società conta 200 dipendenti, molti

dei quali impegnati nella scansione di volumi cartacei presso i centri specializzati.

La più massiccia raccolta digitale della biblioteca è l'archivio web, una sorta di collezione di "fermi immagine" del *World Wide Web* catalogati secondo la data di acquisizione. Per assicurare la stabilità e la sicurezza dei dati archiviati, l'intera collezione ha un mirror nei server della *Bibliotheca Alexandrina* ad Alessandria d'Egitto. L'archivio permette a ricercatori, storici, studiosi e al pubblico in generale il caricamento e lo scaricamento di materiale digitale da e verso i suoi server a costo zero.

Internet Archive ha un budget annuale di circa 10 milioni di dollari, derivanti in massima parte da una varietà di fonti: i profitti dei servizi riguardanti il *web crawling*, collaborazioni varie, sovvenzioni, donazioni, e la Kahle-Austin Foundation, che supporta fortemente l'idea che il software debba essere libero.

Al suo esordio, l'organizzazione ha iniziato ad archiviare lo stesso Internet, un mezzo che aveva appena iniziato a crescere. Come i giornali, il contenuto pubblicato sul web era effimero, ma — a differenza dei giornali — nessuno lo stava salvando. Oggi ci sono più di 20 anni di storia del web accessibile attraverso la *Wayback Machine*, l'interfaccia web utilizzata per l'estrapolazione dei dati, e una partnership con più di 625 biblioteche e altri partner per identificare importanti pagine web attraverso l'*Archive-It Program*.

I dati aggiornati a novembre 2021 mostrano che il catalogo ufficiale, in continua espansione, comprende:

- 630 miliardi di pagine web;
- 33,6 milioni di libri e testi;
- 14 milioni di registrazioni audio (di cui più di 240.000 concerti dal vivo);
- 7,3 milioni tra video, film e trasmissioni televisive (compresi più di 2 milioni di programmi di notizie televisive);
- 4 milioni di immagini;
- 790.000 programmi software.

Poiché *Internet Archive* è una biblioteca, particolare attenzione è data ai libri. Non tutte le persone hanno accesso a una biblioteca pubblica o accademica, quindi per fornire un accesso universale è necessario fornire versioni digitali dei libri.

Nel 2005 è iniziato quindi un programma per digitalizzare i libri — uno dei programmi di digitalizzazione più vasti al mondo — e oggi vengono scansionati 1.000 libri al giorno in 28 località in tutto il mondo. I libri pubblicati prima del 1923 sono disponibili per il download, e centinaia di migliaia di libri moderni possono essere presi in prestito attraverso il sito *Open Library*.

Internet Archive serve milioni di persone ogni giorno ed è uno dei primi 300 siti web del mondo. L'archivio è finanziato attraverso

donazioni, sovvenzioni e fornendo servizi di archiviazione web e digitalizzazione dei libri per i partner. Come per la maggior parte delle librerie, la privacy dei clienti è ritenuta fondamentale e viene protetta non conservando l'indirizzo IP dei lettori e utilizzando per il sito il protocollo HTTPS (*Hypertext Transfer Protocol Secure*).

1.2. Le sezioni di *Internet Archive*

Come abbiamo detto, *Internet Archive* raccoglie miliardi di artefatti culturali in forma digitale. L'archivio è organizzato in sezioni differenti a seconda della natura della risorsa digitale archiviata.

1.2.1. *eBooks and Texts*

Internet Archive offre quasi 34 milioni di libri e testi liberamente scaricabili. Vi è anche una collezione di 2,3 milioni di eBook moderni che possono essere presi in prestito da chiunque abbia un account *archive.org*.

I libri su *Internet Archive* sono offerti in molti formati, inclusi i file DAISY destinati a persone con disabilità di stampa, cioè quelle persone che non possono leggere efficacemente la stampa a causa di una disabilità visiva, fisica, percettiva, dello sviluppo, cognitiva o dell'apprendimento.

L'archivio ha digitalizzato oltre 4 milioni di libri, e tra questi collezioni di monografie, collezioni seriali, materiali d'archivio, mappe, diari e fotografie. La digitalizzazione viene fatta in oltre 33 centri di scansione globali, che si trovano in 4 continenti.

Dal 2005, *Internet Archive* ha collaborato con oltre 1.100 istituzioni bibliotecarie e altri fornitori di contenuti, con i quali ha costruito collezioni digitali. Le partnership includono: Boston Public Library, la Library of Congress e la Lancaster County's Historical Society. Queste collezioni sono digitalizzate da diversi tipi di media, tra cui microfilm e microfiche, pubblicazioni seriali, cioè periodici pubblicati in date programmate, e un'ampia varietà di materiale d'archivio.

Contributi significativi sono arrivati da partner in Nord America (biblioteche americane e canadesi), in Europa e in Asia, e sono quindi presenti materiali in oltre 184 lingue.

1.2.2. *Image*

Questa libreria contiene immagini digitali dell'archivio, che vanno dalle mappe, alle immagini astronomiche, alle fotografie di opere d'arte, alle immagini caricate dagli utenti.

Internet Archive ha già pubblicato oltre 4 milioni di immagini, ma ci si aspetta un incremento ancora più significativo grazie al fatto che ogni nuova immagine pubblicata nei libri scansionati da *Internet Archive* — circa 1.000 ogni giorno — entrerà a far parte dell'archivio.
Il progetto di scansione delle immagini dei libri presenti in *Internet Archive* è opera di Kalev Leetaru, ricercatore presso la Georgetown University. Leetaru è riuscito a modificare il tradizionale sistema di

scansione dei libri OCR (*optical character recognition*, riconoscimento ottico dei caratteri) facendolo funzionare al contrario: se normalmente l'OCR era impostato per concentrarsi sulle parole e scartare tutto ciò che assomigliava ad una immagine, egli ha scritto un programma diverso che ha attraversato gli stessi file, cercando tutto ciò che sembrava un'immagine, e salvandolo in un file JPG separato.

Un altro importante valore aggiunto è che il software crea metadati per ogni immagine, tra cui il nome del libro, l'anno in cui è stato pubblicato, il suo autore, il suo editore, il soggetto, il numero di pagina della foto, e un certo numero di tag che lo descrivono. Inoltre, il software ha anche salvato il testo prima e dopo ogni immagine, per dare un po' di contesto.

Le immagini sono tutte di pubblico dominio e possono essere liberamente utilizzate.

1.2.3. *Moving Image Archive*

Questa libreria contiene più di 7 milioni di filmati digitali caricati dagli utenti dell'archivio, che vanno dai classici lungometraggi, alle trasmissioni quotidiane di notizie, ai cartoni animati e ai concerti.

Questa collezione è gratuita e aperta a tutti. L'obiettivo di digitalizzare questi film e di metterli online è quello di fornire un facile accesso a una ricca collezione di film d'archivio, incoraggiando l'uso diffuso di immagini in movimento in nuovi contesti anche da parte di persone che potrebbero non averle usate prima.

1.2.4. *The Internet Archive Software Collection*

Internet Archive Software Collection è la più grande libreria di software d'epoca e storici esistente al mondo, che fornisce l'accesso immediato a oltre 790.000 programmi, immagini CD-ROM, documentazione e multimedia.
La collezione include una vasta gamma di materiali relativi al software, tra cui *shareware*, *freeware*, video news sui titoli software, *speedrun* di videogiochi, anteprime e promozioni per videogiochi, *replay* di record e abilità di vari generi di videogioco, e l'arte del cinema in tempo reale su motori grafici.
Oltre a questo, il progetto cataloga anche altre risorse informatiche e di gioco come software e manuali hardware, riviste e cataloghi informatici.

Vediamo nel dettaglio alcune collezioni particolarmente rilevanti.

The Old School Emulation Center (TOSEC) è un'iniziativa di retrocomputing dedicata alla catalogazione e conservazione di software, firmware e risorse per microcomputer, mini-computer e console per videogiochi. L'idea principale del progetto è quello di catalogare e controllare vari tipi di software e immagini firmware per questi sistemi. TOSEC cataloga oltre 200 piattaforme di calcolo uniche ed è in continua crescita. Il progetto ha identificato e catalogato oltre 450.000 diverse immagini e set di software, consistenti in oltre 3,60 TB di software, firmware e risorse.

Shareware CD-ROMs raccoglie migliaia di CD-ROM *shareware* del periodo di massimo splendore del CD-ROM — dalla fine degli anni '80 alla metà degli anni 2000 — e fornisce immagini ISO e link all'interno di queste collezioni di software. Con oltre 2.500 dischi ora ospitati, l'archivio consente l'accesso a un'ampia gamma di collezioni storiche, come la curatela di gruppi non più in attività, come i Walnut Creek, e distribuzioni Linux/Unix del passato. Altri CD-ROM includono immagini e musica digitalizzata, set di documentazione e modifiche di giochi. La maggior parte di questi CD-ROM possono essere scaricati come file ISO o CDR, così come è possibile navigare online attraverso l'interfaccia di elenco file di *Internet Archive*.

Classic PC Games è una raccolta che ospita oltre 4.000 giochi classici per PC degli ultimi 25 anni. La collezione contiene programmi *shareware*, *freeware* e *demo* per Microsoft DOS e Microsoft Windows.

1.2.5. *Internet Archive TV News*

Come Internet, anche la televisione è un mezzo effimero. *Internet Archive* ha iniziato ad archiviare programmi televisivi alla fine del 2000, e il primo progetto televisivo pubblico è stato un archivio di notizie televisive riguardanti gli eventi dell'11 settembre 2001.

Nel 2009 l'archivio ha iniziato a rendere le trasmissioni di notizie del panorama televisivo americano ricercabili tramite le didascalie nel *TV News Archive*. Questo servizio consente ai ricercatori e al pubblico di utilizzare la televisione come riferimento citabile e condivisibile, permettendo di:

- cercare le didascalie chiuse dei notiziari televisivi statunitensi;
- visualizzare clip da 2.190.000 spettacoli a partire dal 2009;
- condividere clip show personalizzati e tenere traccia della loro popolarità;
- prendere in prestito un DVD di qualsiasi spettacolo completo.

1.2.6. *Audio Archive*

Questa biblioteca contiene 14 milioni di registrazioni audio che vanno dalla programmazione di notizie alternative, ai concerti dei Grateful Dead, agli spettacoli di Old Time Radio, alle letture di libri e poesie, oltre alla musica originale caricata dai nostri utenti. In particolare, *Live Music Archive* è una comunità impegnata a fornire la più alta qualità concerti dal vivo in formato scaricabile, insieme alla comodità dello streaming on-demand. Nel 2002, *Internet Archive* ha collaborato con *etree.org* — una comunità che raccoglie le registrazioni dal vivo di artisti commerciali — per creare *Live Music Archive* al fine di preservare e archiviare il maggior numero possibile di concerti dal vivo per le generazioni attuali e future.

Tutta la musica presente in questa collezione proviene da artisti che hanno aderito all'iniziativa ed è strettamente non commerciale, sia per l'accesso che per qualsiasi ulteriore distribuzione.

1.3. I progetti più rilevanti di *Internet Archive*

L'*Archive Team* è impegnato in una serie di progetti di grande importanza, sia per quanto riguarda la conservazione degli artefatti culturali della nostra società, sia per l'enorme contributo che questi progetti portano alla democratizzazione dell'accesso alla cultura.

Analizziamoli nella loro specificità.

1.3.1. *Wayback Machine*

La *Wayback Machine*, uno dei più importanti progetti di *Internet Archive*, è l'interfaccia web utilizzata da *Internet Archive* per l'estrapolazione dagli archivi dei dati riguardanti siti web. I siti archiviati rappresentano una sorta di "fermi immagine" raccolti al momento dell'acquisizione delle pagine tramite il software di indicizzazione di *Internet Archive*.

La *Wayback Machine* ha iniziato ad archiviare pagine web, con l'obiettivo di rendere pubblico il servizio cinque anni dopo. Dal 1996 al 2001, le informazioni sono state mantenute su nastro digitale, concedendo occasionalmente l'accesso al database in costruzione solo a ricercatori e scienziati. Nel 2001, l'archivio è stato presentato e aperto al pubblico, con una cerimonia presso l'Università Berkeley in California, e al momento del lancio conteneva già oltre 10 miliardi di pagine archiviate.

L'archivio web di *Internet Archive* contiene oltre 2 petabyte di dati compressi, oltre 150 miliardi di web capture, inclusi contenuti da ogni dominio di primo livello, e oltre 200 milioni di siti web, in oltre 40 lingue. Milioni di siti web con i rispettivi dati — immagini, testo, documenti collegati, ecc. — sono archiviati in questo gigantesco database. Non tutti i siti web sono però disponibili a causa della scelta di molti proprietari di escludere i propri siti dall'indicizzazione.

La *Wayback Machine* permette all'utente di andare "indietro nel tempo" e vedere come erano i siti web nel passato. I suoi fondatori hanno sviluppato la *Wayback Machine* per affrontare lo "svanire" dei contenuti di un sito web ogni volta che viene modificato, o quando un sito web viene chiuso, conservando copie archiviate di pagine web non più online. Il servizio consente agli utenti di vedere le versioni archiviate delle pagine web attraverso un lungo arco temporale, che l'archivio chiama "indice tridimensionale".

Oggi, i dati vengono memorizzati sul grande *cluster* di nodi Linux di *Internet Archive*. I siti possono anche essere "catturati" manualmente inserendo l'URL di un sito web nella casella di ricerca, a condizione che il sito web consenta ai *web crawler* della *Wayback Machine* di entrare e salvare i dati.

Il 30 ottobre 2020, la *Wayback Machine* ha iniziato anche a fare il *fact-checking* dei contenuti.

1.3.2. *Archive-It*

Distribuito per la prima volta nel 2006, *Archive-It* è un servizio di archiviazione web in abbonamento che aiuta le organizzazioni a raccogliere, costruire e preservare collezioni di contenuti digitali. I contenuti sono ospitati e memorizzati presso i data center di *Internet Archive*. Tutto il materiale digitale è conservato in duplice copia — una primaria e una di *backup* — e viene periodicamente indicizzato nell'archivio generale di *Internet Archive*. Una copia dei dati può essere mandata ai sottoscrittori su richiesta.

Attraverso questa applicazione web di facile utilizzo, i partner di *Archive-It* possono raccogliere, catalogare e gestire le loro collezioni di contenuti con accesso 24 ore su 24 e possibilità di ricerca *full text* per il loro uso e quello dei loro clienti.

Più di 240 organizzazioni partner, in 46 degli Stati Uniti, e 15 Paesi utilizzano attualmente *Archive-It*, tra cui archivi statali e biblioteche pubbliche, biblioteche universitarie, istituzioni federali, ONG, musei e organizzazioni culturali.

Oltre alla funzionalità di base di catturare e preservare i contenuti web-based, l'applicazione web *Archive-It* consente agli utenti di aggiungere, importare ed esportare metadati descrittivi e consente la navigazione pubblica e la ricerca *full text* tramite *archive-it.org*. *Archive-It* fornisce anche API e altri strumenti per facilitare integrazioni esterne con siti web locali, repository e servizi di ricerca e conservazione di terze parti.

I partner di *Archive-It* beneficiano dell'accesso a una formazione approfondita, alla documentazione e al supporto tecnico da parte degli archivisti web e degli ingegneri professionisti di *Internet Archive*.

1.3.3. *Building Libraries Together*

Come abbiamo già detto, *Internet Archive* è una delle più grandi biblioteche digitali pubbliche del mondo, con una vasta collezione di cultura umana. Tuttavia, gli utenti dell'archivio caricano solo una piccola percentuale di questi materiali e, per preservare la conoscenza del mondo, il pubblico dovrebbe invece essere maggiormente incoraggiato a contribuire.

L'archivio sta quindi intraprendendo un progetto per rendere il sito *archive.org* più *community-driven*, migliorando gli strumenti che permettono alle persone di caricare, descrivere e organizzare gli elementi. Con questi nuovi strumenti, *Internet Archive* spera di democratizzare la conoscenza, dando alle comunità globali la possibilità di salvare, gestire e condividere gratuitamente i loro tesori culturali.

Quello che *Wikipedia* ha fatto per gli articoli dell'enciclopedia, *Internet Archive* spera di fare per le collezioni di media: dare alle persone gli strumenti per costruire insieme le collezioni di media e renderle accessibili a tutti. Il progetto è sostenuto dalla John S. and James L. Knight Foundation con una sovvenzione di 600.000 dollari.

1.3.4. *Open Library*

Open Library, fra i cui fondatori si annovera anche lo scomparso attivista digitale Aaron Swartz, è una biblioteca digitale nata con lo scopo di raccogliere schede per ogni libro mai pubblicato e di catalogarle in un unico database, in contrapposizione al progetto di digitalizzazione *Google Books*.

Il progetto nasce nel 2007 e include oltre 20 milioni di schede catalografiche e libri digitalizzati nel pubblico dominio, interamente accessibili e scaricabili. *Open Library* è un progetto che si basa su software libero e *open source*, il codice sorgente è interamente accessibile dal sito di riferimento.

A partire da giugno 2010, *Open Library* offre anche un servizio di prestito di eBook svolto in collaborazione col distributore di contenuti digitali statunitense OverDrive e con biblioteche americane. Si tratta di una libreria di prestiti digitali gratuita di oltre 2 milioni di eBook, che possono essere letti in un browser o scaricati per la lettura offline.

1.3.5. *Internet Archive Scholar*

La conservazione della documentazione accademica è stata un punto di preoccupazione fin dall'inizio della produzione della conoscenza. Con le pubblicazioni cartacee, la responsabilità ricadeva principalmente sui bibliotecari, ma lo spostamento verso l'editoria digitale e, in particolare, l'introduzione dell'*open access* (OA, accesso aperto) hanno causato ambiguità e complessità. Di conseguenza, l'accessibilità a lungo termine delle riviste non è sempre garantita, e possono anche scomparire dal web completamente.

Uno studio esplorativo fatto nel 2019 da *arXiv*, un archivio ad accesso libero che ad oggi ospita circa 2 milioni di articoli accademici, ha analizzato il fenomeno delle riviste scomparse. Per l'analisi, sono stati consultati diversi importanti indici bibliografici e sono state tracciate le riviste attraverso la *Wayback Machine* di *Internet Archive*.

È risultato che 174 riviste *open access* sono scomparse dal web tra il 2000 e il 2019 e altre 900 sono a rischio, andando a colpire tutte le principali discipline di ricerca e tutte le regioni geografiche del mondo. I risultati sollevano una forte preoccupazione per l'integrità della documentazione accademica e sottolineano l'urgenza di intraprendere un'azione collaborativa per garantire l'accesso continuo e prevenire la perdita di conoscenze scientifiche.

Per soddisfare questa esigenza, *Internet Archive* ha creato la piattaforma di ricerca scientifica *Internet Archive Scholar*, un indice di ricerca *full text*, che include oltre 25 milioni di articoli di ricerca e altri documenti accademici, conservati sui server di *Internet Archive*. Queste collezioni raccolgono copie digitalizzate degli articoli originali pubblicati dal XVIII secolo fino ai più recenti atti di convegni accademici o di accesso e *pre-prints* estratti dal *World Wide Web*.

Il contenuto di questo indice di ricerca si presenta in tre forme:

– contenuti web pubblici negli archivi web di *Wayback Machine* (*web.archive.org*), provenienti sia dalla collezione storica, che scansionati appositamente per garantire l'accesso a lungo termine ai materiali accademici;
– materiale di stampa digitalizzato proveniente da collezioni cartacee e micro-copie acquistate e scannerizzate da *Internet Archive* o dai suoi partner;
– materiali delle collezioni di *archive.org*, inclusi i contenuti di organizzazioni partner, documenti caricati dal grande pubblico e *mirror* di altri progetti.

Il progetto è ancora in versione alpha, ma potrebbe diventare, quando sarà completamente funzionante, un elemento innovativo nella conservazione e consultazione ricerca accademica.

1.3.6. *National Emergency Library*

Nel mezzo della pandemia di COVID-19 che ha chiuso molte scuole, università e biblioteche, l'archivio ha annunciato il 24 marzo 2020 che stava creando la *National Emergency Library* e che avrebbe quindi rimosso le restrizioni sui prestiti che aveva in atto per 1,4 milioni di libri digitalizzati nella sua *Open Library*, restrizioni che avrebbero limitato il numero degli utenti che potevano accedere ai libri. Normalmente infatti il sito permette solo un prestito digitale per ogni copia fisica dei libri disponibili, con l'uso di un file criptato che diventa inutilizzabile al termine del periodo di prestito.

L'intenzione iniziale era quella di tenere aperta questa biblioteca almeno fino al 30 giugno 2020, e comunque fino alla fine dell'emergenza nazionale di COVID-19 degli Stati Uniti.

Internet Archive ha lanciato la *National Emergency Library* per far fronte ad un bisogno globale e immediato senza precedenti di accesso al materiale di lettura e ricerca, a causa della chiusura delle biblioteche fisiche in tutto il mondo, e ha ritenuto di muoversi in modo corretto dal punto di vista legale, sostenendo che stavano promuovendo l'accesso a risorse inaccessibili, nel rispetto dei principi di *fair use*, e che comunque stavano continuando ad attuare la politica di prestito digitale controllato, come era prima della pandemia, il che significa che le copie prestate venivano ancora criptate e quindi per gli utenti non era più facile di prima creare nuove copie dei libri.

Inoltre, al momento del lancio, *Internet Archive* aveva permesso agli autori e ai titolari dei diritti di presentare richieste di esclusione delle proprie opere dalla *National Emergency Library*, se non ne condividevano l'intento.

Internet Archive, d'altra parte, era già stato criticato da autori ed editori per il suo precedente approccio di prestito e, dopo l'annuncio della nascita della *National Emergency Library*, autori ed editori hanno avanzato pesanti critiche, equiparando l'operato della *National Emergency Library* alla violazione del copyright e alla

pirateria digitale, e accusando *Internet Archive* di usare la pandemia di COVID-19 come pretesto per spingersi oltre i confini del copyright.

Sulla base di tutte queste considerazioni, nel giugno 2020 quattro importanti editori di libri hanno intentato una causa legale contro *Internet Archive*, mettendo in discussione il funzionamento della *National Emergency Library* e la validità del copyright del programma di prestito digitale controllato, e sostenendo che le azioni di *Internet Archive* sono state una violazione intenzionale e di massa del copyright.
A seguito della causa, e prima ancora che si arrivasse alla fase processuale, *Internet Archive* ha chiuso la *National Emergency Library* il 16 giugno 2020, ben prima quindi della fine dell'emergenza nazionale per il COVID-19, ancora in corso.

Come parte della sua risposta alla causa legale portata dagli editori, alla fine del 2020 l'archivio ha lanciato una campagna chiamata *Empowering Libraries*, che ha parlato della causa come di una "minaccia per tutte le biblioteche".

Il processo dovrebbe iniziare nel novembre 2021. I possibili danni non sono ancora stati stabiliti ma, stando a quel che riportano diverse testate d'oltreoceano, potrebbe venir chiesto il pagamento di 150 mila dollari di danni per ogni libro presente nella *National Emergency Library*. Viene chiesto, inoltre, che *Internet Archive* non possa più riprodurre e distribuire i lavori protetti da copyright, e che tutte le attuali copie di questi lavori vengano distrutte.
Il che determinerebbe, in pratica, la completa chiusura dell'intera biblioteca.

1.4. Altri progetti di *Internet Archive*

Vi sono altri progetti minori messi in campo dall'*Archive Team* che meritano di essere ricordati, avendo alcuni di essi particolare attenzione all'impatto sociale del libero accesso alla cultura.

Il *Political TV Ad Archive* è un progetto che offre una raccolta consultabile e condivisibile di spot elettorali del 2016 degli Stati Uniti, affiancata da *fact-checking* e informazioni attendibili. In collaborazione con associazioni giornalistiche affidabili, l'archivio fornisce un servizio gratuito per giornalisti, enti pubblici, docenti universitari e il grande pubblico per contestualizzare gli spot. La prima fase del progetto, che ha riguardato le elezioni primarie più importanti, è stata finanziata dalla donazione di 200.000 dollari fatta dalla *Knight News Challenge*, un'iniziativa della Knight Foundation.
Alla *Challenge* ha collaborato anche il Democracy Fund, che ha fornito circa 50.000 dollari per sostenere la formazione congiunta di giornalisti negli stati primari, in partnership con l'American Press Institute.

L'*Open Content Alliance* (OCA) è uno sforzo collaborativo di un gruppo di organizzazioni culturali, tecnologiche, no-profit e governative di tutto il mondo, che ha l'obiettivo di costruire un archivio permanente di testi digitalizzati e materiale multimediale multilingue.

Il progetto *BookServer* fornisce un'architettura aperta per la vendita, il prestito e la distribuzione di libri su Internet. Costruito su standard aperti, il modello *BookServer* consente a una vasta rete di editori, librai, biblioteche e altri operatori di rendere i loro cataloghi di libri disponibili direttamente ai lettori attraverso i loro computer portatili, telefoni o tablet.

Bookmobile è una libreria digitale mobile in grado di scaricare libri di pubblico dominio da Internet via satellite e di stamparli in qualsiasi momento, ovunque e per chiunque. La *Bookmobile* ha viaggiato attraverso gli Stati Uniti, e versioni di essa sono state costruite e utilizzate anche in Egitto e in Uganda.

Il progetto *Offline Archive* ha l'obiettivo di rendere disponibili le collezioni offline. Poiché la missione centrale di *Internet Archive* è garantire l'accesso universale a tutte le conoscenze, e la mancanza di una connessione a Internet è un ostacolo importante a tale obiettivo, *Internet Archive* ha creato questo progetto che lavora indipendentemente dalla disponibilità di Internet.

Il progetto *Open Community Networking* fornisce internet gratuito, cablato ad alta velocità e wireless ai residenti di San Francisco. Il progetto si è evoluto notevolmente dalla sua nascita nel 1997, e attualmente lavora con la città e la contea di San Francisco per fornire internet gratuito ad alta velocità ai residenti a basso reddito di San Francisco. L'obiettivo per il futuro è di replicare questo modello per fornire lo stesso servizio anche ad altre comunità.

Il progetto *NASA Images* fu creato grazie ad uno *Space Act Agreement* tra *Internet Archive* e la NASA, sottoscritto per rendere accessibili al pubblico gli archivi delle immagini, dei video e dei file audio prodotti dall'agenzia nel corso degli anni attraverso un singolo archivio interamente indicizzato e fruibile tramite ricerche. Il sito web fu lanciato nel luglio del 2008 ed è arrivato a contenere oltre 100.000 file.

Grazie al progetto *Scanning Services*, *Internet Archive* può digitalizzare le collezioni degli utenti e fornire accesso libero e aperto, archiviazione a lungo termine, download illimitati e gestione dei file a vita. *Internet Archive* ha già scansionato più di 600 milioni di pagine con partner come la Library of Congress, lo Smithsonian, la New York Public Library, Harvard e il MIT.

Petabox è stato progettato su misura dal personale di *Internet Archive* per archiviare e processare in modo sicuro un petabyte (un milione di gigabyte) di informazioni. L'obiettivo era quello di realizzare un sistema di storage che fosse a bassa potenza, ad alta densità, scalabile e a basso costo. Questo sistema è ora in uso presso le principali istituzioni accademiche e agenzie governative. *Internet Archive* ospita più di 10 petabyte della tecnologia di archiviazione *Petabox* ed è in costante espansione.

Dalla panoramica appena fatta sulle risorse presenti in *Internet Archive* risulta evidente che la quantità di materiale che viene reso disponibile ogni giorno è estremamente consistente e in costante

crescita. In presenza di una disponibilità di artefatti digitali così esorbitante, l'accesso e l'analisi di questo patrimonio può essere di difficile gestione.

Spesso la navigazione non è semplice o intuitiva, e le potenzialità dell'interfaccia grafica, attraverso la quale si avvia una ricerca e dalla quale vengono restituiti output di visualizzazione dei risultati, sono di estrema rilevanza.

Per evidenziare questo aspetto, nel prossimo capitolo analizzerò alcuni progetti che sono stati sviluppati per differenti tipologie di materiali digitali presenti in *Internet Archive*, e che hanno utilizzato differenti modalità di rappresentazione dei dati.

2. Visualizzare *Internet Archive*

La nostra società diventa ogni giorno sempre più digitalizzata: ne sono esempio lo sviluppo espansivo delle nuove tecnologie e l'infiltrazione di Internet in tutte le sfere dell'attività umana, gli archivi online e i database, i cataloghi delle biblioteche e i numerosi progetti di digitalizzazione delle fonti, così come le riviste scientifiche e i libri in formato elettronico. La quantità di materiale che viene reso disponibile tramite diversi servizi Internet, con connessioni sempre più veloci e maggiori possibilità di accesso, è sempre più consistente.

In questo senso *Internet Archive* ha aperto una nuova dimensione della conservazione e dell'utilizzo del patrimonio digitale. Dal momento della sua creazione, questo archivio pubblico ha iniziato a crescere e ad arricchirsi di materiale multimediale in continua espansione. Lo spazio occupato attualmente è enorme, avendo superato la cifra di 70 petabyte di dati: come abbiamo già detto in precedenza, sono 630 miliardi le pagine web collezionate, 33,6 milioni tra libri e testi, 7,3 milioni i filmati, 14 milioni i file audio, 787 mila software e 4 milioni di immagini.

