# Archivismi

*la serie completa*

*Marco A.L. Calamari a.k.a "Cassandra"*

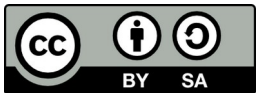
# Table of Contents

Archivismi.....1
Il Dizionario di Cassandra/ Archivismi.....4
Archivismi: l'inizio.....6
Archivismi: upload ed operazioni elementari.....9
Archivismi: il giorno dopo l'upload.....16
Archivismi: l'organizzazione dei documenti in Internet Archive.....20
Archivismi: API, quando il gioco si fa duro.....24
Archivismi: archiviamo Cassandra, parte prima.....28
Archivismi: archiviamo Cassandra, parte seconda.....32
Archivismi: archiviamo Cassandra, parte terza.....36
Archivismi: Cassandra Crossing è per sempre!.....39
Archivismi: Cassandra attraverso i secoli.....42
Archivismi: Cassandra e la miniera.....47
Archivismi: Cassandra tra i ghiacci.....51
Appunti di Archivismi: creare oggetti e correggere errori.....54
Appunti di Archivismi/ Rinfrescare, riordinare, rivedere, rimpolpare.....58
Archivismi: una sessione con Claude.....63
Archivismi: La scomparsa dei media fisici.....68

Scrivere a Cassandra — Twitter — Mastodon
Videorubrica "Quattro chiacchiere con Cassandra"
Lo Slog (Static Blog) di Cassandra
L'archivio di Cassandra: scuola, formazione e pensiero

2 settembre 2026, by Marco A.L. Calamari - v1.4.0

Il volume e tutti i suoi contenuti sono pubblicati sotto licenza libera CC BY-SA 4.0 Deed *Attribuzione - Condividi allo stesso modo 4.0 Internazionale*.




La ridiffusione con qualsiasi mezzo è libera, gratuita **e molto gradita**. Nel caso, una comunicazione a cassandra@cassandracrossing.org sarebbe pure molto apprezzata.

Lo stesso indirizzo può essere utilizzato per contatti ed informazioni.

# Il Dizionario di Cassandra/ Archivismi

*(558) —Come definire un tentativo di portare alla luce le iniziative per la conservazione della cultura? Un tentativo di preservare quello che ognuno pensa che meriti essere preservato? Un atteggiamento corretto verso la cultura, l'infosfera ed il loro futuro?*

*19 novembre 2023*

***Ar-chi-vis-mo**: s.m. (pl. -mi,)*

1. atto di preservazione della cultura, particolarmente di quella in ambito digitale
2. atteggiamento generale orientato alla preservazione di artefatti culturali

*Ohibò* — esclameranno i 24 increduli lettori — *se Cassandra, dopo tanto tempo, formalizza oggi un neologismo, vuol dire che c'è qualcosa di importante all'orizzonte.*

Si, ma non si tratta di una novità, al contrario un discorso portato avanti da tempo, prima in maniera quasi istintiva, poi sempre più strutturata e convinta.

Perché la necessità di preservare ciò che ogni essere umano lascia, prima o poi, in eredità agli altri non può essere gravame solo di particolari categorie di persone, ma prima di tutto responsabilità di ciascuno.

Se guardiamo verso l'orizzonte, non quello davanti a noi, quello dietro, vediamo un percorso continuo di attività di preservazione della cultura, che parte da tradizioni orali e tavolette di argilla per arrivare alla Rete ed alle isole Svalbard; questo percorso passa attraverso di noi e continua in avanti, fino a ed oltre il nostro orizzonte.

Per questo alcune delle prossime esternazioni di Cassandra saranno dedicate a illustrare come anche un normale internauta può indossare la fascia di bibliotecario digitale, se è convinto di avere qualcosa che meriti di essere conservato.

Qualcosa che non sia destinato a rimanere chiuso nei database proprietari delle chat e dei social, foraggio per il capitalismo della sorveglianza, in cui la maggior parte degli internauti “conferisce” le proprie parole, immagini, pensieri.

Qualcosa che non sia nemmeno destinato a perdersi nel cestino della carta, nel cestino digitale delle cancellazioni volontarie, o nel Grande Cestino dei Bit, in cui finiscono i dati degli smartphone rubati, hard disk rotti, chiavette USB illeggibili, PIN crittografici dimenticati ed account cloud spariti od inaccessibili.

Qualcosa che nel prossimo futuro, quando le informazioni vere e quelle false saranno mischiate insieme , si possa più facilmente trovare dentro un’Infosfera sempre più avvelenata dalle false IA .

Stay tuned!

# Archivismi: l’inizio

*(559)—I bibliotecari saranno gli eroi del futuro? Poiché l’inquinamento dell’Infosfera procede a grandi passi, la conservazione della cultura e della vera anima dell’Uomo è sempre più importante. Dobbiamo farlo noi!*

**20 novembre 2023**— Cassandra, i 24 indomiti lettori ormai ben lo sanno, ha sempre avuto a cuore la conservazione della cultura, e per conseguenza anche delle sue profezie.

Per questo si è impegnata, sia come come profetessa, con numerose esternazioni in tema, sia a livello personale, praticando un uso esteso della memorizzazione dei dati su suoi supporti di memoria, ed un selettivo utilizzo di sistemi di archiviazione, come il mai abbastanza ringraziato Archive.org. Addirittura creando all’uopo un neologismo, “Archivismi”.

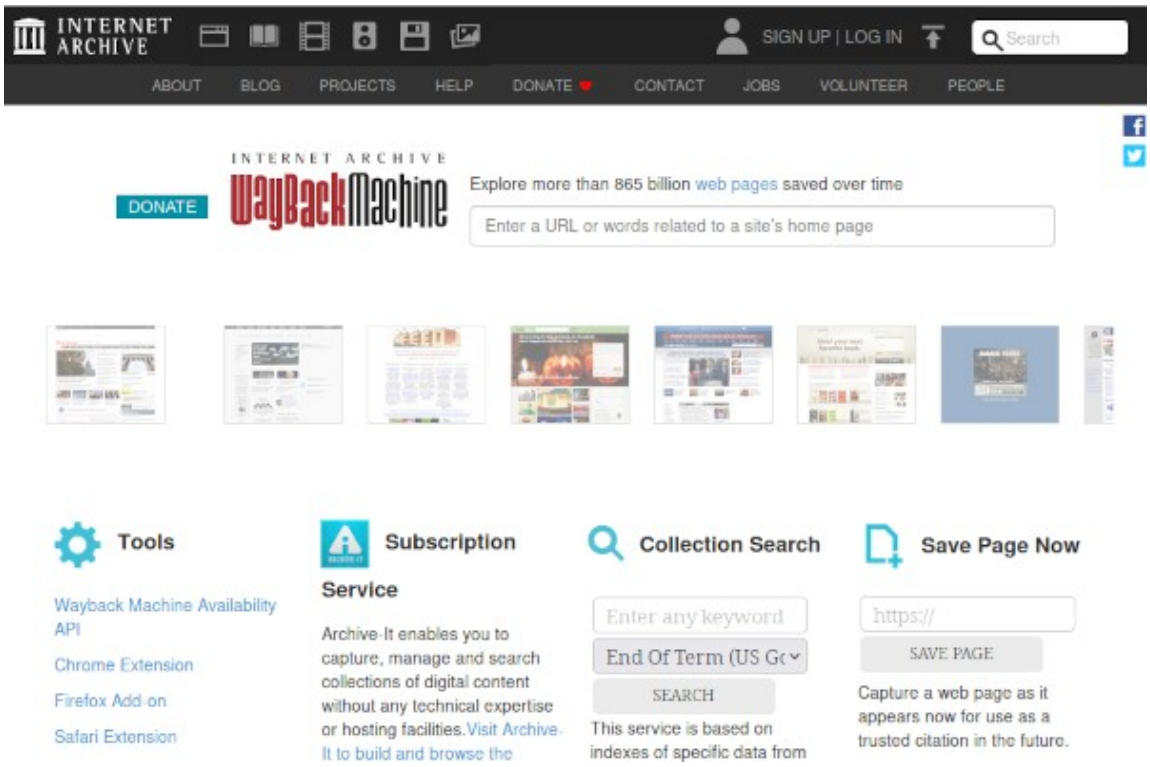
Ed è per questo che Cassandra ha deciso di abbandonare per un po’ l’attualità, pur fitta di temi che dovrebbero essere trattati, a favore di altri temi importantissimi, ma che spesso, mancando di urgenza, finiscono per essere sempre trascurati.

E’ il caso, appunto, dell’archiviazione dei contenuti in rete, a cui pure Cassandra ha già dedicato numerose esternazioni ed addirittura **una lista apposita**.

Ma in questi interventi, i dettagli, gli esempi ed i *TODO* sono stranamente sempre mancati. E’ tempo di rimediare!

Partiamo quindi da una risorsa di uso facile, efficace, gratuita ed alla portata di tutti; senza nessuna sorpresa parliamo di Archive.org, la biblioteca digitale che comprende la ben più nota Wayback Machine, altrimenti detta "*La Macchina del Tempo di Internet*".

Come alcuni hanno già provato, la Wayback Machine (in breve WM) permette, con una semplicissima operazione, di salvare permanentemente una certa pagina di un sito; basta inserire l'indirizzo della pagina da salvare nel campo "Save Page".




Se la pagina fosse già stata salvata in passato, la vostra richiesta ne salverà una ulteriore copia con il contenuto attuale.

L'archiviazione di copie multiple nel tempo, infatti, crea una "storia" della pagina, utilissima per quei siti soggetti a frequenti cambiamenti; questi snapshot vengono spesso usati come prova forense in procedimenti giudiziari, e permettono anche di scoprire "manipolazioni" di pagine contestate, ad esempio quando qualche "manina" "corregge" un articolo giornalistico sperando che nessuno se ne accorga, o quando qualcuno fa proprio "sparire" una pagina diventata scomoda.

Per i siti che sono stati archiviati interamente, è di solito la WM stessa a mandare periodicamente uno dei suoi robot a ri-archiviare le pagine, senza che lo si debba richiedere nuovamente.

Per verificare se una pagina è già stata salvata, basta cercarla tramite il campo di ricerca principale che si trova accanto al logo della WM; questo vi permetterà di accedere anche all'elenco delle copie (snapshot) della pagina che sono state fatte nel tempo.

Se la pagina ricercata non fosse presente nella WM, essa verrà cercata per voi sul web, e se presente vi verrà offerto di archiviarla direttamente.

Questo semplice procedimento può essere facilitato da un'apposita estensione, disponibile per i principali browser; la trovate, sempre nella stessa pagina, nel menu *Tools* a sinistra.

L’estensione è comodissima e velocizza molto le operazioni ripetitive. Potete anche fare in modo che se una pagina non venisse trovata, invece di vedere il classico *404—page not found*, il browser vada a cercarla automaticamente nella WM e, se la trova, visualizzi direttamente la versione più recente archiviata.

Questa iniziale “pillola” di Archivismi volge al termine. Ma prima di salutarci, due cose ed uno spoiler

Prima cosa. La WM e tutto Archive.org girano su una struttura informativa complessa e non ottimizzata per la velocità ma per altri fattori che vedremo, al contarrio dei siti *normali.*

Siate quindi preparati ad aspettare non frazioni di secondo ma qualche secondo. La vostra pagina, che magari nessuno ha mai chiesto prima di voi, deve essere localizzata e recuperata nei meandri della biblioteca digitale di Archive.org. Non è facile destreggiarsi tra 212 Petabyte di dati online.

Seconda cosa. Archive.org è un’organizzazione senza fini di lucro, che vive di contribuzioni volontarie. Chi la usa regolarmente, o la trova utile, od è moralmente d’accordo, dovrebbe considerare doverosa una donazione.

TANSTAAFL … Cassandra ve l’ha già detto tante di quelle volte …

Spoiler. Il cammino per diventare bibliotecario digitale comincia qui; noterete che, sempre nella pagina principale di WM, in alto a destra c’è il link “Sign Up”. Serve per registrarsi come utente. E cominciare a muoversi con garbo, senza fretta e senza cominciare ad archiviare a caso.

Ma questa … questa è un’altra puntata!

# Archivismi: upload ed operazioni elementari

*(560 ) — Continuiamo la nostra esplorazione di Archive.org per iniziare a capire come funziona e come usarlo.*

**26 dicembre 2023**— *Archiviare* non significa memorizzare. *Archiviare* non significa copiare.

*Archiviare*, nel mondo digitale e nel senso più esteso del termine, vuol dire memorizzare in forma significativa un'informazione digitale, e farlo nei formati più opportuni, corredandola dei metadati più adatti e da un ben selezionato insieme di parole chiave di ricerca. E fare questo seguendo le procedure ed i metodi consolidati che generazioni di bibliotecari, oggi *digitali* ma prima anche *analogici*, hanno già predisposto per noi.

In questa terza puntata di *Archivismi* scopriremo che archiviare su Archive.org non è semplice ed immediato come copiare un ebook od un video mp4 su Dropbox, Google Disk od un server Nextcloud.

Alcune operazioni sono, per fortuna, quasi completamente automatiche; come abbiamo visto nella scorsa puntata, archiviare una singola pagina su *The Wayback Machine* è in effetti un'operazione elementare, anche se un po' lenta. Ed in effetti è lenta perché sfrutta una infrastruttura complessa, archiviando la pagina con un meccanismo pensato per consentire anche operazioni molto più sofisticate.

Vediamo di cosa si tratta. Nel database di *Internet Archive* le informazioni sono memorizzate in oggetti. Ad ogni oggetto, al momento della creazione, viene associato un

identificatore univoco. Un oggetto, a tutti gli effetti, può essere rappresentato come una directory, in cui sono contenuti almeno un file di dati ed almeno due file di metadati.

Proviamo a creare un oggetto eseguendo un semplice *upload*, come quelli che si utilizzano per caricare un file in un cloud.

Per proseguire, dovete aver creato il vostro utente di Internet Archive; se non lo avete già fatto, fatelo adesso e poi entrate col vostro utente.

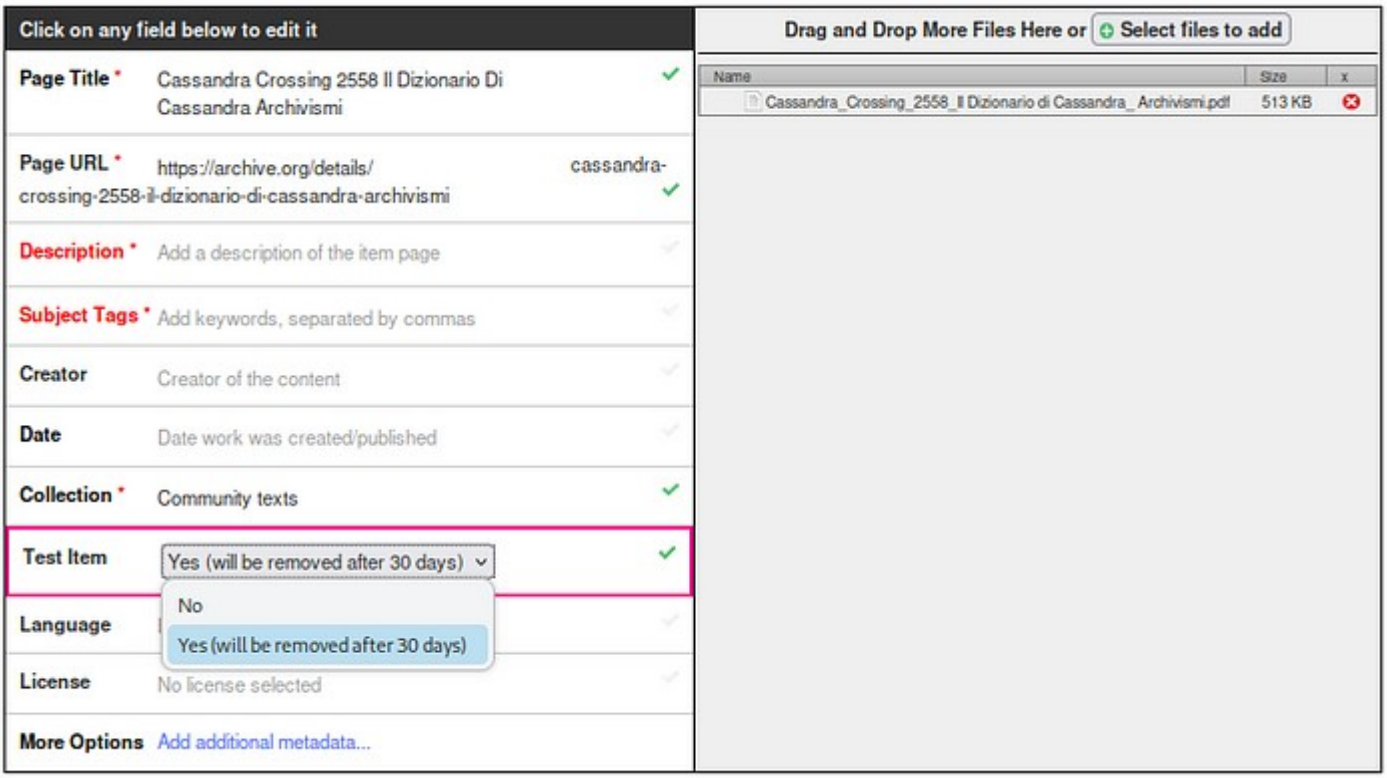
Osservate subito il cuoricino 💗 che sta al centro della barra dei menù; cliccandovi sopra potete effettuare una piccola donazione con qualsiasi mezzo di pagamento abbiate disponibile. Non è ovviamente obbligatorio, i servizi di Internet Archive sono gratuiti, come è giusto che sia in qualunque *biblioteca universale*, ma a loro far funzionare la baracca costa soldi, quindi, come al solito, TANSTAAFL.

Se invece per ora non vi sembra che il servizio di *Internet Archive* valga i vostri spiccioli, procedete pure; probabilmente presto cambierete idea.

Osservate in alto a destra il link UPLOAD; tra parentesi notate, e lo vedremo molte volte, che Internet Archive nasconde i link più importanti nei posti meno visibili, ma deve essere un'*arte oscura* comune tra i bibliotecari digitali …

Se ci cliccate, si apre ovviamente una finestra in cui potete fare il drag&drop di un file od aprire una più pratica finestra di selezione file. Per seguire questo esempio, selezionate un file .pdf, oppure quello che volete voi.

Fatta la selezione, vi si aprirà la finestra più importante in assoluto, quella di *archiviazione*.




Innanzitutto **non infestate Internet Archive con le vostre prove**; anche se è possibile cancellare un oggetto, in realtà questo non viene normalmente rimosso dal database, ma

continua ad esistere per una serie di motivi. Potrebbe essere rimosso successivamente, ma questo richiede un intervento di tipo “sistemistico”. Ci torneremo più avanti.

