600 maestri raccontano la loro vita professionale in video: un progetto di (fully searchable) open data

Gianfranco Bandini
Dipartimento di Formazione, Lingue, Intercultura, Letterature e Psicologia
Università di Firenze
gianfranco.bandini@unifi.it

Andrea Mangiatordi
Dipartimento di Scienze Umane per la Formazione "Riccardo Massa"
Università di Milano Bicocca
andrea.mangiatordi@unimib.it

Abstract[1]

English. Oral History – as in the collection, storage and interpretation of personal accounts – already provided interesting insights to the renovation of school practices, even though this effect is probably still limited. The project presented here is based on the experience of the "Memorie di Scuola" website (https://memoriediscuola.it), a collection of more than 600 interviews to teachers about their professional history. Using Open Source software and YouTube APIs, the authors were able to create a video repository where: 1) hundreds of hours of video interviews to teachers are organized and made accessible online for free, following a "public history" approach; 2) video content is fully searchable thanks to the use of automatic transcription and of a synchronization script.

Italiano. La storia orale – intesa come raccolta, archiviazione e interpretazione di testimonianze – ha dato un contributo interessante, anche se ancora limitato, al rinnovamento della storia della scuola. Il progetto qui presentato si basa su una piattaforma web (https://memoriediscuola.it) che raccoglie ad oggi oltre 600 interviste a maestri che raccontano in video la loro storia professionale. Attraverso l'uso strumenti software Open Source e delle API di YouTube, gli autori descrivono un modello di raccolta, archiviazione e codificazione del parlato che consente di raggiungere due importanti obiettivi: 1) rendere le centinaia di ore di filmato completamente disponibili on line (attraverso un approccio di public history); 2) consentire la libera esplorazione dei video attraverso l'indicizzazione di tutte le parole che sono state pronunciate durante le interviste.

1 Tematiche e obiettivi della ricerca

Il progetto "memorie di scuola" è basato sul Content Management System "WordPress" (https://wordpress.org), una piattaforma sulla quale si basa – secondo le stime dei suoi autori e sviluppatori – circa un terzo dei siti web mondiali. La nostra implementazione, utilizzabile attraverso qualsiasi web browser, è stata adattata a esigenze molto specifiche, in modo da consentire la costruzione di una memoria collettiva della vita scolastica nella scuola primaria in Italia. Al momento raccoglie oltre 600 interviste a maestre e maestri in pensione (o prossimi ad essa) che raccontano con molti dettagli e grande passione la loro vita professionale, a partire dagli anni '40.
Il nostro intento è quello di migliorare la piena disponibilità dei video (che costituiscono un ampio insieme di *open data*) e la loro completa fruibilità attraverso un sistema di indicizzazione analitico. I video, così come le attività formative e didattiche connesse, sono indirizzati in primo luogo agli insegnanti in servizio e in formazione, ma anche a un più ampio pubblico interessato agli aspetti educativi della nostra storia sociale. In questa sede presentiamo quindi l'implementazione di una particolare *feature* del sito web che consente di effettuare una ricerca full-text all'interno di tutte le parole pronunciate in tutti i video che compongono il progetto (e di collegarsi ad essi nel preciso istante durante il quale il soggetto pronuncia la parola cercata).

[1] Gli autori hanno lavorato alla stesura del testo confrontandosi e concordando ogni sua parte. Gianfranco Bandini si è dedicato in particolare alle sezioni 1, 2 e 4; Andrea Mangiatordi ha curato in particolare le sezioni 3 e 3.1.