In presenza di numeri così esorbitanti, l'accesso a questo patrimonio può costituire un problema ed essere di difficile gestione. Uno dei modi più efficaci per aiutare le persone ad affrontare il sovraccarico di informazioni è quello di provare a visualizzarle. In parole povere, questo significa utilizzare grafici, organizzare le informazioni in mappe o anche usare i dati per creare un diagramma interattivo. Mappando i dati visivamente, non solo è più facile assimilare e capire le informazioni importanti, ma è anche possibile individuare più velocemente i modelli chiave, le tendenze significative e le correlazioni convincenti che diversamente sarebbero difficili da scoprire. (Payman, 2018)

In questo capitolo prenderemo quindi in considerazione alcuni casi studio che utilizzano la visualizzazione dei dati per fornire l'accesso a sezioni di *Internet Archive*. I progetti analizzati si caratterizzano per modalità di visualizzazione diverse tra loro, sia per il tipo di dati che vengono rappresentati sia per l'output di visualizzazione.

2.1. *The Deleted City*

The Deleted City è un'archeologia digitale del *World Wide Web*, esploso nel XXI secolo. A quel tempo il web era spesso descritto come un'enorme biblioteca digitale che si poteva visitare, o contribuire a costruire, tramite una homepage. I primi cittadini della rete (o *netizens*) presero sul serio il proprio ruolo e costruirono homepage su se stessi e su argomenti di cui erano esperti.

Questi pionieri trovarono il loro nuovo mondo a *GeoCities*, un servizio di *web-hosting* gratuito. Nella sua forma originale, il sito era modellato come una città nella quale si poteva ottenere un "pezzo di terra" libero per costruire la propria casa digitale, collocando le proprie pagine web in un certo quartiere, in base al tema e al contenuto della homepage. *Heartland* — il quartiere destinato a tutte le cose rurali — era di gran lunga il più vasto, ma c'erano quartieri per la moda, per le arti, per argomenti correlati all'Estremo Oriente, per il tech, per le questioni LGBT, per il vino o per il golf. Ad esempio, siti correlati ai computer venivano collocati in *Silicon Valley*, mentre quelli che si occupano di intrattenimento erano assegnati a *Hollywood*; *Wallstreet* era il quartiere dei siti relativi a finanza e affari e *Vienna* era la comunità degli amanti della musica classica. (Ciociola, 2012)

L'accessibilità simbolica della metafora "urbana", combinata con l'uso libero e semplice del servizio, ha favorito la crescita esponenziale di *GeoCities*: nel giugno 1997 era diventato il quinto sito più popolare del web. Intorno al volgere del secolo, *GeoCities* aveva decine di milioni di *homesteaders*, come gli inquilini digitali venivano chiamati, ed è stato acquistato da *Yahoo!* per tre miliardi e mezzo di dollari. Dieci anni dopo, nel 2009, dato che altre espressioni di Internet — come i social network — avevano preso il sopravvento, e gli *homesteaders* avevano lasciato le loro proprietà vacanti dopo la migrazione a *Facebook*, *GeoCities* è stato chiuso e cancellato.

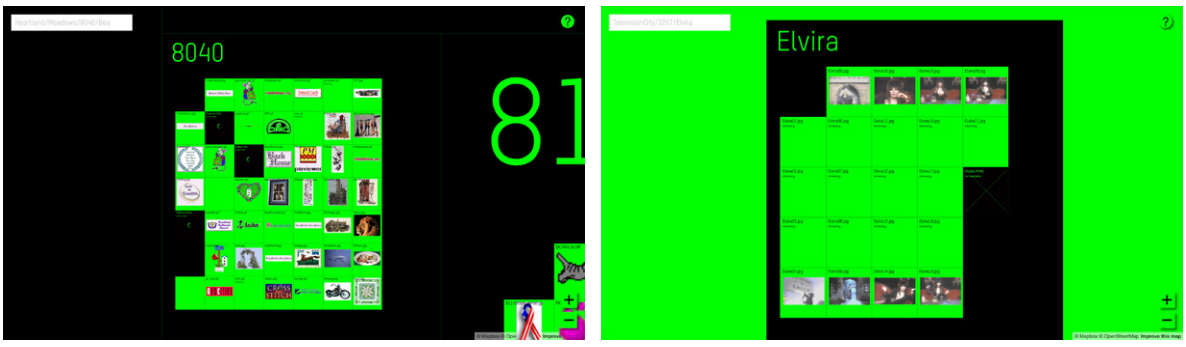


Fig. 2.1 *The Deleted City*: dettagli

In uno sforzo eroico per reservare 10 anni del lavoro collaborativo di 35 milioni di persone, l'*Archive Team* ha fatto un *backup* del sito poco prima che chiudesse, conservando decine di milioni di pagine, e l'artista Richard Vijgen ha creato una visualizzazione interattiva del *backup* da 650 gigabyte di *GeoCities*. È una risposta artistica a una domanda destinata a crescere di importanza, dato che sempre più tanta parte della vita quotidiana viene messa online: come si fa a

presentare una rovina digitale? Il risultato della visualizzazione, che si può osservare sul sito del progetto *The Deleted City*, è una "Pompei digitale", come la definisce Vijgen, cioè uno scavo interattivo che permette di spaziare attraverso un episodio di storia online recente.

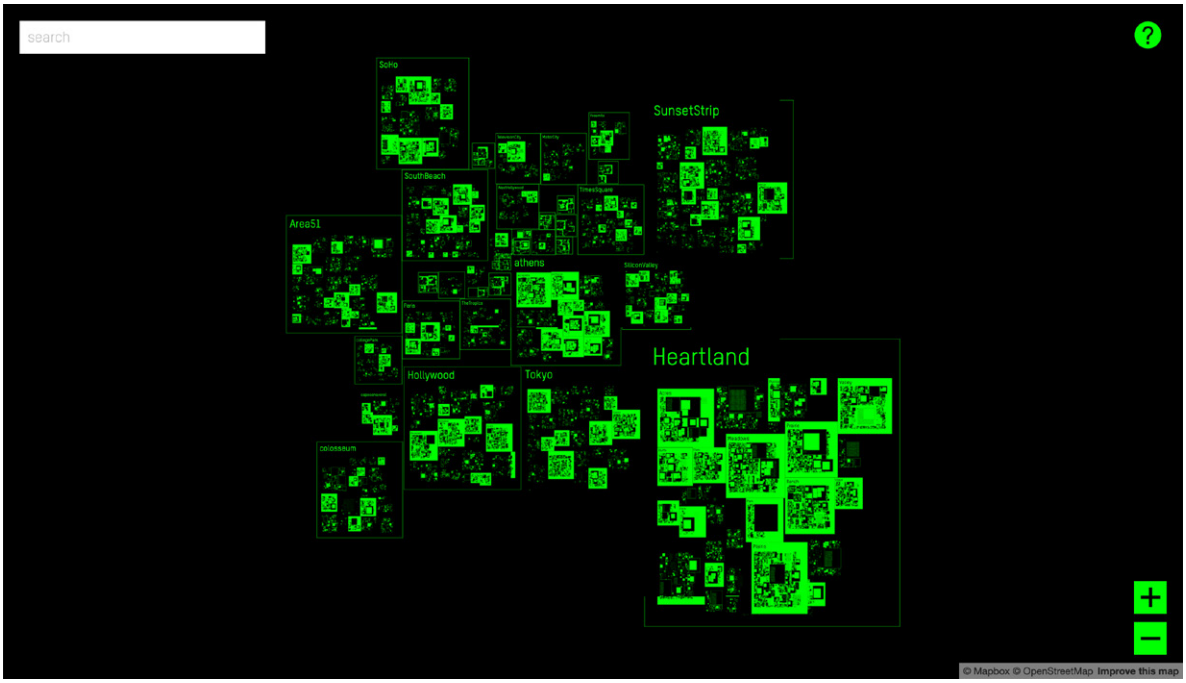


Fig. 2.2 *The Deleted City*: panoramica della "città"

Il design del sito è funzionale e chiaro: su uno sfondo scuro un certo numero di forme quadrate identificano le città e i quartieri, cambiando la trasparenza in linea con la densità della popolazione. I risultati ricordano i classici raggi X, ma anche alcune foto aeree di siti archeologici nascosti, che rendono visibili gli appezzamenti di case antiche, attraverso l'assenza o la presenza di erba a terra e segni di coltura. Navigare questa speciale mappa utilizzando il touch screen permette una *full immersion* in una modalità "primitiva" di espressione sociale attraverso il web. (Ciociola, 2012)

Come osservatori privilegiati di questo enorme sito archeologico digitale, possiamo intravedere bit e pixel, ma soprattutto abbiamo la possibilità di osservare una visione d'insieme, impossibile da vedere nelle complessità di un *backup* di dati illimitato. La mappa completa della città è una visualizzazione dei dati che mostra le dimensioni relative dei diversi quartieri, in base al numero di file che contengono. Aumentando lo zoom sulla città, sempre più dettagli diventano visibili, mostrando alla fine le singole pagine HTML e le immagini in esse contenute. Le immagini che appaiono sulla mappa sono ospitate da *archive.org* e sono disponibili anche tramite la *Wayback Machine*, l'interfaccia web sviluppata da *Internet Archive* per "eseguire la scansione" del web e scaricare tutte le pagine accessibili al pubblico presenti su Internet, grazie alla quale milioni di siti web con i rispettivi dati — immagini, testi, documenti collegati, ecc. — sono stati archiviati.

Il progetto *The Deleted City* è stato esposto a: *The Computer History Museum* (USA) 2016, *The Internet Archive* (USA) 2016, *The Barbican* (UK) 2014, *Los Angeles County Museum of Art* (USA) 2013, *OUDEIS* (FR) 2013, *Counterpath Gallery* (USA), *Cultura Digital* (Brasile) 2012, *IMPAKT Festival* (Paesi Bassi) 2012.

2.2. *Television Explorer*

Nel novembre 2016 *Internet Archive* ha lanciato una nuova visualizzazione cronologica interattiva, il *Television Explorer*. Partner in questo progetto è stato il *GDELT Project — Global Database of Events, Language, and Tone* — che ha creato una piattaforma aperta e libera, che monitora le notizie di televisione, stampa e web di quasi ogni Paese in oltre cento lingue.

Il *Television Explorer* consente di creare rapidamente un cruscotto visivo, che riassume la copertura delle notizie televisive utilizzando i dati del *Television News Archive* disponibili all'interno di *Internet Archive*. Le opzioni di selezione consentono di scegliere il set di dati da cercare, il tipo di dashboard, i *display* specifici da includere nella dashboard, l'arco temporale da considerare e la ricerca da eseguire. Le trasmissioni richiedono 24–48 ore dalla loro messa in onda per essere disponibili per la ricerca, mentre l'indice viene aggiornato ogni 15 minuti.

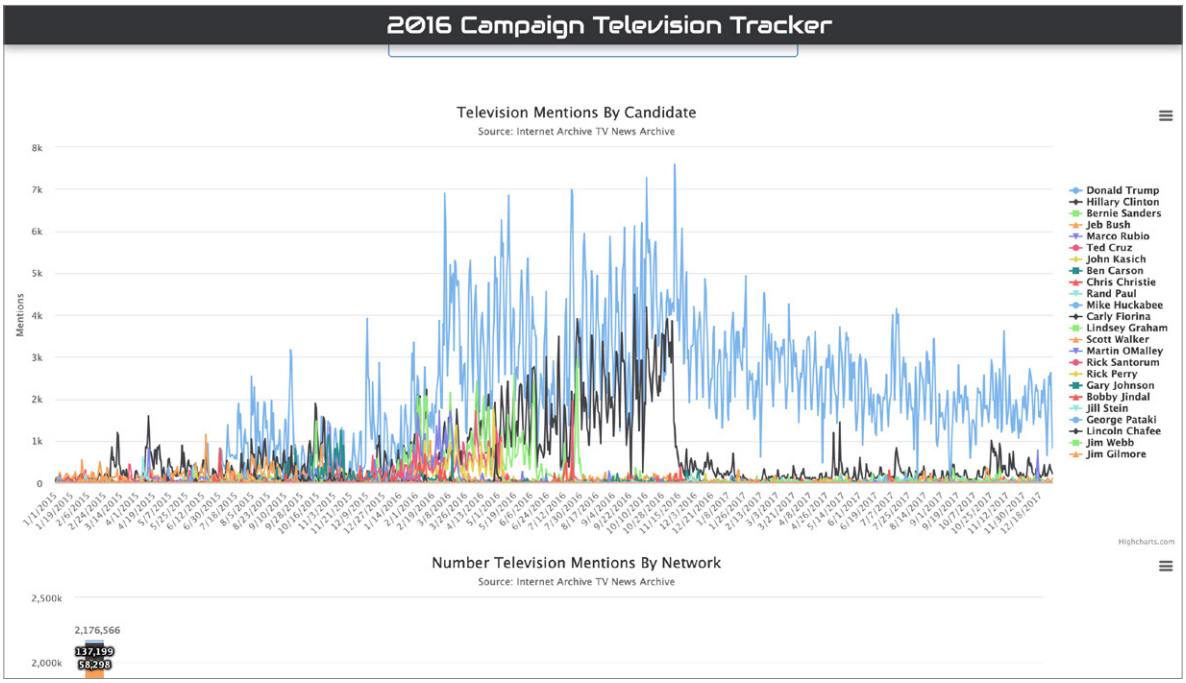


Fig. 2.3 *2016 Campaign Television Tracker*: grafico a linee del numero di menzioni dei candidati

Nel periodo precedente alle elezioni americane del 2016, il *GDELT Project* e *Internet Archive* hanno lavorato a stretto contatto per visualizzare come negli Stati Uniti i programmi televisivi di news avevano coperto la controversa campagna politica del 2016. Per sfruttare il *Television News Archive* nell'analisi del ruolo della televisione in politica, è stato creato un cruscotto — il *2016 Candidate Television Tracker*

— che, su base giornaliera, ha registrato quante volte ogni candidato presidenziale degli Stati Uniti era stato menzionato su ciascuna delle principali reti televisive monitorate dall'archivio.
Questo strumento ha utilizzato i sottotitoli per non udenti di ogni trasmissione, offrendo una linea temporale giorno per giorno, che mostrava gli alti e bassi di chi stava "combattendo" la guerra mediatica. Questo strumento è stato utilizzato da media come *The Atlantic*, *The Washington Post*, *Politico* e *The Guardian*, tra gli altri. (Leetaru, 2016)

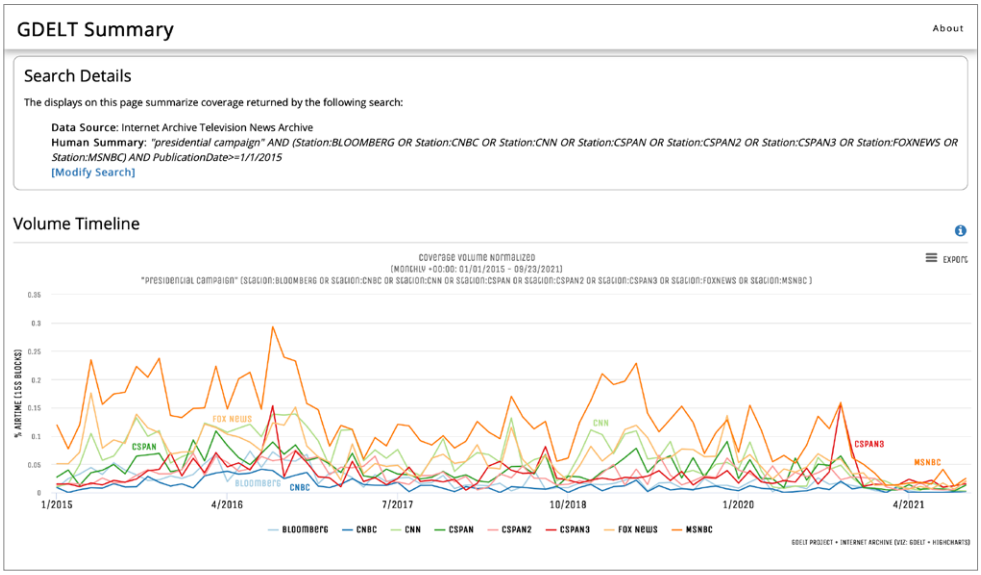


Fig. 2.4 *Television Tracker*: risultati di ricerca per un dato input

Cercando ad esempio la parola "campagna elettorale", si è ottenuta una *timeline* visiva che ha mostrato quanto spesso quella parola era comparsa nelle news televisive americane. Nel periodo pre-elettorale, attraverso questa interfaccia sono stati analizzati quasi due milioni di ore di notiziari televisivi, per un totale di oltre 5,7 miliardi di parole da oltre 150 diverse stazioni, che andavano da luglio 2009 a dicembre 2016.

Television Explorer utilizza anche il grafico *word cloud*, che visualizza una nuvola di parole composta dalle prime 200 più frequenti apparse nelle clip analizzate; questa modalità di visualizzazione è un modo visivo potente che aiuta a comprendere gli argomenti e suggerisce termini di ricerca aggiuntivi per restringere o affinare la ricerca.

Fig. 2.5 *Tag cloud* generata dallo stesso input

Attraverso questi strumenti è stato possibile vedere ad esempio come la CNN ha coperto la campagna presidenziale del 2012 rispetto a quella del 2016 e capire quanto effettivamente quest'ultima sia stata un evento mediatico.
Si è potuto anche vedere quando Edward Snowden ha fatto irruzione sulla scena e come *Wikileaks* ha ottenuto più copertura durante le elezioni presidenziali del 2016 che nel 2010 al suo debutto.

O ancora — per riportare solo alcuni esempi di analisi — si è visto come l'epidemia di ebola del 2014 abbia ricevuto poca attenzione nonostante le migliaia di morti in Africa, diventando un argomento trattato nei notiziari solamente dopo che i primi americani sono stati contagiati.

Nel febbraio 2017 viene rilasciata la versione 2.0 di *Television Explorer*. Rispetto al suo esordio nel 2016, il nuovo *Television Explorer* fa la ricerca in tutte le 163 stazioni televisive che il *Television News Archive* di *Internet Archive* ha monitorato dal 2009, tra le quali troviamo anche stazioni internazionali come Al Jazeera, BBC News e DW, e materiale in lingua spagnola da Univision e Telemundo, anche se non tutte le stazioni sono state monitorate dal 2009 ad oggi.

La nuova versione è stata migliorata per consentire di effettuare ricerche più sofisticate e quindi di tracciare la copertura fatta dai programmi di News televisive in relazione a qualsiasi parola chiave. La più grossa innovazione rispetto alla prima versione sta nel fatto che il nuovo *Television Explorer 2.0*, anziché considerare ogni trasmissione come un unico "contenitore", segmenta ogni trasmissione in una sequenza di blocchi da 15 secondi e cerca le parole/frasi chiave all'interno di ogni blocco, rendendo quindi la ricerca molto più accurata. I risultati vengono rappresentati come percentuale di tempo, sul totale della durata della trasmissione monitorata, in cui ogni stazione ha citato l'argomento oggetto dell'analisi. (*GDELT Project*, 2017)

È stata inserita anche la funzionalità "*context*", che rende possibile eseguire ricerche avanzate utilizzando parole in combinazione con altre parole o frasi, con l'obiettivo di definire anche il contesto in cui le parole chiave sono state utilizzate, e dare quindi la possibilità a giornalisti, ricercatori e *fact-checkers*, ma in generale anche al pubblico, di affinare l'analisi dei dati e di facilitare la comprensione dei fatti.

2.3. *Internet Archive Network Visualization Project*

Nel progetto *Internet Archive Network Visualization Project*, a differenza dei progetti che utilizzano la visualizzazione dei dati per visualizzare l'archivio, *Internet Archive* stesso viene utilizzato per archiviare visualizzazioni di dati, e in particolare quelle relative ai dati dei social media network.

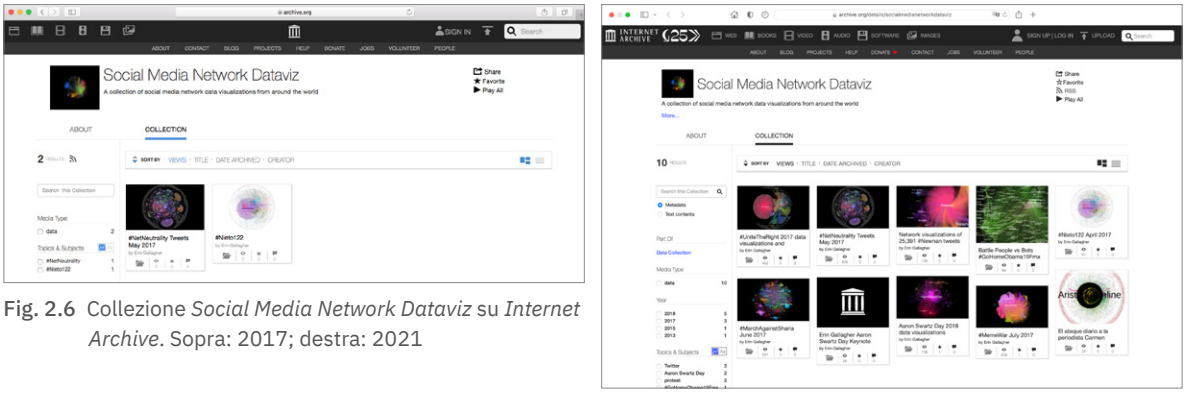


Fig. 2.6 Collezione *Social Media Network Dataviz* su *Internet Archive*. Sopra: 2017; destra: 2021

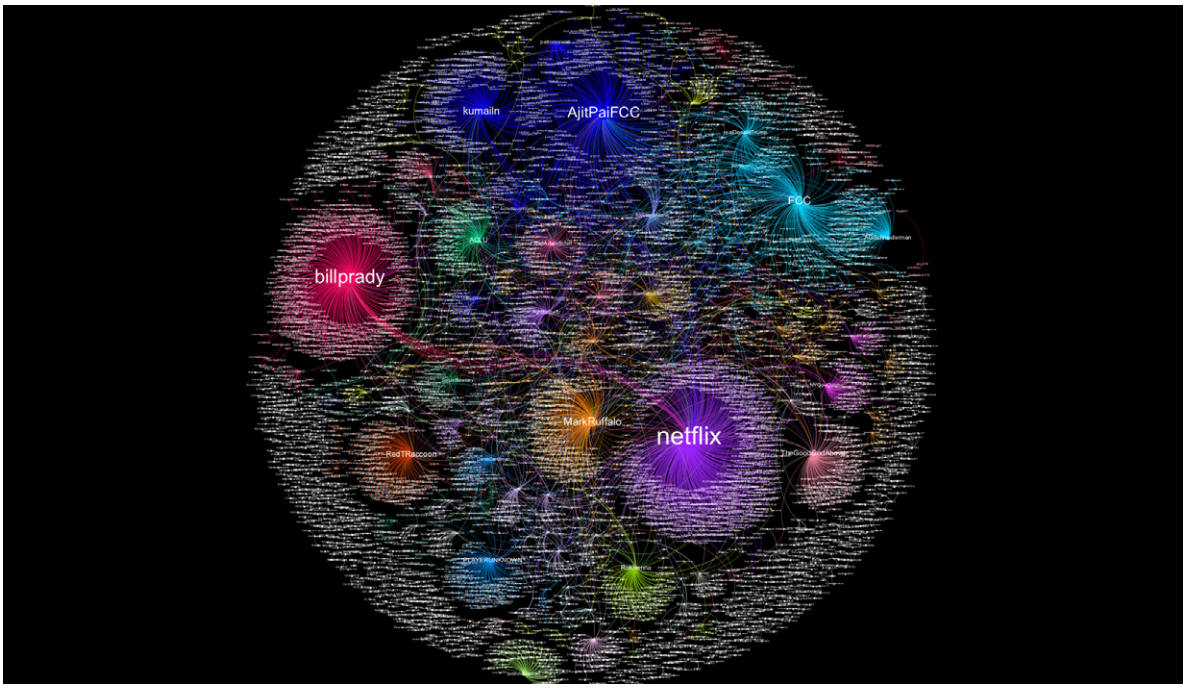
Dal 2017 la visualizzazione dei dati dei social media ha infatti una propria collezione presso *Internet Archive*. Questo lo si deve a Erin Gallagher, social media researcher e multimedia artist, che ha sviluppato un progetto sulle visualizzazioni di reti *Twitter*. Quando la Gallagher ha cominciato il suo lavoro di visualizzazione, alcuni ricercatori messicani avevano già fatto questo tipo di ricerca negli anni precedenti e questo background ha aiutato la ricercatrice nello sviluppo del suo progetto. Gallagher e il collega messicano Alberto Escorcia hanno iniziato a ragionare sulla creazione di un database che raccogliesse modalità diverse di visualizzare di social media, con l'obiettivo di poter confrontare il lavoro e imparare da ricerche simili.

Con l'intervento di Mark Graham, direttore della Wayback Machine di *Internet Archive*, hanno creato la collezione *Social Media Network Dataviz*, una raccolta di visualizzazioni di dati presi dai social media network di tutto il mondo, resa possibile grazie ad una collaborazione di massa con gli utenti di *Internet Archive*.

Il software utilizzato per analizzare l'enorme quantità di dati ottenuta è *Gephi*, un software *open source* di visualizzazione ed esplorazione, attraverso il quale analisti di dati e scienziati possono esplorare e comprendere i grafici, interagendo con la rappresentazione, manipolando strutture, forme e colori per identificare i modelli nascosti. Questo software è stato usato in una serie di progetti di ricerca in ambito universitario, giornalistico e in vari altri campi, per esempio per visualizzare le connessioni globali ai contenuti del *New York Times* o per esaminare la rete di traffico su *Twitter* in occasioni di disordini sociali.

Due casi interessanti analizzati da Gallagher sono *#NetNeutrality Tweets* e *#UniteTheRight Automation and Deleted Tweets*.

Fig. 2.7 *#NetNeutrality Tweets*: network di 4.499 *tweet* inviati dal 21 al 22 novembre 2017



#NetNeutrality Tweets ha analizzato i *tweet user-to-user* con gli hashtag *#NetNeutrality* e *#SaveNetNeutrality* poco dopo che la FCC (Federal Communications Commission) aveva votato per revocare la neutralità della rete nel maggio 2017. Il grafico è codificato a colori per ogni comunità e la dimensione dei nodi è proporzionale alla loro influenza nella rete. Account come @FCC, @AjitPaiFFC (Ajit Pai era presidente della FCC all'epoca dello studio) e @Comcast (il più grande operatore via cavo degli Stati Uniti, terzo fornitore di servizi telefonici domestici e una delle principali aziende mediatiche del mondo) appaiono come nodi più grandi, dato che molte persone hanno menzionato i loro account in relazione alla revoca della neutralità della rete. I grafici finali di questa serie sono stati filtrati per intervallo di gradi per evidenziare i principali *influencer* nella discussione *#NetNeutrality* su *Twitter* al momento della raccolta dei dati.