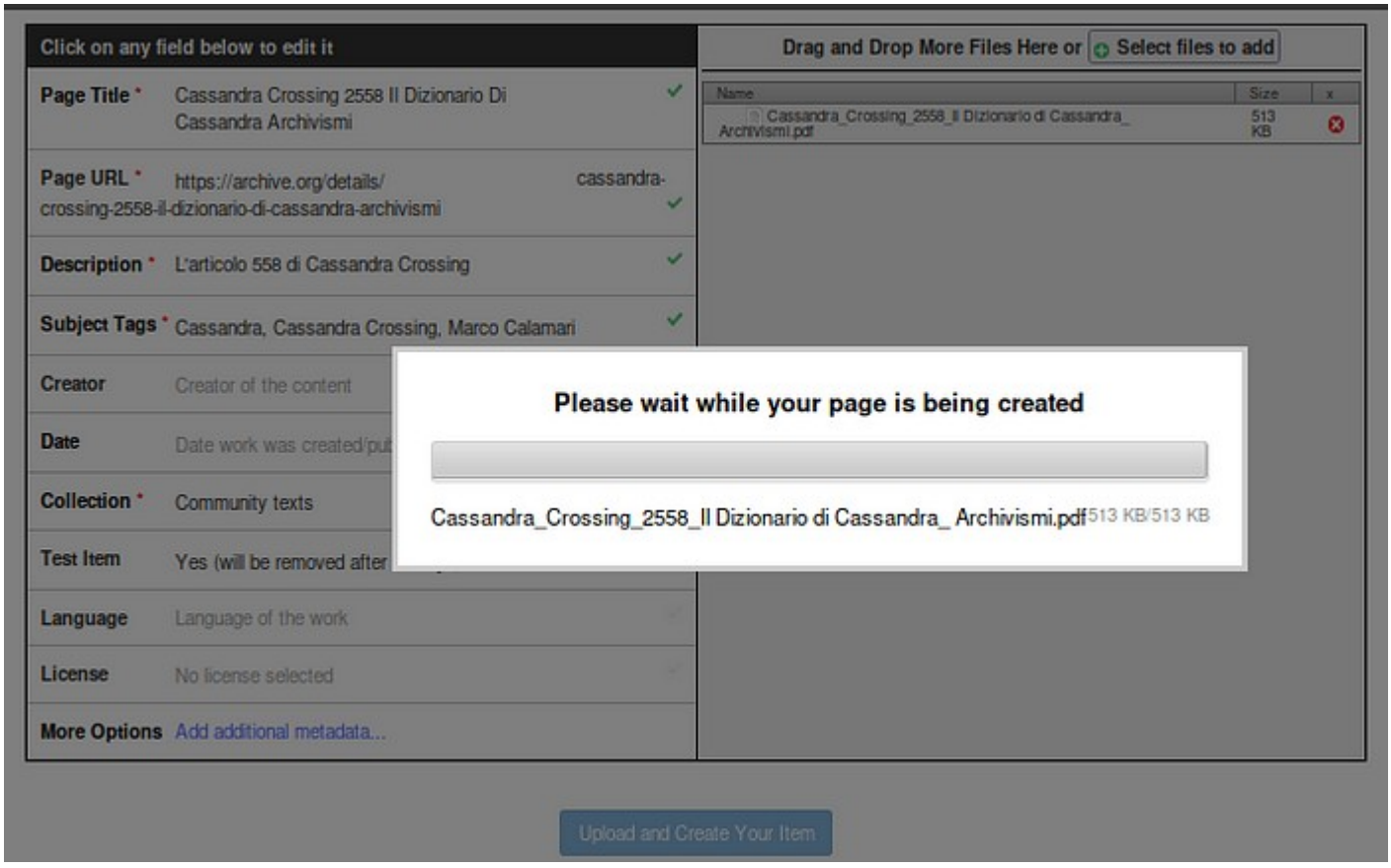
Per praticità è quindi previsto un apposito parametro, che marca un oggetto come *effimero*, e lo cancella automaticamente e completamente dopo 30 giorni. **Avete il dovere di fare i bravi**, quindi usate questo parametro per tutte le archiviazioni di prova, in modo che solo quelle definitive vengano memorizzate permanentemente.

Quindi selezionatelo subito dalla apposita tendina. Osserviamo poi che, ricavandoli dal nome del file, vengono proposti dei valori ragionevoli, come titolo dell’oggetto ed identificatore (URL).

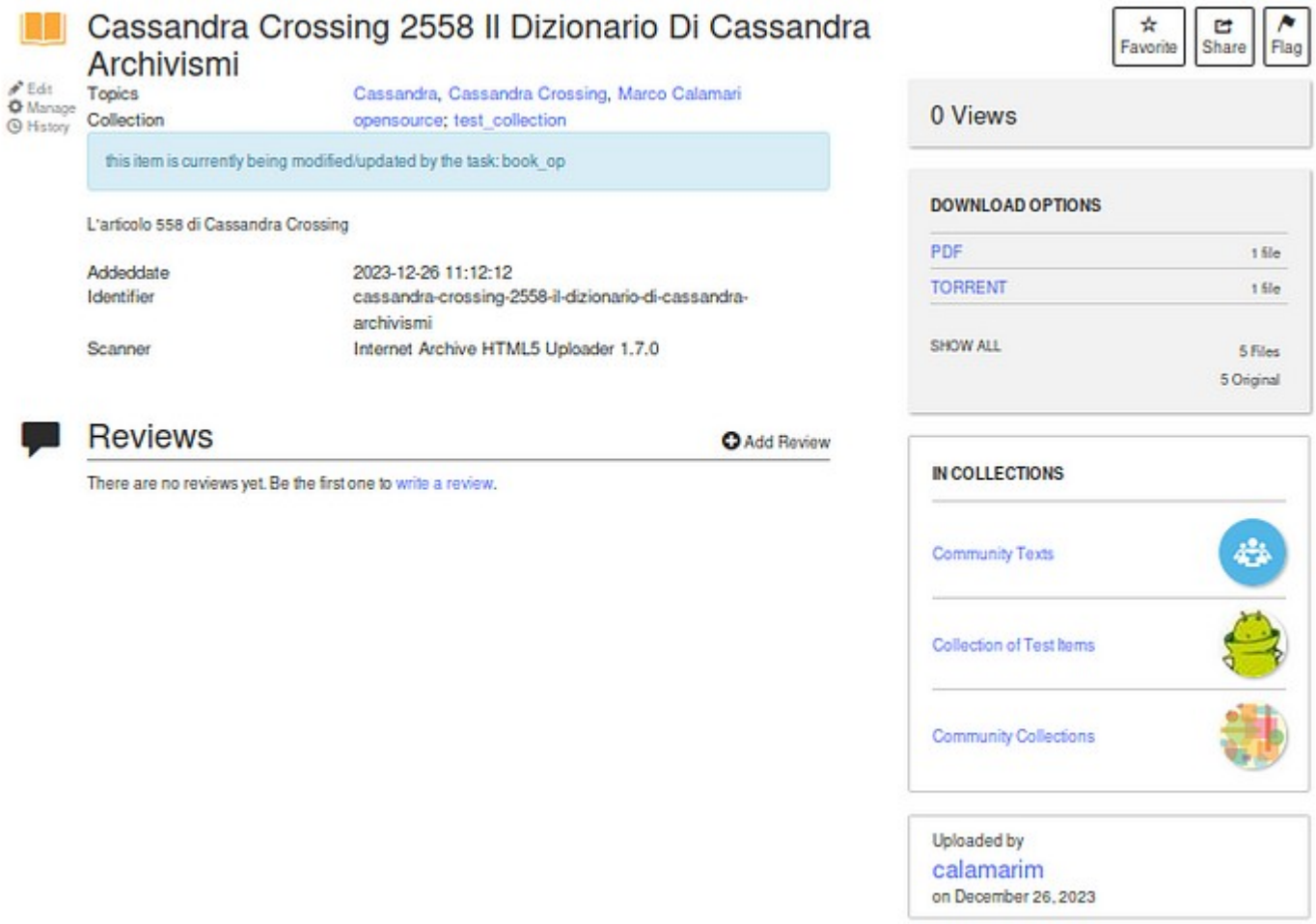
Correggiamoli, se lo riteniamo necessario, ed aggiungiamo una descrizione, delle parole chiave, e quanto altro utile.

Tutti i campi sono facoltativi, ma farne un utilizzo attento e ragionato è il fattore essenziale in una buona campagna di archiviazione. E “*progettarla*” è anche la cosa più difficile, quindi per ora accontentiamoci. Ne riparleremo.

Potete ora cliccare sul pulsante azzurro in basso.




Si apre una normale finestrina di upload che, dopo un tempo che vi parrà troppo lungo (minuti), vi restituirà una *vista* sull’oggetto che avete appena creato. Questa finestra include un browser per l’oggetto appena creato, ma ora vi apparirà probabilmente solo un riquadro bianco, perché in realtà il processo di archiviazione è ancora in corso.




Esaminandolo con attenzione noterete tutta una serie di link cliccabili, ma prima una cosa importante.

A seconda di che browser e sistema operativo utilizziate, navigando indietro con la freccia a sinistra (si può fare tranquillamente) potrebbe accadervi che, oltre a visualizzare la pagina precedente, vi si apra la finestra di download del file; in questo caso potete annullarla/chiuderla tranquillamente e continuare. Quando Cassandra troverà un modo di evitare questo fastidio, certamente ve lo farà sapere.

**Il vostro oggetto di prova non è stato ancora completamente creato**; esiste come identificatore e come informazioni di base, e può perciò già essere utilizzato, ma molte operazioni nel backend di Internet Archive devono ancora essere eseguite, e lo saranno nei prossimi minuti, ore o giorni. Quindi, ancora una volta, pazienza.

Ma di quali operazioni si tratta? Dipende dal tipo di oggetto che avete creato, ed in quale "*collezione*" lo avete inserito. Tralasciamo per ora l'importantissimo aspetto della *collezione,* e concentriamoci sulle operazioni automatiche che sono state schedulate e che vengono o verranno compiute sull'oggetto appena creato. E' possibile esaminarle, utilizzando il link *history* nel microscopico menu in alto a sinistra nella finestra *oggetto*.

My Tasks Not Yet Completed for cassandra-crossing-2558-di-cassandra-archivismi (page drawn: PST: 2023-12-26 03:31:25)

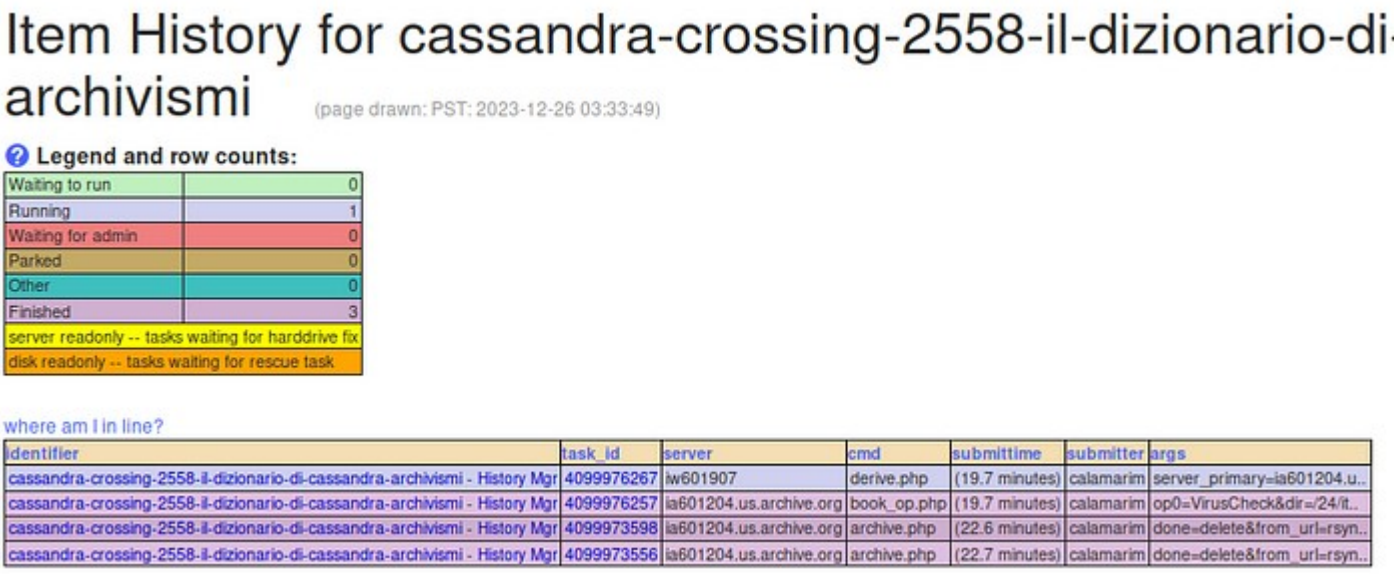
Legend and row counts:

| | |
|---|---|
| Waiting to run | 0 |
| Running | 1 |
| Waiting for admin | 0 |
| Parked | 0 |
| Other | 0 |
| Finished | 0 |
| server readonly -- tasks waiting for harddrive fix | |
| disk readonly -- tasks waiting for rescue task | |

where am I in line?

| identifier | task_id | server | cmd | submittime | submitter | args |
|---|---|---|---|---|---|---|
| cassandra-crossing-2558-il-dizionario-di-cassandra-archivismi - History Mgr | 4099976267 | iw601907 | derive.php | (17.3 minutes) | calamarim | server_primary=ia601204.u.. |

C’è una task in running; si tratta dell’archiviazione dell’oggetto che procede, mentre nella parte bassa della finestra compare, e continua a popolarsi, lo storico delle operazioni eseguite automaticamente sull’oggetto; infatti dopo una mezz’ora apparirà questo.



Item History for cassandra-crossing-2558-il-dizionario-di-archivismi (page drawn: PST: 2023-12-26 03:33:49)

Legend and row counts:

| | |
|---|---|
| Waiting to run | 0 |
| Running | 1 |
| Waiting for admin | 0 |
| Parked | 0 |
| Other | 0 |
| Finished | 3 |
| server readonly -- tasks waiting for harddrive fix | |
| disk readonly -- tasks waiting for rescue task | |

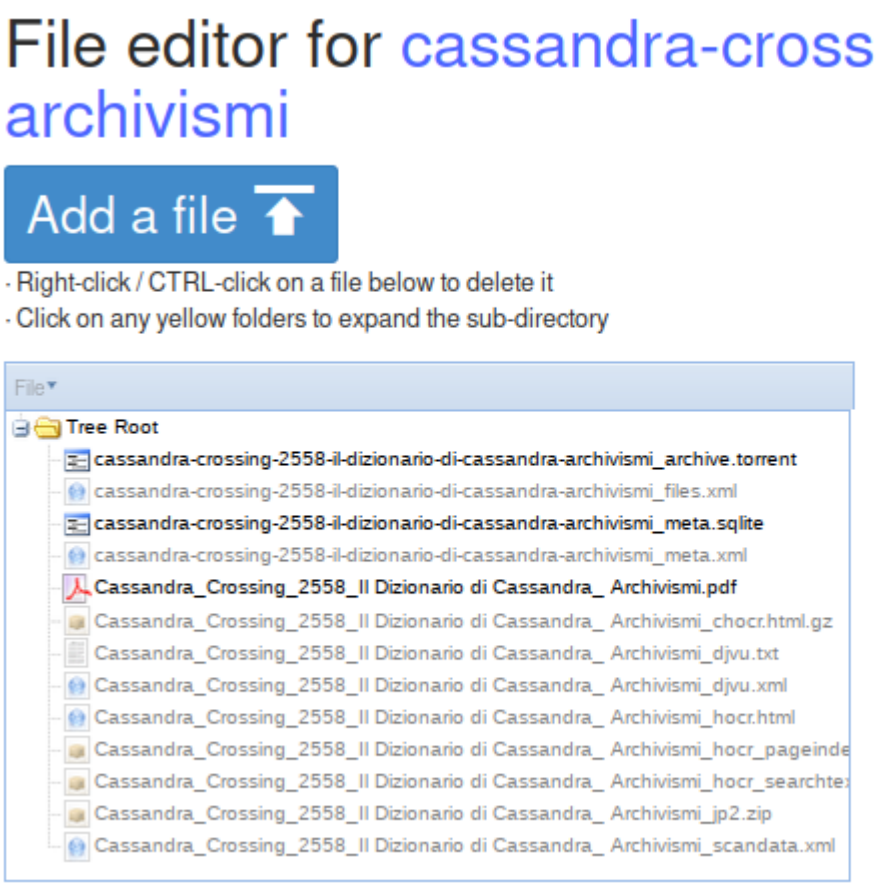
where am I in line?

| identifier | task_id | server | cmd | submittime | submitter | args |
|---|---|---|---|---|---|---|
| cassandra-crossing-2558-il-dizionario-di-cassandra-archivismi - History Mgr | 4099976267 | iw601907 | derive.php | (19.7 minutes) | calamarim | server_primary=ia601204.u.. |
| cassandra-crossing-2558-il-dizionario-di-cassandra-archivismi - History Mgr | 4099976257 | ia601204.us.archive.org | book_op.php | (19.7 minutes) | calamarim | op0=VirusCheck&dir=/24/it.. |
| cassandra-crossing-2558-il-dizionario-di-cassandra-archivismi - History Mgr | 4099973598 | ia601204.us.archive.org | archive.php | (22.6 minutes) | calamarim | done=delete&from_url=rsyn.. |
| cassandra-crossing-2558-il-dizionario-di-cassandra-archivismi - History Mgr | 4099973556 | ia601204.us.archive.org | archive.php | (22.7 minutes) | calamarim | done=delete&from_url=rsyn.. |

Molte cose continueranno a succedere al nostro oggetto nel backend, e ne riparleremo; intanto torniamo nella finestra oggetto, e nel microscopico menù in alto a sinistra clicchiamo su *manage*.

Compariranno due grosse icone; quella di sinistra (importantissima) vi permette di editare i metadati, ma per ora è bloccata dal processo di creazione dell’oggetto in corso, e se ci cliccate vi verrà spiegato perché.

Quella di destra invece permette di editare i file contenuti nell’oggetto, e se la cliccate vi aprirà una vista sulla cartella e sul suo contenuto. A seconda del tempo passato e del file che avere archiviato, troverete contenuti diversi, e molti più file di quelli che vi aspettereste.



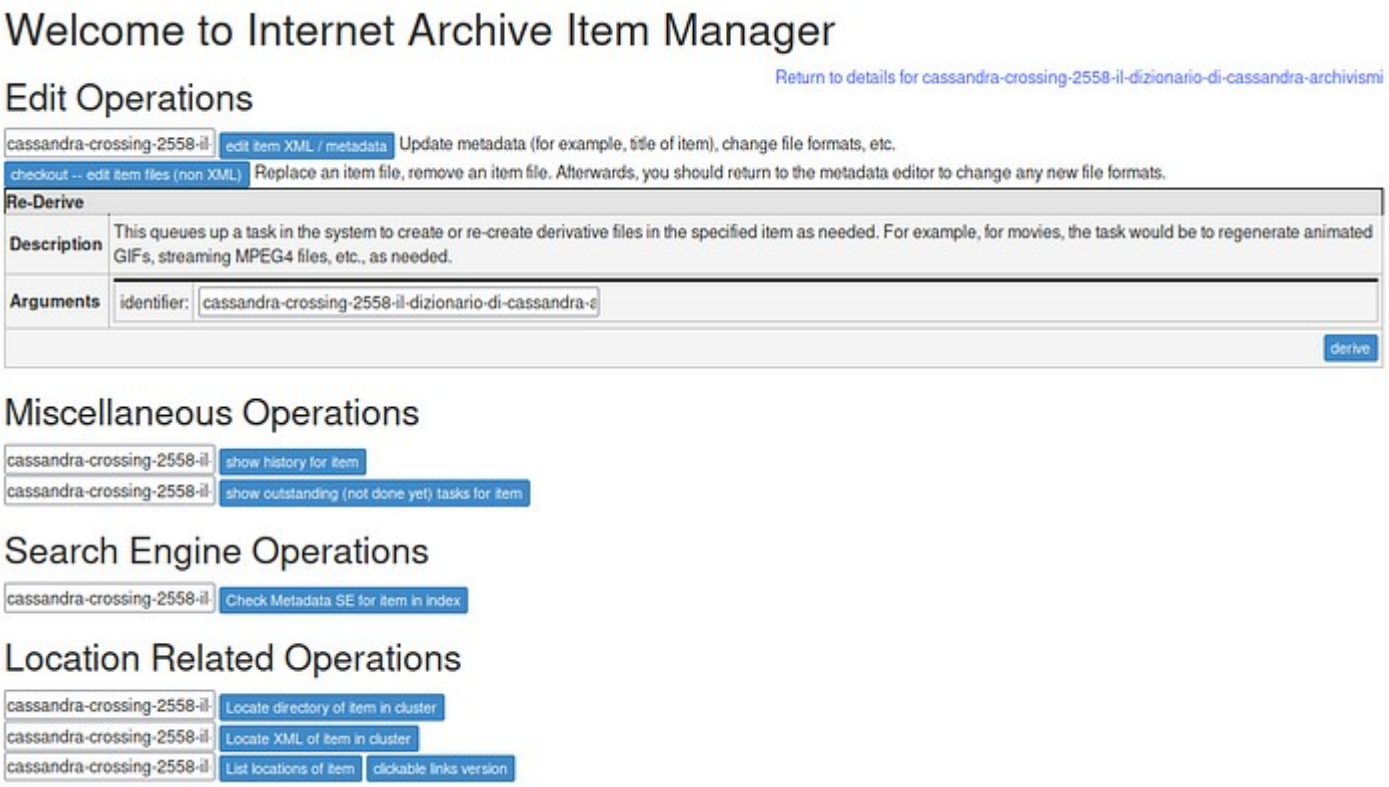

Noterete il file .pdf originale che è stato usato in questo esempio, due file .xml ed uno .sqlite, che contengono informazioni di sistema (e come tante altre cose, ne riparleremo). C'è un nuovo file .torrent, che può già essere scaricato ed utilizzato per fornire un link *torrent*, utile se il file caricato fosse molto grande e lo si dovesse far scaricare da molte persone.