2 Quadro teorico di riferimento

Nel corso del Novecento la storia orale ha affermato, non senza contrasti e opposizioni, la sua legittimità e utilità, soffermandosi soprattutto sulla storia dal basso, degli esclusi dalla storia tradizionale (Gardner, & LaPaglia, 2006). Il settore di studi che si occupa della storia della scuola, all'interno della storia dell'educazione (McCulloch, 2011), ha utilizzato con sempre maggiore convinzione fonti non testuali, come le fotografie o i dipinti. Una piccola parte di queste ricerche ha inoltre scoperto, anche se con un certo ritardo, l'utilizzo delle fonti orali, cioè la raccolta delle testimonianze (Gardner, 2003). La memoria personale ha consentito di spostare l'accento degli studi sulle percezioni e sui sentimenti delle persone, sugli aspetti interiori e comunitari della vita sociale.
La storia dell'educazione e la storia orale, in questa forma congiunta, hanno trovato un campo di nuova e eccezionale sperimentazione nell'ambito della *digital public history,* nel quale il presente progetto si colloca (Bandini, 2017). Nel contesto digitale, l'incrocio di queste diverse tradizioni di ricerca dà la possibilità di potenziare la comune aspirazione a un maggiore contatto tra il mondo accademico e la società, soprattutto per quanto riguarda il rapporto con le professioni educative e di cura (cfr. Depaepe, 2001; Linné, 2001; Vinovskis, 2015). La trasformazione delle classiche fonti storiche in *open data* risponde proprio a questo ambizioso obiettivo.
Il progetto consente, oltre alla piena e completa disponibilità *online* delle testimonianze, di superare una delle principali limitazioni che qualsiasi tradizionale ricerca che fa uso di storia orale si trova a affrontare (cfr. Ritchie, 2011): la numerosità delle testimonianze e la gestione della massa dei dati che possono esserne ricavati (intesi anche come momenti di titubanza del testimone, silenzi, salti temporali, ecc.).
A questo proposito, uno dei punti di discussione metodologica consiste, ad esempio, nel riflettere sulla questione del numero ottimale di testimonianze. In alcuni casi (cfr. l'ampio studio di Johnson & Reuband, 2008), in abbinamento a ricerche di tipo quantitativo, si è stabilito un vero e proprio piano di campionamento, come di consueto nelle ricerche statistiche. In generale, tuttavia, nelle ricerche storiche si sostiene che la ricerca di nuove testimonianze si può arrestare nel momento in cui si satura il campo concettuale che stiamo indagando, cioè quando nuove testimonianze non porterebbero più nulla di nuovo, o insignificanti dettagli, rispetto a quanto già detto.
Nel contesto della storia digitale il panorama risulta ampiamente modificato perché la raccolta delle testimonianze può essere espansa senza eccessiva fatica e scarso dispendio di risorse (cfr. Thomson, 2007). Questo aspetto ci consente di progettare una raccolta di testimonianze aperta, nel senso che può essere incrementata nel tempo e offrire sempre nuove opportunità di conoscenza e formazione. La cassetta degli attrezzi dello storico e l'insieme delle sue fonti vengono così trasportate da uno spazio privato a uno spazio pubblico.
Nell'ottica della *public history*, il contesto digitale può quindi essere progettato non soltanto per la collocazione online di fonti primarie (come sono le interviste), ma per consentire lo sviluppo delle interazioni tra gli utenti. Per superare l'impostazione tradizionale, che trasforma le fonti online in musei statici (per quanto digitali e più facilmente accessibili) c'è bisogno di alcuni strumenti di base, che sono volti a aumentare la significatività delle fonti. La trascrizione automatica dei video, in questo senso, rappresenta un tassello fondamentale della strategia comunicativa, a vantaggio della piena fruibilità da parte degli utenti.