Applicando i filtri presenti in *Gephi* è stato rimosso il "disturbo" della rete per vedere eventuali anomalie che avrebbero potuto disturbare nell'analisi dei dati. Così sono state ad esempio rimosse tutte le interazioni singole tra due soli nodi, lasciando invece le interazioni che rappresentavano connessioni multiple, relative ad account che avevano twittato più volte con *l'hashtag #NetNeutrality*. Quello che la Gallagher ha osservato è che la maggior parte degli account che sembrava avessero amplificato artificialmente i loro *tweet* erano in realtà normali account che twittavano più del solito perché molto preoccupati per la neutralità della rete. (Gallagher, 2017)

Il secondo caso, *#UniteTheRight Automation and Deleted Tweets*, prese il via dal fatto che l'*hashtag #UniteTheRight* era stato potenziato prima dell'omonimo raduno suprematista bianco verificatosi il 12 agosto 2017 a Charlottesville, utilizzando l'automazione — l'attività

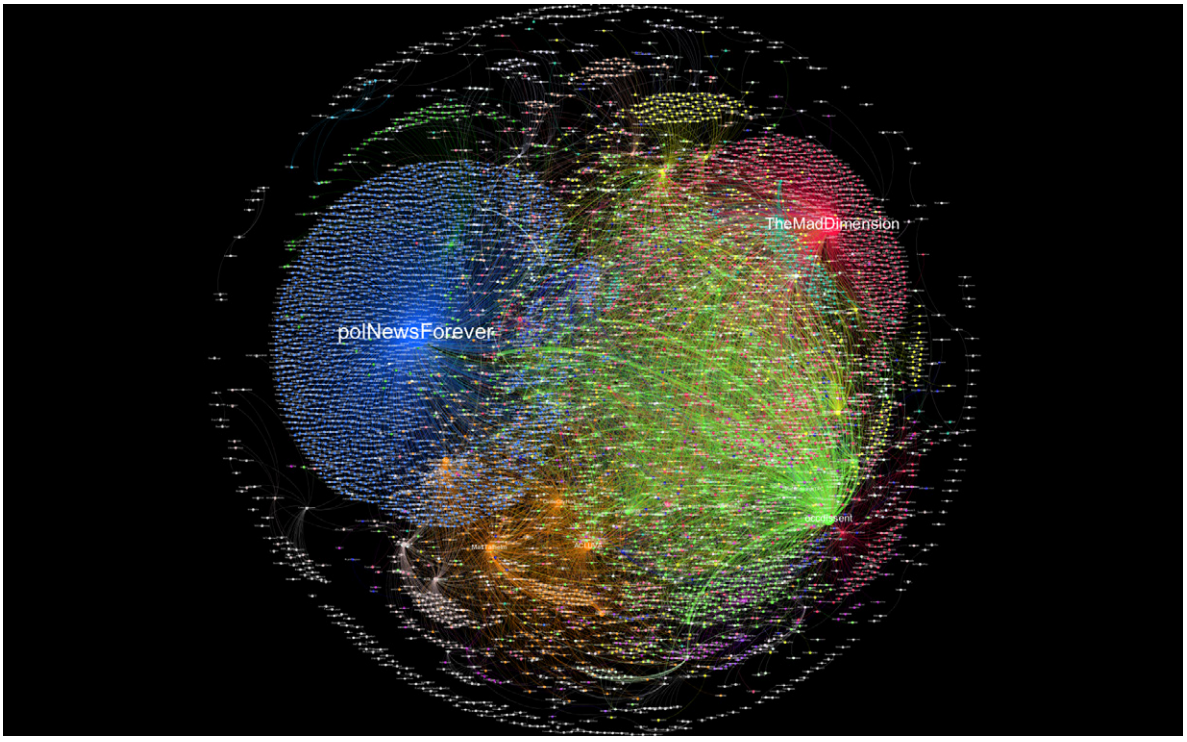


Fig. 2.8 *#UniteTheRight*: network di *tweet* contenenti l'*hashtag #UniteTheRight*

pianificata di *tweeting* e *retweeting* con cui si amplifica artificialmente il numero dei *tweet* — e l'8,2% dei *tweet #UniteTheRight* era stato cancellato nei giorni dopo la protesta, per lo più a causa della sospensione degli account di *Twitter*.

Gallagher ha iniziato a registrare *tweet #UniteTheRigh*t il 7 agosto 2017, cinque giorni prima del raduno, per tenere traccia di come era stato promosso su *Twitter*. A partire dall'11 agosto molti account hanno *twittato* l'*hashtag #UniteTheRight*, mostrando evidenti segni di automazione. Gli utenti più influenti nell'*hashtag* pubblicavano tra i 44 e i 489 *tweet* al giorno, più di quanto umanamente possibile.

I segmenti nei grafici di *#UniteTheRight* sono pesati; quando un account *ritwitta* un altro più volte, i segmenti diventano più spessi. @TheMadDimension (Jason Kessler, organizzatore di *#UniteTheRight*) e @occdissent (Hunter Wallace, pseudonimo di Bradley Dean Griffin, editore del blog *Occidental Dissent*) sono risultati i nodi più influenti nell'*hashtag*, grazie ad una attività automatizzata e pianificata di *tweeting* e *retweeting*.

Per la seconda parte dello studio, *Deleted Tweets*, Erin Gallagher ha collaborato con Ed Summers, lead developer al MITH. Dal 4 agosto al 13 agosto erano stati rilevati 200.113 *tweet* contenenti l'*hashtag #UniteTheRight*. *Twitter* ha però iniziato a sospendere gli account dopo il raduno, quindi una parte di questo dataset è stato rimosso dal social network. Summers è riuscito ad identificare 16.492 *tweet #UniteTheRight* (8,2% del totale) che erano stati cancellati tra il 13 e il 15 agosto.

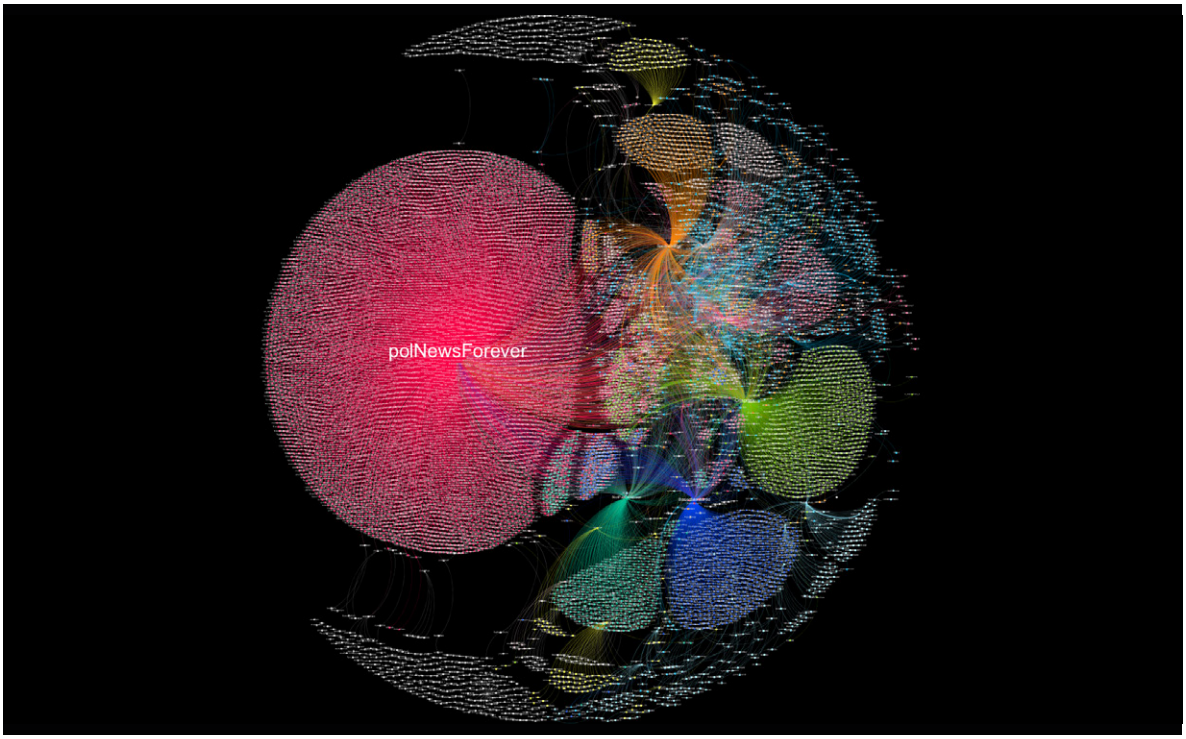


Fig. 2.9 *#UniteTheRight*: network di *tweet* cancellati tra il 13 e il 15 agosto 2017

Le rimozioni erano principalmente il risultato di sospensioni da parte di *Twitter* — probabilmente per aver violato le regole sull'automazione — ma alcuni *tweet* sono stati cancellati dagli utenti stessi, e alcuni utenti avevano protetto i loro *tweet* dopo il raduno, rendendo i propri *tweet* inaccessibili al pubblico. Usando una lista dei *tweet* eliminati, Erin Gallagher ha ricreato una visualizzazione in *Gephi* di come apparivano quei *tweet*, per ricostruire la reale situazione prima che l'operazione di cancellazione fosse messa in atto. (Gallagher, 2017)

2.4. *The Internet Archive's Map of Book Subjects*

Mario Klingemann è un artista tedesco che utilizza algoritmi e intelligenza artificiale per creare e indagare sistemi. È particolarmente interessato alla classificazione e visualizzazione dei dati e allo studio della percezione umana dell'arte e della creatività, ricercando metodi in cui le macchine possano aumentare o emulare questi processi.

Nel 2014 Klingermann rimase affascinato dalla vastità della libreria digitale di libri e testi archiviata in *Internet Archive*, il cui team nel corso degli anni ha reso disponibili milioni di libri digitalizzati sotto forma di pagine scansionate, classificandoli in migliaia di soggetti, fino ad ottenere un archivio di oltre 33 milioni di libri e testi liberamente scaricabili.

Per offrire un modo efficace di visualizzare 2.619.833 di immagini presenti nella collezione di libri di *Internet Archive* nel febbraio 2014, Klingermann ha progettato la *Map of Book Subjects*.

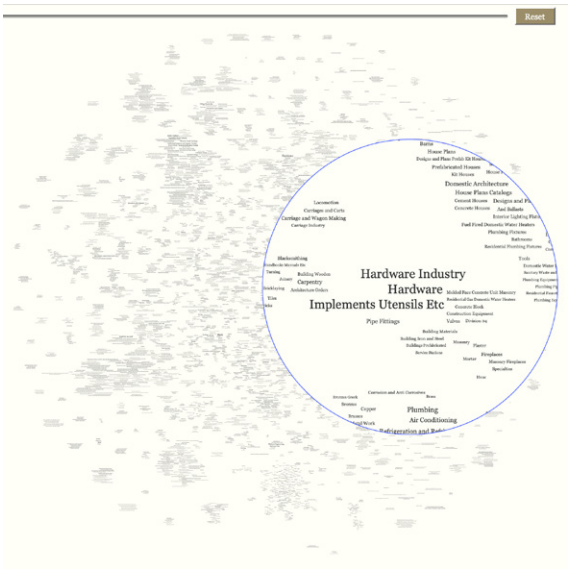


Fig. 2.10 *Map of Book Subjects*: dettaglio selezione "*Hardware*"

Concentrandosi sulle immagini dei libri, l'artista ha mappato 5.500 diversi argomenti, organizzati algoritmicamente in base alle loro relazioni tematiche. La dimensione di ogni argomento presente sulla mappa dipende dalla quantità di immagini che sono disponibili per quell'argomento, e un link ad esso collegato apre la pagina *Flickr* con tutte le immagini relative. Il passaggio sopra il link mette inoltre in evidenza sulla mappa, con un colore diverso, tutti gli argomenti che hanno un collegamento diretto con quello attivo.

Essendo una mappa composta da parole e testi sottostanti, collegati tramite link alle parole di riferimento, è anche possibile usare la casella di ricerca del browser per trovare e localizzare sulla mappa gli argomenti di interesse.

I dati relazionali di questa mappa sono stati generati prima recuperando su *Flickr* tutti i *tag* delle immagini di *Internet Archive* e poi collegando i soggetti che appaiono contemporaneamente su un'immagine. La matrice di similarità risultante, nella quale maggiore è la somiglianza di due oggetti, maggiore è il valore della misura, è stata elaborata usando la tecnica *t-Distributed Stochastic Neighbor Embedding* (t-SNE), che raggruppa gli argomenti in base alla forza della loro relazione, ripulendo automaticamente il layout in modo che nessun blocco di testo si sovrapponga. (Van der Maaten e Hinton, 2008)

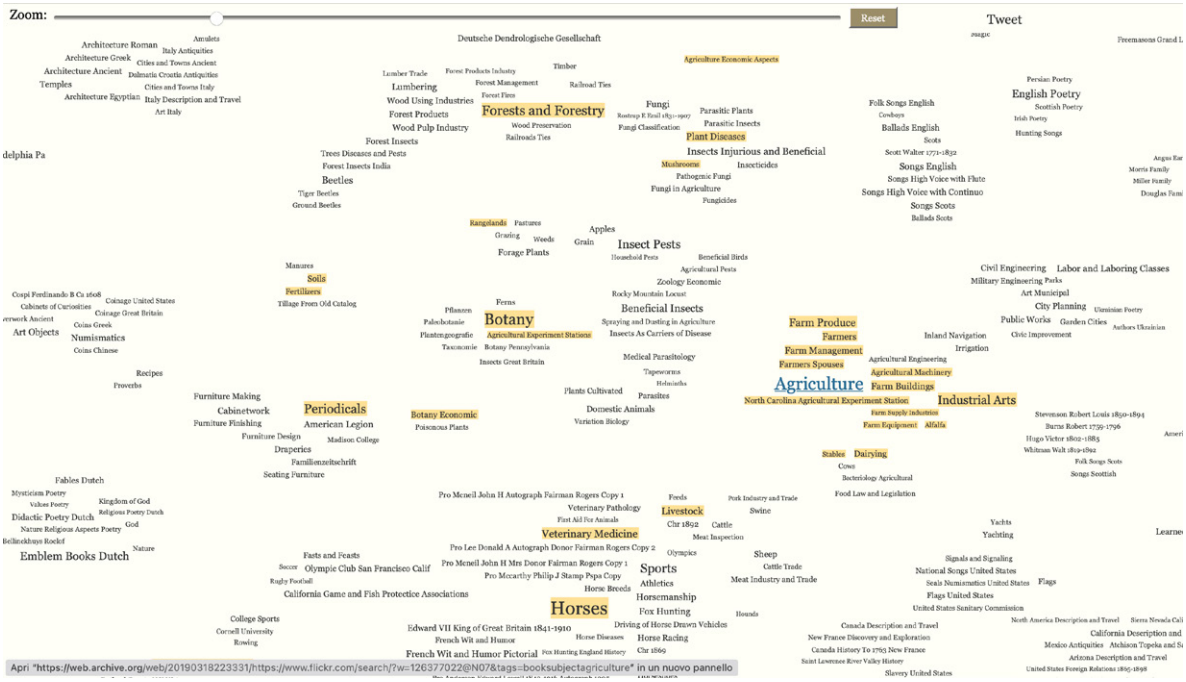


Fig. 2.11 *Map of Book Subjects*: dettaglio selezione "*Agriculture*"

La natura automatica del processo spiega anche alcune stranezze nel layout risultante: a volte, un argomento che è chiaramente parte di un certo cluster è posto lontano da esso, ad esempio "Locomotive" non fa parte di "Ferrovie". Anche alcuni gruppi tematicamente molto vicini tra loro non appaiono vicini sulla mappa, ad esempio "Api" non fa parte del gruppo "Zoologia".

Ciò può essere attribuito al fatto che nessun editore ha creato una singola connessione tra questi soggetti — o probabilmente al fatto che persone diverse sono state coinvolte nella classificazione di quei libri e possono aver usato un diverso insieme di etichette per lo stesso argomento.

Questa tecnica — utilizzata principalmente per l'esplorazione e la visualizzazione di dati ad alta dimensionalità, cioè set di dati costituiti da molte variabili più o meno correlate tra loro — riesce a dare un'idea di come i dati siano disposti in uno spazio ad alta dimensione.

2.5. *Visualizing Digital Collections at Archive-It*

Archive-It è il servizio in abbonamento offerto da *Internet Archive* che consente agli utenti di creare, mantenere e visualizzare collezioni digitali di risorse web. Questo archivio utilizza un'interfaccia in gran parte basata su testo, che supporta la navigazione *drill-down* utilizzando liste di URI — acronimo di *Uniform Resource Identifier*, una sequenza di caratteri che identifica universalmente e univocamente una risorsa, come ad esempio un indirizzo web, un'immagine, un file, un indirizzo di posta elettronica, il codice ISBN di un libro.

Mentre questa interfaccia ha buone capacità di ricerca, non è invece molto efficiente per la navigazione. I siti di una collezione *Archive-It* possono essere organizzati dal curatore della collezione in gruppi per facilitare la navigazione, tuttavia molte collezioni non hanno tali raggruppamenti, oppure hanno raggruppamenti composti da un numero troppo elevato di siti, e diventano quindi difficili da esplorare.

Partendo da questa considerazione, un gruppo di ricercatori del Dipartimento di Computer Science della Old Dominion University in Virginia ha progettato un'interfaccia alternativa per esplorare le collezioni di *Archive-It*, che consiste in nuove visualizzazioni e in una categorizzazione euristica dei siti per rendere le nuove visualizzazioni più significative.

La nuova interfaccia è composta da visualizzazioni multiple — *image plot* con istogramma, *word cloud*, *bubble chart* e *timeline* — che aiutano a fornire una panoramica di ogni collezione e a evidenziare le caratteristiche di fondo della collezione, consentendo all'utente di passare da una visualizzazione all'altra all'interno della stessa collezione.

L'*image plot* con istogramma è un'implementazione di un grafico a barre in pila invertite utilizzato per rappresentare tutti i siti di una collezione in modo visuale. Essendo il grafico diviso in base ai gruppi definiti della collezione, la rappresentazione permette all'utente di esplorare la collezione visualizzando lo *screenshot* dell'ultima versione archiviata di ogni sito.

La rappresentazione invertita permette all'utente di vedere sia i gruppi più grandi che i più piccoli fianco a fianco. Dal momento che è probabile che non tutti i siti possano essere visualizzabili in un'unica finestra, un istogramma ridimensionabile nell'angolo in basso a destra mostra numero di

Fig. 2.12 *Human Rights Collection*: visualizzazione con *image plot* e *word cloud*

siti in ogni gruppo, dando all'utente una panoramica della distribuzione dei siti nei gruppi.

La seconda visualizzazione, la *word cloud*, è stata pensata per "potenziare" il grafico a barre di immagini in pila. Il software utilizzato è *Wordle*, un'applicazione che genera *word cloud* a partire da un testo fornito dall'utente, nella quale le parole che compaiono più frequentemente vengono messe in maggior risalto. Il passaggio del mouse su qualsiasi immagine presente nell'*image plot* rivela in sovrapposizione una *word cloud* che riassume il contenuto discusso nel sito. Questa rappresentazione aiuta la comprensione integrando la rappresentazione visiva fornita dal grafico ad immagini e, attraverso l'analisi di diverse *word cloud*, permette all'utente di cogliere rapidamente le idee chiave della collezione.

Un'altra visualizzazione offerta dalla nuova interfaccia è la *bubble chart*. Questa visualizzazione fornisce un rapido riepilogo statistico della raccolta, mostrando ogni gruppo presente nella collezione come una bolla, dove la dimensione della bolla rappresenta il numero di siti in ogni gruppo. Il numero totale dei siti, le copie visualizzate e archiviate, e l'arco temporale in cui è stata costruita la collezione sono sempre visibili all'utente appena sotto alla *bubble chart*. Ogni bolla è collegata all'elenco dei siti che compongono tutti i gruppi presenti su *Archive-It*, consentendo all'utente di utilizzare i gruppi come filtro rapido per visualizzare le collezioni.

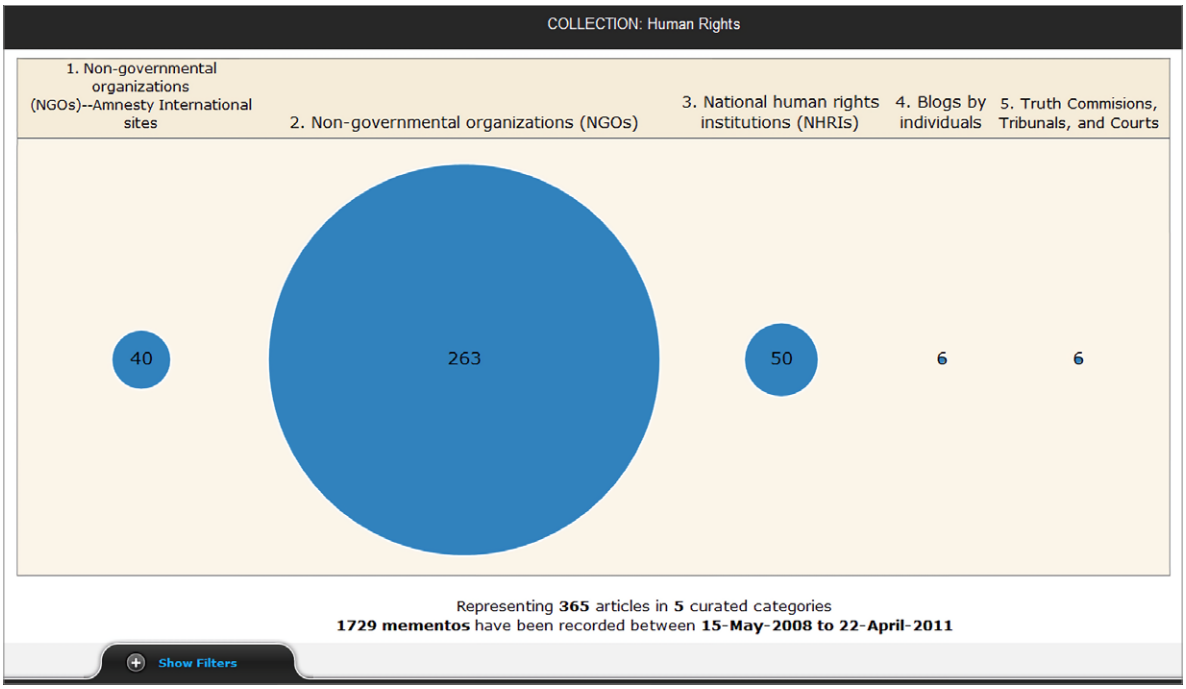


Fig. 2.13 *Human Rights Collection*: visualizzazione con *bubble chart*

Le visualizzazioni fino a qui descritte forniscono il riepilogo dei contenuti presenti in una collezione (*image plot*) e un rapido riepilogo statistico dei numeri di una collezione (*bubble chart*). Tuttavia, i curatori possono essere anche interessati a scoprire come si è sviluppata la collezione nel tempo, per correlare gli eventi storici con l'organizzazione della collezione.

Il gruppo di ricerca della Old Dominion University ha quindi pensato ad una visualizzazione della *timeline* per focalizzarsi sulla variabile temporale. In questa visualizzazione, ogni sito è rappresentato come una singola linea orizzontale, la cui lunghezza indica l'arco temporale nel quale sono state acquisite tutte le copie archiviate del sito. Grazie alla *timeline* si può vedere quali siti sono nella collezione fin dall'inizio, e quante copie del sito sono state regolarmente archiviate nel tempo; ogni punto rappresenta una copia salvata del sito, quindi tanto più il sito è stato salvato tanto più la linea è densa. Questo suggerisce l'alta importanza di questi siti nella collezione e che forse sono quelli che l'utente dovrebbe esplorare prima.

Con questa modalità di visualizzazione è anche possibile cogliere un cambiamento nella composizione della collezione, deducendo dall'improvvisa fine di molte linee sul lato destro della *timeline* il radicale cambiamento dei contenuti della collezione.

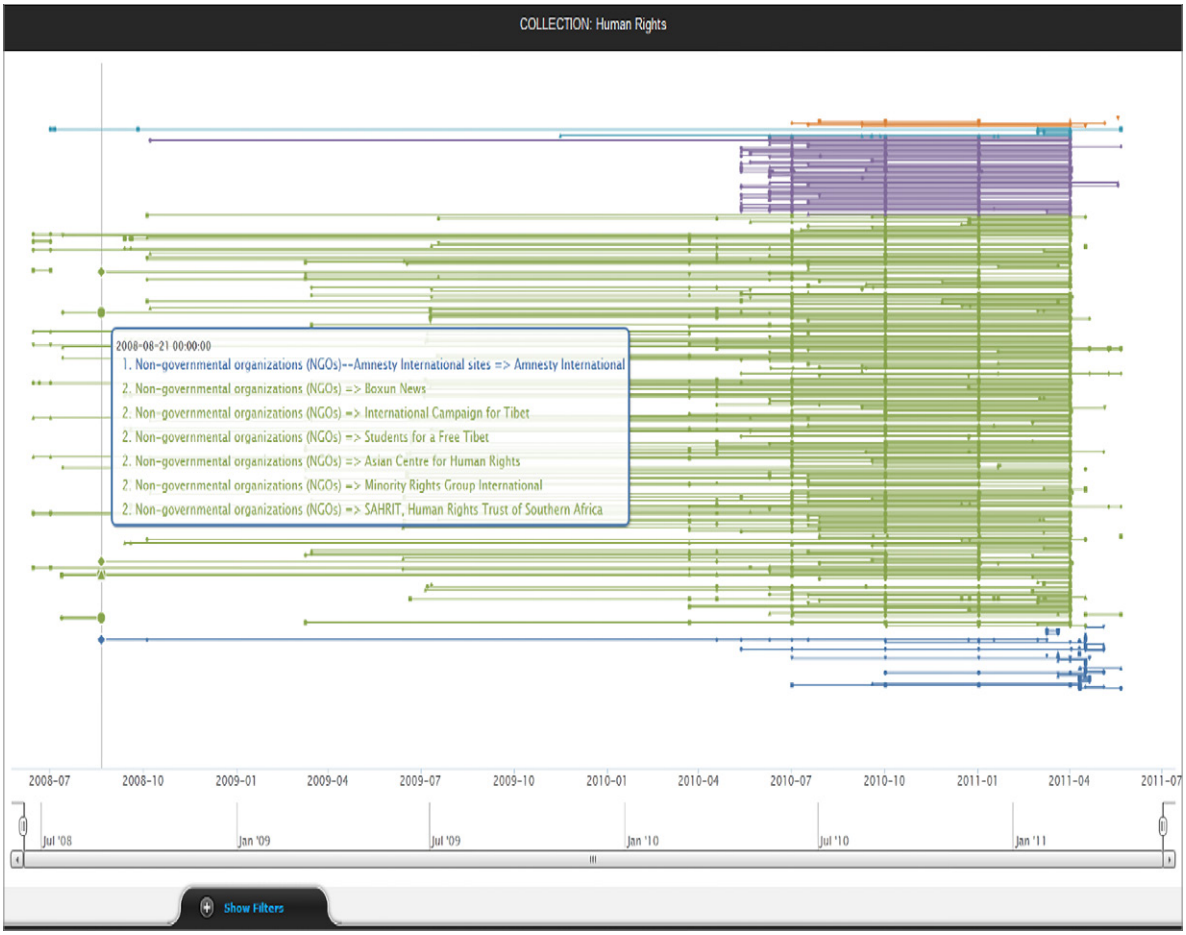


Fig. 2.14 *Human Rights Collection*: visualizzazione con *timeline*

Un altro aspetto importante della ricerca parte dalla constatazione fatta dal gruppo di lavoro che per diverse collezioni di *Archive-It* non era stata fatta una ripartizione in gruppi dei siti in esse presenti, rendendo difficile l'esplorazione da parte dell'utente. Hanno quindi studiato un'opzione per esplorare le collezioni usando una categorizzazione basata sull'euristica.

Si è già accennato all'utilità dell'utilizzo degli URI (*Uniform Resource Identifier*) per la classificazione delle pagine web. Similmente, le regole per la categorizzazione euristica sono basate sull'*hostname* che si trova all'interno degli URI.
Alcuni esempi delle regole di categorizzazione utilizzate sono:

- se l'*hostname* contiene *facebook*, *twitter* o *wiki*, il sito va messo nel gruppo "Social Media";
- se l'*hostname* indica un sito di notizie, ad esempio *bbc*, *cnn* o *nytimes*, il sito va messo nel gruppo "Siti Web News";
- se l'*hostname* contiene *blog* o *wordpress*, il sito va messo nel gruppo "Blog";
- se l'*hostname* contiene *youtube* o *dailymotion*, il sito va messo nel gruppo "Video".

Sono state anche sviluppate regole basate sul nome del dominio di primo livello (*top level domain*, TLD) di un sito web. Ad esempio, tutti i siti con un TLD *.gov* sono stati raggruppati come "Siti Web Governativi" e quelli con il TLD *.edu* sono stati raggruppati come "Siti Web per l'Istruzione".

La categorizzazione basata sull'euristica è utile non solo nell'organizzare le collezioni, ma anche nell'aiutare gli utenti a capire quali fonti e quali tipi di media contribuiscono di più a una collezione.

2.6. Osservazioni conclusive sui casi studio

Dall'analisi dei casi studio relativi a modalità diverse di visualizzazione dei materiali presenti in *Internet Archive*, derivano alcune riflessioni che mettono in luce i punti di forza dei progetti, ma anche le limitazioni ad un possibile utilizzo su materiali diversi da quelli per cui sono stati pensati.

Il risultato ottenuto dal progetto *The Deleted City* è quello di aver reso possibile la conservazione di decine di milioni di pagine web, caricate sul sito di *GeoCities* in 10 anni da più di 35 milioni di persone, e quindi di essere riuscito a dare testimonianza di un'esperienza passata molto significativa che ha caratterizzato i primi anni di sviluppo del *World Wide Web*.

La visualizzazione interattiva del progetto offre la possibilità di avere una visione d'insieme della città digitale — altrimenti impossibile da vedere in un *backup* di dati così vasto — e, assegnando ai quartieri dimensioni proporzionali alla numerosità dei file che contengono, consente di mappare le aree d'interesse degli utenti dell'epoca ed eventualmente di farne anche un'analisi sociologica.
La specificità del progetto e della visualizzazione stessa — la città, i quartieri — limitano però la versatilità di questa modalità, rendendo difficile ipotizzarne l'utilizzo per visualizzare materiali archiviati di diverso tipo.