Ci sono infine diversi file, in parte ancora indicati in grigio ed inaccessibili, che *testimoniano* le operazioni ancora in corso che Internet Archive sta facendo per voi, e che dipendono dal tipo di file che avete archiviato.

Ad esempio, dal nostro file .pdf verrà creato automaticamente un file di solo testo, contenente appunto tutto il testo presente nel pdf. Sempre nel caso di un pdf verrà creato un indice della pagine. Se si fosse invece trattato di un file video, tra le altre cose sarebbe stata creata una directory contenente 255 thumbnail, uniformemente estratti da tutta la lunghezza del video, che possono essere usati per visualizzarlo come oggetto *video* (ad esempio in una *timeline*). Altri file verranno creati, ma ci avviciniamo alla fine di questa intensa puntata.

Perché *… questa è un'altra storia*.

Ma un'ultima cosa. Sempre dal peculiare *micromenù* in alto a sinistra della finestra *oggetto* si può accedere al link che apre la finestra dell'*item manager*, in cui è possibile gestire l'oggetto creato in molteplici aspetti.




Alcuni tra i 24 indomiti lettori, i più interessati e dotati di tempo ed iniziativa, potranno partire da qui o dalle altre finestre che abbiamo visto per un’esplorazione *in solitario*, che potrà durare tantissimo e portarli molto lontano.

A questi arditi Cassandra raccomanda di dotarsi di un po’ di Python e di confidenza con le API; a questo fine suggerisce di utilizzare la molto ben organizzata pagina di help e consegna loro questo prezioso link alla documentazione sviluppatori di Internet Archive.

Come esempio, se voleste sviscerare l’argomento dei file creati automaticamente durante un upload, potreste leggere questo articolo dell’help.

Gli altri aspetteranno invece che Cassandra, *lento pede*, compia questa esplorazione per loro od insieme a loro.

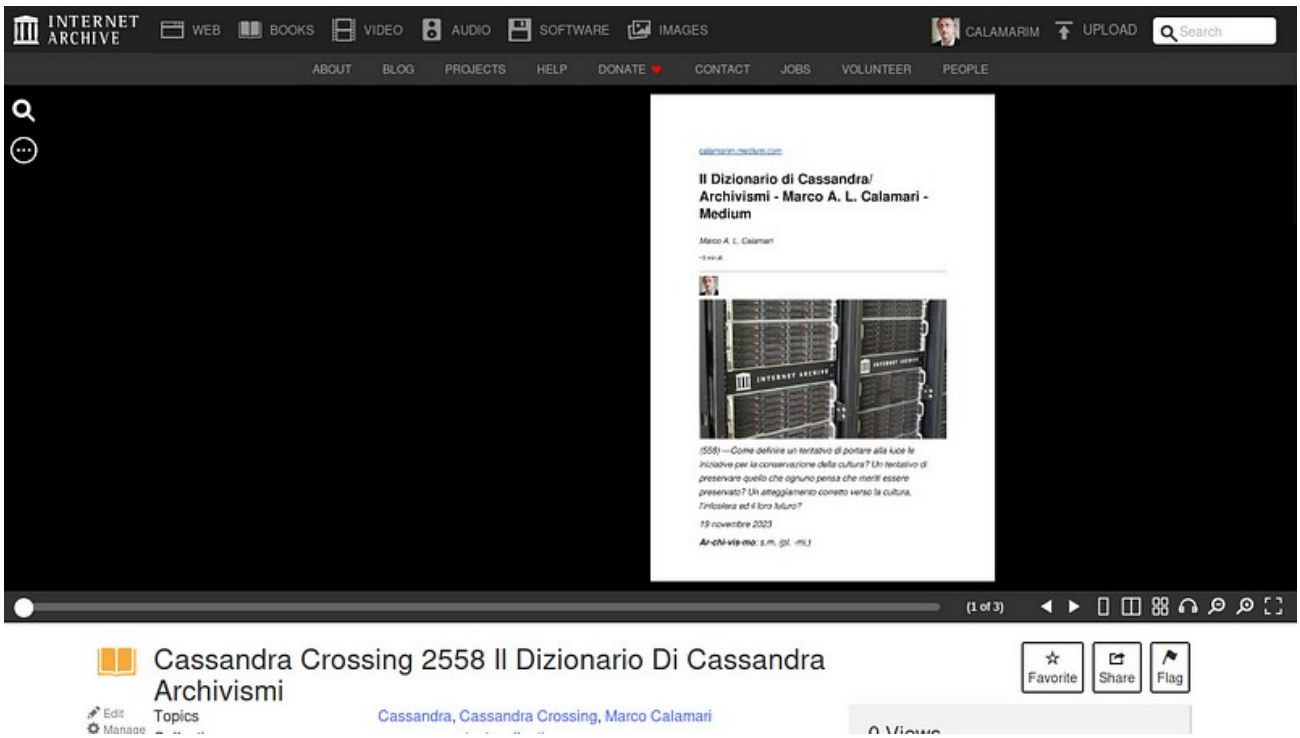
*Stay tuned* per la prossima puntata di “*Archivismi*”.

# Archivismi: il giorno dopo l’upload

*(561 )—Ieri abbiamo fatto il nostro primo upload e ne abbiamo visto i risultati. Ma oggi è cambiato qualcosa?*

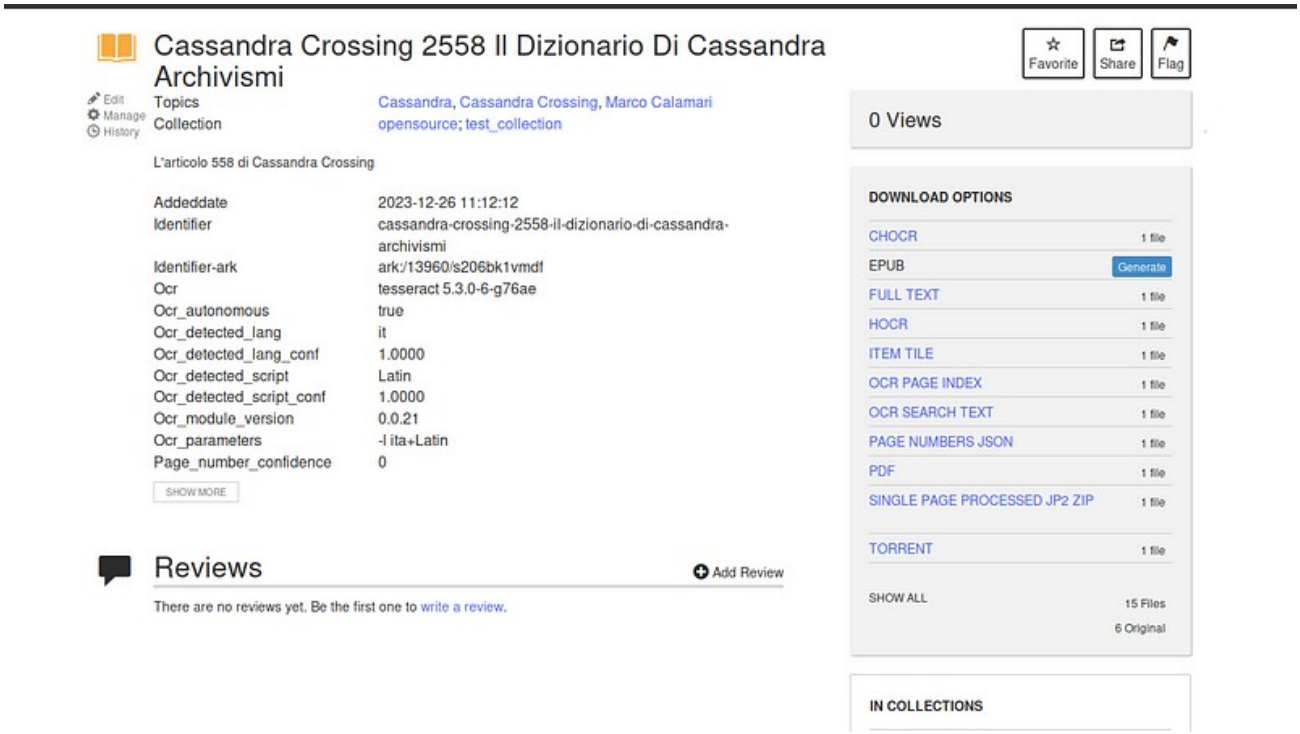
**27 dicembre 2023**—Nella scorsa puntata Cassandra ha cercato di raccontarvi una parte del funzionamento di Internet Archive. Abbiamo appena scalfito la superficie delle sue caratteristiche, e per non annoiarci abbiamo provato ad archiviare il file .pdf di un articolo di Cassandra, ed a descrivere cosa succedeva.

Ci siamo così resi conto di aver avviato un processo tanto complesso quanto lento, ma per fortuna completamente automatico. Tanto lento che dopo più di mezz’ora non si era ancora concluso. Tornando oggi sulla pagina del documento, troviamo il *browser di oggetti* di Internet Archive attivo, ed il processo che si è completato.




E’ possibile sfogliare rapidamente le pagine, farle leggere ad una voce molto robotica, e selezionare parti di testo su qualsiasi pagina. Sembrano cose da poco, considerando che la sorgente era un pdf “moderno”, ottenuto direttamente da un documento Libreoffice, ma in effetti l’apparentemente “semplice” pdf è stato scomposto in una quantità di file, alcuni dei quali non avevamo ancora analizzato.

---




Anche solo dai nomi, possiamo facilmente capire che un qualche processo OCR di riconoscimento dei caratteri è stato eseguito automaticamente. Questi file, alcuni dei quali vengono usati dal *browser di oggetti* di Internet Archive, permettono a quest’ultimo di visualizzare il documento.

A questo punto qualcuno degli informatissimi 24 lettori sbotterà “*Ma tutto questo è assolutamente banale, lo si poteva fare anche con Acrobat Reader, senza tutto questo ambaradan.*” Il caro lettore ha ragione sul fatto specifico, ma torto sulla questione più generale. Si, perché archiviando il pdf moderno di 3 pagine abbiamo in realtà usato un cannone per ammazzare una zanzara, perdipiù gracilina e malata.

Ora è arrivato il momento di provare a scatenare tutta la potenza archiviativa di *Internet Archive*. Per questo Cassandra ha sfruttato un lavoro di archiviazione che attendeva il suo alter-ego Marco Calamari. Si trattava di archiviare un centinaio di numeri di una piccola rivista, uscita negli ultimi 30 anni ed esclusivamente in formato cartaceo.

Erano già stati raccolti i file .pdf generati dai vari programmi di impaginazione elettronica usati per realizzare la rivista, e per fortuna conservati come sottoprodotto. Erano state anche realizzate, artigianalmente ed in vari modi, le scansioni dei primi numeri cartacei, anche questi in formato pdf, ma ovviamente non ricercabili, essendo le pagine delle “*fotografie*”.

Tutto questo materiale, anche se già in formato digitale, avrebbe richiesto un tempo lunghissimo per essere messo insieme, allineato e pubblicato in un formato ricercabile e riutilizzabile, particolarmente in ambiti di archiviazione “seria”.

Infatti il vero, grosso problema non era quello di creare una collezione di file pdf, ma quella di archiviarla in maniera utile, ricercabile e consultabile. Altrimenti, come spesso accade, questi file, pur faticosamente raccolti, sarebbero prima o poi finiti dimentcati in una chiavetta in fondo ad un cassetto, od in un angolo di cloud commerciale, effimero e dove nessuno (tranne i GAFAM) li avrebbe potuti trovare ed utilizzare.

Ma è bastato mettere insieme i 75 file di vario formato e contenuto in un unico pdf, usando l’utilissimo software libero Pdftk, realizzando così un pdf unico di quasi 1 terabyte, ed uploadare quest’ultimo su Internet Archive, esattamente come avevamo fatto per l’articoletto di 3 pagine. Anche questo file è stato preso in carico dal sistema e “tritato” per tutta la notte; stamani era già disponibile.

Tutte le anomalie e le differenze erano state risolte automaticamente, ed un documento di 662 pagine, contenente l’intera raccolta della rivista, era disponibile, rapidamente sfogliabile, selezionabile, ricercabile e ascoltabile, ed era stato creato con un impegno di pochi minuti di tempo.

Se aggiungiamo a questo il fatto che il documento è stato archiviato in maniera ridondante in più datacenter, e si trova in una in una biblioteca digitale che lo mette a disposizione di chiunque, liberamente ricercabile e visualizzabile, la cosa diventa quasi stupefacente, anche senza aggiungere che è disponibile pure in formato ebook (.epub) e che se necessario può essere ulteriormente “lavorato” per altri scopi.

Giusto per descrivere in linea di massima cosa è stato prodotto durante l’archiviazione, il pdf originale è stato diviso in pagine, prima di tutto per velocizzarne la visualizzazione. Ciascuna pagina è costituita da un file pdf in un formato particolare, una immagine di sfondo, la scansione della pagina originale, più un layer di testo selezionabile, sovrapposto alla pagina e generato sottoponendo ad OCR la scansione stessa.

La cosa veramente notevole è che il sistema è stato in grado di gestire correttamente un misto di file pdf con differenti strutture interne, da semplici scansioni a pdf strutturati, e di riportarli tutti ad un minimo comune multiplo costituito dai pdf a strati delle singole pagine.

Beh, se tutto questo vi sembrasse poco, è perché questa serie di articoli non è adatta a voi; è invece adatta ai futuri *bibliotecari digitali* che, per caso o per fortuna, siano capitati su queste paginette. Ma potreste ancora cambiare idea.

*Stay tuned* per la prossima puntata di “*Archivismi*”.

# Archivismi: l’organizzazione dei documenti in Internet Archive

*(562 )—Completiamo la descrizione di come Internet Archive organizza i documenti, e di come il sito permette di utilizzarli*

**28 dicembre 2023**—Nella scorsa puntata siamo riusciti ad archiviare documenti ,anche grandi ed in formati eterogenei, e convertendoli durante il processo in modo da averli disponibili in più formati digitali, riutilizzabili per gli scopi più diversi.

Ma per poter dire di aver realmente *archiviato* un documento bisogna anche averlo inserito in un più vasto corpo di documenti, a sua volta dotato di indici e metodi di vario tipo per organizzare e ricercare i documenti e le informazioni in essi contenute.

Facile quindi cogliere l’importanza di sapere *a priori* come una biblioteca digitale già esistente permette di organizzare i propri dati, adeguandosi ad utili e ben studiati standard comuni.

L’architettura di *Internet Archive* è tanto semplice quanto potente.

**Il primo livello dell’architettura è l’oggetto**, che può essere creato e successivamente modificato in vari modi; un oggetto è tipicamente un singolo documento. Se l’oggetto viene creato da un utente registrato e collegato, all’utente viene assegnato il ruolo di amministratore dell’oggetto, che può quindi modificarlo, arricchirlo di ulteriori file di dati e nuovi metadati, e così via. Se l’oggetto viene invece creato in forma anonima da un utente non registrato o non collegato, ad esempio utilizzando la Wayback Machine, non può più

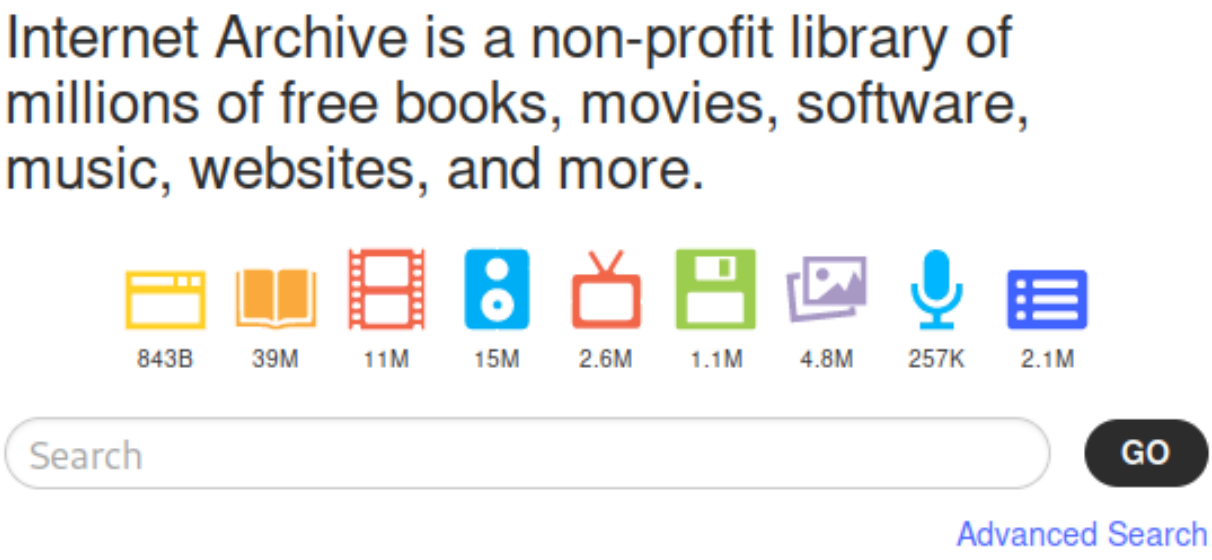
essere modificato da chi lo ha creato, ma solo dagli amministratori di Internet Archive, dietro specifica richiesta da inoltrare via email, formattata con specifici template.

**Il secondo (ed ultimo!) livello di architettura è la collezione (Collection)**. Una collezione è un oggetto di tipo particolare, formato solo da riferimenti ad altri oggetti. Come tutti gli oggetti è dotato di suoi propri metadati, ma può essere creato solo dagli amministratori di Internet Archive dietro specifica richiesta di un utente registrato, utente che deve possedere certi requisiti, elencati nelle policy di creazione delle collection. Una collezione può contenere altre collezioni come sotto-collezioni. L'utente che si è fatto creare ed assegnare la collezione la può amministrare, inserendoci gli oggetti di cui è il creatore, ad esempio quelli che ha uploadato.

Quando un oggetto viene creato, viene assegnato per default ad una collezione; se l'oggetto è creato in maniera anonima o direttamente da un utente tramite upload, viene assegnato automaticamente ad una collezione che potremmo definire "di sistema".

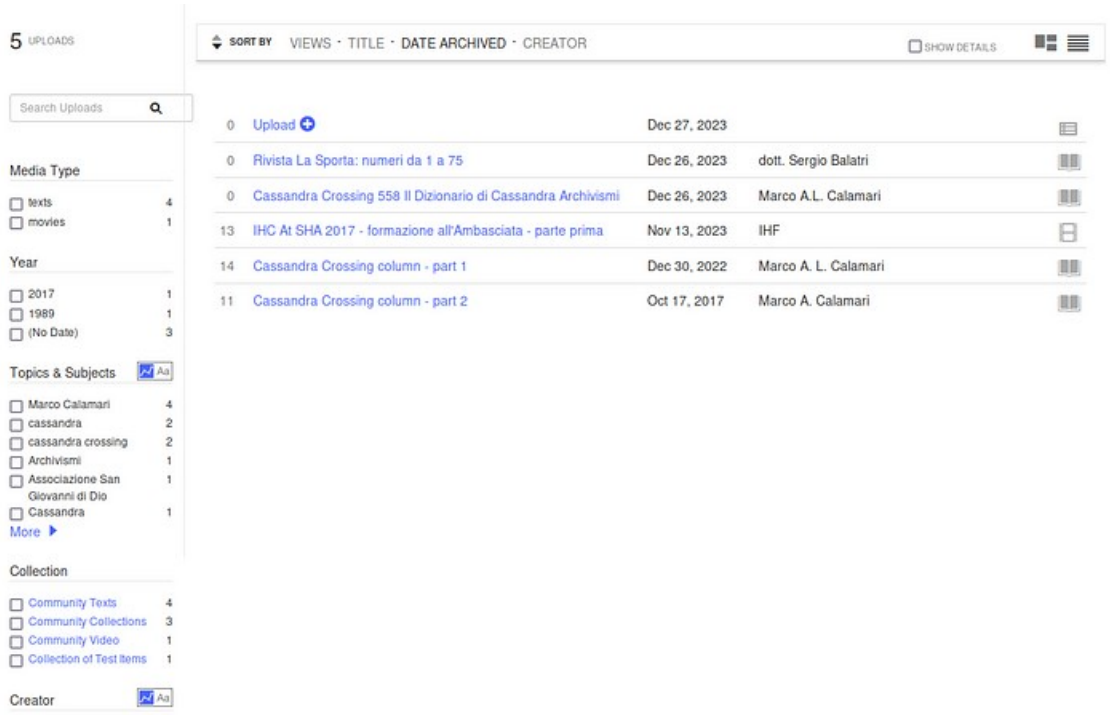
Ad esempio i documenti che abbiamo creato nelle precedenti puntate, come si può vedere esaminando i metadati nella finestra dell'oggetto o tramite il metadata editor, sono stati assegnati per default alla collezione "*opensource*". Ricorderete che il file dell'articolo usato è stato da noi specificatamente marcato come *oggetto effimero* e destinato ad essere cancellato dopo 30 giorni. Esaminando i suoi metadati, si può notare che è stato assegnato anche alla collezione *test_collection.* Un processo automatico, evidentemente, "spazzola" tutti gli oggetti assegnati a questa collezione e rimuove definitivamente quelli più vecchi di 30 giorni.