3 Metodologia

La raccolta delle testimonianze video presenta delle particolari caratteristiche rispetto alla raccolta di documenti testuali (per esempio diari o autobiografie). Il *repository* "memorie di scuola" attualmente supera le 5.000 ore di filmato e rende oggettivamente molto difficile sia l'ascolto integrale, sia una completa attività di indicizzazione manuale basata sull'inserimento di *tags* (per esempio quelli contenuti in TESE, Thesaurus Europeo dei Sistemi Educativi). Nella prospettiva della *public history* il numero delle ore è inoltre destinato a crescere ancora: in questa situazione la messa a disposizione del pubblico, con una piccola serie di *tags* di indicizzazione, di fatto non consente l'accesso alla ricchezza documentaria contenuta nei video che diventa quasi casuale.
Mantenendo la prospettiva fin qui esposta, il progetto si è indirizzato a cercare di produrre un archivio video liberamente accessibile attraverso l'uso di risorse sostenibili – per lo più software Open Source – e l'automazione di una serie di operazioni che permettono di facilitare la ricerca all'interno della grande mole di contenuti raccolti. Il servizio online YouTube (https://www.youtube.com), famoso per essere tra i principali

repository gratuiti di contenuti video online, consente la trascrizione automatica del parlato dalla lingua italiana al fine di produrre sottotitoli. La quantità di errori di trascrizione è altamente variabile e dipendente da molteplici fattori, primo tra tutti la qualità della traccia audio (cfr. Alberti e altri, 2009). Tuttavia, un sistema in grado di estrarre i sottotitoli creati automaticamente dal servizio e di inserirli in un database permette di rendere le interviste esplorabili e ricercabili in modo simile a un corpus testuale. Questo sistema, pur non presentando un'innovazione dal punto di vista degli algoritmi e del software, consente di ottenere un significativo vantaggio rispetto ai sistemi attuali e offre un modello di funzionamento facilmente replicabile e applicabile a grandi masse di dati. Bisogna ricordare che quando sono in gioco grandi quantità di interviste, la trascrizione umana (ovviamente più accurata dei sistemi automatici) solo in rarissimi casi riesce a raggiungere il risultato: è il caso, quasi unico nel suo genere, della grande raccolta di testimonianze condotta dalla Shoah Foundation.[2]

3.1 Struttura del Software e flussi di lavoro

L'architettura della soluzione software messa in atto per sostenere il progetto Memorie di Scuola consta di servizi e applicazioni Open Source che si interfacciano con il servizio proprietario YouTube per l'hosting dei materiali video e per la produzione di trascrizioni rese disponibili nella forma di sottotitoli e in formati diversi. In particolare, lo *stack software* utilizzato è basato su:

- Un web server dotato in grado di eseguire codice scritto in linguaggio PHP – nel caso specifico del progetto è stato utilizzato il software NginX (https://www.nginx.com/), ma a questo livello non ci sono requisiti particolarmente stringenti;
- Un database MySQL (https://www.mysql.com/);
- Il CMS WordPress (https://wordpress.org);
- Il plugin Meks Video Importer per WordPress (https://wordpress.org/plugins/meks-video-importer/), che permette la ricerca di video disponibili pubblicamente su YouTube e l'importazione automatica del contenuto testuale della loro descrizione;
- Il plugin Advanced Custom Fields per WordPress (https://wordpress.org/plugins/advanced-custom-fields/), che facilita l'inserimento di metadati ai contenuti dei post;
- Il software YouTube Transcript/Subtitle API (https://github.com/jdepoix/youtube-transcript-api), uno script in grado di estrarre i sottotitoli da video YouTube pubblicamente disponibili;
- Lo script "YouTube transcript to WordPress" (https://github.com/andreamangia/youtube-transcript-to-wp), progettato specificamente per automatizzare le operazioni.

Il software si inserisce nel workflow descritto di seguito e lo sostiene, minimizzando la necessità di intervento umano ma rendendola comunque possibile in una fase successiva di revisione dei contenuti:

- L'intervistatore effettua l'intervista e esegue l'upload su YouTube, aggiungendo al video una semplice descrizione testuale;
- Un operatore del sito web www.memoriediscuola.it importa il video, nella forma di un nuovo post WordPress, con l'aggiunta di tag ed eventuali altri termini tassonomici, la verifica della congruenza tematica e l'indicazione di dati quali il nome dell'intervistatore, il luogo;
- L'operatore del sito web esegue lo script "YouTube transcript to WordPress" indicando il numero identificativo del post WordPress generato al punto precedente. Lo script si occupa di:
  - Estrarre in formato JSON la trascrizione dell'audio;

[2]La Shoah Foundation gestisce il Visual History Archive, disponibile all'indirizzo https://sfi.usc.edu/vha, con lo scopo di documentare e rinforzare l'empatia verso le memorie di persone che hanno vissuto genocidi e altri drammi.