Possiamo riscontrare un limite analogo anche nel progetto *Television Explorer*. Con questo progetto *Internet Archive* ha introdotto una nuova visualizzazione cronologica interattiva che, analizzando attraverso un cruscotto visivo i dati di *Internet Archive Television News Archive*, riesce non solo a rappresentare la copertura fatta dai programmi di news televisive trasmessi negli Stati Uniti in relazione a qualsiasi parola chiave, ma anche a definire il contesto in cui le parole chiave sono state utilizzate.
La visualizzazione, che si avvale tanto di grafici a linee quanto di *word cloud*, consente di cogliere in modo visivamente efficace l'interesse dell'audience rispetto a determinati temi o eventi, così come di comparare la copertura fatta dalle diverse emittenti televisive, permettendo di affinare l'analisi dei dati e di facilitare la comprensione dei fatti.
Se certamente *Television Explorer* offre la possibilità di un'analisi avanzata dei dati e di una efficace rappresentazione visiva, risulta però evidente che si tratta una soluzione studiata ad hoc per l'archivio delle news televisive, e quindi specifica di un particolare tipo di contenuti, che non può trovare un possibile utilizzo con altri tipi di materiali presenti in *Internet Archive*.

Considerazioni analoghe sulla specificità possono essere fatte anche per il caso studio relativo al progetto *Internet Archive Network Visualization*, il cui obiettivo è stato l'analisi delle interazioni tra gli utenti dei social media in relazione a determinati topic di discussione e specifici *hashtag*.

Nel progetto di *Internet Archive's Map of Book Subjects* è stata realizzata una mappa, composta da parole e testi sottostanti, collegati tramite link alle parole di riferimento, utilizzando una tecnica (t-SNE) che raggruppa gli argomenti in base alla forza della loro relazione.
Per rappresentare la mappa è stata progettata una nuova interfaccia, completamente separata da quella di *Internet Archive*, capace di esplorare e visualizzare dati ad alta dimensionalità, cioè costituiti da molte variabili più o meno correlate tra loro, che riesce a dare un'idea di come i dati siano disposti in uno spazio ad alta dimensione.
Il fatto però che la nuova interfaccia sia separata da quella che *Internet Archive* utilizza per la navigazione e visualizzazione dei propri archivi non rende possibile far dialogare le due interfacce e quindi non può essere integrata per l'utilizzo su altri set di dati.

Il progetto sviluppato per esplorare e rappresentare le collezioni di *Archive-It* consiste in nuove visualizzazioni e in una categorizzazione euristica dei siti per rendere le visualizzazioni più significative. La nuova interfaccia è composta da diversi tipi di grafici — *image plot* con istogramma, *word cloud*, *bubble chart* e *timeline* — che aiutano a dare una panoramica di ogni collezione e ad evidenziarne le caratteristiche di fondo, consentendo all'utente di passare da una visualizzazione all'altra all'interno della stessa collezione.
Le modalità proposte rispondono ad obiettivi diversi: le visualizzazioni con *image plot* e *word cloud* forniscono il riepilogo dei contenuti presenti nella collezione, la visualizzazione tramite *bubble chart* consente un rapido riepilogo statistico dei numeri di una collezione,

mentre con la *timeline* è possibile visualizzare lo sviluppo di una collezione nel tempo. Inoltre tutte queste visualizzazioni possono, a differenza di quelle viste negli altri progetti, funzionare efficacemente anche con altri tipi di materiali digitali presenti in *Internet Archive*.

Quello che invece accomuna tutti i casi studio analizzati è che, pur essendo le interfacce utilizzate visuali, in nessun caso l'immagine degli artefatti presi in considerazione viene usata nella visualizzazione dei risultati.

La pratica di utilizzare l'immagine dell'artefatto per rappresentare se stesso è chiamata visualizzazione diretta. Nel prossimo capitolo tratteremo questa modalità di visualizzazione, analizzandone le caratteristiche e mettendo in evidenza le differenze rispetto alla classica visualizzazione delle informazioni.

Fig. 3.1 Lev Manovich, fotografia di Evgeniya Gorobets

3. Visualizzazione delle informazioni e visualizzazione diretta

Dall'analisi di alcuni progetti di visualizzazione degli oggetti digitali catalogati in *Internet Archive*, abbiamo visto come possano essere utilizzati diversi tipi di interfacce e diverse modalità di rappresentazione visiva. Abbiamo anche notato che le immagini degli artefatti oggetto della ricerca raramente vengono utilizzate nella visualizzazione dei risultati.
In questo capitolo, riprenderemo il tema della visualizzazione delle informazioni per poi introdurre il metodo della visualizzazione diretta, analizzando alcuni progetti in cui è stato applicato.

Per fare questo, utilizzeremo un importante studio di Lev Manovich (*What is visualisation?*, 2011), diventato punto di riferimento per chi si occupa di visualizzazione delle informazioni. Manovich, teorico leader della cultura digitale e pioniere nell'uso dei *big data* per studiare la cultura visiva, fornisce importanti spunti che riassumeremo e analizzeremo per comprendere l'evoluzione che nel XX secolo ha caratterizzato la visualizzazione delle informazioni.

3.1. La visualizzazione delle informazioni

Navigando nel web possiamo trovare moltissimi progetti sofisticati di visualizzazione, creati da scienziati, designer, artisti e studenti. La visualizzazione delle informazioni — vale a dire la rappresentazione visuale avanzata di dati, numeri, processi e informazioni — è un linguaggio capace di trasformare creativamente i dati e trasportarli in un campo più semplice, quello visivo, a cui siamo maggiormente sensibili. Come suggerisce Lev Manovich, introducendo il suo studio *What is visualisation?*, attraverso un equilibrio ragionato tra estetica e funzionalità, tra ricchezza di dati e velocità di lettura, l'*infovis* — abbreviazione comune per visualizzazione delle informazioni — acquisisce un importante potere cognitivo e consente di avere una visione d'insieme efficace e immediatamente suggestiva di quello che si vuole rappresentare.

Nonostante la crescente popolarità degli *infovis*, non è però facile trovare una definizione che funzioni per tutti i tipi di progetti creati oggi, e che allo stesso tempo li separi chiaramente da altri settori correlati, come la visualizzazione scientifica e la progettazione dell'informazione.

Lev Manovich dà una prima definizione della visualizzazione delle informazioni come mappatura tra dati discreti e rappresentazione visiva, ma precisa subito che questa definizione non copre tutti gli aspetti della *infovis*, come ad esempio le distinzioni tra visualizzazione statica, dinamica e interattiva — essendo naturalmente quest'ultima la più importante oggi. Ad esempio, per la maggior parte dei ricercatori della *computer science* la visualizzazione dell'informazione è la comunicazione di dati astratti attraverso l'uso di interfacce visive interattive e della computer grafica, grazie alle quali è più semplice comprendere concetti complessi.

Di fatto, non importa se stiamo guardando una visualizzazione stampata su carta o una disposizione dinamica di elementi grafici sullo schermo del computer, che abbiamo generato utilizzando un software interattivo e che può cambiare in qualsiasi momento: in entrambi i casi l'immagine con cui stiamo lavorando è il risultato di una mappatura.

Mentre alcuni ricercatori distinguono in modo rigido la visualizzazione delle informazioni dalla visualizzazione scientifica, in quanto quest'ultima utilizza dati numerici mentre la prima utilizza dati non numerici, come testi e reti di relazioni, nella realtà molti progetti *infovis* usano i numeri come dati primari e — anche quando si concentrano su altri tipi di dati — usano comunque ancora spesso alcuni dati numerici. Ad esempio, la visualizzazione tipica della rete può utilizzare sia i dati sulla struttura della rete (quali nodi sono collegati tra loro), sia i dati quantitativi sulla forza di queste connessioni (ad esempio, quanti messaggi vengono scambiati tra i membri di un social network).

Un esercizio che Manovich suggerisce per fare un po' di chiarezza è quello di inserire i due tipi di visualizzazioni nella ricerca di immagini di *Google*, confrontando i risultati. Noteremo che la maggior parte delle immagini restituite dalla ricerca di "visualizzazione delle informazioni" è bidimensionale e utilizza grafica vettoriale (ad es. punti, linee, curve e altre forme geometriche semplici), mentre la maggior parte delle immagini restituite dalla ricerca di "visualizzazione scientifica" è invece tridimensionale (forme 3D solide o volumi fatti da punti 3D). I risultati di queste ricerche suggeriscono che i due ambiti differiscono effettivamente, non perché necessariamente utilizzano tipi diversi di dati, ma perché privilegiano tecniche e tecnologie visive diverse.

La visualizzazione scientifica e la visualizzazione delle informazioni provengono da culture diverse — scienza e progettazione, rispettivamente. La visualizzazione scientifica si è sviluppata negli anni '80 insieme al campo della computer grafica 3D, mentre la visualizzazione delle informazioni ha avuto il suo sviluppo negli anni '90, insieme all'ascesa del software di grafica 2D per desktop e all'adozione di PC da parte dei designer. La sua popolarità è cresciuta negli anni 2000 grazie a due fattori chiave: la facile disponibilità di grandi set di dati tramite interfacce di programmazione applicativa (tradotto dall'inglese *application programming interface*, API) fornite

dai principali servizi di social network dal 2005, e nuovi linguaggi di programmazione ad alto livello progettati specificatamente per la grafica e le librerie software per la visualizzazione.

Ma anche le visualizzazioni delle informazioni che venivano fatte manualmente prima dell'avvento dei computer avevano la stessa idea di base, e cioè mappare alcune proprietà dei dati in una rappresentazione visiva. Mentre quindi la disponibilità di computer ha portato allo sviluppo di nuove tecniche di visualizzazione, il linguaggio visivo di base di *infovis* è rimasto lo stesso come era nel XIX secolo, e cioè punti, linee, rettangoli e altre primitive grafiche.

3.2. I principi base della visualizzazione delle informazioni: riduzione e spazio

Lev Manovich dedica una parte importante del suo articolo ai due principi chiave che hanno caratterizzato la pratica della visualizzazione delle informazioni, dalle sue origini nel XVIII secolo fino ad oggi, e cioè riduzione e spazio.

Parlando di riduzione, osserviamo che l'*infovis* utilizza primitive grafiche come punti, linee rette, curve e semplici forme geometriche per rappresentare gli oggetti di dati e le relazioni tra di loro, qualunque sia il tipo di dati. Impiegando primitive grafiche — o, per usare il linguaggio dei media digitali contemporanei, grafiche vettoriali — l'*infovis* è in grado di rivelare modelli e strutture negli oggetti di dati che queste primitive rappresentano.

Le nuove tecniche di riduzione vennero introdotte nella prima parte del diciannovesimo secolo, quando molti studiosi cominciarono ad utilizzare le statistiche per cercare le "leggi della società". Ciò comportava inevitabilmente una sintesi e una riduzione. Non stupisce quindi che, al fine di rappresentare tali sintesi di dati, la maggior parte dei metodi grafici che sono oggi standard siano stati inventati in quel periodo: grafici a barre e a torta, istogrammi, grafici a linee e grafici a serie temporale, grafici di contorno, e così via.

Oltre alla riduzione, vi è un secondo elemento che le tecniche di visualizzazione hanno in comune, e cioè che tutte usano variabili spaziali — posizione, dimensione, forma e, più recentemente, curvatura delle linee e movimento — per rappresentare le differenze chiave nei dati e rivelare i modelli e le relazioni più importanti. L'utilizzo dello spazio è dunque il secondo principio base della pratica di visualizzazione delle informazioni.
Possiamo affermare che l'*infovis* privilegia le dimensioni spaziali rispetto ad altre dimensioni visive. In altre parole, mappiamo le proprietà dei dati che ci interessano attraverso la topologia e la geometria. Altre proprietà meno importanti degli oggetti sono comunque rappresentate, ma attraverso altre dimensioni visive, come toni, ombreggiature, colori o trasparenza degli elementi grafici.

La maggior parte dei metodi di visualizzazione delle informazioni, dalla seconda parte del Settecento fino ad oggi, segue lo stesso principio, e cioè quello di riservare la disposizione nello spazio (layout)

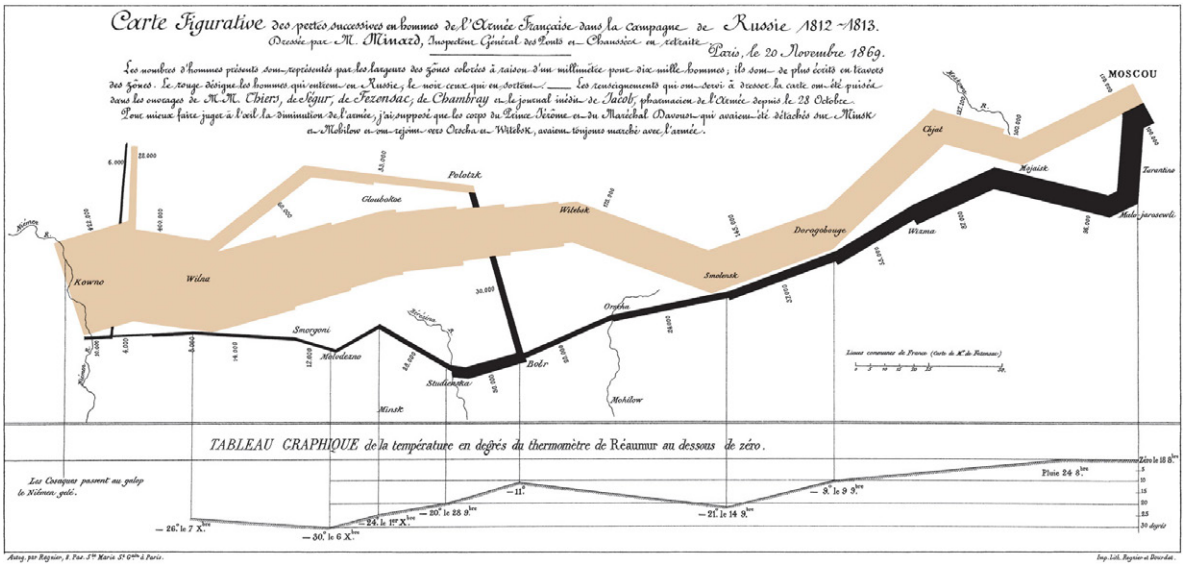


Fig. 3.2 Charles Joseph Minard, *Campagna di Russia di Napoleone*, 1869

per le più importanti dimensioni dei dati, e utilizzare altre variabili visuali per le restanti dimensioni. Questo principio può essere trovato in diverse visualizzazioni, che vanno dalla geniale rappresentazione della *Campagna di Russia di Napoleone* realizzata da Charles Joseph Minard nel 1869, al recente *The evolution of The Origin of Species* di Stefanie Posavec e Greg McInerny del 2009.

L'utilizzo delle variabili spaziali — quali distanze tra gli elementi e le loro posizioni, forma, dimensioni, linee, curvatura — codifica le differenze quantitative tra gli oggetti e le loro relazioni e aiuta a dare un senso ai dati. Nel grafico a dispersione, ad esempio, se alcuni punti formano un *cluster*, ciò implica che gli oggetti di dati corrispondenti hanno qualcosa in comune; se invece abbiamo due *cluster* distinti, ciò implica che gli oggetti cadono in due classi diverse.

Altri tipi di variabili — come l'utilizzo di colori, modelli di riempimento o diversi livelli di saturazione — aiutano inoltre a dividere visivamente gli elementi grafici in gruppi. In altre parole, queste variabili non spaziali funzionano come etichette di gruppo, che aggiungono leggibilità alla visualizzazione senza aggiungere nuove informazioni.

Nel suo studio, Manovich si interroga sui motivi che spingono i progettisti di visualizzazioni — siano essi gli inventori dei grafici del diciottesimo secolo o i milioni di persone che oggi usano quei grafici — a privilegiare le variabili spaziali rispetto ad altri tipi di mappature visive. In altre parole, perché il colore, il tono, la trasparenza e i simboli sono utilizzati per rappresentare aspetti secondari dei dati, mentre le variabili spaziali sono riservate alle dimensioni più importanti?

Egli ipotizza che questo sia legato al fatto che chi progetta una visualizzazione segue la percezione visiva umana che, nel dare senso ad una scena, privilegia la disposizione nello spazio di parti della scena rispetto ad altre proprietà visive. Questo, a sua volta, si lega probabilmente al fatto che ogni oggetto occupa una parte unica dello spazio; per il cervello umano è quindi fondamentale essere in grado

di segmentare un mondo 3D in oggetti spazialmente distinti, con identità distinte. Alcuni tipi di oggetti possono anche essere identificati con forme 2D uniche — per esempio un albero con tronco e rami o un essere umano con testa, torso, braccia e gambe.

Riconosciuto il ruolo cruciale che le variabili spaziali hanno nella nostra esistenza quotidiana, bisogna tuttavia tenere anche in considerazione i tempi di evoluzione delle tecnologie di rappresentazione visiva. L'utilizzo del colore nelle visualizzazioni di informazioni è diventato la norma solo negli anni '90, quando le persone hanno iniziato ad usare i computer per progettare e presentare le visualizzazioni. La stampa a colori è infatti notevolmente più costosa rispetto all'uso di un singolo colore, pertanto il costo decisamente più alto ha probabilmente contribuito in modo significativo al fatto che, negli ultimi due secoli, siano state privilegiate le variabili spaziali.

I due principi chiave che abbiamo analizzato — riduzione dei dati e ruolo privilegiato delle variabili spaziali — se pure non tengono conto di tutte le possibili visualizzazioni prodotte negli ultimi 300 anni, tuttavia sono sufficienti per separare la visualizzazione delle informazioni da altre tecniche e tecnologie per la rappresentazione visiva (mappe, incisione, disegno, pittura ad olio, fotografia, film, video, radar, risonanza magnetica, spettroscopia a infrarossi), e danno ad *infovis* una identità relativamente unica.

3.3. Visualizzazione senza riduzione: la visualizzazione diretta

I significati della parola "visualizzare" includono "rendere visibile" e "dare un'immagine mentale". Questo implica che fino a quando non visualizziamo qualcosa, questo qualcosa non ha una forma visiva: diventa un'immagine attraverso un processo di visualizzazione. L'*infovis* prende dati che non sono visivi e li mappa in un dominio visivo.

Tuttavia Lev Manovich suggerisce l'ipotesi che alcune nuove tecniche di visualizzazione e alcuni progetti sviluppati dalla metà degli anni '90 — anche se comunemente considerati come *infovis* — rappresentino uno sviluppo sostanzialmente nuovo nella storia delle tecnologie rappresentative, o almeno un nuovo metodo di visualizzazione ad ampio raggio.

Consideriamo una tecnica chiamata *tag cloud* o *word cloud*. La tecnica è stata resa popolare dal sito web di photo hosting *Flickr* nel 2005, e oggi la si può trovare su numerosi siti web e blog. Una *tag cloud* mostra le parole più comuni in un testo, dando alle parole stesse la dimensione corrispondente alla loro frequenza nel testo. Potremmo usare un grafico a barre con etichette di testo per rappresentare le stesse informazioni — e questo funziona addirittura meglio se le frequenze delle parole sono molto simili — ma se le frequenze rientrano in un intervallo più ampio, non dobbiamo mappare i dati in una rappresentazione visiva come le barre, mentre possiamo variare la dimensione delle parole stesse per rappresentare le loro frequenze nel testo.

Fig. 3.3 *Tag cloud generata con la parola chiave "big data"*

La *tag cloud* esemplifica un metodo ampio che può essere chiamato visualizzazione multimediale, e cioè la creazione di nuove rappresentazioni visive partendo dagli oggetti visivi reali. Invece di rappresentare testo, immagini, video o altri media attraverso nuovi segni visivi come punti o rettangoli, le visualizzazioni multimediali costruiscono nuove rappresentazioni dal supporto originale: il testo rimane testo, le immagini rimangono immagini. (Manovich, 2011)

Manovich chiama questo metodo "visualizzazione diretta", o visualizzazione senza riduzione. Nella visualizzazione diretta, i dati sono riorganizzati in una nuova rappresentazione visiva che conserva la sua forma originale. Di solito, questo comporta qualche trasformazione dei dati, infatti ad esempio la *tag cloud* applica la modifica della dimensione del testo a un piccolo numero di parole più frequentemente utilizzate. Tuttavia, a differenza delle *infovis* classiche in cui gli oggetti multimediali sono sostituiti da nuovi oggetti, come linee rette o curve e semplici forme geometriche, che comunicano solo le proprietà selezionate di questi oggetti (ad esempio, barre di diverse lunghezze che rappresentano le frequenze delle parole), nel caso della visualizzazione diretta la riduzione che viene applicata è di tipo quantitativo piuttosto che qualitativo, e quindi quando creiamo visualizzazioni direttamente dagli oggetti di dati viene conservato un insieme molto più ricco delle loro proprietà.

Se proiettiamo il concetto della visualizzazione diretta in modo retroattivo nella storia, possiamo trovare tecniche precedenti che usano la stessa idea. Un buon esempio è l'indice di un libro. Guardando infatti l'indice, si può rapidamente vedere se particolari concetti o nomi sono importanti nel libro: questi avranno più inserimenti, mentre concetti meno importanti occuperanno una singola riga. Poiché la vecchia tecnica dell'indice si basava sulla tecnologia tipografica utilizzata per la stampa dei libri, ogni carattere era disponibile solo in un numero limitato di dimensioni; l'idea che si potesse mappare con precisione la frequenza di una particolare parola sulla base della dimensione del carattere era contro-intuitiva e quindi non è stata nemmeno ipotizzata.
Al contrario, la tecnica della *tag cloud* è un'espressione tipica di quello che possiamo chiamare il "*software thinking*" — cioè le idee che esplorano le capacità fondamentali del software moderno. La tag cloud sfrutta le capacità del software di variare i parametri di una rappresentazione e di controllarli utilizzando dati esterni. I dati possono venire da un esperimento scientifico, da una simulazione matematica, dal corpo della persona in un'installazione interattiva, dal calcolo di alcune proprietà dei dati, e così via.

La rapida crescita del numero e della varietà dei progetti di visualizzazione, delle applicazioni software e dei servizi web dalla fine degli anni '90 è stata resa possibile dai progressi delle capacità di computer grafica, sia hardware (processori, RAM, display), che software (librerie grafiche *C* e *Java*, *Flash*, *Processing*, *Flex*, *Prefuse*, ecc.). Questi sviluppi non solo hanno favorito la diffusione della visualizzazione delle informazioni, ma hanno anche cambiato radicalmente la sua identità mettendo in primo piano l'interattività, l'animazione e anche visualizzazioni più complesse, che riescono a rappresentare connessioni tra molti più oggetti. E proprio questi stessi progressi hanno anche reso possibile l'approccio della visualizzazione diretta.

Prima dell'avvento delle applicazioni software, la visualizzazione di solito comportava una prima fase di quantificazione dei dati, e quindi una seconda fase di rappresentazione grafica dei risultati. Il software invece ha dato accesso alla manipolazione diretta degli artefatti multimediali senza necessità di quantificarli, suggerendo che poteva esserci un modo diverso di rappresentare le informazioni visive.

Dopo averci guidati verso il concetto di visualizzazione diretta, Lev Manovich prende in esame tre noti progetti — *Cinema Redux*, *Preservation of Favoured Traces* e *Listening Post* — per chiarire ulteriormente da cosa ha preso spunto la tecnica della visualizzazione diretta e in cosa si differenzia dalla visualizzazione "classica" delle informazioni.

Cinema Redux, creato nel 2004 dall'interactive designer Brendan Dawes e acquisito per la collezione permanente del Museum of Modern Art a New York nel 2008, crea un'unica "distillazione visiva" di un intero film. Dawes ha scritto un programma di elaborazione per campionare un film al ritmo di un fotogramma al secondo e per scalare ogni fotogramma a 8 per 6 pixel. Il programma organizza questi

fotogrammi in una griglia rettangolare, nella quale ogni riga rappresenta un minuto di pellicola, composta da 60 fotogrammi, ciascuno preso ad intervalli di un secondo. Il risultato è un'impronta unica di un intero film, nata dall'acquisizione di tanti momenti sparsi nel tempo riuniti poi in un unico "momento visivo" per creare qualcosa di nuovo, una rappresentazione unica del film nel suo complesso, come un oggetto fisico e tangibile che prende l'immagine in movimento e la traduce in una singola immagine. Questa tecnica mostra istantaneamente l'atmosfera, la fotografia e l'illuminazione presenti in film quali *Vertigo*, *Taxi Driver*, *Deliverance*, *Gone With The Wind* e *Jaws* (Dawes, 2004).

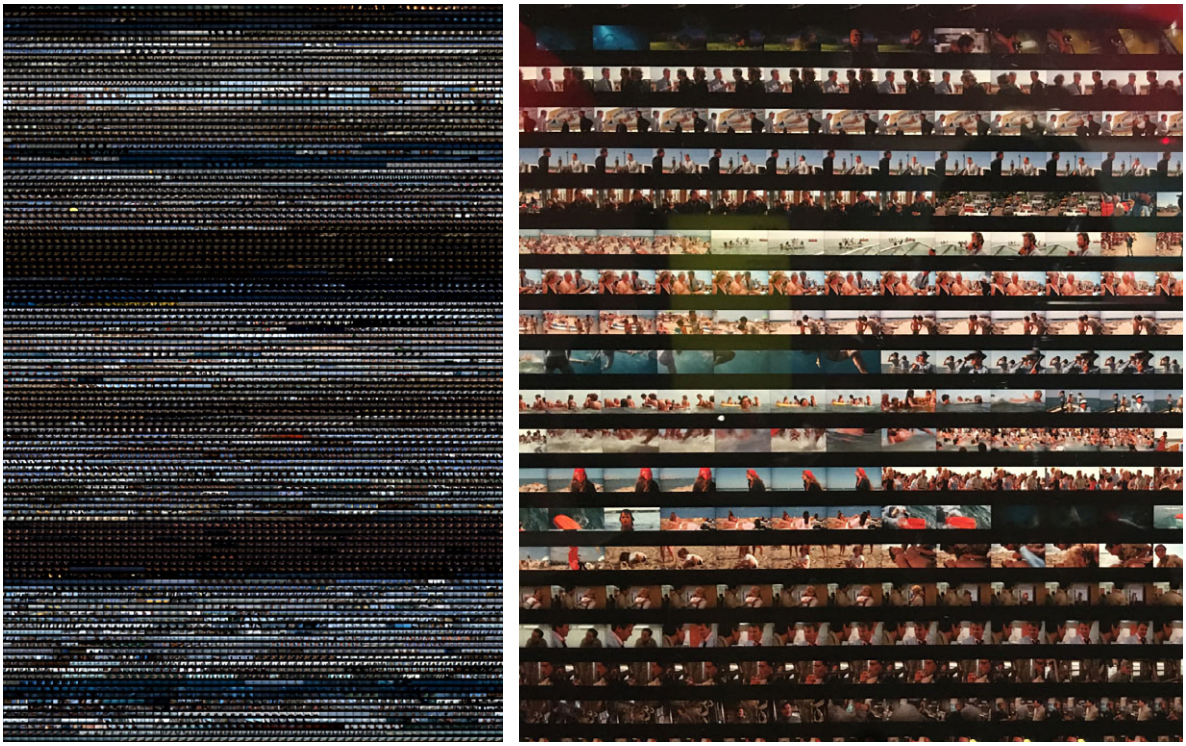


Fig. 3.4 Cinema Redux. Sinistra: campionatura del film *Jaws* (*Lo squalo*); destra: dettaglio

Anche se Dawes avrebbe potuto facilmente continuare questo processo di campionatura e rimappatura — per esempio, rappresentando ogni fotogramma attraverso il suo colore dominante — ha scelto invece di utilizzare gli effettivi fotogrammi presi dal film, in scala ridotta. La visualizzazione risultante rappresenta un compromesso tra i due possibili estremi: preservare tutti i dettagli del manufatto originale o astrarre completamente la sua struttura. Livelli più elevati di astrazione potrebbero rendere più visibili gli schemi nella cinematografia e nella narrativa, ma d'altra parte potrebbero anche allontanare ulteriormente lo spettatore dall'esperienza del film. Di contro, rimanere più aderenti al manufatto d'origine consente di conservare il dettaglio originale e l'esperienza estetica, ma può non riuscire a rivelare alcuni dei pattern sottostanti.

Ciò che è importante nel contesto che stiamo analizzando non sono tanto i parametri particolari che Dawes ha usato per *Cinema Redux*, quanto invece il fatto che egli ha reinterpretato la pratica predefinita di visualizzazione considerandola variabile. Come abbiamo già detto, in precedenza i designer di *infovis* mappavano i dati in una nuova rappresentazione schematica composta da primitive grafiche. Con i computer, un progettista può ora scegliere di utilizzare primitive grafiche, o le immagini originali esattamente come sono, o qualsiasi formato tra le une e le altre. Così, mentre il titolo del progetto *Cinema Redux* richiama l'idea di riduzione, nella sua realizzazione vediamo in realtà un processo di "espansione" dalle primitive grafiche tipiche (punti, rettangoli, ecc.) verso oggetti di dati reali (fotogrammi cinematografici).

Un altro esempio di visualizzazione diretta ci è dato da *Preservation of Favoured Traces* di Ben Fry. Questo progetto web è un'animazione interattiva del testo completo dell'opera di Charles Darwin *On the Origin of Species* (*L'origine delle specie*).