Esiste uno pseudo "terzo livello" di organizzazione che è solo di "presentazione", e viene costruito dai creatori del sito assegnando gli oggetti a collezioni particolari ed utilizzandole poi per generare specifiche pagine sul sito di Internet Archive, per favorire un accesso rapido ed estemporaneo a certe categorie di informazioni. Queste sono, ad esempio, le icone che si trovano in home page e sulla barra dei menu del sito.

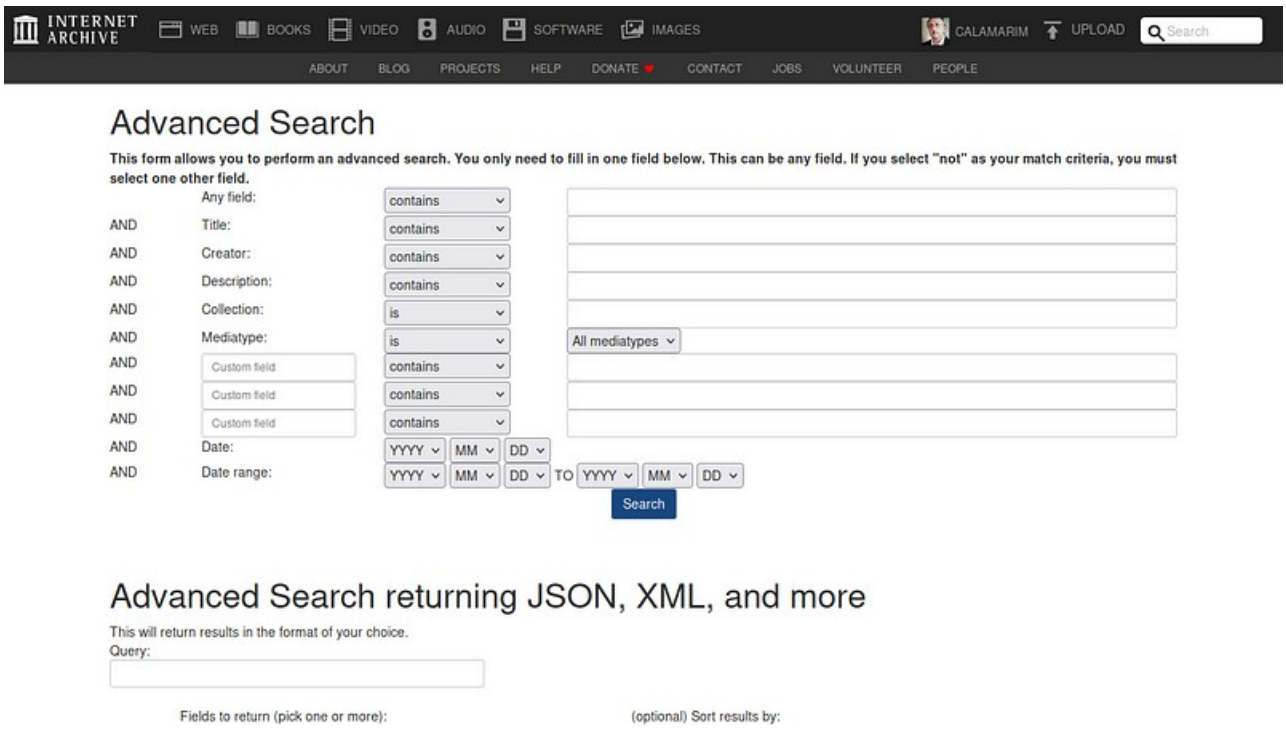


Il sito di Internet Archive ha un'aria un po' "farraginosa" e retrò. In effetti però, appena preso un minimo di confidenza, si rivela un meccanismo abbastanza utile e potente per trovare documenti di interesse od avere spunti di cose nuove, che sono di solito collezioni molto accedute.

In realtà, comunque, le informazioni di interesse si trovano, come è facile immaginare trattandosi di una biblioteca, tramite le funzioni di indicizzazione e ricerca, rese disponibili in vari modi sul sito. Ad esempio, visualizzando i propri upload, nella parte sinistra dello schermo si ha accesso ad una serie di categorie di selezione pertinenti, simili a quelle di Amazon.




Quando necessario, si può accedere direttamente alla funzione di ricerca tramite il box "*Search"* in alto a destra nel sito. Si può accedere alla funzione di ricerca completa cliccando dentro il box stesso e selezionando "*advanced search*".




Ed anche per oggi è tutto. *Stay tuned* per la prossima puntata di “*Archivismi*”.

# Archivismi: API, quando il gioco si fa duro

*(563) — Oggi ci sposteremo su un differente piano di utilizzo di Internet Archive, quello della "programmazione" via API; ma prima dovremo obbligatoriamente parlare di doveri e responsabilità degli utenti di Internet Archive.*

**29 dicembre 2023** — Nelle ultime due puntate (è da oggi disponibile una lista completa degli articoli di "Archivismi") ci siamo occupati di un'archiviazione elementare su *Internet Archive*; archiviare un singolo file ci ha comunque aperto una parte significativa del sistema che abbiamo davanti, e delle potenti funzionalità che ci mette a disposizione.

Molto, molto altro rimane da mostrare, anche solo per le operazioni di archiviazione manuali. Prossimamente descriveremo e realizzeremo una vera campagna di archiviazioni, raccontando le minuzie ed i problemi spiccioli che distinguono i casi reali dagli esempi che troviamo sui manuali.

Ma oggi tratteremo di un argomento già accennato di sfuggita in una puntata precedente, e che porta la potenza archivistica che *Internet Archive* mette a disposizione dei suoi utenti ad un nuovo livello. Parliamo ovviamente della possibilità di "programmare" le operazioni su *Internet Archive*.

Non ci vuole un genio per immaginare che un servizio come *Internet Archive* esiste perché ha alle spalle un piccolo esercito di programmatori che scrivono, mantengono e fanno evolvere una base di software dedicato. E per inciso, per fomentare la mai estinta "*Classifica dei migliori linguaggi di programmazione*", anche in *Internet Archive* *Python* la fa da padrone!

Ma torniamo all'argomento di oggi.

In breve: sì, è possibile usare Internet Archive usando script o veri programmi che automatizzano le operazioni di archiviazione che decidiamo di realizzare.

E sì, questo viene realizzato "*esponendo una API*". Per il conforto dei non programmatori, significa semplicemente che è possibile automatizzare le operazioni da compiere utilizzando degli script o dei veri e propri programmi, che eseguono, ovviamente via Rete, precise chiamate a delle funzionalità elementari di Internet Archive, definite appunto in una API — *Interfaccia per la Programmazione Applicativa*.

Non ci sarebbe bisogno di dire altro, semplicemente di fornire nuovamente il link al *Portale degli Sviluppatori di Internet Archive*, e lasciare che chiunque abbia mai *trafficato*, anche solo realizzando uno script .bat per il DOS, scopra ed utilizzi la potenza delle API di Internet Archive.

Ma no, un minimo di indicazioni e raccomandazioni preliminari sono comunque necessarie, prima di fare anche solo un piccolissimo esempio.

In primis, **Internet Archive non pone limiti predefiniti a quello che un utente può fare dei servizi che vengono forniti**; ad esempio non limita a priori la quantità di informazioni che possono essere archiviate.

Ma nessuna realtà esposta al pubblico può essere "*indifesa*", visto che una percentuale di imbecilli, profittatori e delinquenti esistenti al mondo è presente anche tra gli utenti di Internet Archive.

Come la storia della Rete ha più volte dimostrato, realtà collaborative di grandi dimensioni, ad esempio Wikipedia, riescono a sopravvivere e svilupparsi solo se gestite come un ibrido tra democrazia imperfetta e tirannia illuminata. *Internet Archive* non fa eccezione.

E' per questo che alcune risorse, come ad esempio le *Collezioni*, vengono centellinate e fornite solo a richiesta. Una serie di amministratori di vario livello supervede e controlla infatti il funzionamento e l'utilizzo di Internet Archive, e tiene in riga, bacchetta od espelle gli utenti *disfunzionali*. **Una tale presenza non deve essere vista come un problema od un limite, ma come una risorsa**; infatti gli amministratori hanno il ruolo principale di aiutare tutti gli utenti ad utilizzare *Internet Archive*.

Gli amministratori sono tuttavia una risorsa preziosa e scarsa; **mandare una email agli amministratori**, quando non direttamente previsto dalle procedure (ad esempio per la creazione di una Collection) **deve essere vista come ultima risorsa**, da utilizzare solo dopo un'attenta lettura della documentazione e dell'help in linea, molte prove, una ricerca nel blog e perché no, anche sui normali motori di ricerca. **Mi raccomando!**

Ma non si era detto che avremmo *programmato* qualcosa? Verissimo, e passiamo subito alla pratica. E per partire da qualcosa di semplice ed innocuo, ipotizziamo di aver trovato una serie di cose che ci interessano, ad esempio parecchi numeri di una rivista, e di volerli scaricare in maniera veloce, affidabile, e che non richieda operazioni manuali ripetitive.

E per semplicità, faremo il tutto da linea comandi, senza utilizzare direttamente le API e quindi senza dover scrivere un vero programma in Python o simili; ci basterà scaricare il programma Python "*ia*" ed utilizzarlo. *ia* è un programma già "pseudo-compilato", cioè

scritto in un “linguaggio” intermedio detto *Python Bytecode*, che è portabile su qualsiasi piattaforma abbia un ambiente Python3 installato.

**L’utilizzo di una versione di Linux, Debian, Ubuntu etc., è vivamente consigliato**. Potete anche utilizzarlo in una macchina virtuale Virtualbox o VMWare su qualsiasi computer.

Dovrebbe anche funzionare l’ambiente WSL di Windows, ma qui Cassandra non procede oltre ed abbandona gli arditi che volessero cimentarsi; anzi, eventualmente aspetta da loro dei feedback a riguardo per integrare questo articolo.

Quindi torniamo con Cassandra alla sua amata Debian, ed installiamo e configuriamo *ia* con la procedura che troviamo qui. Ma anche un semplice

```
sudo apt install internetarchive
```

è sufficiente. Miracoli di Debian …

In breve, su un computer dove sia installato l’ambiente Python3 si deve scaricare dove preferiamo, oppure installare, il comando ia, renderlo eseguibile, ed infine lanciarlo con il parametro *configure* per associarlo al nostro utente (avete creato il vostro utente, vero?).

E’ tutto pronto; come primo esempio con il seguente comando possiamo scaricare il solo pdf originale del nostro articolo di esempio, che avevamo caricato la scorsa puntata.

```
$ ./ia download cassandra-crossing-2558-il-dizionario-di-cassandra-archivismi
— no-directories — format=”Text PDF”

cassandra-crossing-2558-il-dizionario-di-cassandra-archivismi:
downloading Cassandra_Crossing_2558_Il Dizionario di Cassandra_
Archivismi.pdf: 100%|█| 513k/513k [00:00<00:00, 709kiB/s
```

Ma se avessimo voluto scaricare tutto l’oggetto, file derivativi inclusi, avremmo potuto scrivere ancor più semplicemente

```
$ ./ia download cassandra-crossing-2558-il-dizionario-di-cassandra-archivismi
```

Avremmo così ottenuto una directory con lo stesso nome dell’identificatore dell’oggetto, contenente tutti i file da cui è formato. Lo stesso procedimento funziona anche per scaricare una intera collezione, o parti di essa. Un’altra raccomandazione, **calcolate prima quanto è grande la selezione che avete fatto**; su *Internet Archive* ci sono oggetti di dimensioni enormi.

Per avere aiuto, oltre che consultare la guida online, basta dare i comandi

```
$ ./ia help

$ ./ia help download

$ ./ia help upload
```

Terminiamo con altre raccomandazioni in ordine sparso.

Se caricate nuovi oggetti, è meglio usare il metodo con foglio elettronico in formato CSV, di cui trovate un esempio qui o nella guida. In questo modo avrete sempre sotto controllo tutti

i parametri insieme. Dare tutti i parametri da linea comandi può essere complesso e si possono facilmente commettere errori.

Quando creerete i vostri oggetti, **inseriteli sempre nella collezione** *test_collection*, come è mostrato anche nel foglio di esempio. I motivi li abbiamo già spiegati.

Quando invece inserirete i vostri primi oggetti *definitivi*, non inserite tra i parametri la collection, lasciando quella di default *opensource*. Buona sperimentazione!

Ed anche per oggi è tutto. *Stay tuned* per la prossima puntata di “*Archivismi*”.

# Archivismi: archiviamo Cassandra, parte prima

*(564)—Oggi cambiamo lato della medaglia; niente tecnica, raccontiamo una storia vera.*

**31 dicembre 2023**—Nelle ultime tre puntate abbiamo lavorato su Internet Archive, ma solo con esempi semplici.

Archiviare vuole però spesso dire archiviare una quantità di materiali diversi, con uno scopo finale. Ed in questi casi non ci sono esempi semplici che bastino; il diavolo sta sempre nei dettagli, e le informazioni più utili si apprendono ascoltando storie ed esperienze reali.

Ecco che oggi Cassandra vi racconterà una storia vera, tuttora non conclusa, e parlerà solo di dettagli che non hanno a che fare direttamente con Internet Archive, ma con le fasi preliminari una campagna di archiviazione generica, in cui il lavoro più lungo è ritrovare, raccogliere e soprattutto preparare il materiale per l'archiviazione vera e propria.

E cosa di meglio che raccontare la **campagna di archiviazione di Cassandra Crossing**? Si, era da tempo che Cassandra metteva da parte pezzi destinati ad essere archiviati. Ma andiamo con ordine.

Le origini di Cassandra Crossing risalgono al lontano 2003, la pubblicazione regolare (beh, quasi regolare….) inizia invece nel 2005 su Punto Informatico. Prosegue poi su altre testate come Zeusnews.it, talvolta in parallelo. Si estende anche su carta e in video.

I materiali disponibili erano dei tipi più svariati; file di testo con e senza accenti, file di word processor di tipi diversi, file pdf e chi più ne ha più ne metta. Tanti file sono ovviamente andati semplicemente persi.

Fu così che parecchi anni fa Cassandra cercò il modo di recuperare, omogeneizzare e *centralizzare* tutto il *corpus* di Cassandra.

Come in tutte le cose, conviene buttarsi a capofitto in un lavoro, mapensare, programmare, fare e poi cercare una via ancora migliore. Dopo diversi tentativi, Cassandra ha provato Medium.com, un *social specializzato* per scrittori od aspiranti tali. Oltre a fornire un punto unico, in cui scrivere con un discreto editor online ed immagazzinare gli articoli, Medium.con è dotato di una ottima funzionalità di importazione di testo da qualunque sito, anche con pagine piene di pubblicità od effetti vari.

E' dotato di una funzionalità di esportazione dei dati utente, che salvava i singoli articoli in formato in HTML.

Fu così che Cassandra *centralizzò* l'archivio su Medium.com, non senza aver dedicato molto tempo a ritrovare, con i motori di ricerca, i link ai vecchi articoli, mai archiviati in locale o comunque *perduti*.

Ma la soluzione non era soddisfacente per vari motivi, a cominciare dal fatto che gli articoli erano in un cloud, e peggio ancora in quello che sostanzialmente era un social, con tutti gli aspetti deleteri che Cassandra odia e vi racconta spesso.

E così Cassandra decise di iniziare ad archiviare Cassandra Crossing su Internet Archive. E visto che si partiva da un archivio completo in formato omogeneo, sembrava dovesse essere una passeggiata. "Madornale errore", come usa dire Jack Slater.

Infatti l'omogeneità necessaria non è solo una questione di formato, ma soprattutto di struttura interna e di omogeneità delle informazioni memorizzate dei file degli articoli.

Partiamo dalla cosa più semplice: i nomi dei file. Ovviamente Medium.com utilizza una sua filosofia, e forma il nome dalla data di pubblicazione (non quella originaria, ma quella su Medium.com), aggiungendo un identificativo binario ed una derivazione del titolo.

Qualcosa tipo

*2023–12–29_Cassandra-Crossing—Archivismi—l-organizzazione-dei-documenti-in-Internet-Archive-e83b9e3b9cca.html*

Ora, è pur vero che i file si rinominano anche a mano, ma si tratta di un lavoro improbo quando i file sono centinaia o migliaia. Automatizzare diventa indispensabile. Per fortuna in Linux sono disponibili linguaggi di scripting potenti e librerie che hanno del miracoloso.

Si riesce quindi a rinominare abbastanza facilmente i file togliendo, aggiungendo e riordinando informazioni. Paradossalmente la cosa più difficile è stata inserire automaticamente il numero dell'articolo all'inizio del nome del file.

Per fortuna Cassandra, che talvolta è metodica, aveva l'abitudine di scrivere il numero dell'articolo all'inizio del sottotitolo, mettendolo tra parentesi tonde. Con qualche piccola alchimia di espressioni regolari è stato così possibile estrarlo automaticamente ed utilizzarlo per costruire un più "*umano*" nome di file come

*562_Cassandra-Crossing—Archivismi—l-organizzazione-dei-documenti-in-Internet-Archive.html*

Poi è stato necessario elaborare i file, ripulirli e convertirli in formati bene archiviabili.

Il primo passo necessario è stato ripulire i file html da una immane quantità di tag nascosti, totalmente inutili per definire il testo ma necessari per garantire le funzionalità del sito di Medium.com. Infatti, come tutti i social, Medium.com implementa le funzioni di esportazioni al minimo sindacale richiesto dal (sempre sia lodato) GDPR, e quindi produce dati completi si, ma non adatti per essere facilmente riutilizzati.

La soluzione migliore che Cassandra ha trovato è stata quella di convertire l'html in formato markdown, filtrare delle linee che non contenevano informazioni utili e riconvertirlo nuovamente in html. Questo piccolo miracolo è stato possibile grazie alle librerie di conversione documentale Pandoc, coadiuvate dalle normali utilità unix come grep.

Ora che i file sono ripuliti ed hanno un nome umano sussiste ancora il problema delle immagini incluse nei file. Infatti le immagini non vengono esportate con gli altri dati, e gli url delle immagini puntano tutti ai server di Medium.com, che quindi, malgrado tutto il lavoro fatto, ha ancora *in pugno* una parte importante degli articoli.

E' necessario quindi convertire le immagini remote in immagini inline, dentro lo stesso codice html, codificandole in base64. Questo processo, concettualmente semplice, deve di solito essere svolto a mano per ogni singolo file ed url; per fortuna esiste il modo di farlo automaticamente, tramite il parametro—*self-contained*, aggiunto al comando Pandoc di riscrittura dell'html.

Per l'archiviazione, il formato principale scelto è comunque il pdf, che non ha questo problema perché convertendo l'html in pdf le immagini vengono inserite direttamente nel file.

Per non farsi mancare niente, sempre grazie ai miracoli di Pandoc, Cassandra ha potuto convertire in maniera semplicissima in pdf tutti i formati già prodotti, l'html di partenza, il markdown e l'html semplificato, scegliendo poi il migliore.

Il risultato, per ora, lo trovate qui.