    - Trasformare ciascun frammento della trascrizione (in genere, ma non sistematicamente, corrispondente a una breve frase) in un elemento HTML contenente l'indicazione del momento temporale in cui il testo viene pronunciato nel video;
    - Salvare l'intera trascrizione come valore per un campo personalizzato di WordPress associato al post contenente il video.
- L'utente effettua una ricerca libera all'interno del repository, che è stato opportunamente configurato per consentire la ricerca anche all'interno dei campi personalizzati, normalmente ignorati dal motore di ricerca interno di WordPress;
- Il sito web elenca tutti i video che contengono la parola ricercata. Accedendo alla pagina di ciascun video è possibile raggiungere il momento in cui una frase è pronunciata attraverso un click sulla porzione di trascrizione corrispondente.

La selezione degli strumenti operativi che rendono possibile il procedimento è dunque stata pensata per favorire la replicabilità totale dell'esperienza in qualunque sistema basato su WordPress, indipendentemente da elementi quali il tema grafico utilizzato o dalla necessità di modificare l'architettura del database sul quale poggia il sistema. Non è previsto supporto in questa fase per altri CMS, ma non è da escludere che le stesse funzionalità e la stessa logica di lavoro possano essere applicate anche altrove, data la presenza di diversi layer basati su tecnologie standard e interoperabili.

4 Risultati attesi e interventi futuri

Il progetto ha come obiettivo principale la costruzione di un set di *open data* costituiti da testimonianze video, liberamente accessibili sul web e esplorabili in profondità attraverso gli strumenti di ricerca del parlato sopra descritti.
Questo obiettivo, del resto, è propedeutico a molte azioni formative che possono essere svolte proprio grazie alla possibilità di trovare all'interno dei video esattamente ciò che stiamo cercando, siano esse parole generiche o identificatori di luogo o di persona.
L'insieme delle trascrizioni si presta inoltre a successive analisi e esplorazioni con software di data mining (per esempio T-Lab) o di categorizzazione dei testi (per esempio Nvivo), all'interno del paradigma di ricerca della Grounded Theory. Per quanto l'accuratezza delle trascrizioni possa essere migliorata, già allo stato attuale i testi dei video appaiono ben comprensibili e di grande aiuto per effettuare delle analisi approfondite. Nel caso di uno studio volto alla categorizzazione dei testi, un ulteriore passaggio manuale di correzione potrebbe portare a un corpus assestato e corretto in tempi ragionevolmente brevi. Tenendo conto che i software di analisi dei testi sono di fatto degli strumenti semi-automatici, questo tipo di operazione risulta essere di carattere ordinario. Bisogna inoltre considerare che il presente progetto, in modo del tutto automatico (attraverso il citato plugin Meks Video Importer per WordPress), si gioverà dei miglioramenti nel riconoscimento del parlato che verranno implementati nella piattaforma YouTube. Lo sviluppo dei sistemi di Intelligenza Artificiale ha dato prova, in questi ultimi anni, di consentire dei progressivi e tangibili miglioramenti nella comprensione del parlato, a partire dalla lingua inglese (riconosciuta nelle sue molte varianti di pronuncia).
L'insieme delle trascrizioni, infine, può costituire la base per una piattaforma partecipativa che permetta di dare avvio a progetti di crowdsourcing di correzioni e integrazioni alle trascrizioni automatiche, aumentando ulteriormente la sostenibilità del progetto; inoltre, in linea con l'approccio di public history fin qui adottato, sarà possibile dare la possibilità agli utenti di apporre commenti ai video, commenti che a loro volta verranno trascritti e inseriti in una mappa esplorabile per concetti chiave.
Dal punto di vista tecnico è possibile pensare in tempi brevi all'integrazione di altre due tecnologie Open Source. La prima di queste è la piattaforma di ricerca Apache Solr (https://lucene.apache.org/solr/): questa mette a disposizione un motore di indicizzazione più raffinata rispetto alla semplice ricerca testuale disponibile nel CMS WordPress e permetterebbe di avere risultati di ricerca rispondenti a query di tipo diverso (con operatori booleani, stemming, proximity search). La seconda tecnologia potenzialmente utilizzabile è Hypothes.is (https://web.hypothes.is/), software Open Source di annotazione di pagine web, che renderebbe possibile appunto l'annotazione dei contenuti, in modalità aperta e collaborativa da parte di ricercatori diversi.