Spesso pensiamo alle idee scientifiche, come la teoria dell'evoluzione di Darwin, come a nozioni fisse accettate come finite. In realtà, il libro di Darwin si è evoluto nel corso delle sei edizioni, che l'autore ha curato e aggiornato durante la sua vita. La prima edizione inglese era di circa 150.000 parole e la sesta è di ben 190.000 parole. Nelle modifiche al testo ci sono affinamenti, aggiunta di dettagli e perfino cambiamenti nelle idee stesse.
Seguendo le sei edizioni come guida, Fry utilizza diversi colori per mostrare lo sviluppo e le modifiche apportate da Darwin mentre sviluppava ulteriormente la sua teoria. Con il procedere dell'animazione, vediamo l'evoluzione del testo del libro, da un'edizione all'altra, con frasi e passaggi cancellati, inseriti e riscritti (Fry, *Fathom Information Design*, 2009).

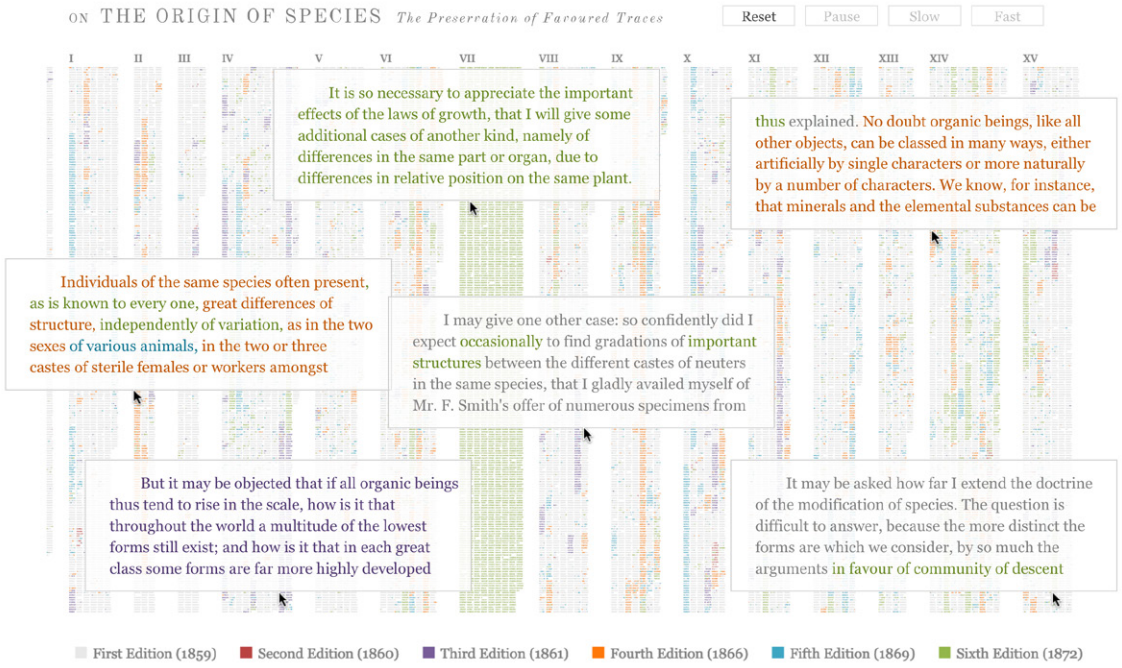


Fig. 3.5 *Preservation of Favoured Traces*: visualizzazione interattiva

In contrasto con le tipiche visualizzazioni animate di informazioni, che mostrano alcune strutture spaziali che cambiano costantemente la loro forma e dimensione nel tempo, riflettendo i cambiamenti nei dati, nel progetto di Fry la forma rettangolare contenente il testo completo del libro di Darwin rimane sempre la stessa; ciò che cambia è il suo contenuto. Questo ci permette di vedere come nel corso del tempo lo schema di aggiunte e revisioni diventi sempre più intricato, via via che i cambiamenti di tutte le edizioni si accumulano.

Possiamo sostenere che *Preservation of Favoured Traces* comporta una certa riduzione dei dati. Infatti, data la risoluzione tipica dei monitor del computer e la larghezza di banda web, Fry non era in grado di mostrare contemporaneamente tutto il testo effettivo del libro. Le frasi sono quindi rese come piccoli rettangoli di diversi colori. Tuttavia, in qualunque momento dell'animazione, abbiamo accesso al testo completo del libro di Darwin: quando si passa infatti il mouse su qualsiasi parte dell'immagine, una finestra pop-up mostra il testo reale completo. Ed è proprio il fatto che tutto il testo del libro di Darwin è in questo modo facilmente accessibile all'utente che ci suggerisce di considerare questo progetto come un esempio di visualizzazione diretta.

Un progetto estremamente particolare, che si avvale di una installazione computerizzata, è *Listening Post* di Ben Rubin e Mark Hansen (2002–2005). *Listening Post* raccoglie in tempo reale i post generati da partecipanti inconsapevoli seduti dietro schermi invisibili da qualche parte nell'universo web, dentro alle *chatroom*.
I post vengono estratti utilizzando vari parametri, impostati dagli autori, e vengono poi trasmessi attraverso un *display wall* composto da oltre 200 piccoli schermi LED, spogliati dei loro involucri e posizionati in una griglia verticale che piega delicatamente ad arco verso il centro della stanza. Tre semplici panchine, poste di fronte alla griglia, permettono agli spettatori di monitorare visivamente e acusticamente in tempo reale l'attività delle *live chatroom*.
Una sequenza di *loop* presenta sei atti diversi tra loro — per esempio, in un atto le frasi si muovono attraverso la parete in un modello ondulatorio; in un altro, le parole appaiono e scompaiono in un modello a scacchiera. Ogni atto ha anche il proprio ambiente sonoro, guidato da parametri estratti dal testo, che viene accompagnato da una vocalizzazione periodica di parole o frasi specifiche ad opera del computer. (Hebron, 2008)

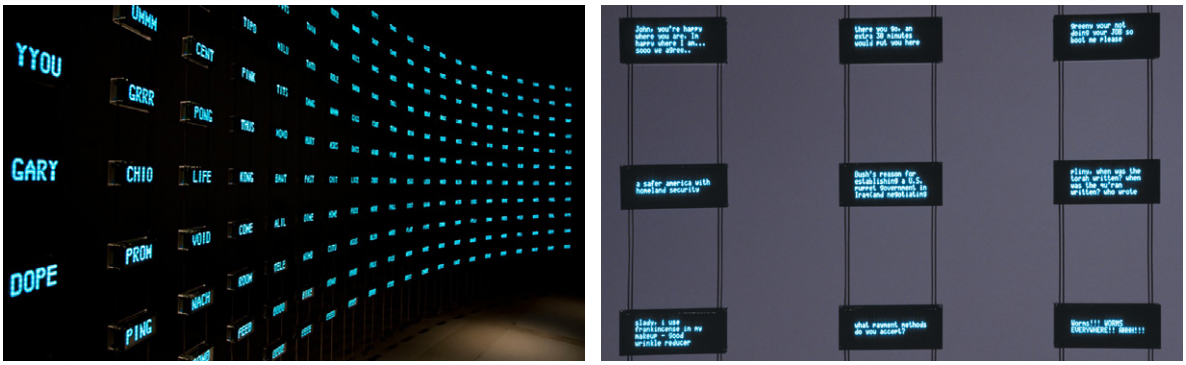


Fig. 3.6 Listening Post: fotografie dell'installazione

Qualcuno potrebbe obiettare che *Listening Post* non è una visualizzazione perché i modelli spaziali non dipendono dai dati, ma sono prestabiliti dagli autori. Bisogna però tenere presente che, mentre i layout sono pre-organizzati, i dati in questi layout non lo sono: essi sono infatti il risultato del data mining in tempo reale del web. Così, se è vero che i frammenti di testo sono visualizzati in layout predefiniti (onda, scacchiera, ecc), il risultato complessivo è comunque sempre unico perché il contenuto di questi frammenti è sempre diverso. Per questo motivo *Listening Post* può essere considerato una perfetta rappresentazione della visualizzazione diretta.

Attraverso l'analisi di questi tre progetti, Manovich ci ha mostrato che, per evidenziare i modelli nei dati, non è sempre necessario ridurre gli oggetti di dati, rappresentandoli attraverso elementi grafici astratti, e non è neppure necessario riassumere i dati, come si fa nella grafica statistica, ad esempio con l'utilizzo di istogrammi. Nello stesso tempo, questo non implica che una visualizzazione diretta debba mostrare il 100% dei dati originali. E infatti solo *Preservation of Favoured Traces* utilizza tutti i dati disponibili, mentre nei progetti *Cinema Redux* e *Listening Post* viene applicata la tecnica del campionamento.

Ciò che è importante, e che accomuna i tre progetti, è che gli elementi di queste visualizzazioni non sono il risultato della rimappatura dei dati in un nuovo formato di rappresentazione, ma sono invece gli stessi oggetti di dati originali selezionati dal set di dati completo.

Possiamo dunque concludere che, se pure una gran parte delle *infovis* "*mainstream*" continua ad utilizzare primitive grafiche, tuttavia si è manifestata un'altra tendenza grazie alla quale vediamo la realizzazione di progetti in cui i dati sono già visivi — testi, fotogrammi di film, copertine di riviste.

In altre parole, questi progetti creano nuove rappresentazioni visive dei dati visivi originali, senza tradurli in segni grafici, attraverso il metodo della visualizzazione diretta. Come abbiamo visto in *Cinema Redux*, *Preservation of Favoured Traces* e *Listening Post*, alcuni dei progetti di *infovis* più noti degli ultimi 15 anni hanno seguito un approccio di visualizzazione diretta.

È anche possibile che il principio centrale della visualizzazione dell'informazione durante i suoi primi tre secoli, e cioè la riduzione dei dati alle primitive grafiche, sia stata una scelta legata alle tecnologie grafiche disponibili all'epoca. Analoga considerazione può riguardare l'aver privilegiato le variabili spaziali rispetto ad altri parametri visivi.
Lo sviluppo della computer grafica ha portato infatti nuove abilità che consentono di controllare con precisione colore, trasparenza, texture e qualsiasi altro parametro visivo di qualsiasi parte di un'immagine, e questo permette di utilizzare questi parametri non-spaziali per rappresentare le dimensioni chiave dei dati. Questo approccio, già comune nella visualizzazione e geo-visualizzazione scientifica e medica, si è diffuso anche nella visualizzazione delle informazioni.

3.4. La visualizzazione diretta nei progetti di Lev Manovich

Come abbiamo visto, la visualizzazione diretta consiste in visualizzazioni — di immagini o di dati video — che fanno uso dei dati nel loro formato visibile originale.

Lev Manovich e i collaboratori del suo *Software Studies Initiative Lab* (oggi *Cultural Analytics Lab*) sono stati in gran parte responsabili della promozione e dello sviluppo della visualizzazione diretta per le scienze umanistiche digitali, attraverso la visualizzazione delle collezioni di media culturali.

Ispirati dai progetti artistici che hanno aperto la strada all'approccio di visualizzazione diretta, nonché dalla altissima risoluzione e dalle funzionalità in tempo reale dei sistemi interattivi di super-visualizzazione come *HIPerSpace*, sviluppato presso il California Institute for Telecommunication and Information, dove si trova il suo laboratorio, Manovich e il suo gruppo hanno lavorato su tecniche e software per consentire l'esplorazione interattiva di grandi set di dati culturali visuali. Alcune delle visualizzazioni create utilizzano la stessa strategia di *Cinema Redux*, che abbiamo analizzato in precedenza, organizzando un grande set di immagini in una griglia rettangolare, e utilizzando display ad altissima risoluzione che spesso permettono di includere tutto il 100% dei dati, invece di doverli campionare. (Manovich, 2011)

Molti sono i progetti di visualizzazione diretta di media culturali ad opera di Manovich, tra cui *Mapping Time* (2009), un montaggio ordinato cronologicamente di ogni copertina della rivista *Time* dal 1923 al 2009, *Manga Style Space* (2010), una visualizzazione contenente oltre 1 milione di pagine *manga* ordinate in base alle loro caratteristiche visive, *Phototrails* (2013), una serie di grafici di immagini contenente ciascuno migliaia di foto *Instagram* ordinate per caratteristiche di colore di base, *Selfiecity* (2014–2015), che attraverso i selfies indaga l'auto-rappresentazione di Instagram in sei città in tutto il mondo, e *Visual Earth* (2017), il primo studio ad analizzare la crescita della condivisione di immagini su *Twitter* in tutto il mondo in relazione alle differenze economiche, geografiche e demografiche.

In questi e in molti altri progetti, Manovich ha dimostrato la potenza di questo metodo che, inizialmente applicato alle scienze umanistiche digitali e ai media culturali, può essere utilizzato per rappresentare tutti i dati di un'immagine. (Crockett, 2016)

Ogni progetto ha una sua specificità e l'analisi dei progetti più significativi ci può aiutare a mettere a fuoco ulteriormente le potenzialità della visualizzazione diretta.

3.4.1. *Mapping Time*

Il progetto *Mapping Time*, sviluppato nel 2009 da Manovich insieme a Jeremy Douglass, presenta un'analisi di visualizzazione delle 4.553 copertine di tutti i numeri della rivista *Time* pubblicati tra

il 1923 e il 2009, in un arco temporale di 86 anni. I bordi rossi, distintivi della rivista, che incorniciato la copertina sono stati eliminati per focalizzare l'attenzione sulle immagini ed evidenziare i cambiamenti nel corso del tempo. Questo progetto è composto da due distinte visualizzazioni.

Nella prima visualizzazione, *Grid*, le copertine appaiono in ordine di pubblicazione, disposte in una griglia da sinistra a destra e dall'alto verso il basso. *Mapping Time*, così come abbiamo già visto in *Cinema Redux*, mette a confronto i valori delle variabili spaziali per rivelare i pattern nel contenuto, nei colori e nelle composizioni delle immagini. Tutte le immagini sono visualizzate alla stessa dimensione, disposte in una griglia rettangolare secondo la loro sequenza originale. Essenzialmente, questa visualizzazione diretta usa una sola dimensione, con la sequenza di immagini raccolte in un certo numero di righe per rendere più facile vedere i modelli, senza dover scansionare visivamente immagini molto lunghe.

Ecco alcuni modelli evidenziati da questa visualizzazione:

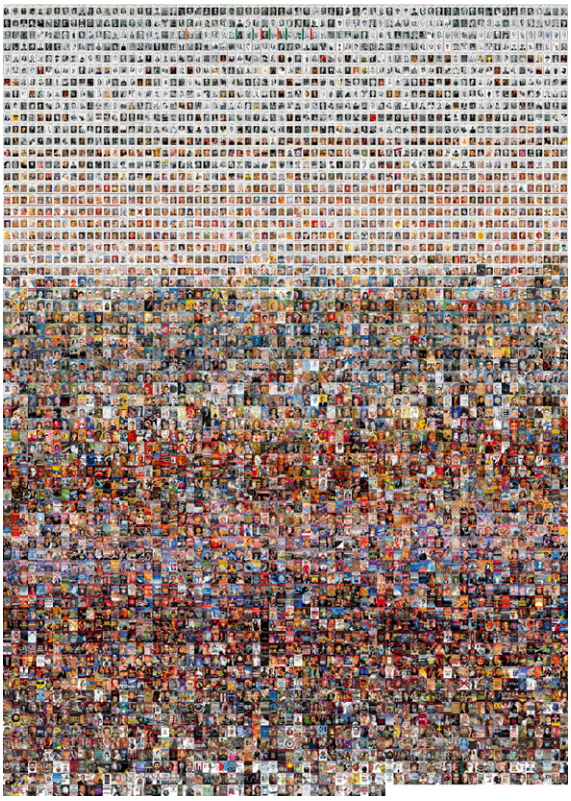


Fig. 3.7 *Mapping Time*: *Grid*

- media: negli anni '20 e '30 le copertine di Time utilizzano per lo più la fotografia. Dopo il 1941, la rivista passa ai dipinti. Negli ultimi decenni del secolo la fotografia gradualmente torna di nuovo a dominare. Dagli anni '90 si assiste all'emergere di un linguaggio visivo contemporaneo basato sull'utilizzo del software che combina fotografia manipolata, elementi grafici e tipografici;
- contenuto: inizialmente la maggior parte delle copertine sono ritratti di individui ambientati su sfondi neutri. Nel corso del tempo, gli sfondi dei ritratti si trasformano in composizioni che rappresentano concetti. Successivamente, queste due diverse strategie si uniscono: i ritratti ritornano a sfondi neutri, mentre i concetti sono ora rappresentati da composizioni che includono sia oggetti che persone — ma non individui particolari;
- colore vs. bianco e nero: il passaggio dal primo bianco e nero a copertine a colori pieni avviene gradualmente, con un periodo molto lungo di coesistenza di entrambi i tipi;
- tonalità: distinti "periodi di colore" appaiono in fasce — verde, giallo/marrone, rosso/blu, giallo/marrone di nuovo, giallo, e giallo più chiaro/blu negli anni 2000;

– luminosità: le variazioni di luminosità, cioè la media dei valori di tutti i pixel in scala di grigi per ogni cover, seguono un modello ciclico simile;
– contrasto e saturazione: entrambi aumentano gradualmente nel corso del XX secolo. Tuttavia, dalla fine degli anni '90, questa tendenza è invertita: copertine recenti hanno meno contrasto e meno saturazione;
– la visualizzazione rivela anche un importante "meta-pattern": quasi tutti i cambiamenti sono graduali: ciascuna delle nuove strategie di comunicazione emerge lentamente nel corso di un certo numero di mesi, anni o addirittura decenni.

Nella seconda visualizzazione, *Timeline*, le linee temporali di immagini unidimensionali vengono girate in 2D, con la seconda dimensione che comunica informazioni supplementari. L'asse orizzontale è utilizzato per posizionare le immagini nella sequenza originale: le 4.353 copertine sono disposte da sinistra verso destra, e ogni copertina è posizionata in base alla data di pubblicazione.

La strategia utilizzata in questa visualizzazione si basa sulla tecnica del grafico a dispersione. Tuttavia, un normale grafico a dispersione riduce i dati visualizzando ogni oggetto come un punto, mentre con la visualizzazione diretta vengono utilizzati i dati nella loro forma originale. Il risultato è un nuovo tipo di grafico, che è letteralmente fatto da immagini: ecco perché è opportuno chiamarlo *image graph*. (Manovich, 2011)

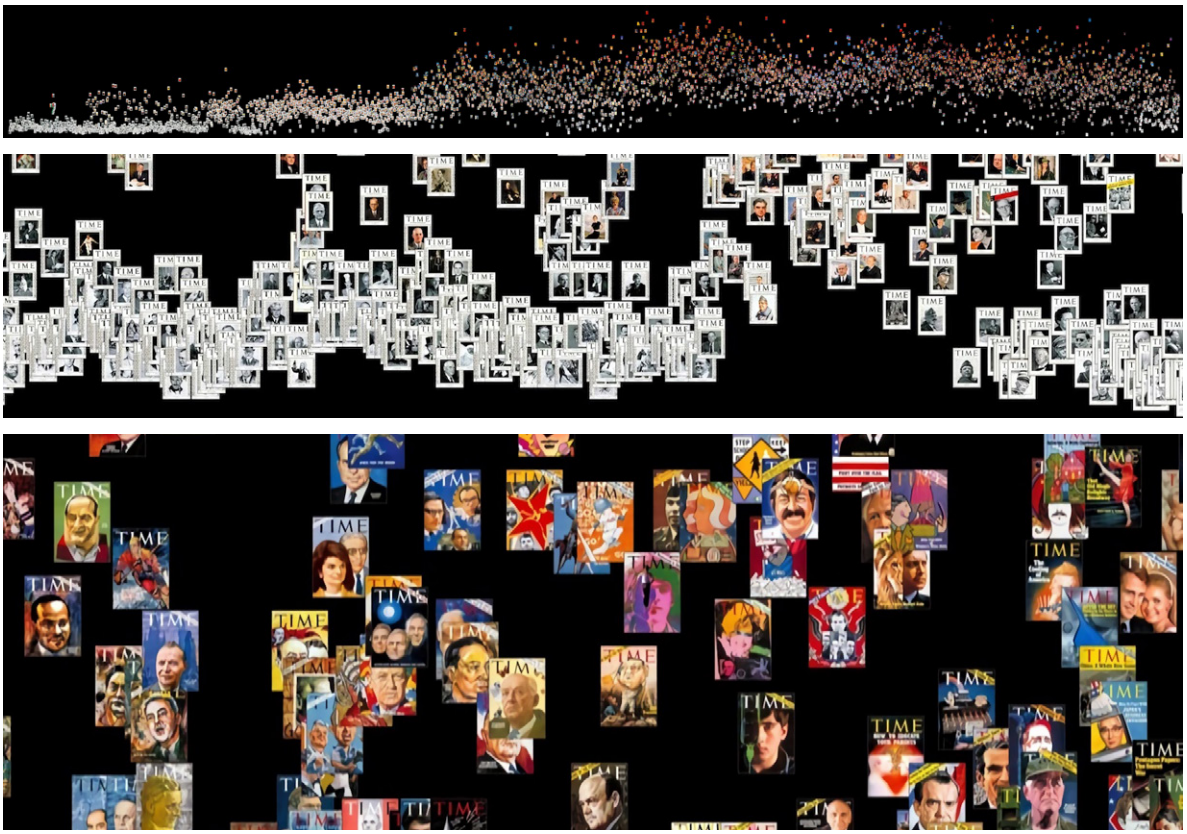


Fig. 3.8 *Mapping Time*: *Timeline*; dettagli

Le posizioni sull'asse verticale rappresentano nuove informazioni — in questo caso, la saturazione media, cioè l'intensità percepita dei colori di ogni copertina, che è stata misurata con il software di analisi delle immagini. Tale mappatura è particolarmente utile per mostrare variazioni dei dati nel tempo. La visualizzazione in timeline rivela infatti un certo numero di modelli temporali: è visibile l'era di stampa pre-colore sulla sinistra, un gruppo di brevi primi esperimenti nella stampa a colori e poi il graduale passaggio dal bianco e nero alle copertine a colori, con entrambi i tipi che coesistono per un certo numero di anni.

Si può osservare che la luminosità e la saturazione seguono un modello ciclico di aumento e riduzione, con notevoli picchi e cali che diventano evidenti solo per periodi di un decennio o più. A prescindere dalla curva generale, ci sono eccezioni estreme: immagini luminose e disegni tenui che fluttuano sopra o sotto la nuvola di copertine che si sviluppa lungo l'arco temporale.

Possiamo vedere come la saturazione dei colori aumenti gradualmente durante la pubblicazione di *Time*, raggiungendo il suo picco nel 1968. Forse non è sorprendente vedere che l'intensità — o l'aggressività? — dei mass media aumenta gradualmente fino alla fine degli anni '60, come risulta dai cambiamenti nella saturazione e nel contrasto. Ciò che è inaspettato, tuttavia, è che fin dall'inizio del XXI secolo, questa tendenza è invertita: le copertine ora hanno meno contrasto e meno saturazione.

3.4.2. *One million manga pages*

One million manga pages è il primo progetto applicato a studi umanistici digitali che utilizza l'analisi e la visualizzazione di immagini digitali per lo studio di una massiccia collezione di immagini, e cioè appunto un milione di pagine *manga*.

Il mondo legato ai *manga* è molto attivo: non appena nuovi libri *manga* vengono pubblicati, i fan li acquistano, scansionano le pagine, traducono il testo in altre lingue e distribuiscono tramite siti web le immagini digitali delle pagine tradotte. Nel fare questo, inseriscono anche pagine aggiuntive, contenenti crediti di gruppo, commenti e *fan art* originale. Questo processo è indicato come *scanlation*, termine inglese derivante dalla fusione delle parole *scan* e *translation*. L'archivio online più popolare delle *scanlations* è *onemanga.com*.

Nell'autunno 2009 Lev Manovich, Jeremy Douglass, William Huber, Tara Zepel hanno scaricato da questo sito 883 serie di *manga* contenenti 1.074.790 pagine uniche. I dati includono le serie più popolari — provenienti da Giappone, Corea e Cina — come *Naruto* (1999–oggi, 8.835 pagine) e *One Piece* (1997–oggi, 10.562 pagine), insieme anche a serie più brevi che sono apparse negli anni 2000 e sono state pubblicate solo per 2–3 anni.

Manovich e i suoi collaboratori hanno utilizzato un sistema software da loro progettato, e installato su un supercomputer presso il National Department of Energy Research Center (NERSC), che utilizzando

tecniche computazionali ha permesso di analizzare sistematicamente il linguaggio visivo delle pagine inserite dai fan in versioni scansionate e tradotte e anche di studiare le differenze visive tra le pagine delle pubblicazioni originali giapponesi e quelle delle traduzioni ufficiali in inglese. (Manovich, Douglass, Huber, Zepel, 2011)

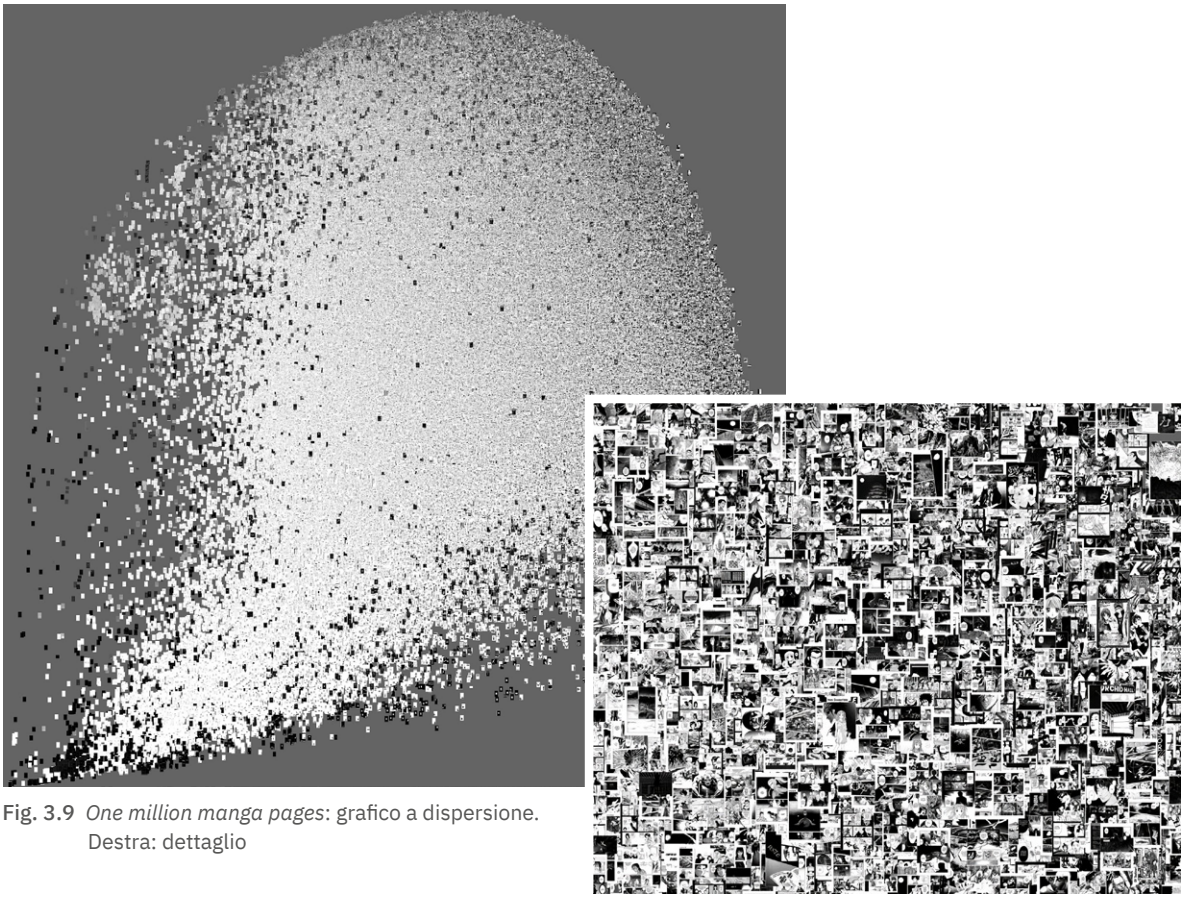


Fig. 3.9 *One million manga pages*: grafico a dispersione. Destra: dettaglio

Per poter inserire tutte le pagine in una singola immagine di grandi dimensioni (l'originale è di 44.000 per 44.000 pixel, ridotta a 10.000 per 10.000 per la pubblicazione su *Flickr*), le immagini sono state tutte rese in scala di grigio, malgrado alcune — come ad esempio tutte le copertine dei volumi — siano a colori.

La visualizzazione dei risultati mostra il set di dati completo (1 milione di pagine di *manga*) organizzate in base alle loro caratteristiche visive in uno spazio 2D, in cui l'asse X corrisponde alla deviazione standard dei valori in scala di grigi dei pixel in una pagina, mentre l'asse Y corrisponde all'entropia misurata su tutti i valori in scala di grigi dei pixel in una pagina.