Concludendo, un paio di giornate "piene" di lavoro hanno portato a questo script bash di 39 righe, certamente non ottimale né privo di errori, che qui comunque commenteremo, giusto per rendere l'idea. Capirlo a grandi linee è sufficiente. Ma se vi servisse, riutilizzarlo sarebbe per voi un bel risparmio di tempo.

```
# Procedura per la preparazione all'archiviazione articoli
#
# inizializzazioni varie
_base=”./tuttocassandra_elaborazione/”
_base2=”./posts/”
_base3=”./markdown/”
_base4=”./temp/”
_base5=”./html/”
_base6=”./pdf/”
_temp=”temp.txt”
#
# cambio directory di lavoro, creazione directory e
pulizia file
cd “${_base}”
mkdir markdown html temp pdf
rm markdown/* html/* temp/* pdf/*
cd “${_base2}”
_dfiles=”*”
rm “${_temp}”
#
# inizio loop principale
for f in $_dfiles
do
#
# estrazione del numero dell’articolo
g=`grep -Eo -m 1 ‘\([0–9]+\)’ $f | tr -d ‘()’`
g=”000"$g
g=`echo $g | rev | cut -c 1–3 | rev`
h=`echo $f | cut -d ‘_’ -f2- | rev | cut -d ‘-’ -f2-| rev`
#
# formazione del nuovo nome del file e copia col nuovo
nome
i=$g”_”$h
echo “$i”
cp $f “../$_base4${i}.html”
#
# conversione in formato markdown, ripulitura e
riconversione in html
pandoc -f html -t markdown “../”$_base4$i”.html” > “$
{_temp}”
grep -v “^:::” “${_temp}” |sed -e ‘s|{#.*}||g’ > “../”$
{_base3}$i”.md”
pandoc — self-contained -f markdown -t html “../”$
{_base3}$i”.md” > “../”$_base5$i”.html”
pandoc — pdf-engine=xelatex -f markdown -t pdf
“../”$_base3$i”.md” > “../”${_base6}$i”.pdf”
#
# pulizia e fine ciclo
done
rm -rf “${_temp}” “../$_base4”
```

Ed anche per oggi è tutto. *Stay tuned* per la prossima puntata di “*Archivismi*”.

# Archivismi: archiviamo Cassandra, parte seconda

*(565)—Dopo aver preparato i pdf non ci sono più scuse, dobbiamo archiviare il nostro primo articolo di Cassandra Crossing.*

**1 gennaio 2024**—Nelle precedenti puntate di *Archivismi* abbiamo raccontato le caratteristiche principali di Internet Archive e caricato un semplice documento di esempio. Successivamente ci siamo dati l'ambizioso obbiettivo di uploadare l'*opera omnia* di Cassandra, ed abbiamo faticosamente preparato il materiale necessario nei formati e struttura più opportuni.

Non ci sono più scuse; è il momento di iniziare a caricare il primo documento di Cassandra Crossing, con tutte le cosette ed i metadati al posto giusto!

Dobbiamo quindi cimentarci davvero con *ia* e, visto che dovremo caricare centinaia di documenti, non farlo direttamente con la linea comandi, caricando un file per volta e scrivendo tutti i parametri ed i metadati su una lunghissima linea comandi.

Molto meglio impratichirsi fin da subito con i *bulk upload*, che si realizzano fornendo ad *ia* un unico parametro, cioè il nome di un foglio elettronico in formato CSV, in cui inseriremo i dati necessari (e li modificheremo tantissime volte per rimediare ad inevitabili errori).

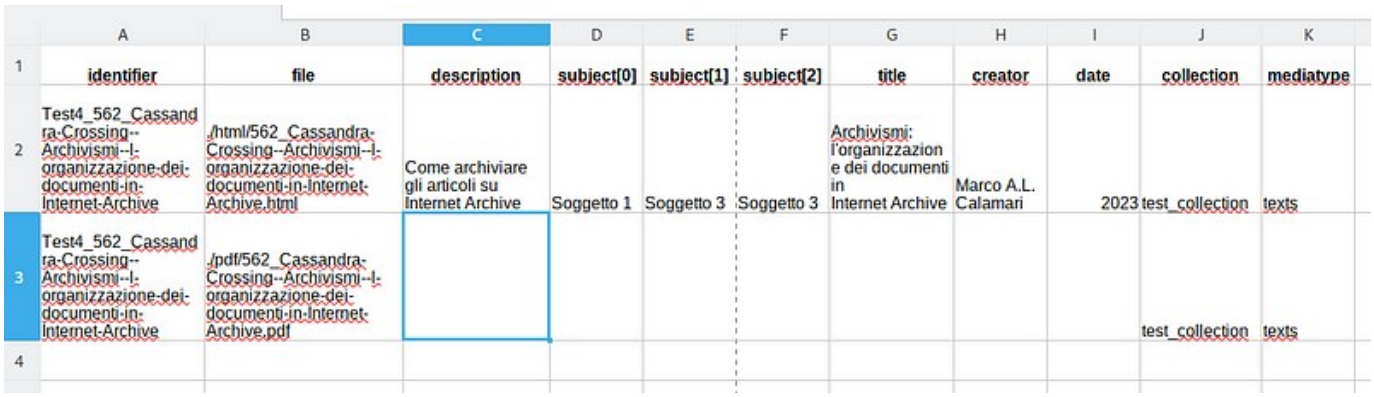
Il comando per fare ciò è semplicemente

```
ia metadata—spreadsheet=metadata.csv
```

Il lavoro vero sarà riempire il foglio elettronico finale con migliaia di righe di dati, ma facciamo un passo alla volta e carichiamo un solo oggetto, per cui un file di tre righe basterà.

Il nostro primo documento conterrà due file tra quelli generati per l'archiviazione, il *pdf* come documento principale e l'*html entrocontenuto* come secondo file; aggiungeremo anche un *minimo sindacale* di metadati, e l'identificativo verrà scelto uguale al nome dei file, tolta l'estensione.

Insomma, dopo molti, molti tentativi ecco il foglio …



| | A | B | C | D | E | F | G | H | I | J | K |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | identifier | file | description | subject[0] | subject[1] | subject[2] | title | creator | date | collection | mediatype |
| 2 | Test4_562_Cassandra-Crossing--Archivismi--l-organizzazione-dei-documenti-in-Internet-Archive | ./html/562_Cassandra-Crossing--Archivismi--l-organizzazione-dei-documenti-in-Internet-Archive.html | Come archiviare gli articoli su Internet Archive | Soggetto 1 | Soggetto 3 | Soggetto 3 | Archivismi: l'organizzazione dei documenti in Internet Archive | Marco A.L. Calamari | 2023 | test_collection | texts |
| 3 | Test4_562_Cassandra-Crossing--Archivismi--l-organizzazione-dei-documenti-in-Internet-Archive | ./pdf/562_Cassandra-Crossing--Archivismi--l-organizzazione-dei-documenti-in-Internet-Archive.pdf | | | | | | | | test_collection | texts |
| 4 | | | | | | | | | | | |

Sembra facile, ma c'è voluta mezza giornata di lavoro, per avere il primo inserimento soddisfacente. Minuzie apparentemente insignificanti ma in realtà diaboliche hanno richiesto un sacco di tempo per prove e controprove. Ve ne racconto qualcuna qui, sperando così di farvi risparmiare tempo prezioso.

**uno**—quando salvate un foglio elettronico in formato CSV, che vuol dire "*valori separati da virgole*" non fidatevi della vostra applicazione. In certi casi, qui in Italia, l'applicazione potrebbe decidere di usare non la virgola ma il punto e virgola, e voi non ve ne accorgerete subito. Giuro, è successo!

**due**—disabilitate, nell'applicazione con cui state gestendo il foglio elettronico, tutti gli strumenti di autocorrezione; altrimenti il programma deciderà certamente di sostituire qualcosa per *il vostro bene*. Nel mio caso ha deciso di sostituire due segni meno consecutivi, presenti nei nomi di file, con un "*trattino lungo*", una modifica praticamente invisibile, anche da linea comandi. Questo ha portato all'inspiegabile messaggio di errore di *file non trovato*, ed ha rese necessarie alcune dozzine di prove, con relativi arrampicamenti sugli specchi. Non riferisco qui le parole che sono state pronunciate quando il problema è stato finalmente localizzato!

**tre**—state molto attenti quando inserite i valori nei campi. Un singolo spazio bianco prima o dopo il valore può non farlo interpretare, ed avere effetti imprevisti. Uno spazio all'inizio di " *test_collection*" ha ad esempio impedito l'assegnazione corretta dell'oggetto alla *collection di test*, destinata, come già sapete, ad abilitare la cancellazione automatica dopo 30 giorni. In più considerate che non è possibile assegnare esplicitamente l'oggetto a

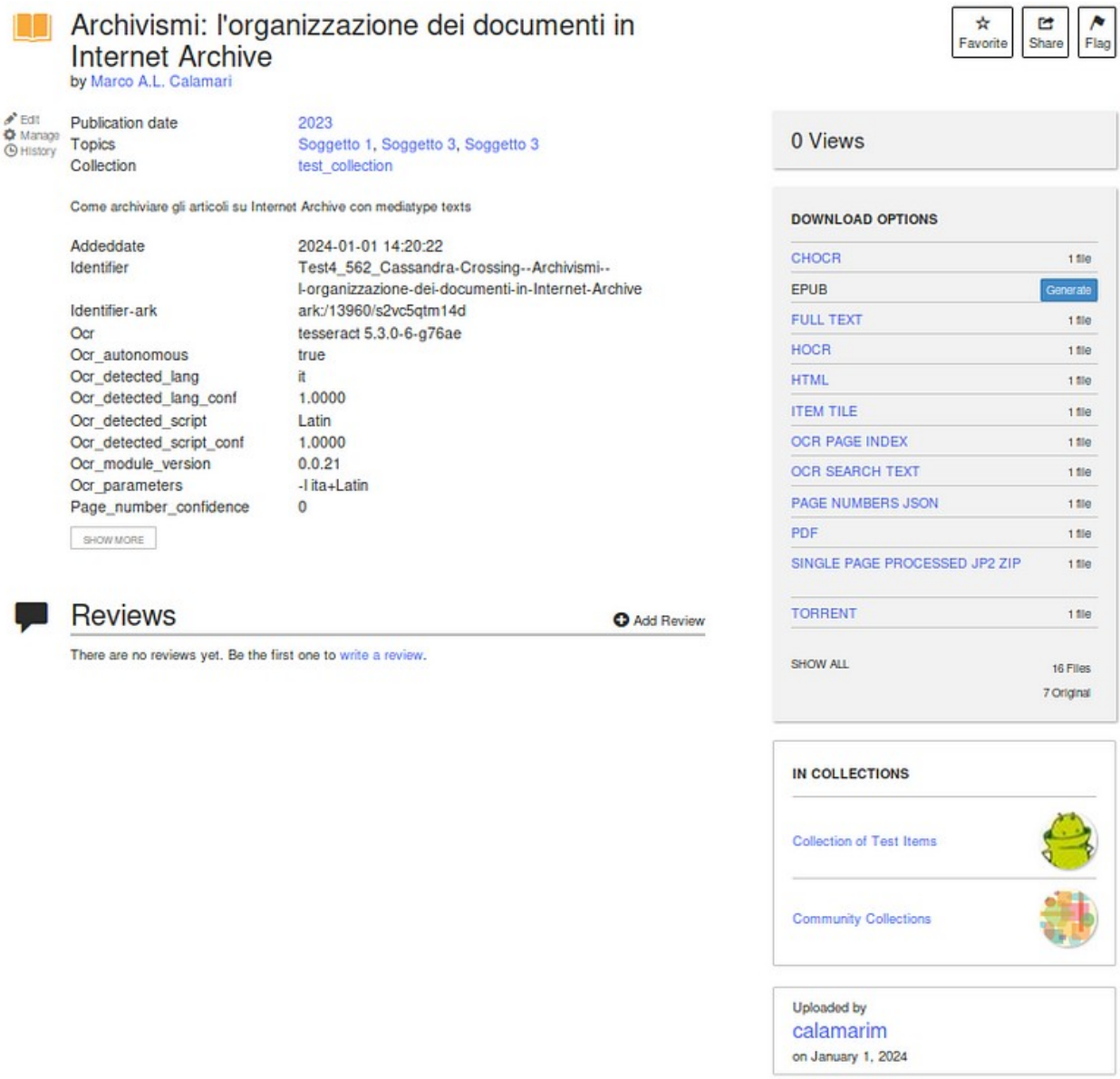
collezioni pubbliche come “*opendata*”, ma bisogna accettare la selezione automatica che verrà operata dal sistema.

**quattro**—inserite nel foglio la colonna *mediatype*, quando i documenti sono testuali (txt, html, pdf, etc.). Usate il valore, “*texts*” altrimenti il sistema assegnerà automaticamente il valore “*data*” e questo avrà effetti collaterali insidiosi. Ad esempio il *browser di oggetti* non vi farà sfogliare le pagine, malgrado tutti i file derivati necessari siano stati creati correttamente. Il *mediatype*, contrariamente alla grande maggioranza dei parametri, non può più essere modificato, ma è necessario cancellare e rigenerare l’oggetto.

**cinque**—cancellare un oggetto non è un’operazione istantanea, ma richiede minuti o decine di minuti prima che l’effetto si propaghi in tutte la parti dell’interfaccia del sito. Non merita cancellare da linea comandi con *ia*; è decisamente più pratico farlo dalla pagina *My Upload*. Ricaricate spesso la pagina, e se notate cose strane, provate anche a svuotare la cache del browser.

**sei**—la comparsa di un oggetto appena creato nella finestra *My Upload* è, stranamente, abbastanza veloce, ma scatena tutte le operazioni “*derivative*”, che a loro volta generano gli altri file in tempi variabili ma abbastanza lunghi. Questo vuol dire, ad esempio, che il *browser di oggetti* non sarà in grado di farvi sfogliare le pagine prima di una mezz’ora, e che la funzionalità di ricerca interna al *browser di oggetti* sarà attiva solo dopo parecchie ore.

**Però, alla fine, che soddisfazione …**




Ed anche per oggi è tutto. *Stay tuned* per la prossima puntata di “*Archivismi*”.

# Archivismi: archiviamo Cassandra, parte terza

*(566) — E' tempo di concludere; parte il mass uploading di Cassandra Crossing!*

**5 gennaio 2024** — Nelle precedenti puntate di *Archivismi* abbiamo raccontato come funziona, a grandi linee, una archiviazione "*vera*" su Internet Archive. "*Vera*" perché non si tratta di caricare una directory di file, ma di creare veri oggetti archivistici, corredati di tutti i file ed i metadati necessari per definire l'oggetto, e renderlo utile e fruibile. Ed i metadati, credeteci o no, sono di gran lunga la cosa più difficile e più utile.

Quindi, innanzitutto, per archiviare la nostra rubrica preferita, è stato necessario chiedersi *cosa* archiviare, oltre al classico PDF. La scelta è stata quella di aggiungere un file HTML entrocontenuto ed un file in formato MARKDOWN, quest'ultimo utile per ulteriori elaborazioni che fossero necessarie. Alcuni articoli parlavano inoltre di libri o pubblicazioni libere, ed in questi pochi casi anche il pdf della pubblicazione è stato inserito nell'oggetto.

Bene, detto questo, è stato necessario crearli, questi benedetti 1686 file. I file markdown, html e e pdf sono stati generati in completo automatismo a partire dai file html degli articoli esportati da Medium.com, grazie agli strumenti preparati nelle puntate precedenti che erano pronti all'uso, elaborando i dati di input esportati da Medium.com. Tutto semplice, quindi?

**Ovviamente no**. In questi appunti di viaggio, la vostra profetessa preferita vi racconterà le ulteriori peripezie incontrate nel suo viaggio.

**Uno**: i dati da Medium.com contenevano ancora degli errori. La tipologia più comune e più dolorosa era l'errata costruzione del nome del file, creato rilevando automaticamente il numero dell'articolo. Questo per due ragioni principali. La prima è che alcuni articoli erano semplicemente numerati in maniera errata. La seconda è che i file contenevano si il

numero dell'articolo, ma non solo nel testo, anche nella intestazione creata automaticamente da Medium.com. Intestazione che una volta creata non veniva poi più aggiornata; indovinate da dove veniva preso il numero dell'articolo?

**Due**: la creazione del foglio elettronico, avendo i file ben creati e rinominati è stata semplice. Aver conservato ogni run di upload in un nuovo foglio è stato utilissimo per localizzare gli errori e ritornare sui propri passi. Anche conservare il log delle esecuzioni di *ia* è stato utilissimo per estrarre gli errori.

**Tre**: *aggiustando* la numerazione degli articoli in alcuni casi si è persa la corrispondenza tra nome dei file ed identificativo dell'oggetto. Infatti, mentre i file ed i metadati si possono modificare, aggiungere e cancellare, non è possibile modificare l'identificativo del l'oggetto, una volta creato. E quando si lancia nuovamente la procedura di generazione file, se cambia la numerazione cambiano anche alcuni nomi di file. Per generare i successivi fogli per il caricamento è stato necessario tenere conto di questo, e operare esaustive verifiche di *allineamento* tra identificatori e nomi dei file. Certo, la tentazione di correggere tutto e rilanciare daccapo le procedure era forte. Ma l'automazione totale non è il fine, ma solo un mezzo. Risparmiare tempo, **facendo comunque le cose per bene,** è il vero fine.

**Quattro**: il primo bulk upload del solo file PDF è stato fatto per 10 oggetti. Si è poi atteso che le varie alchimie automatiche di Internet Archive si compissero, e si è esaminato attentamente il risultato. A livello di metadati questo ha portato a modificare le scelte per renderli più utili.

**Cinque**: Si è poi fatto il bulk upload dei rimanenti 552 pdf, creando così tutti gli oggetti. Gli oggetti, ed in particolare gli identificatori, in tutte le successive operazioni che abbiamo fatto non sono mai variati. Durante questo primo vero bulk upload si sono generati messaggi di errore di *mancata creazione,* perché l'operazione in corso era stata identificata come spam, come questo

*error uploading 186_Cassandra-Crossing — L-Internet-senza-Rete.pdf: Please reduce your request rate. — Your upload of 186_Cassandra-Crossing — L-Internet-senza-Rete from username pippo@pluto.paperino appears to be spam. If you believe this is a mistake, contact info@archive.org and include this entire message in your email.*

Detto fatto, ho contattato via email l'help desk che, forse perché sono un utente di vecchia data nonché *donatore* regolare, in poche ore mi ha tolto qualche evidente limitazione antispam. I successivi inserimenti non hanno più dato nessun problema.

**Sei**: Sono stati eseguiti due ulteriori bulk upload separati, uno per i file markdown ed uno per gli html. Sono state necessarie solo due colonne nei fogli elettronici; identificatore e file. I metadati sono stati assegnati al momento della creazione dell'oggetto, quindi del primo bulk upload. Se dovessero essere cambiati in massa, sarà necessario effettuare "*bulk correction*".

**Sette**: si sono appunto editati i metadati in bulk, inserendo la descrizione (presa dal sottotitolo) e la data di pubblicazione. Ambedue queste colonne di dati sono state generate con una versione modificata della procedura già vista, partendo dai file markdown, estraendo il campo con una regular expression, aggiungendo, ripulendo e correggendo i campi mancanti od errati a mano, e poi copiando i range giusti nel foglio elettronico per il

bulk upload. Malgrado le “*standardizzazioni*” delle precedenti fasi di redazione e manipolazione dei file degli articoli, per sistemare le discrepanze c’è voluta più di mezza giornata.

**Otto**: E qualche altra ora c’è voluta per esaminare sul dito di Internet Archive l’elenco degli articoli ordinati per data e vedere che dentro ci fosse quello che ci deve essere. Anche qui qualche piccolo errore è emerso, ma solo di data. Solo in un caso i titoli e le date erano ambedue invertiti, ma per fortuna anche questi sono metadati, quindi facilmente correggibili. Ma è stata anche una soddisfazione ripercorrere venti anni di lavoro in poche ore!

Ed anche per oggi è tutto, perché il lavoro di revisione è davvero stancante. Le conclusioni ed i commenti li riserviamo per la prossima e finale puntata di questa prima campagna di “*Archivismi*”.

# Archivismi: Cassandra Crossing è per sempre!

*(567)—Cassandra Crossing è per sempre! Alla fine di questo lungo percorso, la rubrica è al sicuro su Internet Archive e, fino a quando questa degnissima istituzione durerà, i 24 lettori ed i loro figli e nipoti, se riterranno ne valga la pena, avranno il tempo per decidere se leggerla, ed un luogo dove trovarla.*

**6 gennaio 2024**—Nelle precedenti puntate di *Archivismi* abbiamo descritto l'archiviazione della rubrica Cassandra Crossing, dal numero 0 al 566, su I*nternet Archive*. Per non tediare ulteriormente i 24 irriducibili lettori che ci avessero seguito fino a questo punto, riassumiamo due ultimi importanti dettagli.