Ringraziamenti

Si ringrazia sentitamente il prof. Gianfranco Crupi per gli utili consigli e suggerimenti; Inclusive Cloud S.r.l.s. per aver messo a disposizione competenze e infrastrutture informatiche; si ringraziano vivamente tutti i maestri che hanno cortesemente messo a disposizione il racconto della loro vita professionale e tutti gli studenti del corso di laurea in scienze della formazione primaria (università di Firenze) che hanno raccolto, con pazienza e serietà, le testimonianze in video.

Bibliografia

Agneta Linné. 2001. Myths in Teacher Education and the Use of History in Teacher Education Research. *European Journal of Teacher Education* 24(1): 35-45, DOI: 10.1080/02619760120055871

Christopher Alberti, Michiel Bacchiani, Ari Bezman, Ciprian Chelba, Anastassia Drofa, Hank Liao, Pedro Moreno, Ted Power, Arnaud Sahuguet, Maria Shugrina, and Olivier Siohan. 2009, April. An audio indexing system for election video material. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. 4873-4876. IEEE.

Gianfranco Bandini. 2017. *Educational Memories and Public History: A Necessary Meeting*. In: Cristina Yanes-Cabrera, Juri Meda, Antonio Viñao (eds.). *School Memories. New Trends in the History of Education*. 143-156. Springer International Publishing, Cham, Switzerland. ISBN:978-3-319-44062-0

Marc Depaepe. 2001. A Professionally Relevant History of Education for Teachers: Does it Exist? Reply to Jurgen Herbst's state of the art article. *Paedagogica Historica*. 37(3): 629-640, DOI: 10.1080/0030923010370305

Philip Gardner. 2003. Oral history in education: teacher's memory and teachers' history. *History of Education.* 32(2): 175-188.

James B. Gardner and Peter S. LaPaglia. 2006. *Public History: Essays from the Field*, 2nd edition. Krieger , Malabar, Florida. ISBN:978-157-524244-6.

Eric A. Johnson and Karl-Heinz Reuband. 2008. *La Germania sapeva. Terrore, genocidio, vita quotidiana. Una storia orale*. Mondadori, Milano. (orig. ed. 2005)

Gary McCulloch. 2011. *The Struggle for the History of Education*. Routledge, Abingdon, UK. ISBN:978-0-415-56535-6.

Simonetta Polenghi, Gianfranco Bandini. 2016. The history of education in its own light: signs of crisis. Potential for growth. *Espacio Tiempo y Educacion*. 3(1, January-July 2016): 3-20.

Donald A. Ritchie. 2011. *The Oxford handbook of oral history*. Oxford University Press.

Alistair Thomson. 2007. Four Paradigm Transformations in Oral History. *The Oral History Review*. 34(1): 49–70.

Vinovskis, Maris A. (2015). *Using Knowledge of the Past to Improve Education Today: US Education History and Policy-Making*. Paedagogica Historica, 51(1-2), 30-44, DOI: 10.1080/00309230.2014.9977