Inoltre, poiché le pagine sono sovrapposte una sopra l'altra, in realtà non si vede un milione di pagine distinte: la visualizzazione mostra una distribuzione di tutte le pagine con esempi tipici che appaiono nell'ultima pagina in alto della sovrapposizione. (Manovich e Douglass, 2010)

Le pagine nella parte inferiore della visualizzazione sono le più grafiche, hanno la minore quantità di dettagli; le pagine in alto a destra hanno molti dettagli e texture. Le pagine con il contrasto più alto sono a destra, mentre le pagine con il minor contrasto sono a sinistra. La visualizzazione mostra inoltre, nella parte centrale della "nuvola" di pagine, quali scelte grafiche sono più comunemente utilizzate dagli artisti *manga*, mentre quelle che appaiono molto più raramente sono nella parte inferiore e sinistra. Tra questi quattro estremi, troviamo ogni possibile variazione stilistica.

Come possiamo vedere, lo spazio stilistico di *manga* non ha cluster distinti. La visualizzazione permette di descrivere questo spazio molto meglio delle categorie linguistiche discrete. Questo suggerisce che il concetto di base di "stile", che presuppone che si possa dividere un insieme di artefatti culturali in un piccolo numero di categorie discrete, potrebbe non essere appropriato quando si considerano grandi set di dati culturali. Nel caso del set di un milione di pagine *manga*, troviamo infatti variazioni grafiche praticamente infinite e quindi, se si cerca di dividere questo spazio in categorie stilistiche discrete, qualsiasi tentativo sarà arbitrario. Per descrivere queste variazioni meglio quindi usare la visualizzazione e/o modelli matematici.

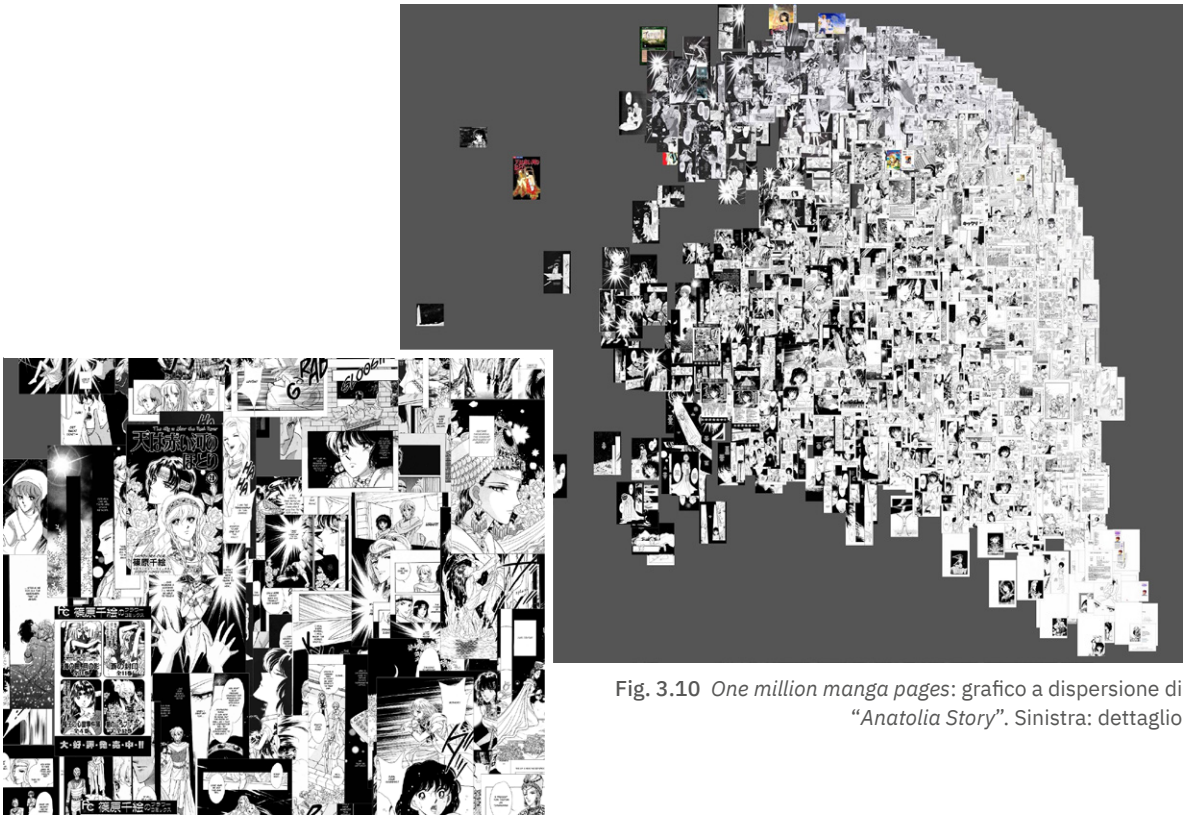


Fig. 3.10 *One million manga pages*: grafico a dispersione di "*Anatolia Story*". Sinistra: dettaglio

Cosa succede se prendiamo in considerazione i singoli titoli *manga*? All'interno di un singolo titolo vi può essere una tale variabilità grafica che nemmeno qui è possibile usare il concetto di "stile". Questo complica l'idea di poter dividere tutti i titoli *manga* considerati dal progetto in una serie di categorie stilistiche. Abbiamo un esempio nella serie *Anatolia Story*, le cui 879 pagine sono organizzate secondo le caratteristiche visive di luminosità (asse X) ed entropia (asse Y).

Se nell'analisi dei dati, oltre a queste, si prendono in considerazione anche altre caratteristiche di ordine superiore — quali contenuto, composizione, convenzioni visive del manga per la resa dei personaggi, i loro volti, sfondi, ecc. — i risultati rivelano la presenza di stili distinti nel campione ampio e mostrano anche una maggiore coerenza stilistica nei singoli titoli *manga*. (Manovich, Douglass, Huber, Zepel, 2010)

3.4.3. *Phototrails*

Quando nel 2004–2005 c'è stata l'esplosione dei social media, sono state sollevate nuove domande che riguardano i nuovi modi per descrivere il mondo. Una gran parte dei media culturali contemporanei viene creata, modificata e condivisa utilizzando software, ma come si possono esplorare i dati visivi di social media che contengono miliardi di fotografie condivise da centinaia di milioni di persone? Come è possibile individuare i modelli culturali più ampi contenuti in questo enorme universo visivo? E come tutte queste fotografie riflettono le caratteristiche dei luoghi dai quali arrivano?

Il progetto di ricerca *Phototrails*, nato per rispondere a queste domande, punta i riflettori sul mondo delle fotografie provenienti dai social media, analizzando in particolare *Instagram*, la popolare applicazione di condivisione di foto scattate tramite dispositivi mobili, che offre ai suoi utenti un modo per caricare foto e per applicare diversi strumenti di manipolazione, noti come filtri, al fine di trasformare l'aspetto di un'immagine (colore, contrasto, saturazione e così via), per poi condividerla istantaneamente con altri.

Quando nel giugno del 2013, a soli tre anni dal lancio di *Instagram*, Lev Manovich, Nadav Hochman (Dipartimento di Storia dell'Arte dell'Università di Pittsburgh) e Jay Chow (*Software Studies Initiative Lab*) hanno collaborato a questo progetto, l'applicazione aveva già oltre 130 milioni di utenti registrati che avevano condiviso quasi 16 miliardi di foto da tutto il mondo. I ricercatori hanno scelto di utilizzare visualizzazioni ad alta risoluzione che mostrano set di immagini completi per consentire l'esplorazione sia dei metadati delle foto (date di *upload*, filtri utilizzati, coordinate spaziali), che dei modelli creati dal contenuto delle fotografie, come pure delle singole fotografie, volendo dimostrare che la visualizzazione delle foto su più livelli, spaziali e temporali, può portare ad intuizioni culturali, sociali e politiche sia su scala planetaria che a livello delle singole città, e persino dei singoli individui. (Manovich e Hochman, 2013)

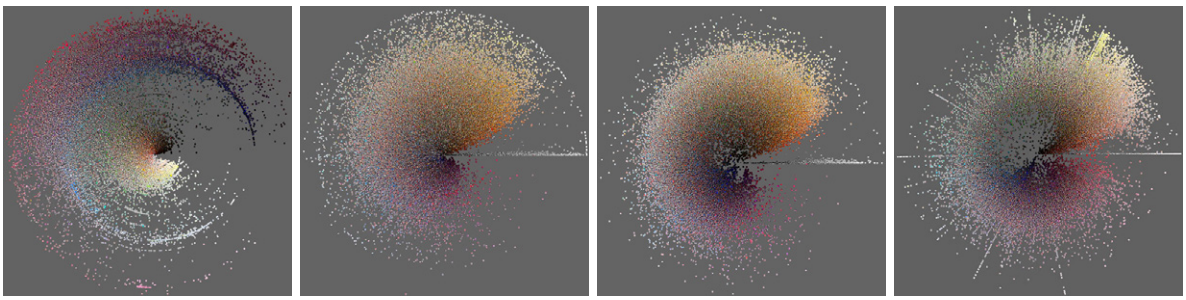


Fig. 3.11 *Phototrails*, *Global Cities' Visual Signatures*. Da sinistra a destra: New York, San Francisco, Tokyo, Bangkok

Lavorando su un campione di più di 2,3 milioni di foto *Instagram* condivise pubblicamente da oltre 300.000 persone in un periodo di quattro mesi — provenienti da 13 città come New York, San Francisco, Tokyo e Bangkok — e misurando molteplici attributi visivi come tonalità, luminosità, texture, i ricercatori hanno scoperto che ogni città ha la sua firma visiva unica su *Instagram*. Essi hanno anche trovato, all'interno del loro campione, differenze nell'uso dei filtri *Instagram*: la percentuale di foto a cui gli utenti Instagram hanno applicato filtri varia tra il 68 e l'81 percento; le città con la più alta percentuale di foto filtrate sono Tel Aviv, Londra e San Francisco, mentre la città con la più bassa percentuale è New York. (Manovich, Hochman e Chow, 2013)

Dal punto di vista tecnico, attraverso le API fornite dai più diffusi servizi di media sharing, sono stati rilevati milioni di foto condivise pubblicamente, insieme ai loro metadati. Ogni immagine è stata analizzata e quindi, utilizzando strumenti software, sono stati individuati e comparati gruppi di immagini, al fine di trovare i modelli sottostanti.

Per esplorare le foto generate dagli utenti, nel progetto sono stati utilizzati vari layout e tecniche di visualizzazione prendendo in analisi diversi aspetti dei dati, per presentarli in modi nuovi e rivelatori.

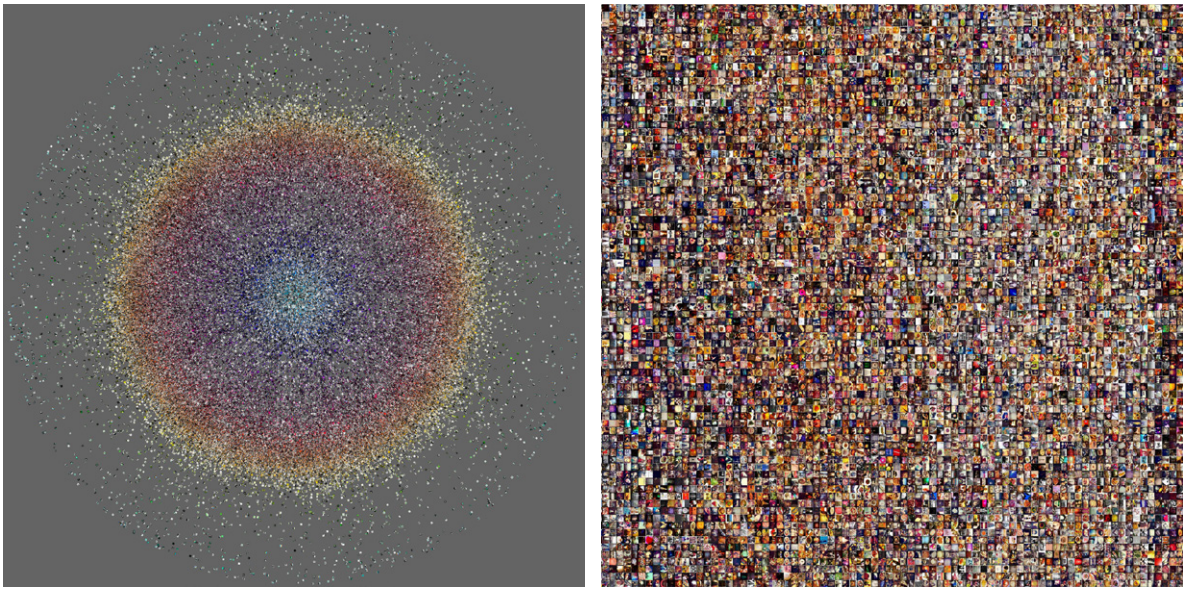


Fig. 3.12 *Phototrails*. Sinistra: visualizzazione radiale di 120.000 fotografie di sei città su *Instagram* in una settimana, in ordine di cronologico e di tinta media; destra: montaggio di 53.498 fotografie di Tokyo in ordine cronologico

Tra le diverse visualizzazioni, quelle radiali — particolarmente utili per studiare i cambiamenti temporali — posizionano le foto lungo un cerchio usando date di upload, coordinate geografiche o attributi visivi. Un parametro controlla l'angolo (posizione lungo il perimetro), l'altro parametro controlla il raggio (quanto è lontana una foto dal centro). Queste forme di visualizzazione compatte permettono di esplorare i dati su più dimensioni, presentando ad esempio le immagini organizzate secondo le loro caratteristiche visive (tonalità, luminosità, texture, ecc.) e in relazione a quando e dove sono state scattate. Il tutto in un'unica visualizzazione.

Le visualizzazioni a montaggio sono griglie di immagini "cucite" tra di loro. La visualizzazione di foto generate dall'utente secondo la posizione e l'ora (data di caricamento) rivela il particolare "ritmo visivo" di ogni luogo, mentre le caratteristiche visive multiple (tonalità, contrasto, saturazione, ecc.) ne determinano la "firma visiva", mostrando variazioni spazio-temporali nel colore, nel numero di foto condivise e nelle affinità visive. Ad esempio, le immagini possono essere organizzate in base al momento di caricamento, catturando così il "ritmo visivo" di un luogo nel tempo — ad esempio l'alternarsi del giorno e della notte che si manifesta nel cambiamento delle caratteristiche visive delle foto. Se invece si organizzano le foto scattate in un determinato periodo di tempo e in un determinato luogo in base alla luminosità media, alla tonalità media o alla saturazione dei colori di ogni foto, verrà rivelata una "firma visiva" caratterizzata dalle preferenze visive dominanti.

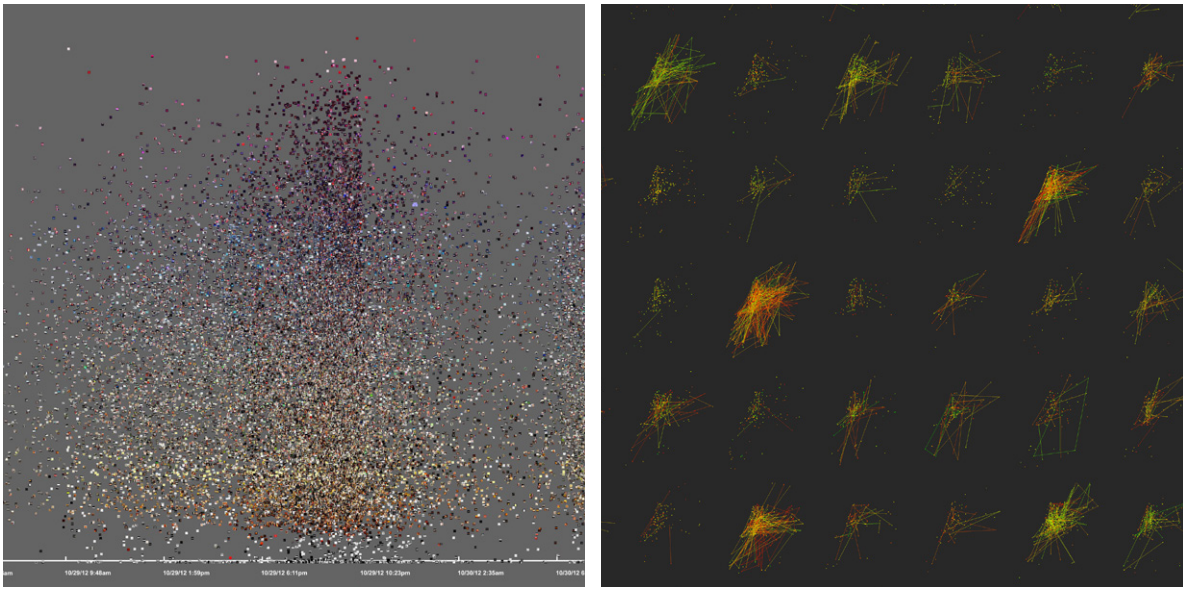


Fig. 3.13 *Phototrails*. Sinistra: *photoplot* di 34.993 fotografie di Brooklyn durante l'uragano Sandy in ordine cromatico e cronologico; destra: visualizzazione con punti e linee di 289 utenti di Tel Aviv secondo l'orario di caricamento

Le visualizzazioni *photoplot* utilizzano il principio del grafico a dispersione ma, invece di tracciare punti, il software *Image Plot*, progettato da Manovich e i suoi collaboratori, traccia singole immagini. Queste visualizzazioni permettono di vedere i modelli di cambiamento delle caratteristiche visive delle foto nel tempo e nei luoghi. Zoomando su una particolare città e in un determinato periodo, le foto caricate sui social media possono essere utilizzate anche per la lettura locale di attività sociali e culturali. In altre parole, anziché aggregare i contenuti digitali generati dai singoli utenti con lo scopo di mappare la società, dove gli individui e i loro diari media diventano invisibili, viene fatta una lettura più dettagliata dei dati. Visualizzando in diversi modi un set di foto caricate su *Instagram* dagli utenti di una determinata città durante i giorni di particolari eventi, è così possibile individuare profili culturali, sociali e politici di specifici luoghi durante particolari eventi, sociali o naturali.

Le visualizzazioni con punti e linee sono comunemente utilizzate per rappresentare grandi quantità di dati in una singola visualizzazione "condensata" di una città o di un Paese. I punti sono colorati utilizzando una sfumatura da verde a rosso per indicare diversi momenti della giornata: verde per la mattina, giallo per il pomeriggio, rosso per la sera. Quando due punti-foto sono stati fatti nella stessa ora, viene tracciata tra loro una linea. I risultati rivelano centri di concentrazione nella città in particolari ore, così come diversi tipi di utenti: alcuni scattano molte foto in una sola zona, altri si muovono rapidamente in tutta la città. Diversi utenti non scattano mai più di una foto all'ora, mentre altri scattano molte foto in brevi periodi di tempo. Alcuni infine scattano altre foto al mattino presto, mentre altri scattano solo durante la tarda serata. (Manovich e Hochman, 2013)

L'analisi di queste visualizzazione permette di mappare i comportamenti tipici degli individui e può fornire un profilo sociologico del luogo in cui essi vivono.

I tre progetti di visualizzazione diretta che abbiamo appena analizzato mostrano non solo che, mappando i dati visivamente, è più facile cogliere le informazioni importanti ed individuare più velocemente le tendenze e le correlazioni significative, ma anche che l'utilizzo nella visualizzazione degli artefatti nella loro forma visiva originale rende la lettura dei risultati più efficace ed immediata.

La forza di questa modalità di visualizzazione sta anche nel fatto che può essere utilizzata per rappresentare diversi tipi di materiali digitali: scansioni di immagini o di libri, foto digitali, ma anche fotogrammi di film, come visto in *Cinema Redux*, o ancora i messaggi di *Listening Post*.

Queste considerazioni sulla visualizzazione diretta suggeriscono la possibilità di implementare l'attuale interfaccia di *Internet Archive* inserendo nuove modalità di visualizzazione che, nel garantire l'analisi avanzata dei dati, consentano una rappresentazione visiva più efficace e possano essere utilizzate per i diversi tipi di materiali digitali presenti in *Internet Archive*.

Nel prossimo capitolo presenteremo un'ipotesi di implementazione in questo senso.

4. Implementazione di nuove funzioni su *Internet Archive*

Internet Archive contiene un numero esorbitante, sempre in aumento, di artefatti digitali. Abbiamo visto nei capitoli precedenti che non sempre la navigazione è semplice o intuitiva, e che non sempre la visualizzazione degli oggetti è la più adeguata a rappresentare il risultato della ricerca.

Dall'analisi dei casi studio fatta nel Capitolo 2 abbiamo capito come alcune interfacce siano state progettate in modo specifico per lo scopo del progetto e non siano quindi mutuabili per un uso più generalizzato su qualunque tipo di materiale presente in *Internet Archive*.

Abbiamo anche osservato che, nei casi in cui l'interfaccia non è stata progettata *ad hoc*, le interfacce di esplorazione utilizzate, pur essendo visuali, non mostrano l'immagine degli artefatti nella visualizzazione dei risultati della navigazione.

Partendo da queste evidenze, il mio progetto propone un'implementazione dell'attuale interfaccia di *Internet Archive* con l'obiettivo di superare i limiti individuati e migliorare sia la navigazione che la visualizzazione dei risultati.

Tra i vari tipi di materiali digitali presenti all'interno di *Internet Archive*, ho scelto di utilizzare l'immagine come oggetto principale del mio progetto perché è un mezzo estremamente immediato e non è stato preso in considerazione nei casi studio analizzati in precedenza.

4.1. Lo stato dell'arte

Una panoramica dello stato attuale dell'interfaccia di *Internet Archive* può essere utile per individuare i suoi punti critici.

La collezione Image è una libreria di immagini caricate dagli utenti di *Internet Archive* che contiene ad oggi più di 4 milioni di immagini, centinaia di sottocollezioni e decine di oggetti di altri media type (testi, video, audio, software). All'interno della collezione sono presenti mappe geografiche, immagini astronomiche e fotografie di opere d'arte. Molte di queste immagini sono scaricabili liberamente.

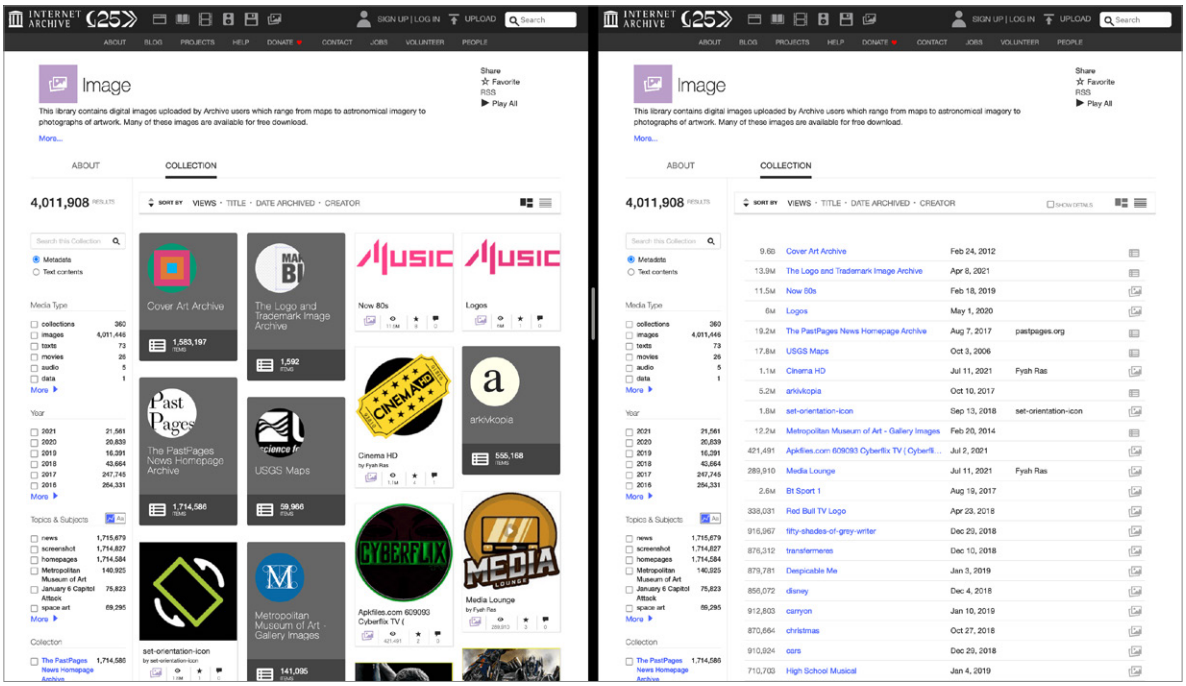


Fig. 4.1 Collezione *Image*. Sinistra: *thumbnails view*; destra: *list view*

Nella figura possiamo vedere le due modalità attualmente disponibili su *Internet Archive* per mostrare gli elementi all'interno di una collezione, in questo caso in ordine decrescente di visualizzazioni accumulate.
A sinistra, la *thumbnails view* mostra piccole immagini di anteprima degli oggetti all'interno della collezione, oltre che ai loro nomi, autori, media type dell'oggetto, numero di visualizzazioni e di commenti. Per le sottocollezioni è indicato solamente il numero di oggetti all'interno.
A destra, la *list view* mostra un elenco testuale con i nomi degli oggetti (preceduti dal suo numero di visualizzazioni), la data di caricamento dell'oggetto, l'autore e un'icona che rappresenta il media type dell'oggetto.
Non è insolito che gli oggetti in *list view* appaiano in ordine sfalsato; ad esempio il secondo oggetto per numero di visualizzazioni, la collezione *The PastPages News Homepage Archive*, è in quinta posizione. Questo è un problema generale di *Internet Archive* e si presenta in qualunque contesto, come si può vedere in **fig. 4.2**: l'oggetto evidenziato ha il numero più alto di visualizzazioni (più di 550 milioni) ma appare in nona posizione nell'elenco.

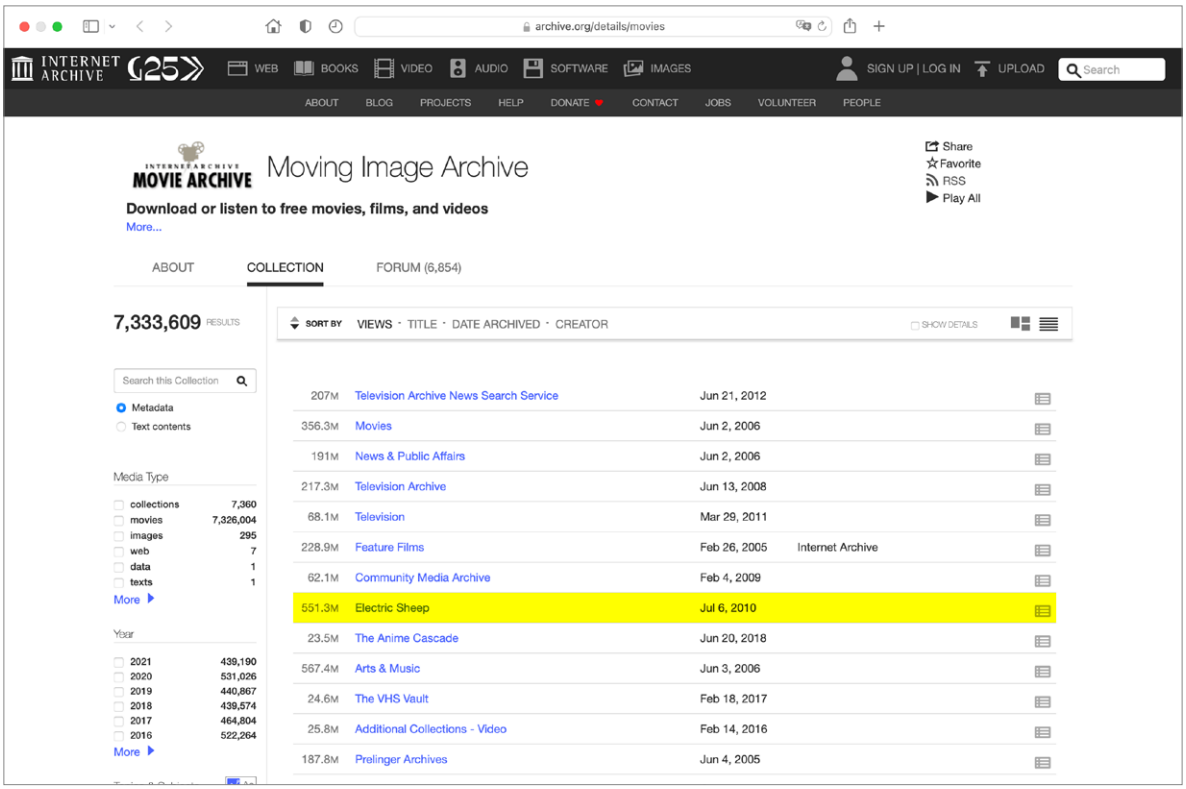


Fig. 4.2 Collezione *Moving Image Archive*

Sebbene l'interfaccia corrente funzioni bene per la gran parte degli utilizzi di *Internet Archive*, non è sufficiente per rappresentare grandi insiemi di immagini, specialmente se si intende porre l'accento sui legami tra le immagini e i loro metadati, e sulle differenze tra gli stessi.