**Uno**: dopo l'archiviazione iniziale, gli oggetti sono stati "*arricchiti*" inserendo il file "*originale*" degli articoli. Decidere quale dovesse essere il *file originale* non è stato banale ma, dopo attenta riflessione, si è scelto il file html ottenuto da medium.com, opportunamente rinominato. Questo file, che quindi rappresenta *il Verbo*, non verrà mai aggiornato. Deciso questo, tutti gli altri file, in giro per laptop, server, cloud o dischi, *da oggi divengono solo copie degli originali o file di lavoro.* **E non è un dettaglio da poco.**

**Due**: Alcuni altri oggetti, una mezza dozzina in tutto, che recensivano libri o traducevano articoli, sono stati ulteriormente *arricchiti*, inserendovi una copia del libro o dei testi in

lingua originale. Il *browser di oggetti* di *Internet Archive* permette di *sfogliare* anche tutti i pdf così aggiunti. Tutti gli altri file, originali od aggiunti, dovranno invece essere scaricati, come d'uso, dall'elenco file in basso a destra nella finestra dell'oggetto.

Ora è doveroso *tirare le somme* del lavoro concluso.

**In primis**, rispetto alle attese, la fase di apprendimento e quella di bulk upload sono durate meno del previsto, circa tre-quattro giornate piene in tutto. La fase di correzione errori e raffinamento dell'archiviazione è invece durata molto, molto più del previsto, circa tre giornate.

Tuttavia l'esperienza accumulata permette adesso di eseguire aggiornamenti, anche massivi, in poche decine di minuti; anche il caricamento di un nuovo set di 4 articoli ha richiesto meno di un quarto d'ora.

Messo duramente alla prova, *Internet Archive* si è rivelato uno strumento davvero utile ed efficiente. Per questo Cassandra torna per l'ennesima volta a ricordare che Archive.org è un'organizzazione senza fini di lucro, che vive di contribuzioni volontarie. Chi la usa regolarmente, o la trova utile, od è moralmente d'accordo, **dovrebbe considerare doverosa una donazione**. *TANSTAAFL …*

**In secundis**, tutto questo lavoro era veramente utile e necessario? Cassandra da parte sua non ha dubbi ma, per motivare la scelta, le è necessario distinguere il punto di vista di un autore da quello di un normale utente della Rete.

**Per un autore** sono certamente importanti la diffusione e la conservazione del proprio lavoro. Per quanto attiene la diffusione, Cassandra a suo tempo ha compiuto riguardo ai social una scelta molto radicale quanto ben nota; usa alcuni social solo per "pubblicare" i suoi articoli , ma non li usa per discuterli, diffonderli o *spingerli* in altro modo. Se mai vi fosse del valore nei fili di parole messi insieme da Cassandra, sarà questo ad alimentarne la diffusione. Per dare eventualmente il tempo a questo lento processo di poter avvenire, l'archiviazione duratura su *Internet Archive* è certamente una condizione necessaria.

**Per un cittadino della Rete** interagire con la Cultura (si, con la maiuscola) dovrebbe essere una occupazione a tempo pieno. Anche solo da semplice fruitore, contribuire correttamente a preservarla e diffonderla è non solo possibile ma doveroso.

Conoscete qualche opera digitale o digitalizzabile che meriti di essere conservata? Contribuite a farlo, ad esempio raccogliendola, arricchendola di dati ed archiviandola in maniera durevole.

Avete una competenza su qualche cosa specifica, sapete scrivere correttamente in una lingua (l'italiano, ad esempio) e, quando serve, avete un minimo di autodisciplina? Realizzate od ampliate una pagina di Wikipedia. Archiviate le cose migliori che avete scritto con la stessa cura che avete impiegato per realizzarle.

E non fermatevi qui. Esistono altre oasi di conservazione, altre biblioteche elettroniche, altri gruppi di persone che si dedicano alla conservazione della cultura ed alla salute dell'*infosfera*, proprio mentre le *false IA* le stanno inquinando con false informazioni. Supportatene una, c'è tanto bisogno anche di questo

E per questa *campagna di archiviazione* abbiamo finito.

Ma gli *Archivismi*, quelli no, quelli non finiscono mai. Date un'occhiata al sito; Cassandra ha già delle idee…

E d'altra parte, parafrasando *Conan il Barbaro*, si potrebbe aggiungere che "*C'è sempre un'altra storia…*"

# Archivismi: Cassandra attraverso i secoli

*(568)—Cassandra Crossing non si accontenta, vuole arrivare più lontano e vuole sopravvivere non per decenni ma per secoli o millenni. Ce la può fare?*

**10 gennaio 2024**—Nelle precedenti 10 puntate di *Archivismi* abbiamo descritto la prima *campagna di archiviazione*; quella della rubrica Cassandra Crossing su I*nternet Archive*. E' stato un lungo percorso, poiché siamo partiti dallo studio della struttura di Internet Archive, seguito la preparazione dei dati, realizzato qualche decina di righe di script per automatizzare il tutto, eseguito gli upload veri e propri, ed infine la ripulitura dei dati e la correzione degli errori nei metadati.

Oggi invece introdurremo la terza campagna di archiviazione di Cassandra Crossing.

"*Ohibò*—dirà qualcuno dei più informati 24 lettori—*la terza campagna? Ma dove ci hai raccontato la seconda?*"

Giustissimo, la seconda non l'ho raccontata perché è stata troppo facile e veloce. La seconda campagna consisteva nell'archiviazione dei 106 video di *Quattro Chiacchiere con Cassandra* su *Internet Archive*. Cassandra ha deciso di non parlarne perché è appena finita, ed ha richiesto solo 20 minuti di preparazione del foglio elettronico di bulk upload e circa un'ora di caricamento. E' pur vero che avevamo maturato una preziosa esperienza precedente, che i metadati inseriti sono elementari e che i dati di partenza erano già ben strutturati, ma una cosa così semplice e veloce non poteva meritare una pur breve

esternazione di Cassandra. Per cui la butto li, andatevi a vedere il risultato, e passiamo davvero alla terza campagna di archiviazione, che ve lo anticipo, sarà ben più stuzzicante.

Dobbiamo però, come Cassandra vi ha ormai abituato, raccontare un po' di storia. Veramente assai più di un po', visto che non si tratta di partire dall'alba di Internet, nemmeno dall'alba dei computer, ma addirittura dall'alba della scrittura, il che vuol dire riavvolgere il nastro, così all'ingrosso, di 5 millenni abbondanti. E' da quella remota epoca che è giunto fino a noi il primo archivio di informazioni omogenee, scritto in caratteri cuneiformi su circa 4.000 tavolette di argilla. Se consideriamo la tavoletta di argilla come supporto informativo, potremmo dire che le tavolette di Uruk si sono rivelate molto durevoli, facendo impallidire tutti i moderni supporti informatici.

E' pur vero che innumerevoli altre tavolette di argilla non hanno superato, come le loro più famose 4000 colleghe, il lungo viaggio fino a noi, ma comunque l'efficacia del supporto rimane notevole.

I rotoli di pergamena si sono rivelati poco meno durevoli; i più antichi superano infatti di poco i duemila anni, e la durata "media" della pergamena, conservata in condizioni ideali, è stimata intorno ai mille anni.

Alcuni papiri sono giunti a noi dall'antico Egitto e quindi sono durati anche loro per millenni, ma in condizione estremamente particolari (tombe sigillate nel deserto). Nei climi europei ed in condizioni di conservazione ideali hanno invece una durata stimata intorno ai 300 anni. Vale la pena di notare che la scomparsa della pergamena come supporto per le informazioni è dovuto proprio all'avvento del papiro, più economico, più facile da scrivere, più leggibile ma meno durevole.

L'avvento della carta ha ulteriormente peggiorato le cose; se alcuni volumi del passato hanno superato molti secoli, tutta la produzione moderna di carta ha una durata limitata a pochi decenni, con casi estremi come certi tascabili degli anni '90 o la carta di giornale, che bastava lasciare al sole per vederla letteralmente sbriciolarsi. Colpa di addittivi chimici e sbiancanti, usati per migliorarne l'aspetto, e di processi di lavaggio inefficienti.

Possiamo riassumere che c'è stato un progresso continuo tra un supporto e l'altro che ha prodotto costi minori, prestazioni migliori e durate peggiori. D'altra parte sostituire supporti inorganici ed incombustibili con supporti organici e combustibili non poteva che peggiorare la durata delle informazioni ivi registrate.

In campo informatico non c'è una esperienza storica così lunga. Inizia solo dagli anni '50 del secolo scorso, con le schede perforate (e per inciso ne ho un pacchetto in perfetto stato di conservazione in un cassetto, perforate per la tesi nel 1980).

I supporti informatici, magnetici od ottici, hanno avuto performance assai meno brillanti. A parte l'obsolescenza tecnologica intrinseca delle periferiche di lettura/scrittura che diventano introvabili o non funzionanti, che rende illeggibili anche supporti che sarebbero ben conservati, persino i nastri magnetici ed i cd-rom, che vantavano durate di 30 anni, si sono in realtà rivelati molto più cagionevoli del previsto. Una campagna di trasferimento dati eseguita di persona da CD-R di meno di venti anni conservati in condizioni ideali ha portato a quasi il 10% di supporti con problemi più o meno gravi di lettura.

La triste verità è che lo sviluppo dell’informatica moderna ha sempre privilegiato la riduzione del costo unitario dei supporti, la densità delle informazioni ivi registrate, la velocità di accesso alle informazioni stesse, senza porre una equivalente cura alla durata dei supporti stessi.

E questo può essere sufficiente per spiegare come mai la durata dei supporti, a partire dai 20–30 anni degli anni ’60, non sia migliorata ma anzi sia semmai peggiorata. Non stiamo infatti parlando di sistemi dotati di ridondanza ed algoritmi di correzione; questi sistemi devono essere dinamici, consumano energia e sono soggetti comunque a problemi di sicurezza informatica e di scarsa resilienza alle catastrofi.

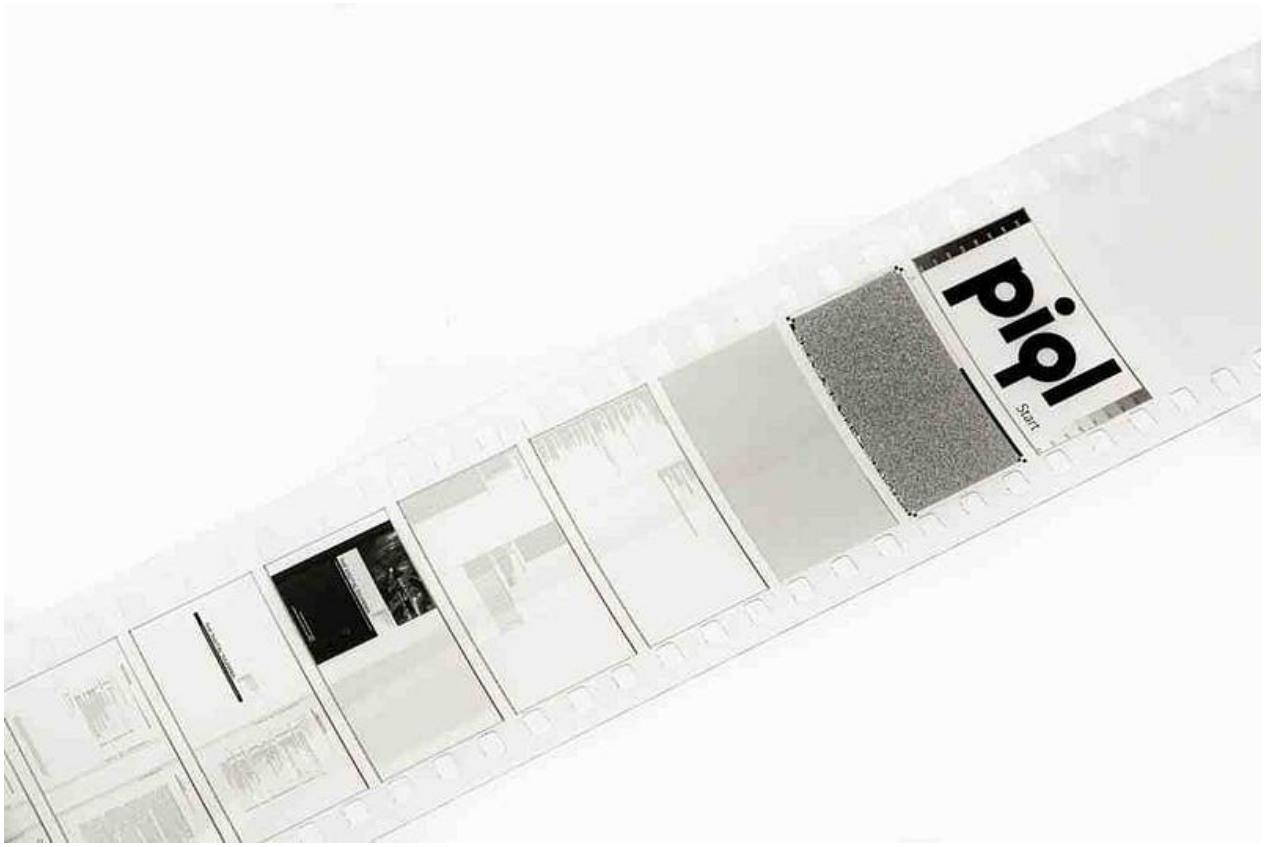
Quello che serve sono supporti che conservino in maniera affidabile le informazioni per la loro stabilita e durata intrinseche, ed in maniera completamente passiva, senza consumare energia, né direttamente, come una stringa di dischi in RAID che deve essere alimentata e funzionante per essere stabile, né indirettamente, a causa di processi produttivi costosi e/ la necessità di impianti attivi di conservazione, come condizionamento/riscaldamento per la stabilizzazione della temperatura.

E servono anche supporti in cui la rappresentazione dei dati non sia così “lontana” dalla percezione degli utenti. La maggior parte delle unità di lettura/scrittura di dati digitali producono supporti sui quali i dati sono impercettibili con mezzi normali, e possono essere rivelati solo con un particolare tipo di unità hardware.

Ambedue queste caratteristiche sono presenti nella soluzione che attualmente, garantisce i tempi di conservazione più lunghi tra i prodotti disponibili sul mercato. E, curiosamente, ma forse non per caso, si tratta di una tecnologia abbastanza vecchia, a cui sono state apportati alcuni miglioramenti. Parliamo delle pellicole fotografiche “normali”, cioè all’alogenuro di argento, ed in particolare di quella utilizzata dalla Piql, una azienda norvegese, insieme al macchinario per registrarvi informazioni digitali.

Il formato della pellicola, che è un prodotto commerciale, è il normale 35 millimetri, il supporto usato è un tipo di poliestere, e la gelatina e l’emulsione hanno ovviamente caratteristiche particolari. La durata di questa pellicola, opportunamente impressionata e sviluppata, è stimato poter arrivare a 500 anni, conservata a temperatura ambiente ed in condizioni ottimali.

La scrittura dei dati sulla pellicola, che alla fine è comunque una normale pellicola fotografica, può avvenire in vari modi, sia visuali che codificati.




Dati analogici come immagini e microfilm possono essere inserite normalmente. I dati digitali vengono invece codificati in fotogrammi simili a dei QR code che contengono ciascuno un blocco di dati.

Il fatto che la codifica sia “visiva” rende possibile eseguire la decodifica, noto il metodo di codifica, anche senza le apparecchiature originali, usando un oggetto che esegua scansioni ad alta risoluzione ed un computer, dotato di un opportuno software, che riassembli le scansioni nei file digitali originali.

Alla fine circa un chilometro di pellicola viene inserito in un contenitore appositamente progettato per una lunga conservazione,

Il periodo di conservazione viene ulteriormente esteso diminuendo la temperatura di conservazione …

… ma per oggi siamo già andati un po' *lunghi*, e quindi qui ci aggiorniamo alla prossima puntata di Archivismi.

# Archivismi: Cassandra e la miniera

*(569)—Archiviare per dei secoli richiede tecnologie poco comuni però tutto sommato semplici. Ma dove, esattamente, può essere realizzato un tale archivio?*

**12 gennaio 2024**—Nelle 10 puntate della prima campagna di *Archivismi* abbiamo raccontato l'archiviazione di 566 numeri di *Cassandra Crossing* su Internet Archive, che tra l'altro ieri l'ha anche promossa a Collezione; la seconda campagna, quella di archiviazione dei 106 video di *Quattro Chiacchiere con Cassandra* è stata invece appena accennata nella precedente puntata perché troppo semplice e veloce. Siamo stati davvero bravi!

Abbiamo poi *raccontato* la tecnologia di registrazione digitale più durevole oggi sul mercato, accennando anche al fatto che la durata certificata a temperatura ambiente può essere ulteriormente estesa abbassando la temperatura di conservazione.

Come si possono conservare delle bobine di pellicola fotografica, ben protette dentro contenitori appositamente progettati, e poi sigillate in buste di materiale protettivo, a temperature molto al di sotto della nostra temperatura ambiente di circa 20 gradi?

Spoiler: la soluzione non è quella di dotarsi di grossi frigoriferi, ma di trovare un'adatta "*temperatura ambiente*".

Per fortuna, non c'è bisogno di essere pionieri; basta seguire quello che hanno fatto i pionieri di un diverso tipo di *archiviazione*, di cui molti non hanno mai sentito parlare.

E ancora una volta Cassandra deve chiedere pazienza ai 24 irriducibili lettori, perché è di nuovo necessario riavvolgere il nastro (qui potremmo dire la pellicola), anche se solo di una quarantina di anni. E non di archiviazione di dati dovremo parlare, ma di *archiviazione* di semi; sì, semi e campioni genetici.

Nel 1984, la Nordic Gene Bank creò un impianto di sicurezza per lo stoccaggio di semi in una miniera di carbone dismessa nelle isole Svalbard. Il permafrost (il terreno permanentemente gelato), le infrastrutture disponibili e la cooperazione con la compagnia carboniera Store Norske Spitsbergen Kullkompani permisero la creazione di una struttura che avrebbe conservato una raccolta di semi in un contenitore d'acciaio all'interno della miniera di carbone n. 3 a Longyearbyen, miniera che si inoltra per 300 metri nel permafrost della montagna.

Nel 2001 fu stipulato il *Trattato internazionale sulle risorse fitogenetiche per l'alimentazione e l'agricoltura* (ITPGRFA), che prevedeva l'istituzione di un sistema mondiale comprendente regole per l'accesso e la condivisione generalizzata dei benefici di tali risorse.

Tuttavia uno studio nel 2004 rivelò che il permafrost—che mantiene una temperatura costante di circa -3,5°C—non era ottimale per la conservazione del patrimonio genetico; inoltre lo stoccaggio dei semi in una miniera di carbone esposta a un livello elevato di gas idrocarburi non era geneticamente sicuro.

Il governo norvegese valutò allora la creazione di una struttura più adatta e, nell'ottobre 2004, si impegnò a finanziare e realizzare lo *Svalbard Global Seed Vault*, realizzando una costruzione scavata nel permafrost privo di carbone, dotata di un impianto di raffreddamento attivo per abbassare ulteriormente la temperatura fino a -18 °C, cioè alle condizioni standard per le banche genetiche.

Il *Global Seed Vault* in questa nuova struttura è stato inaugurato il 26 febbraio 2008; ancora oggi tuttavia molti pensano che esso si trovi invece nella miniera di carbone abbandonata, e non in una struttura nuova, scavata appositamente. Questo tour virtuale vi permette di visitare la nuova struttura.

*Ma allora se alle Svalbard ci sono solo semi*—diranno i 24 infastiditi lettori—*dove sono i dati?*

Risposta facile. Ricordate che il primo deposito di semi realizzato nel 1980 si trovava nella miniera di carbone n. 3 a Longyearbyen? Bene, con la creazione della nuova struttura la miniera è tornata sfitta, ed una piccola azienda norvegese, creata apposta dalla già nominata Piql, ha pensato bene di rilevarla e di creare il primo deposito di dati nell'Artico, l'**Arctic World Archive**. Uno yuppie direbbe *Tecnologia + logistica = servizio innovativo*.