Le piccole dimensioni delle anteprime degli oggetti in *thumbnails view* non consentono un'osservazione soddisfacente o un confronto delle immagini informato e consapevole, mentre queste ultime scompaiono completamente in *list view* a favore del puro dato numerico.

Per colmare queste lacune, ho progettato alcuni prototipi di interfacce interattive, più precisamente nuove modalità di visualizzazione dei risultati di *Internet Archive*, al fine di dare più visibilità all'immagine, pur mantenendo un certo livello di conformità con l'estetica e il linguaggio grafico preesistente dell'archivio per un'implementazione *seamless*, senza soluzione di continuità.
La scelta di non progettare una nuova interfaccia *ex novo*, ma di progettare una serie di viste che si inseriscono nell'attuale cornice di *Internet Archive*, è motivata dalla volontà di avere un approccio incrementale alla progettazione. Ho ritenuto utile mantenere la maggior parte delle funzionalità dell'attuale interfaccia, perché già perfettamente funzionanti e funzionali all'esplorazione dei risultati di ricerca.

Alla luce di quanto visto nei capitoli precedenti, il metodo di visualizzazione che ho scelto di utilizzare nell'implementazione dell'interfaccia attuale di *Internet Archive* è quello della visualizzazione diretta, che rappresenta i risultati di ricerca attraverso la forma visiva originale degli artefatti stessi.

4.2. Processo e metodo di lavoro

Dopo aver individuato i punti da affrontare, ho pensato a diverse situazioni in cui un metodo di visualizzazione diverso potrebbe migliorare la lettura delle immagini sull'archivio.

Per poter ottenere delle simulazioni soddisfacentemente realistiche, ho utilizzato i software *WebScraper* e *ParseHub* per fare *scraping* di dati relativi ad alcune collezioni su *archive.org*, dando priorità a quelle composte per la maggior parte da immagini.
Una di queste collezioni è *Magazine Art: Food and Beverages*, contenente alcune centinaia di fotografie e scansioni di immagini pubblicitarie di riviste del secolo scorso.

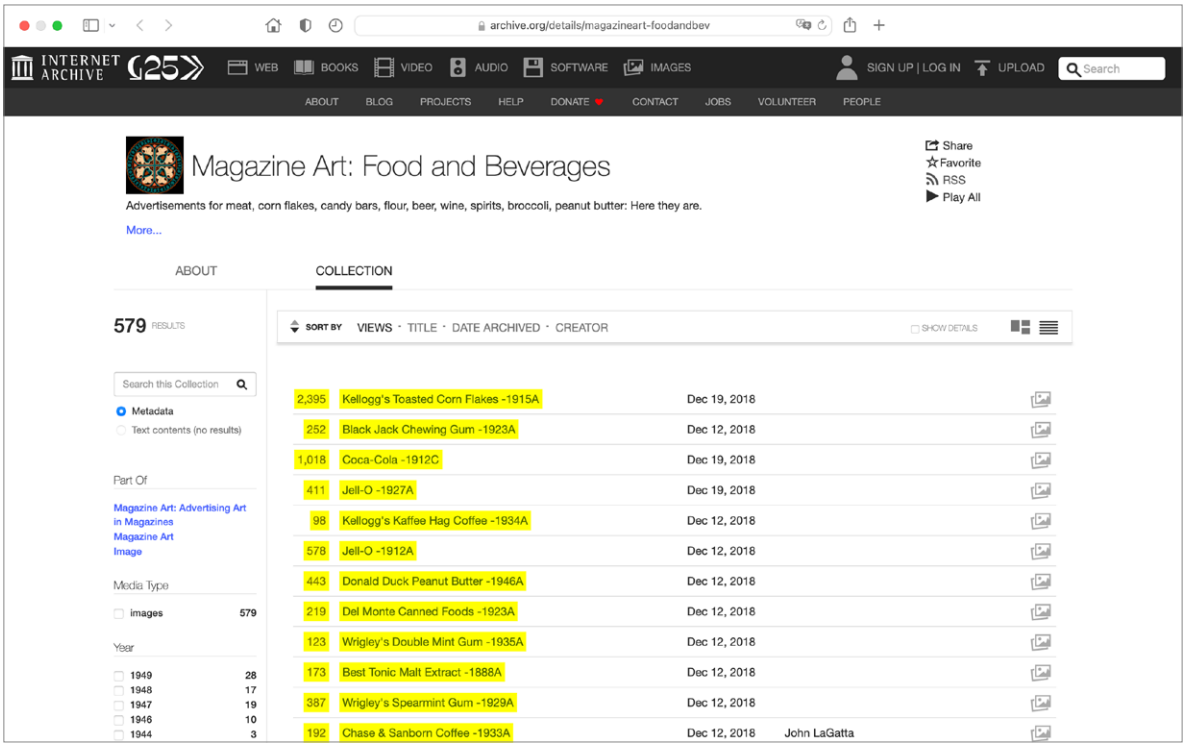


Fig. 4.3 Collezione *Magazine Art: Food and Beverages* in *list view* in ordine descrescente di visualizzazioni

Per ogni collezione considerata sono state eseguite operazioni ricorsive di estrazione sia dalla pagina di dettaglio delle collezioni (in questo caso *https://archive.org/details/magazineart-foodandbev*) sia dalle pagine di dettaglio dei singoli oggetti (ad esempio *https://archive.org/details/KelloggsToastedCornFlakes1915A*) secondo le seguenti variabili:

- numero di visualizzazioni;
- nome, titolo;
- autore, artista, fonte;
- anno di creazione, di pubblicazione;
- URL della pagina di dettaglio;
- URL dell'immagine.

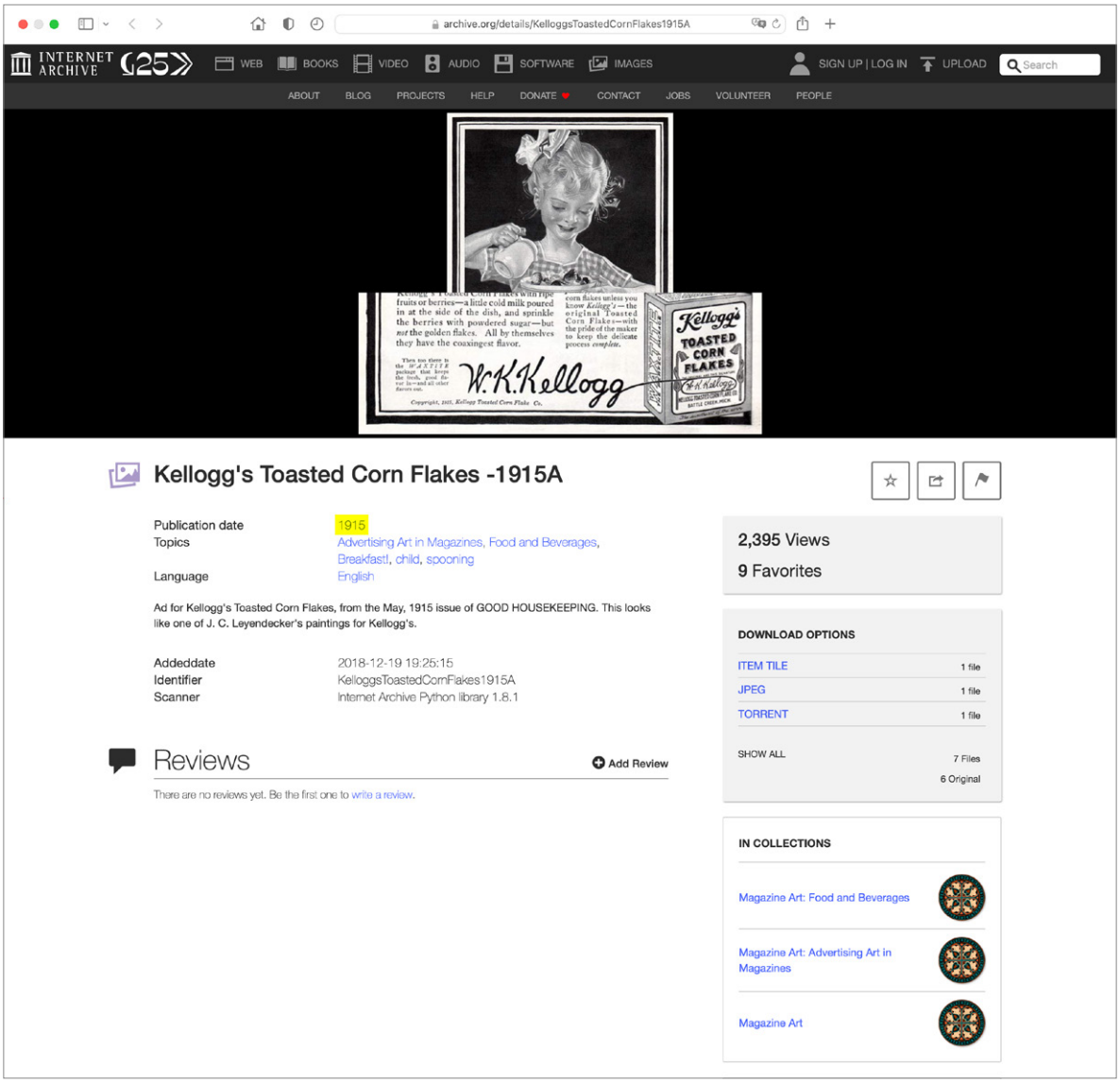


Fig. 4.4 Pagina di dettaglio dell'oggetto *Kellogg's Toasted Corn Flakes -1915A*

Ho preso in considerazione tre macrocollezioni principali — *NASA Image Exchange Collection*, *Magazine Art* e *Album Covers* — da cui ho ottenuto diversi dataset.

Per la prima collezione ho filtrato i risultati usando 9 diversi valori del datapoint "*Creators*", corrispondenti ad altrettanti centri di ricerca e sviluppo gestiti dalla NASA. Per ogni centro spaziale ho ricavato i dati dei primi 75 oggetti più visualizzati, per un totale di 675 membri.

Per la collezione *Magazine Art* ho considerato le prime 10 categorie di immagini secondo la tipologia di prodotto rappresentato. Per ogni sottocollezione ho estratto i dati dei primi 75 oggetti per numero di visualizzazione, per un totale di 750 membri.

Infine, ho disposto le immagini della collezione *Album Covers* secondo 2 valori diversi, cronologico e di visualizzazioni, per un totale di 125 membri.

4.3. Visualizzare le immagini su *Internet Archive*

Nei prototipi sviluppati, l'attenzione è rivolta in particolare al numero di visualizzazioni degli oggetti caricati su *Internet Archive*, che ritengo molto interessanti e utili per l'individuazione di tendenze nell'utilizzo dell'archivio da parte dei suoi utenti.

4.3.1. *Alluvial view*

La prima nuova modalità di visualizzazione è la *alluvial view*, che prende il suo nome dal diagramma alluvionale, un'evoluzione del sistema a coordinate parallele e un tipo di diagramma di flusso in cui i blocchi rappresentano *cluster* di nodi e i flussi tra i blocchi rappresentano cambiamenti nelle strutture di rete in un certo periodo di tempo.

Questa vista ha l'obiettivo di mostrare la relazione tra immagini e metadati (in questo caso particolare, gli artisti collegati agli album e il numero di visualizzazioni degli album) in una collezione. Attraverso il diagramma alluvionale è possibile sia osservare il rapporto tra immagini e metadati, sia visualizzare la popolarità di ogni immagine attraverso il numero di visualizzazioni.

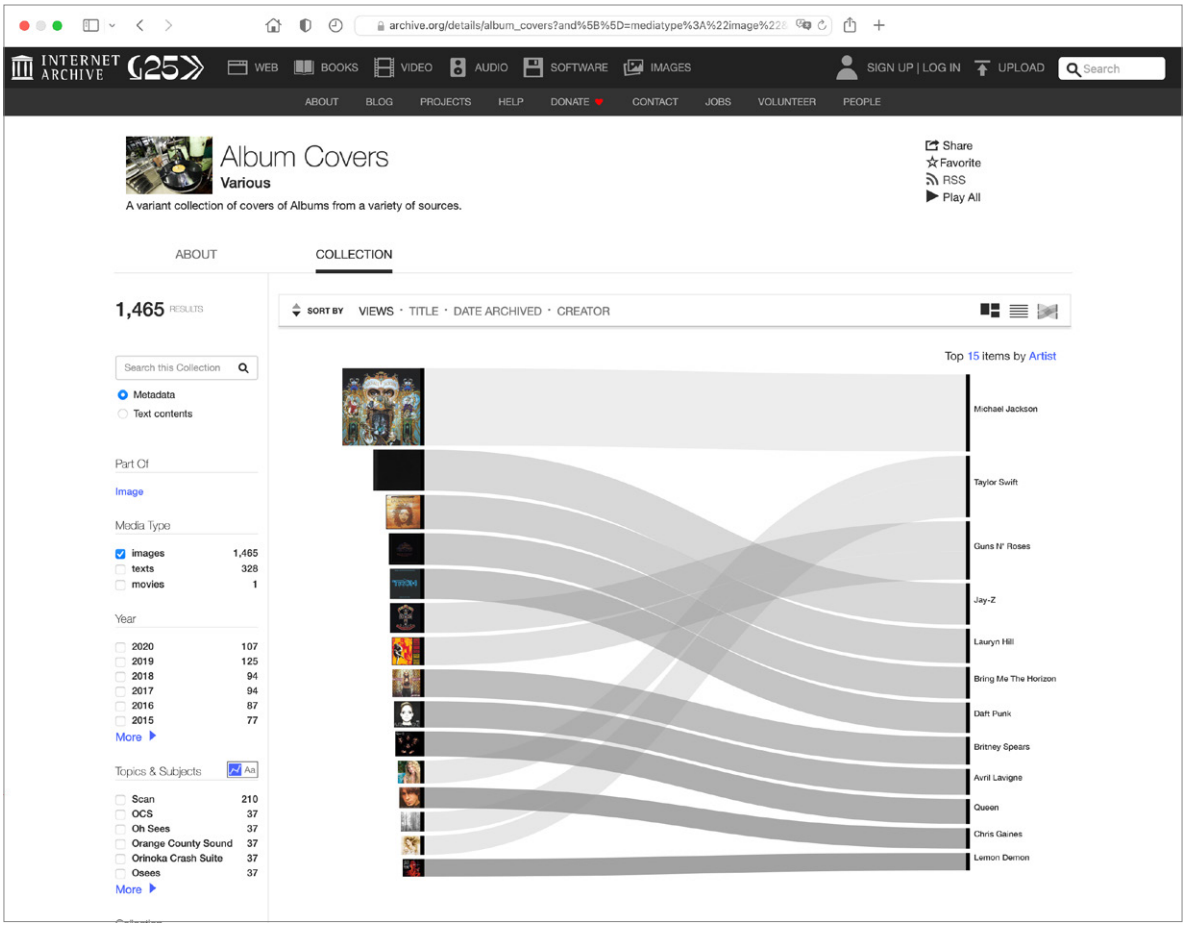


Fig. 4.5 Collezione *Album Covers* in *alluvial view*

In **fig. 4.5** possiamo vedere la collezione *Album Covers*, dedicata alle immagini di copertina di album musicali, in *alluvial view*.

A sinistra, le prime 15 illustrazioni di album più visualizzate su *Internet Archive* sono disposte in colonna; le altezze delle immagini — così come le altezze dei blocchi a loro corrispondenti e lo spessore dei flussi provenienti da essi — sono proporzionali al numero di visualizzazioni accumulate da ogni oggetto. Nella colonna a destra appaiono invece i nomi degli artisti contenuti nei metadati degli oggetti-album. In presenza di più album di uno stesso artista, i flussi provenienti dai blocchi-album a sinistra convergono nello stesso blocco-artista a destra. Il numero di oggetti visualizzati è regolabile dall'utente per ottenere diagrammi più estensivi.

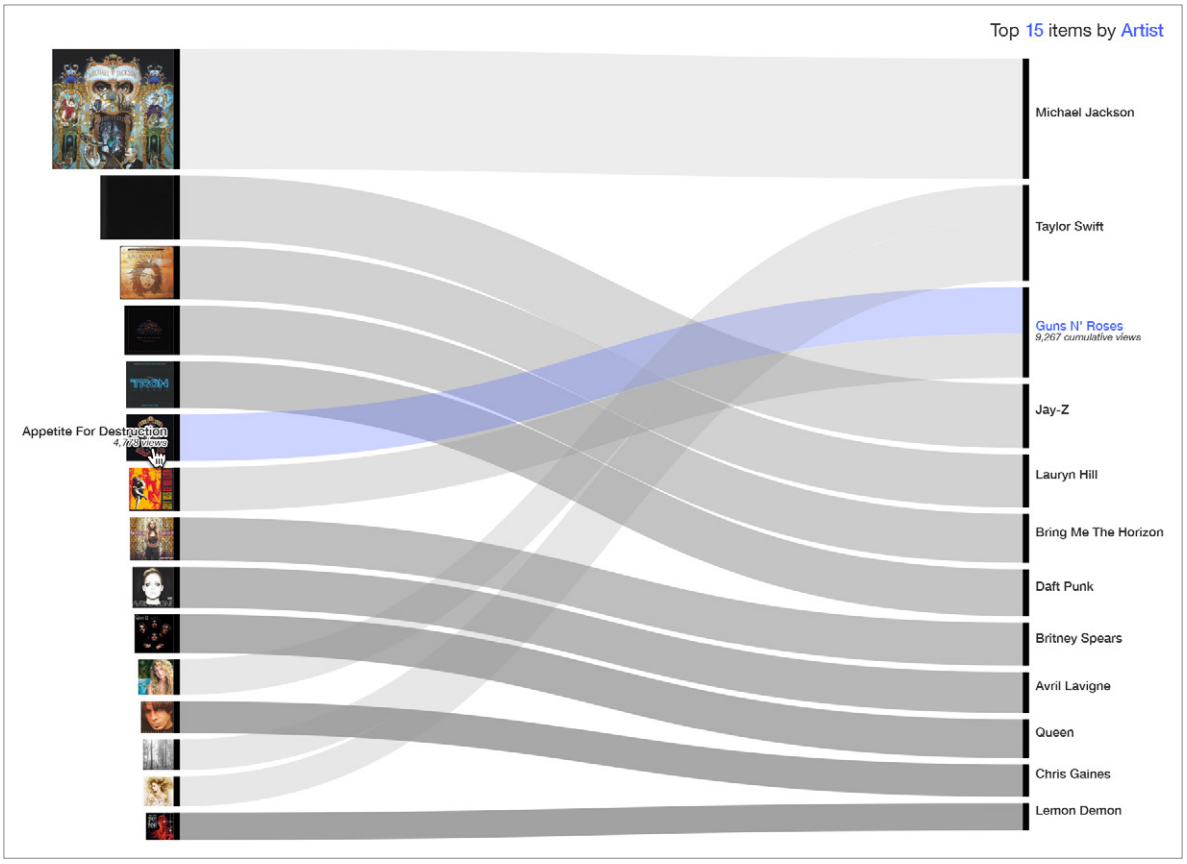


Fig. 4.6 Utilizzo del mouse per selezionare l'album *Appetite For Destruction* con conseguente evidenziazione della stringa "Guns N' Roses" e flusso corrispondente

Il mouse è il mezzo principale per l'interazione con la *alluvial view*, in quanto è possibile accedere a più informazioni sugli oggetti con un click o portando il cursore del mouse sopra di essi.

Il passaggio del mouse su un album a sinistra mostra il numero di volte che è stato visualizzato, l'artista corrispondente e il numero di visualizzazioni cumulative per quell'artista. Il flusso proveniente dall'album e il nome dell'artista si colorano di blu. La pagina di dettaglio del singolo oggetto è raggiungibile cliccando l'oggetto.

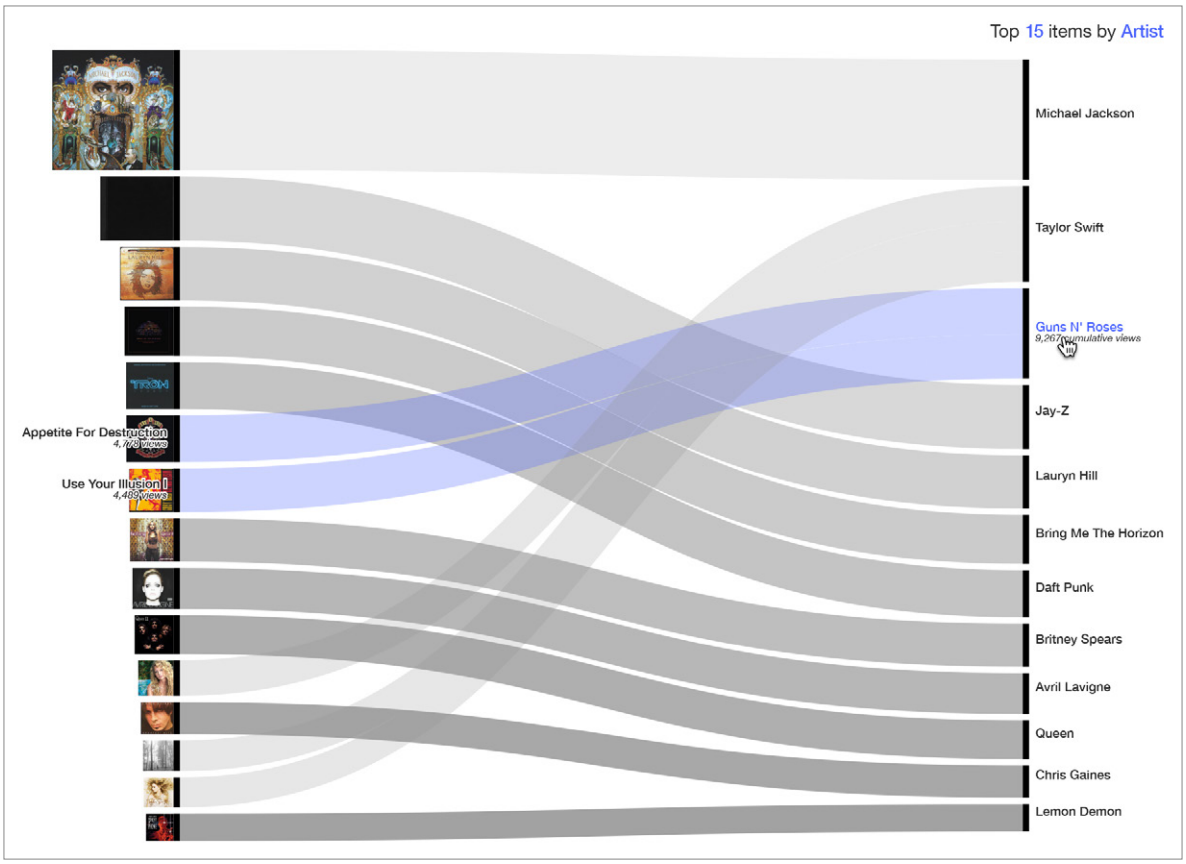


Fig. 4.7 Utilizzo del cursore per selezionare la stringa "Guns N' Roses" con conseguente evidenziazione degli album *Appetite For Destruction* e *Use Your Illusion I* e flussi corrispondenti

Lo stesso gesto compiuto su un artista a destra rivela le sue visualizzazioni complessive, tutti gli album ad egli associati e il loro numero di visualizzazioni. Come nell'altra interazione, i flussi provenienti dagli album e il nome dell'artista vengono evidenziati in blu. Facendo click su un artista si può accedere a una nuova pagina di ricerca dedicata, con nuove parole chiave per limitare i risultati ad oggetti-album che contengono l'artista selezionato nei loro titoli o metadati.

La modalità *alluvial view* può essere applicata ad altri tipi di materiali, come ad esempio la collezione di un museo di opere d'arte in relazione agli artisti o i luoghi di provenienza delle opere.
Un'altra collezione ben rappresentabile in *alluvial view* è la *NASA Image Exchange Collection*, ponendo nella colonna di destra i nomi dei centri spaziali.

4.3.2. *Timeline view*

Considerando la stessa collezione di immagini di copertina di album, passiamo alla seconda nuova modalità di visualizzazione, la *timeline view*. Prende il suo nome dal grafico a bolle su linea del tempo, un tipo di visualizzazione dati composta.

La *timeline view* confronta la popolarità (misurata attraverso le visualizzazioni) di una serie di immagini poste su una linea del tempo. Maggiore è il numero delle visualizzazioni, maggiore è la dimensione delle bolle che contengono le immagini.

L'intervallo di tempo analizzato può essere regolato dall'utente con il selettore in alto a sinistra.

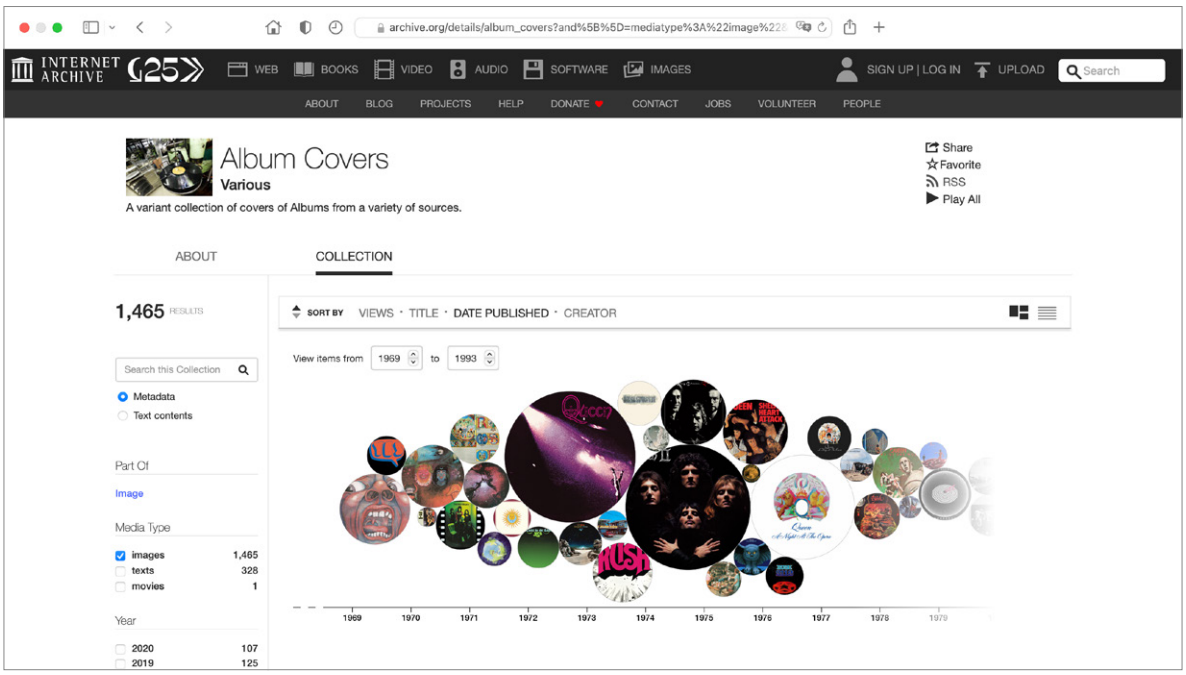


Fig. 4.8 Collezione *Album Covers* in *timeline view*

Nella figura possiamo vedere il dataset in *timeline view*, in particolare le immagini di album rilasciati dal 1969 al 1993, contenute in bolle di dimensioni proporzionali al numero di visualizzazioni accumulate.

Se il diagramma è troppo largo per la finestra del browser o del dispositivo, è possibile usare la rotellina del mouse, i tasti freccia della tastiera o il *click and drag* per scorrere a destra e a sinistra, andando così avanti e indietro nella linea del tempo.

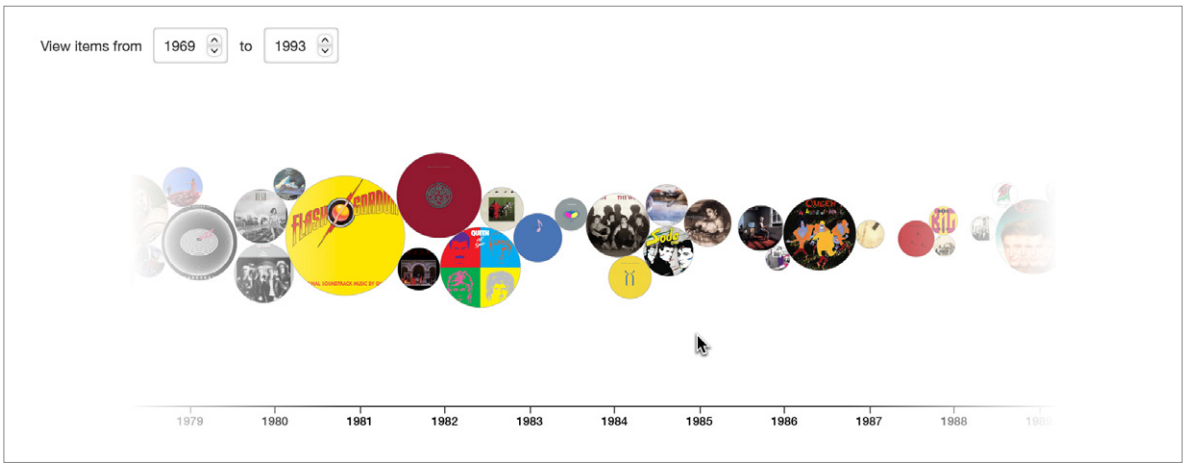


Fig. 4.9 Scorrimento in avanti della linea del tempo

Al passaggio del mouse su qualunque immagine di oggetto all'interno del diagramma appaiono più informazioni sull'oggetto corrispondente, in particolare il nome, l'autore, la data di rilascio e il numero di visualizzazioni; gli oggetti non selezionati sfumano leggermente per facilitare la lettura del diagramma e la visibilità dell'oggetto selezionato. Cliccandolo si può accedere alla sua pagina di dettaglio.