Certo, il look avveniristico da bunker del Global Seed Vault qui non c'è; il look è più simile a quello della miniera di *Indiana Jones ed il Tempio maledetto*, con in più un tocco di Cronache del Dopobomba.

Ma laggiù, in fondo ad una galleria resa praticabile da puntelli e reti metalliche antinfortunistiche, occhieggia un container di acciaio inossidabile …

… pieno di contenitori avvolti in quella che sembra stagnola, ma che in realtà sono buste sigillate. Nella maggior parte di queste buste è custodita la prima campagna di archiviazione di Github.

Il Github Archive Program nel 2000 ha archiviato in 186 contenitori di pellicola una copia di tutti i progetti attivi (incluso quello del sito di e-privacy!) e li ha immagazzinati nella miniera n.3, battezzando l'iniziativa Arctic Code Vault; successivamente c'è stata una ulteriore campagna di archiviazione, ed una successiva è prevista in data non ancora fissata.

*Ma Cassandra dove è finita*—interloquisce nervosamente il più indisciplinato del 24 lettori—*è tutto interessante, ma veniamo al punto!*

Beh, il punto … sarà nella prossima puntata di Archivismi.

# Archivismi: Cassandra tra i ghiacci

*(570)—Abbiamo visto che l'archiviazione a prova di secoli tra i ghiacci esiste davvero. Ma come può fare Cassandra per "congelare" le sue esternazioni?*

**12 gennaio 2024**—Nelle 10 puntate della prima campagna di *Archivismi* abbiamo raccontato dell'archiviazione di 566 numeri di *Cassandra Crossing* su Internet Archive. Nelle successive due puntate abbiamo raccontato storia e tecnologie che rendono possibile l'archiviazione a lungo termine nell'Artico, con periodi di conservazione stimati tra i 500 ed i 1000 anni.

Resta da trattare il punto più importante; come fare per archiviare laggiù.

La buona notizia è che è semplice e relativamente economico, quella cattiva è che bisogna capire ed adeguarsi ad un processo tanto lento quanto "alieno", e quindi apparentemente *innaturale* nella sua lentezza se non lo si conosce nei dettagli.

Si trova spiegato, in maniera un po' dispersiva, sul sito dell'Arctic World Archive. Ne ricapitoliamo qui le fasi principali, per aver chiaro il processo. Per archiviare ledelle informazioni è necessario:

- Aprire un account sul portale AWA, che è gratuito per i primi 45 giorni, poi 9 Euro/mese.
- Creare un *film virtuale*, caricarvi i file e folder da archiviare; se necessario caricare anche i metadati, sia standard che personalizzati in caso di esigenze particolari. I dati resteranno sempre disponibili sul cloud del portale AWA, per tutto il tempo in cui l'account sarà attivo.
- Finalizzare il film, pagando l'importo con carta di credito (139 Euro per un film da 1 GB).

- Attendere il successivo turno di deposito dei film nell’archivio e la relativa cerimonia, che per ovvi motivi climatici e di distanza avvengono di rado, tipicamente una-due volte l’anno (ma tanto non abbiamo fretta perché lavoriamo per i secoli a venire). Alla cerimonia si può assistere da remoto o, se avete il tempo ed i soldi per un viaggio complicato ma affascinante, anche di persona. E se avete ancora più soldi potete far organizzare un deposito ed una cerimonia quando volete, a vostro esclusivo uso e consumo.

A questo punto, se le vostre esigenze di archiviazione sono terminate, l’account può addirittura essere chiuso. Si perde in tal caso la possibilità di accedere ai dati nel cloud, e non si fornisce più un piccolo sostegno economico all’Archivio.

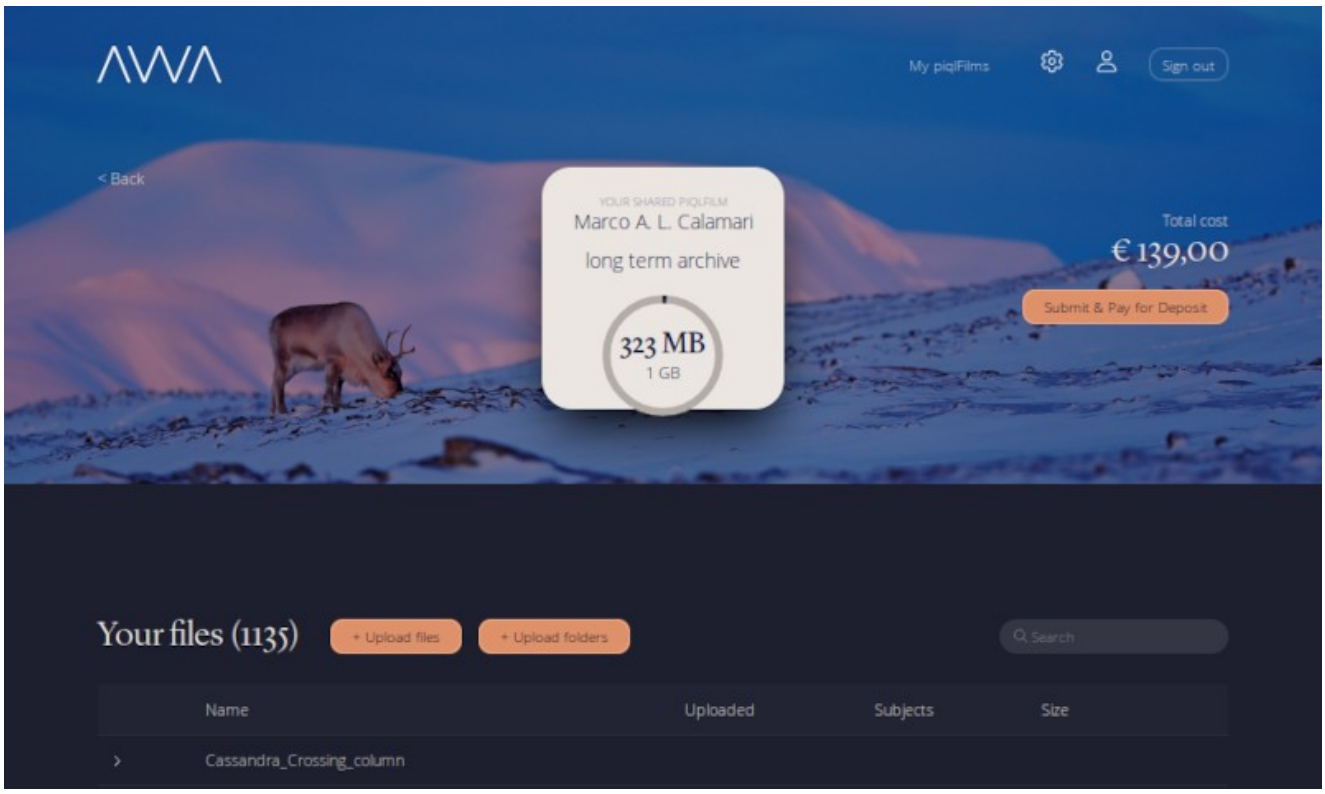
C’è da tener presente che, data la natura del progetto, salvo diversi accordi i dati archiviati diventano, in prospettiva, pubblici. **La memoria del pianeta**. Se questo non fosse quello che vi serve, possono comunque essere presi accordi ad-hoc. Comunque, ricordate che in questi casi la crittografia è sempre la vostra amica migliore.

Nel caso che invece abbiate esigenze molto maggiori o particolari, esplorate gli account di fascia superiore, ed eventualmente contattate l’azienda via email.

E cosa ha fatto, per adesso, Cassandra? Ha incaricato il suo alter ego nel mondo materiale di eseguire queste attività.

Così il malcapitato ha dovuto:

- Creare un account minimale (45 giorni gratis, poi 9 Euro/mese)
- Creare il film più piccolo possibile, di 1 GB, con durata “eterna”. I film virtuali “piccoli” vengono scritti tutti insieme su un singolo film fisico, che cuba 120GB. Potete comprarne anche uno intero tutto vostro, nel qual caso potete anche gestirlo fisicamente. Oltre a trovare cose utili con cui riempirlo, ricordatevi però che dovrete scucire circa 9000 Euro.




- Inserire gli articoli di Cassandra Crossing, prelevati direttamente da Internet Archive. Tuttavia ho il forte sospetto che abbia inserito qualche file personale, anzi “romantico”, insieme a quelli della rubrica.
- Informarsi sulla data del prossimo deposito, non ancora fissata ma prevista a giugno.
- Inserire i dati della carta di credito e mettere il dito sul tasto “invio”.

Poi fermarsi, perché ci sono ancora due mesi di tempo per utilizzare la prossima data di deposito, ed io e lui dobbiamo ancora decidere come finire di riempire il film, che è ancora per due terzi vuoto.

Avete qualche suggerimento? Qualche cosa da inserire nello spazio libero del film di Cassandra? Fatecelo sapere, scrivendo a Cassandra od a Marco.

Cassandra ringrazia chi ha avuto la pazienza di seguirla fino qui e annuncia la sospensione del racconto della terza campagna di Archivismi.

Ma gli Archivismi invece continuano; non devono mai fermarsi!

# Appunti di Archivismi: creare oggetti e correggere errori

*(614) — Anche archiviare conoscenza per il futuro può dare problemi; gli Appunti di Archivismi vi possono far risparmiare tempo e karma.*

**30 marzo 2025** — Non sono solo rose e fiori quando si lavora su Archive.org. E' un ambiente software particolare, complesso, e gestito da poche persone, quasi tutti volontari.

Per questo motivo, per utilizzarlo aldilà della banale memorizzazione di un singolo file od una singola pagina, è bene avere chiari i fondamentali.

Per chi vuole sapere tutto e subito, consiglio una visita alla pagina della documentazione per gli "sviluppatori", che è molto ben strutturata; non ha una collocazione indipendente, ma si trova nel Blog di Internet Archive.

Nel blog postano coloro che lavorano per Internet Archive, e vi si trovano piccoli gioielli di utilità e di sintesi, tipo questo post dove è descritta la struttura degli oggetti memorizzati su Internet Archive.

Ma per sintetizzare al massimo, Cassandra vi regalerà il distillato dai suoi errori e dai rimedi che ha trovato.

Prima di tutto una raccomandazione; **registratevi subito come utenti.**

E' ovviamente utile per avere più facilmente accesso agli amministratori, per chiedere più banda per gli inserimenti, etc.

Ma sopratutto, se creaste un oggetto in forma anonima, cioè senza esservi collegato come utente registrato, non lo potreste più modificare, visto che per fare questo occorre essere l'utente creatore dello stesso.

E veniamo ai fondamentali. Internet Archive può essere visto come un enorme singolo database. Gli oggetti che contiene possiedono dei metadati, e tra questi tre sono particolarmente importanti.

**L' identificatore** — tutti gli oggetti di Internet Archive sono simili, ovvero sono dello stesso tipo. Ciascun oggetto possiede un'identificatore (un nome univoco) ed almeno un file. Una volta creato, l'oggetto non può essere cancellato, né può esserne modificato l'identificatore.

Se durante la creazione non viene specificato l'identificatore, esso è generato a partire dal nome del primo file caricato. Dopo la creazione, a qualsiasi oggetto possono essere aggiunti altri file od altri metadati.

Non è quindi possibile "recuperare" un identificatore che è stato utilizzato per creare un oggetto, cancellandolo come si farebbe con un semplice file. Infatti, se necessario, un oggetto può essere reso invisibile, ma non può essere mai cancellato.

Morale: un identificatore utilizzato "male" è perso per sempre. Sappiatelo!

**Il mediatype** — Il mediatype di un oggetto creato dall'interfaccia web viene normalmente dedotto dal tipo del primo file caricato; se l'oggetto invece viene creato da linea comandi usando l'interfaccia IA (che sta per Internet Archive, non altro), è possibile definire il mediatype a piacere, scegliendolo tra i vari tipi previsti.

Se caricate un file che non viene riconosciuto come tipo, al relativo mediatype verrà assegnato il valore più generico possibile, “*data*”, ed il file non potrà essere visualizzato nei visualizzatori standard di Internet Archive.

Il *mediatype*, una volta definito, non può essere più modificato dal proprietario. Se, eccezionalmente, ci sono validi motivi per farlo (tipo, ne avete creati 1000, tutti sbagliati — a Cassandra è capitato!), si può scrivere una email all'amministratore per chiedergli di cambiarlo.

Si deve ovviamente allegare una lista degli identificatori degli oggetti ai quali cambiare il mediatype, oppure una espressione di ricerca che in tutta Internet Archive selezioni solo quegli oggetti.

Se avete scritto con sufficiente gentilezza e precisione, e magari siete già noti per aver versato in precedenza qualche obolo, in un tempo variabile da qualche ora a qualche settimana il vostro desiderio verrà esaudito. Gli archivisti non hanno mai fretta, visto che lavorano per l’eternità.

**La collezione (collection)** — Non è completamente vero che in Internet Archive tutti gli oggetti sono dello stesso tipo; oltre agli oggetti normali, ne esiste un secondo tipo, detto “C*ollezione*” (collection).

Un oggetto Collezione è semplicemente un “contenitore” di altri oggetti, che non può essere creato dagli utenti normali. Viene usato, ad esempio, per riunire assieme tutti i numeri di una rivista, che siano già stati archiviati come oggetti separati.

Una nuova *collection* deve obbligatoriamente essere richiesta agli amministratori, sempre via email; la richiesta deve provenire da un utente registrato, che abbia già creato un buon numero di documenti. Se i documenti sono già stati creati prima della richiesta di creazione della collection, è necessario richiedere anche l’assegnazione degli oggetti preesistenti alla nuova collection. Altrimenti dovrete farlo voi, assegnando agli oggetti la nuova collezione come collezione aggiuntiva.

Al momento della creazione, infatti, ad ogni oggetto viene assegnata una collezione primaria; se non ne viene assegnata una esplicitamente, in base al suo mediatype

l’oggetto viene inserito nelle collection “community texts” se è un documento, “community videos” se è un video, e così via.

Una volta assegnata, la collezione primaria non può più essere modificata dall'utente. L’utente può eccezionalmente chiederne, come al solito e con i soliti mezzi, una modifica all'amministratore. Tuttavia ce ne è raramente bisogno, perché l’utente può assegnare collection aggiuntive agli oggetti che ha creato. Un oggetto può infatti appartenere a più collezioni.

Qui, ad esempio, trovate tutte le collection di Cassandra, mentre qui trovate tutti gli oggetti da lei creati.

**Infine Cassandra vi rinnova una raccomandazione**. Non inquinate Internet Archive con le vostre prove. All’atto della creazione è possibile attivare un metadato che crea un oggetto *effimero*, che viene poi cancellato automaticamente e completamente dopo 30 giorni.

**Avete il dovere di fare i bravi**, quindi usate questo parametro per tutte le archiviazioni di prova, in modo che solo quelle definitive vengano memorizzate permanentemente. Potete ora consultare di nuovo l’intero processo di creazione di un oggetto rileggendo questo articolo, oppure ripassandovi l’intera serie.

A Cassandra non resta che augurarvi “buone archiviazioni”!

# *Appunti di Archivismi/ Rinfrescare, riordinare, rivedere, rimpolpare*

*(639)—Quando si crea un'archiviazione di informazioni se ne deve curare anche aggiornamento, manutenzione e correzioni. E se si è usato Internet Archive è bene conoscere potenzialità, problemi e trucchi*

**26 Agosto 2025**—L'iniziativa "*Archivismi*" ha lasciato a Cassandra l'onere di occuparsi (con gioia) di tutto quanto archiviato.

E se l'archiviazione nell'*Arctic World Archive* non permette, e quindi neppure richiede, nessuna manutenzione, essendo laggiù possibile solo effettuare una nuova archiviazione, altrettanto non si può dire per le archiviazioni fatte su Internet Archive.

Si tratta infatti di archiviazioni "vive", come in una biblioteca vera, si possono inserire nuovi libri, si può "cancellare" un libro danneggiato, si possono aggiornare i libri con un'edizione successiva qualitativamente migliore, si possono aggiungere allegati ad un libro, si possono ampliare o correggere collezioni di libri.

Come in precedenti articoli di Archivismi, anche questo sarà formato da singoli punti monotematici; se quello che vi è descritto vi dovesse capitare, avrete così una soluzione già pronta in mano. E comunque anche solo leggerli serve anche a capire come muoversi.

**Zero**: Grazie al lavoro fatto, **Cassandra ha scoperto una grossa falla nel processo di estrazione dei dati che Medium.com offre**, cosa che solo per fortuna non ha prodotto serie conseguenze nell'archiviazione di Cassandra Crossing.

Quando si esportano i dati di un account di Medium.com, si ottiene, come in altri social, una serie di directory, ciascuna delle quali contiene informazioni di un certo tipo, il tutto eventualmente con una pagina html di indice.

Questo era il caso di Medium.com; scompattando l'estrazione, una della directory conteneva tutti gli articoli in formato html. Si trattava però di un html pessimo, pieno di tag utili solo al funzionamento del loro sito, che lo rendevano illeggibile e difficilmente elaborabile.

Quindi, prima di produrre i file per l'archiviazione, era necessario "ripulire" l'html, e solo poi generare i file markdown, html e pdf destinati all'archiviazione finale.

Tutto questo era stato eseguito in maniera automatica da un singolo script bash, tramite l'impiego di normali utility unix e di Pandoc, un programma di conversione di formati testuali estremamente potente, che usa LaTeX come formato intermedio. Modificare lo script per altri sistemi operativi dovrebbe essere fattibile, ma lasciamo volentieri l'esercizio al lettore, che nel frattempo potrebbe anche cadere vittima della tentazione di passare ad un sistema operativo libero.

Cassandra vi aveva già fornito e spiegato lo script nelle precedenti puntate di Archivismi, in particolare questa e questa.

L'ultima versione, come quella di allora, non è perfetta ma sempre un work-in-progress; Cassandra ve lo riallega nuovamente, per i piccoli ma importantissimi cambiamenti che ha apportato in questi mesi (e questa volta non c'è bisogno di convertire i caratteri!).

```
# Procedura per la preparazione all'archiviazione degli articoli
# di Cassandra Crossing
#
# inizializzazioni varie
_base="./"
_base2="./posts/"
_base3="./markdown/"
_base4="./temp/"
_base5="./html/"
_base6="./pdf/"
_base7="./post2/"
_base8="./md2/"
_temp="temp.txt"
_temp2="temp2.txt"
_temp3="temp3.txt"
#
# cambio directory di lavoro, creazione directory e pulizia file
cd "${_base}"
mkdir markdown html temp pdf post2 md2
rm ./markdown/* ./html/* ./temp/* ./pdf/* ./post/temp*  ./post2/* ./md2/*
```

```
cd "${_base2}"
rm "${_temp}"  "${_temp2}"
_dfiles="*"
#
# inizio loop principale
for f in $_dfiles
do
rm "${_temp}"
#
# estrazione del numero dell'articolo
g=`grep -Eo -m 1 '\([0-9]+\)' $f | tr -d '()'`
g="000"$g
g=`echo $g | rev | cut -c 1-3 | rev`
h=`echo $f | cut -d '_' -f2- | rev | cut -d '-' -f2-| rev`
#
# formazione del nuovo nome del file e copia col nuovo nome
i=$g"_"$h
echo "---> Identifier: $i"
echo "$i" >> "${_temp3}"
cp $f "../$_base4${i}.html"
cp $f "../$_base7${i}.html"
#
# conversione in formato markdown, ripulitura e riconversione in html


# traduzione del file html da Medium.com in markdown con risorse incorporate
pandoc --embed-resources=true -f html -t markdown  "../"$_base4$i".html" > "$
{_temp}"


# pulizia del markdown da porcherie html provenienti da Medium.com
grep -v "^:::" "${_temp}" |sed -e 's|$i".md"


# traduzione del markdown embedded in html embedded
pandoc -V geometry:margin=2cm --embed-resources --standalone -f markdown -t html
"../"${_base3}$i".md"> "../"${_base5}$i".html"


# traduzione html in pdf, con size, pagina, link e margini, ma sposta le
immagini grandi in fondo ed ha ancora un riassunto in testa
pandoc --metadata title="" -V papersize:a4 -V fontsize=12pt -V colorlinks   -V
geometry:mrgin=2cm --embed-resources --standalone --pdf-engine=xelatex -f html
-t pdf "../"$_base5$i".html"  > "../"${_base6}$i".pdf"
```

```
echo `cat "../"${_base3}$i".md" | tr "\n" " " | sed -e "s/.*\*(//" | sed -e
"s/\*.*//"` >> "${_temp2}"

#

# pulizia e fine ciclo

done

#rm -rf "${_temp}" "../$_base4"
```

All’inizio Cassandra aveva deciso di archiviare il file html originali provenienti da Medium.com come fonte “originale” degli articoli, considerando gli altri file prodotti da questi come file “derivati”. Si era accorta però che il file originali non contenevano le immagini, ma solo dei link che puntavano a server di Medium.com, cosa abbastanza comune.