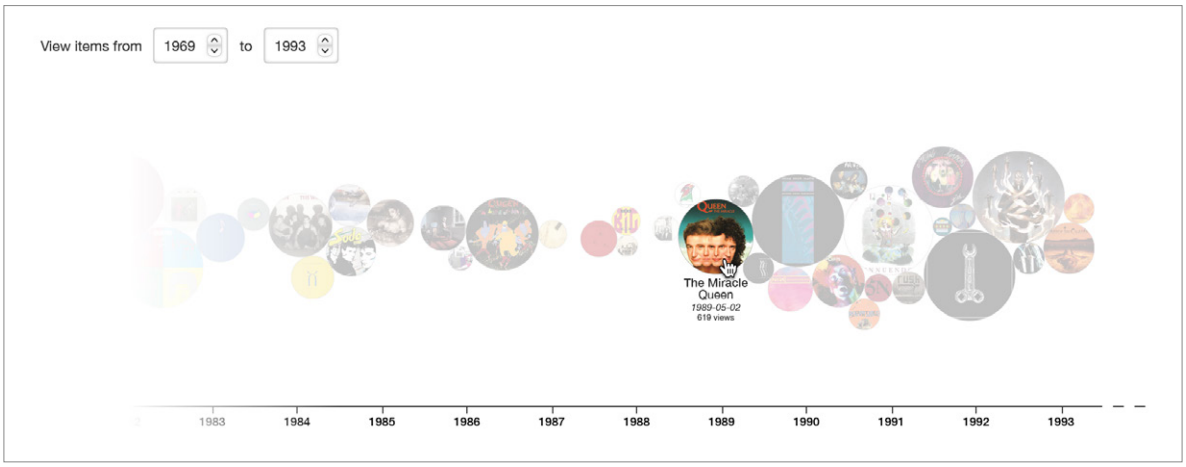


Fig. 4.10 Utilizzo del cursore per evidenziare l'album *Miracle* dei Queen

La modalità *timeline view* è utilizzabile per rappresentare altri materiali in ordine cronologico, come ad esempio una collezione di opere d'arte create dallo stesso artista nel corso della sua carriera o le molteplici versioni dello stesso software, come si può vedere nella collezione *Operating System CD-ROMs*.

4.3.3. *Bubble view*

Per dimostrare la terza nuova modalità di visualizzazione, la *bubble view*, è necessario utilizzare una nuova collezione di oggetti.

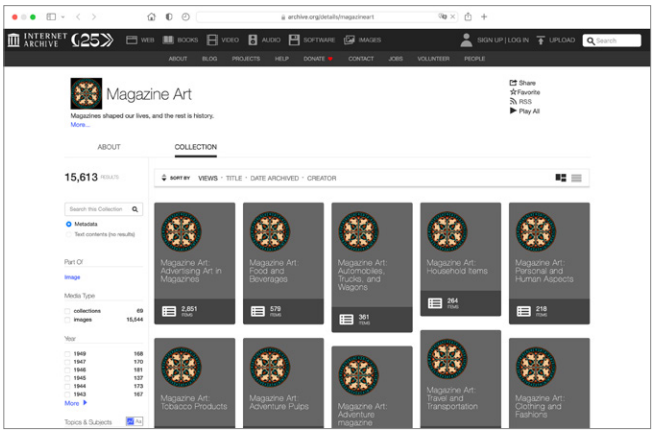


Fig. 4.11 Collezione *Magazine Art* in *thumbnails view* in ordine decrescente di visualizzazioni

La collezione *Magazine Art* è dedicata a scansioni e fotografie di annunci che sono stati pubblicati nel periodo tra il 1835 e il 1970. Sono presenti decine di sottocollezioni che suddividono e categorizzano ulteriormente gli oggetti in *Magazine Art* in base al tipo di prodotto o servizio raffigurato.

Come già accennato in precedenza, la modalità *thumbnails view* mostra solamente piccole icone per le sottocollezioni e non prevede immagini di anteprima degli elementi al loro interno, perciò è difficile rappresentare e confrontare adeguatamente gli insiemi. Per rimediare a questa mancanza, la *bubble view* si serve dei primi oggetti all'interno di ogni sottocollezione, come in **fig. 4.12**.

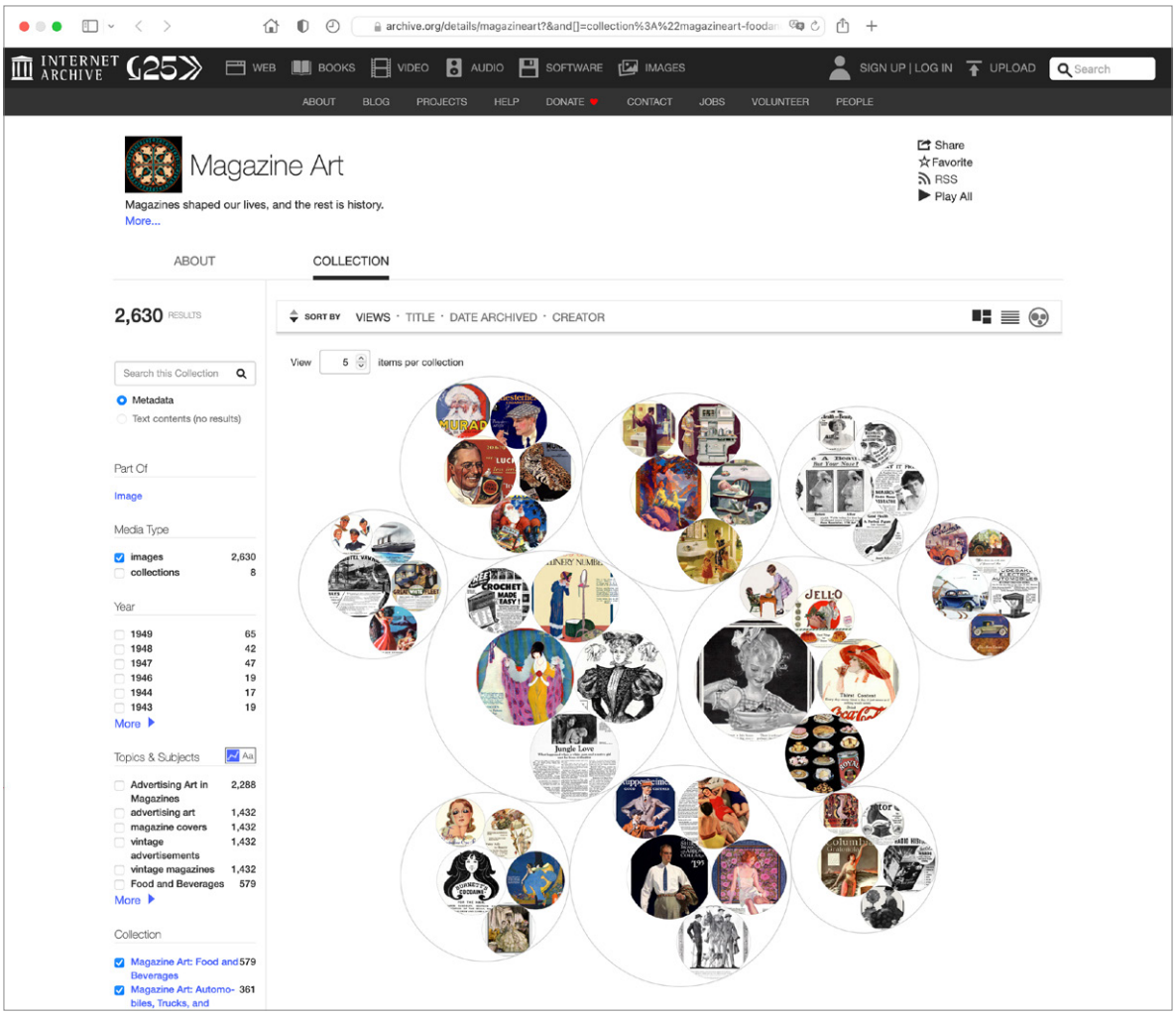


Fig. 4.12 Collezione *Magazine Art* in *bubble view*: analisi dei primi 5 oggetti di ogni collezione secondo le loro visualizzazioni

Le immagini degli oggetti sono contenute in bolle tangenti tra di loro di dimensioni proporzionali al numero di visualizzazioni da esso accumulato; oggetti appartenenti alla stessa collezione sono racchiusi in cerchi più grandi, tangenti a quelli interni. L'utente può regolare il numero di oggetti visualizzati per ogni collezione con il selettore in alto a sinistra.

Fig. 4.13 Dettaglio del selettore, "*Vedi [5] oggetti per collezione*"

Ponendo il cursore del mouse all'interno di un cerchio-collezione appare il nome della collezione e il numero di immagini al suo interno, mentre le collezioni non selezionate diventano meno visibili. Cliccando l'area del cerchio si può accedere alla pagina di dettaglio della collezione.

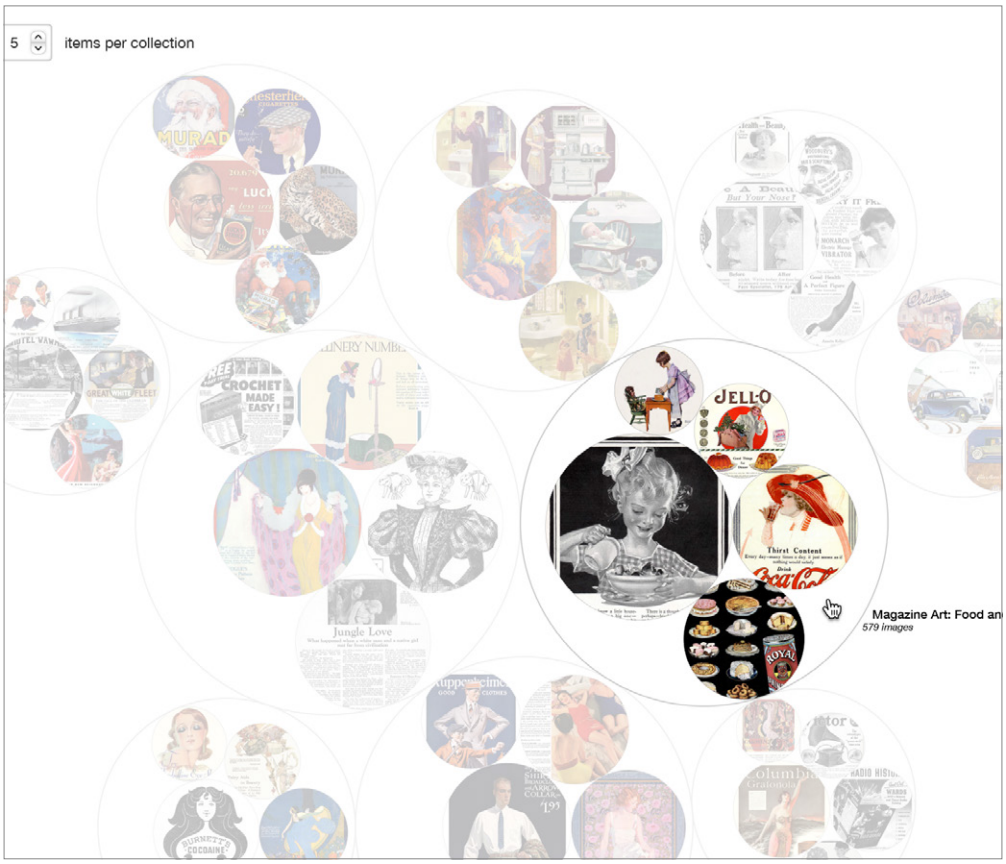


Fig. 4.14 Utilizzo del cursore per evidenziare la collezione *Food and Beverages*

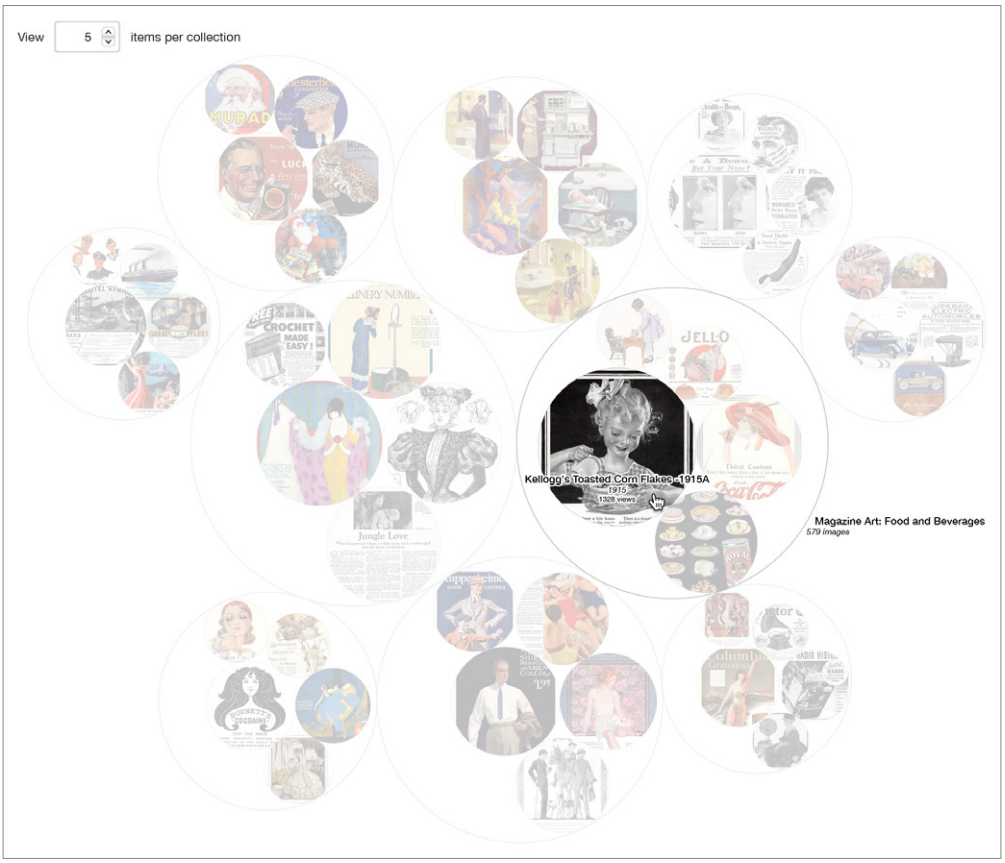


Fig. 4.15 Utilizzo del cursore per evidenziare l'oggetto *Kellogg's Toasted Corn Flakes -1915A* all'interno della collezione *Food and Beverages*

È possibile passare il mouse sopra uno degli oggetti all'interno di una collezione per evidenziarlo e rivelare il numero di visualizzazioni da esso accumulate. La pagina di dettaglio dell'oggetto è raggiungibile con un click.

La modalità *bubble view* è utilizzabile anche per altri tipi di materiali divisibili in sottogruppi o categorie, come per esempio la collezione *NASA Image Exchange Collection*.

In sintesi, il *mock up* ha riguardato tre nuovi metodi di vista dei risultati di ricerca di *Internet Archive*:

- l'*alluvial view*, che permette di unire il numero di visualizzazioni di oggetti con metadati condivisi, per esempio l'autore, all'interno di una collezione;
- la *timeline view*, che impiega un grafico a bolle sovrapposto a una linea del tempo per indicare gli oggetti più visualizzati in una collezione e disporli in ordine cronologico, con possibilità da parte dell'utente di muoversi lungo la linea temporale;
- la *bubble view*, che utilizza categorie preesistenti per suddividere gli oggetti di una collezione e confrontare il numero cumulativo di visualizzazioni.

Conclusioni

L'analisi del contesto sempre in evoluzione in cui si inserisce *Internet Archive* e lo studio di alcuni progetti di visualizzazione già realizzati sui materiali dell'archivio hanno permesso di individuare alcuni aspetti critici presenti nell'attuale interfaccia grafica utilizzata da *Internet Archive*, suggerendo nuove opportunità di visualizzazione degli artefatti digitali non ancora esplorate. In particolare, è stata riscontrata la necessità di progettare delle visualizzazioni che rendessero conto della varietà del materiale visivo contenuto all'interno dell'archivio.

L'approfondimento delle tematiche legate alla visualizzazione delle informazioni e l'attribuzione alla visualizzazione diretta di un ruolo centrale nel progetto hanno ispirato la progettazione di nuove implementazioni dell'interfaccia grafica.

Nello specifico, l'implementazione proposta in questa tesi ha inteso fornire una risposta progettuale alla necessità di semplificare l'accesso alla sempre crescente quantità di materiali digitali presenti in *Internet Archive*, rendendo più intuitiva e immediata la selezione dei parametri necessari per effettuare la ricerca e offrendo nuove modalità di visualizzazione più innovative ed efficaci. In particolare, sono state progettate viste che mettono in primo piano gli oggetti digitalizzati, e che usano i loro metadati per offrire percorsi esplorativi nell'archivio.

Le tre viste presentate nell'ipotesi di progetto rispondono ad obiettivi diversi.

L'*alluvial view* mostra le relazioni tra gli oggetti di una collezione e i metadati e consente di misurare la popolarità di ogni immagine attraverso il numero di visualizzazioni. Con la *timeline view* è possibile visualizzare la popolarità di una serie di oggetti sovrapponendo un grafico a bolle, le cui dimensioni sono proporzionali al numero di visualizzazioni, ad una linea del tempo, la cui lunghezza può essere regolata dall'utente. Infine la *bubble view*, utilizzando categorie preesistenti, permette di suddividere le collezioni in gruppi di oggetti e di confrontare il numero cumulativo di visualizzazioni di ogni gruppo.

Le diverse modalità di visualizzazione sono state progettate tenendo conto anche del fatto che i materiali presenti in *Internet Archive* sono di diversi tipi, perseguendo quindi l'obiettivo di poter rappresentare visivamente qualunque tipologia di materiale digitale.

Le soluzioni proposte in questa tesi di ricerca risultano capaci da un lato di sfruttare l'ampiezza dell'archivio e dall'altro di valorizzare anche visivamente la specificità e la ricchezza degli artefatti digitali presenti in *Internet Archive*.

Bibliografia

Cap. 1 *Internet Archive*

Internet Archive
<https://archive.org/>

Di Corinto, Arturo
"Io leggo digitale", l'iniziativa italiana (e non solo) per leggere da casa gratis, sezione *Italian Tech*, *La Repubblica*, 2020
<https://www.repubblica.it/tecnologia/2020/04/17/news/_io_leggo_digitale_l_iniziativa_italiana_per_leggere_da_casa_gratis-254305776/>

Porro, Gabriele
Wikipedia e Internet Archive stanno costruendo una grande biblioteca digitale, 2019
<https://www.wired.it/internet/web/2019/11/05/wikipedia-internet-archive/>

Wikimedia Italia
Libri liberi e biblioteche di emergenza: da Internet Archive a Wikisource, sezione *Blog e News*, 2020
<https://www.wikimedia.it/libri-liberi-e-biblioteche-di-emergenza-da-internet-archive-a-wikisource/>

Giannicchi, Federico
Internet Archive o Archive.org: come funziona e come recuperare un sito internet, *WebHero*, 2020
<https://www.webhero.it/seo/internet-archive-org/>

Kramer, Anna
The internet is splitting apart. The Internet Archive wants to save it all forever, *Protocol*, 2021
<https://www.protocol.com/internet-archive-preserving-future>

Romano, Aja
A lawsuit is threatening the Internet Archive — but it's not as dire as you may have heard, *Vox*, 2020
<https://www.vox.com/2020/6/23/21293875/internet-archive-website-lawsuit-open-library-wayback-machine-controversy-copyright>

Rijtano, Rosita
Editori contro l'Internet Archive: stop al prestito libero di ebook, *La Repubblica*, 2020
<https://www.repubblica.it/tecnologia/sicurezza/2020/06/16/news/editori_contro_l_internet_archive_stop_al_prestito_libero_di_ebook-259354027/>

Jones, Josh
The Internet Archive Will Digitize & Preserve Millions of Academic Articles with Its New Database, "Internet Archive Scholar", *Openculture*, 2020
<https://www.openculture.com/2020/09/internet-archive-scholar.html>

Kwon, Diana
More than 100 scientific journals have disappeared from the Internet, *Nature*, 2020
<https://www.nature.com/articles/d41586-020-02610-z>

Mandić, Slobodan

Internet Archive e nuove tipologie di fonti storiche, *Diacronie* N° 8-4, documento 7, 2011
<http://journals.openedition.org/diacronie/3561>

Taei, Payman

How to Transform Boring and Dry Reports with Data Visualization, *Towards Data Science*, 2018
<https://towardsdatascience.com/how-to-transform-boring-and-dry-reports-with-data-visualization-81fd908bcc62>

Ciociola, Chiara

The Deleted City, social web archeology, *Neural*, 2012
<http://neural.it/2012/01/the-deleted-city-social-web-archeology/>

Vijgen, Richard

The Deleted City 3.1, 2017
<http://deletedcity.net/>

GDELT Project

GDELT 2.0 Television API Debuts!, 2018
<https://blog.gdeltproject.org/gdelt-2-0-television-api-debuts/>

2016 Campaign Television Tracker, 2015
<http://television.gdeltproject.org/cgi-bin/iatv_campaign2016/iatv_campaign2016>

Leetaru, Kalev; Watzman, Nancy

New Research Tool for Visualizing Two Million Hours of Television News, *Internet Archive Blogs*, 2016
<http://blog.archive.org/2016/12/20/new-research-tool-for-visualizing-two-million-hours-of-television-news>

Gallagher, Erin

#NetNeutrality Tweets May 2017, *Internet Archive*, 2017
<https://archive.org/details/NetNeutralityTweetsMay2017>

#NetNeutrality hashtag visualizations, *Medium*, 2017
<https://erin-gallagher.medium.com/NetNeutrality-hashtag-visualizations-f7ec8325e6eb>

#UniteTheRight 2017 data visualizations and datasets, *Internet Archive*, 2017
<https://archive.org/details/UTRdata>

#UniteTheRight Automation and Deleted Tweets, *Medium*, *Documenting DocNow*, 2017
<https://news.docnow.io/UniteTheRight-automation-and-deleted-tweets-9dbf4b641755>

Internet Archive Network Visualization Project, *Medium*, 2017
<https://erin-gallagher.medium.com/internet-archive-network-visualization-2b2c53570b85>

Klingermann, Mario

The Internet Archive's Map of Book Subjects, *Internet Archive*, 2020
<https://web.archive.org/web/20201112002859/http://incubator.quasimondo.com/internetarchive/InternetArchiveBookSubjectsMap.html>

Yau, Nathan
Map of book subjects on Internet Archive, *FlowingData*, 2014
<https://flowingdata.com/2014/10/20/map-of-book-subjects-on-internet-archive/>

Van der Maaten, Laurens; Hinton, Geoffrey
Visualizing Data using t-SNE, *Journal of Machine Learning Research-9*, 2008
<http://lvdmaaten.github.io/publications/papers/JMLR_2008.pdf>

Violante, Andre
An Introduction to t-SNE with Python Example, *Towards Data Science*, 2018
<https://towardsdatascience.com/an-introduction-to-t-sne-with-python-example-5a3a293108d1>

Padia, Kalpesh; Alnoamany, Yasmin; Weigle, Michele C.
MS Thesis - Visualizing Digital Collections at Archive-It, 2012
<https://ws-dl.blogspot.com/2012/08/2012-08-10-ms-thesis-visualizing.html>

ResearchGate, *Visualizing digital collections at Archive-It*, *Web Science and Digital Libraries Research Group*, 2012
<https://www.researchgate.net/publication/262396061_Visualizing_digital_collections_at_archive-it>

Cap. 3 Visualizzazione delle informazioni e visualizzazione diretta

Manovich, Lev
What is visualisation?, *Visual Studies*, 26:1, 36–49, 2011
<https://doi.org/10.1080/1472586X.2011.548488>

Media Visualization: Visual techniques for exploring large media collections; Gates, Kelly (a cura di), *The International Encyclopedia of Media Studies* Vol. VI: *Media Studies Futures*, 2011
<http://manovich.net/index.php/projects/media-visualization-visual-techniques-for-exploring-large-media-collections>

Dawes, Brendan
Cinema Redux ™: Creating a visual fingerprint of a movie, 2004
<https://brendandawes.com/projects/cinemaredux>

Somerset House
Artist Brendan Dawes selects his top 5 sources of inspiration, *Big Bang Data*, 2016
<http://bigbangdata.somersethouse.org.uk/artist-brendan-dawes-selects-his-top-5-sources-of-inspiration/>

Fry, Ben
The Preservation of Favoured Traces, *Fathom Information Design*, 2009
<https://fathom.info/traces/>

Hebron, Micol; Hansen, Mark; Rubin, Ben
Listening Post, *X-TRA Journal*, 2008
<https://www.x-traonline.org/article/mark-hansen-and-ben-rubin-listening-post>

Crockett, Damon
Direct visualization techniques for the analysis of image data: the slice histogram and the growing entourage plot, *International Journal for*

Digital Art, 2016
<https://journals.ub.uni-heidelberg.de/index.php/dah/article/view/33529>

Manovich, Lev; Douglass, Jeremy
Timeline: 4535 Time Magazine Covers, 1923-2009, *Cultural Analytics Lab*, 2009
<http://lab.culturalanalytics.info/2016/04/timeline-4535-time-magazine-covers-1923.html>

Mapping Time, *Softwarestudies.com*, 2009
<https://www.flickr.com/photos/culturevis/4038907270/in/set-72157624959121129/>

Manovich, Lev; Douglass, Jeremy; Huber, William; Zepel, Tara
Exploring One Million Manga Pages with Supercomputers and HIPerSpace, *Cultural Analytics Lab*, 2010
<http://lab.culturalanalytics.info/2010/11/one-million-manga-pages_14.html>

Understanding scanlation: how to read one million fan-translated manga pages, *Image and Narrative*, 2011
<http://www.imageandnarrative.be/index.php/imagenarrative/article/view/133>

Manovich, Lev; Douglass, Jeremy
Manga Style Space, *Softwarestudies.com*, 2010
<https://www.flickr.com/photos/culturevis/5109394222/>

manga.pages.1_million.Xstdev_Yentropy.jpeg_medium, *Softwarestudies.com*, 2010
<https://www.flickr.com/photos/culturevis/4497385883/in/set-72157624959121129/>

4. Implementazione di nuove funzioni su *Internet Archive*

Internet Archive
Image (filtro *media type*: immagini); Kaplan, Jeff e utenti di *Internet Archive* (a cura di), dal 2010
<https://archive.org/details/image?and[]=mediatype%3A%22image%22>

Moving Image Archive; Kaplan, Jeff e utenti di *Internet Archive* (a cura di), dal 2005
<https://archive.org/details/movies>

Magazine Art: Food and Beverages; Rossi, Alexis e utenti di *Internet Archive* (a cura di), dal 2018
<https://archive.org/details/magazineart-foodandbev>

Kellogg's Toasted Corn Flakes -1915A; @jakej (a cura di), 2018
<https://archive.org/details/KelloggsToastedCornFlakes1915A>

Magazine Art; Rossi, Alexis e utenti di *Internet Archive* (a cura di), dal 2018
<https://archive.org/details/magazineart>

NASA Image Exchange Collection; @BonnieReal, Williamson, Greg e utenti di *Internet Archive* (a cura di), dal 2010
<https://archive.org/details/nasaimageexchangecollection>

Album Covers; Scott, Jason e utenti di *Internet Archive* (a cura di), dal 2019
<https://archive.org/details/album_covers?&sort=-views>

Album Covers (ordine cronologico); Scott, Jason e utenti di *Internet Archive* (a cura di), dal 2019
<https://archive.org/details/album_covers?&sort=date>

Operating System CD-ROMs; Scott, Jason e utenti di *Internet Archive* (a cura di), dal 2019
<https://archive.org/details/operatingsystemcds>

Magazine Art (filtro *media type*: immagini); Rossi, Alexis e utenti di *Internet Archive* (a cura di), dal 2018
<https://archive.org/details/magazineart?and[]=mediatype%3A%22image%22>

Ringrazio i miei genitori, Lorenza e Stefano, e le mie sorelle, Marta e Camilla, per il costante incoraggiamento, sostegno e infinita pazienza.

Ringrazio anche il mio relatore, che con la sua disponibilità e cortesia è stato un'ottima guida.