Aveva quindi fatto in modo di archiviare anche una seconda versione di HTML che, grazie ai miracoli di Pandoc, conteneva le immagini come risorse codificate rot64.

Durante l’ultima campagna di aggiornamento della collezione degli articoli di Cassandra Crossing, la vostra profetessa preferita si è tuttavia accorta **che molti degli articoli originali erano troppo piccoli per la quantità di testo che avrebbero dovuto contenere.**

Ha constatato, non senza una buona dose di raccapriccio e di rabbia verso chi “bara” nei confronti dei propri utenti, che **intere parti di testo, e spesso l’intero articolo, non erano contenute nel file ma venivano anche esse caricate come risorse esterne dai server di Medium.com.**

Insomma, nei file “originali” non solo le immagini ma anche intere parti di testo degli articoli non c’erano, ma erano, per fortuna, state incluse successivamente dal processo di conversione realizzato da Cassandra.

Andreottianamente, Cassandra non crede che queste “caratteristiche” siano state inserite per sbaglio.

**Morale:** Cassandra ha deciso che il file “originale” degli articoli sarà un file markdown con le immagini embedded, e che il file html di Medium.com continuerà ad essere allegato solo per motivi “storici” come file accessorio.

Questo richiederà una riarchiviazione completa di tutti i file principali di tutti articoli, anche se non richiederà la creazione di nuovi oggetti e nemmeno la modifica dei metadati. Gli automatismi di Internet Archive e lo script di archiviazione dovrebbero permettere di risolvere il problema con un paio di ore di lavoro.

**Uno:** il formato dei pdf per l’archiviazione e la leggibilità dei suoi contenuti sono probabilmente la cosa più importante per l’utente.

La prima archiviazione di Cassandra Crossing su Internet Archive era stata fatta dopo aver superato molti ostacoli, e con una conoscenza molto limitata di Pandoc e LaTeX.

Per questo i pdf prodotti erano ben lungi da essere ottimali, perché l’uso di Pandoc non permetteva di cambiarne le caratteristiche in maniera semplice, come si sarebbe fatto con qualunque elaboratore di testi. I primi pdf avevano i seguenti problemi:

- [margini ridicolmente larghi;]
- [font eccessivamente piccoli;]
- [la prima pagina di ogni articolo aveva una bruttissima intestazione, contenente l'ID Internet Archive del documento.]
- [gli articoli più recenti iniziavano con una doppia copia di titolo e sottotitolo;]
- [i link erano inclusi nel pdf e cliccabili, ma la loro veste grafica era assolutamente indistinguibile dallo scritto normale, quindi per trovarli ed utilizzarli, bisognava passarci sopra il cursore per vedere quando cambiava da freccia a manina;]

Risolvere tutti questi apparentemente banali problemi ed ottimizzare alcuni altri aspetti, ha richiesto un minimo di studio di Pandoc e molte, ma davvero molte prove; le modifiche si sono alla fine tradotte nell'aggiunta di pochi parametri sulla linea comandi di pandoc.

Tuttavia la risoluzione di questi problemi ha comportato dover lasciare maggiormente la "mano libera" a LaTeX, che ha preso l'autonoma ed inarrestabile decisione di spostare le immagini troppo grandi in posizione fuori testo.

**Cassandra leva perciò il suo grido di aiuto** a chi potesse aiutare a risolvere il problema, partendo per esempio dal confronto dei file html e pdf di un articolo. Grazie anticipatamente!!!

Cosa dire per tirare un po' le somme di questa esternazione? Probabilmente la cosa più importante, almeno per chi deve archiviare molti dati, è la **raccomandazione di stare estremamente attenti ai dati che vi vengono forniti dalle applicazioni cloud, al loro formato, alla loro completezza ed alla presenza di risorse remote**.

Anche Cassandra, notoriamente pignola, ci stava cascando.

**Due:** la gestione delle correzioni necessarie ad alcuni degli oggetti preesistenti, creati in maniera errata (pochissimi, per fortuna), ha richiesto abbastanza tempo e non è ancora stata completata; per questa ragione Cassandra la rimanda al prossimo numero di questa piccola saga.

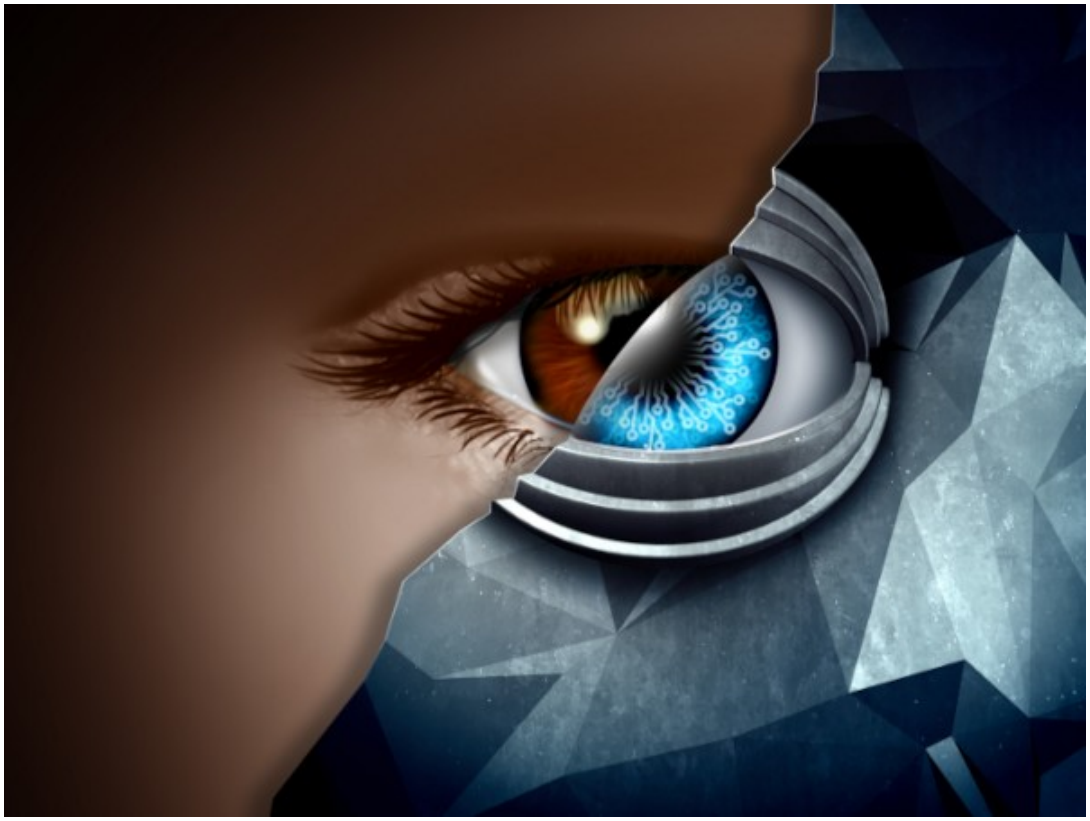
Resta comunque a disposizione dei suoi lettori per consiglli o discussione al suo abituale indirizzo di email cassandra@cassandracrossing.org.

Enjoy!

Scrivere a Cassandra—Twitter—Mastodon
Videorubrica "Quattro chiacchiere con Cassandra"
Lo Slog (Static Blog) di Cassandra
L'archivio di Cassandra: scuola, formazione e pensiero

***Licenza d'utilizzo****: i contenuti di questo articolo, dove non diversamente indicato, sono sotto licenza Creative Commons Attribuzione—Condividi allo stesso modo 4.0 Internazionale (CC BY-SA 4.0),* tutte le informazioni di utilizzo del materiale sono disponibili a questo link.

# *Archivismi: una sessione con Claude*



*(663) - Usare gli LLM non è il male assoluto; se dovete manipolare linguaggio sono tool adatti. L'ultima sessione di Archivismi li ha utilizzati, e questa è la storia.*

**21 aprile 2026** - Mantenere, migliorare ed aggiornare un'archiviazione periodica, anche piccola come la rubrica di "Cassandra Crossing", è un processo senza fine. Man mano che le cose archiviate aumentano, l'entropia della Rete le complica. Escono nuovi articoli, la formattazione inevitabilmente cambia, alcuni editori iniziano a pubblicarli, altri cessano, i link alle pubblicazioni cambiano o spariscono, qualcuno ripubblica, talvolta cambiando la licenza…

Mentre l'archiviazione di un batch di nuovi articoli può essere molto automatizzata, fino a diventare una faccenda da mezz'ora, gestire il cambiamento e la complessità del passato diventa difficile.

Prendiamo un esempio semplice il sito di Cassandra Crossing.

Generato a mano nel 2009 (all'inizio usando Notepad, oggi con Gedit o Bluefish), minimalista, fatto di solo HTML statico con un unico foglio di stile CSS, si è arricchito di tutto quanto ho salvato in più di venti anni di pubblicazioni di Cassandra Crossing. Gli articoli sono sparsi in Rete in vari posti, e con previdenza, fatica ed autodisciplina la struttura di un articolo è stata mantenuta perlopiù costante; titolo, sinossi e testo con link, formattazione semplice stile markdown, rare illustrazioni.

Tuttavia i link e le informazioni di singoli articoli erano talvolta non standard, incomplete o con errori. Anche la memorizzazione degli articoli come file locali sul mio laptop e sul sito ed i relativi indici aveva questi problemi.

Di recente l'aver fatto, nell'ambito della campagna Archivismi, una archiviazione completa su archive.org ha permesso, come sottoprodotto, di rendere più omogeneo e corretto anche il sito.

La settimana scorsa ho deciso una ulteriore "campagna" di miglioramento del sito e di archiviazione degli ultimi articoli, sia sul sito che su Internet Archive (IA); avevo messo giù qualche obbiettivo:

- definire in maniera inequivoca ed esatta la numerazione e la data di pubblicazione degli articoli (sembra impossibile, ma in due decenni si sbagliano anche quelli);
- controllare la validità e l'univocità degli identificativi di IA;
- inserire sul sito tutti i link con cui ogni articolo è stato pubblicato (cosa difficile, lunga e da fare a mano);
- recuperare ed inserire le sinossi (sottotitolo) nell'archiviazione su IA e sul sito;
- creare una nuova pagina contenente le librerie tematiche di articoli, convertendo quelle presenti su Medium.com, e formattandole in maniera "ricca" con sinossi e link di pubblicazione.

Ora, sembrano tutti compiti semplici, ma sono abbastanza sfrangiati da essere di difficile automazione usando lo scripting.

Di questo fatto sulla raccolta degli articoli ad essa dedicati se ne danno ampi esempi; per questo motivo, dovendo accedere a risorse in linguaggio sia formale che naturale, perdipiù sparse nel web, usare un LLM che potesse anche accedere a risorse in Rete era un opzione interessante, visto che non di programmazione pura né di questioni agentiche si trattava, ma di "semplici" manipolazioni di linguaggi.

Per cui, invece di armarsi di infinita pazienza ed olio di gomito, ho per la seconda volta attivato seriamente un account di un LLM, questa volta per par condicio non ChatGPT (che mi aveva molto deluso) ma Claude, ed usando la versione PRO (20 €/mese) per avere un contesto di dimensioni adeguate rispetto al volume di informazioni da trattare.

Da tutta questa attività, che è costata una giornata piena di lavoro, ma che ha anche generato il risultato che volevo, e che potete trovare nelle pagine Liste ed Articoli di cassandracrossing.org, ho ricavato come sottoprodotto anche questa atipica ma spero utile nuova puntata di Archivismi.

Iniziamo da qualche nota tecnica.

Sapere cosa si acquista in termini di token, di dimensione del contesto e di altre risorse, quali l'utilizzo di tool esterni, è praticamente impossibile. Anche un attentissimo esame dei vari tipi di account e delle loro caratteristiche non fornisce nessun dato affidabile. Questo accade per vari motivi, incluso il fatto che i modelli e le risorse che questi utilizzano sono in continua evoluzione, sia tecnica che "commerciale".

Infatti le risorse che si utilizzano lavorando con gli LLM sono fornite, dal punto di vista monetario, gratuitamente od in forte perdita. Anche gli account di livello Enterprise , facendo alcuni conti, fanno si fatturato, ma sono sempre venduti sottocosto, generando non profitti ma perdite.

Questo causa una perenne “vaghezza” delle caratteristiche degli account utilizzabili in un dato momento; ma peggio ancora, le caratteristiche di un account possono cambiare senza preavviso e senza nessuna comunicazione, in pochissimo tempo, addirittura durante una sessione.

In breve, è impossibile sapere cosa si compra.

Si aggiunga a questo il problema di lavorare con un’infrastruttura di cloud computing che deve bilanciare e limitare i carichi in ogni momento. Un problema grosso quando si deve fare un vero lavoro.

Comunque alla fine ho aperto l’account Anthropic Claude PRO, ed ho utilizzato il modello generalista al momento più avanzato, Opus 4.6, senza mai cambiarlo per tutto il lavoro.

Per non finanziare nessun membro dell’”Impero del Male”, ho anche utilizzato il Codice del Consumo con il diritto di recesso entro 14 giorni, la lettera di recesso l’ho fatta scrivere da Claude e l’ho passata direttamente al chatbot di Claude; ha funzionato tutto velocemente e molto bene, almeno da questo punto di vista.

Veniamo alla parte più interessante, gli appunti di questo viaggio.

Sono partito da dati già abbastanza strutturati, cioè da un foglio elettronico contenente un elenco semicompleto degli articoli, con numero, titolo e link di prima pubblicazione; per alcuni articoli avevo anche uno o due link alternativi. L’idea era far estrarre dall’LLM la data di pubblicazione e la sinossi, in modo da poter poi generare un pagina di elenco articoli più ricca e con i link sia ai file locali del sito che alle varie pubblicazioni (storicamente sono 7 testate diverse).

Successivamente, sono partito dalle liste tematiche che avevo generato su Medium.com, piattaforma che sto abbandonando per vari ed in parte ovvi motivi, per generare una nuova pagina del sito, contenente le varie liste, con il relativo titolo e la sinossi della lista, e contenente anche, come sezione espandibile, un elenco degli articoli, con numero, link, titolo e sinossi dell’articolo, tutto ben formattato.

Ma per prima cosa dovevo estrarre le informazioni mancanti, prima di tutto data di pubblicazione e sinossi, e poi anche l’identificativo di IA.

Il problema della data era relativamente semplice, perché fin dalla notte dei tempi avevo quasi sempre incluso il numero dell’articolo tra parentesi tonde all’inizio della sinossi, e la data di pubblicazione, in formato con mese in lettere, all’inizio dell’articolo. Infatti la maggior parte di queste informazioni ero già riuscito a recuperarle con degli script bash dai sorgenti degli articoli, come spiegato nelle prime puntate di “Archivismi”.

Ho quindi aperto una sessione, fornendo tre file: il foglio elettronico con i dati incompleti, e le due pagine html del sito che contenevano gli elenchi degli articoli.

Con il primo prompt ho cominciato a spiegare dove recuperare le informazioni mancanti, e subito mi sono reso conto che praticamente tutti i siti bannavano le richieste che l’LLM faceva.

Per aggirare questo problema ho individuato fonti più precise per le informazioni: ho fornito i file degli articoli in formato html, e la struttura della collection di Cassandra Crossing, che

ho realizzato su archive.org, in formato json, recuperandolo manualmente via API e caricandolo nella sessione dell'LLM.

In questo modo, dopo diversi prompt, sono riuscito ad ottenere una nuova versione del foglio elettronico ben numerato, contenente la maggior parte delle date, delle sinossi e degli identificativi di IA. Ma restavano una trentina (su 650) sinossi mancanti, e malgrado parecchi altri tentativi, costati una strage di token, ne sono rimaste 15 non recuperate.

Pragmaticamente, ho quindi semplicemente completato a mano il foglio elettronico a forza di copia/incolla e l'ho caricato in una nuova sessione, come informazione principale.

Ho poi iniziato a generare la pagina delle liste. Per lo stesso motivo. non potendo far leggere le liste di articoli su Medium.com, le ho prese dal download di un salvataggio dei dati dell'account Medium.com, e li ho caricati come file.

Le liste erano 35, e stranamente per alcune, come successo per gli articoli, non si riusciva a recuperare la sinossi di lista.

Il problema era che l'LLM subiva una caduta fuori contesto delle prime informazioni che recuperava, prima di essere riuscito a trovare le ultime.

Aprendo ancora una nuova sessione, e fornendo oltre al resto anche la pagina html delle liste già parzialmente generata, riuscivo a farmi completare la pagina HTML.

Bellina, con un html abbastanza ordinato, ma notavo che alcune delle sinossi erano molto banali, non sembravano scritte da me.

Un controllo a campione rivelava che l'**LLM, senza comunicarlo in nessun modo, aveva "generato" le sinossi che non riusciva a recuperare "riassumendo" in qualche modo il testo degli articoli!!**

"Interrogando" il fedifrago, si otteneva una parziale confessione del fattaccio, ma nessuna collaborazione.

Per sistemare le cose si è creata una nuova sessione, fornendo il foglio elettronico e la pagina "parzialmente falsificata". Con qualche insistenza, e fornendo nuovamente i file delle liste, sempre prelevati dal salvataggio precedente, ho infine ottenuto il risultato oggi in linea.

Con operazioni simili, e sempre partendo dall'ormai completissimo foglio elettronico e dalla pagina html preesistente, si è ottenuta la pagina degli articoli, che contiene i link ai file ed a tutte le pubblicazioni di ogni articolo.

Essendo un maniaco della precisione, e volendo continuare in futuro a poter modificare a mano l'HTML, ho esaminato i sorgenti HTML delle due pagine, trovando varie "licenze poetiche" realizzate dall'LLM, nonché parecchie inutili complicazioni.

Ho quindi nuovamente usato l'LLM, a parità di aspetto della pagina, per far modificare e riformattare il sorgente della pagina stessa, eliminando la formattazione inline, come gli operatori "font" sparsi per ogni dove, riconducendo le formattazioni al CSS esistente, e formattando il sorgente stesso in maniera leggibile. Anche qui è stato lungo riuscire a scrivere prompt che mantenessero il contesto, differenziando con precisione la formattazione del sorgente da quella della pagina HTML.

Altra strage di token; Claude ha confuso ripetutamente le due cose, scrivendo arrosti fantasiosi, prima di generare i sorgenti come li volevo.

Infine ho inserito i link degli articoli su archive.org, che sono stati "sintetizzati" dall'LLM partendo nuovamente dal file JSON della collection; è stato però necessario "sistemare" a mano 6 di essi, perché per varie peripezie di archiviazione, i loro ID avevano una formattazione diversa dalla stragrande maggioranza degli altri, poiché su IA un ID creato male non si può né cancellare né riutilizzare..

Riassumendo: in totale una giornata piena di lavoro ed una serie di esperienze interessanti su come lavora il modello che è stato utilizzato.

A consuntivo, il modello scriveva ed eseguiva script in bash e python, anche una dozzina per ogni prompt. Gli script in Python facevano in realtà uso di sofisticati moduli standard Python per l'elaborazione di testi e la codifica in HTML. il LLM si limitava a suddividere in passi il lavoro richiesto, e realizzare gli script per ogni passo, che poi eseguivano il lavoro vero. Interessante il fatto che dopo ogni script, di solito ne veniva girato un altro che doveva verificare il lavoro del primo!

La possibilità che la chat di Claude fornisce di seguire i vari passi di "ragionamento", ed all'interno dei passi vedere le azioni pratiche, ed esaminare (anche a posteriori in una chat archiviata) gli script che vengono generati ed utilizzati, è stato molto interessante ed istruttivo.

Un ultima nota. In almeno tre occasioni durante il lavoro mi è stata chiesta l'autorizzazione di poter leggere e scrivere file sul mio computer e lanciare comandi, inveci di caricare e scaricare i file.

Indovinare la risposta corretta è lasciato agli intuitivi 24 lettori come esercizio finale.

Abbiamo finito per oggi; appuntamento alla prossima puntata di "Archivismi".

Scrivere a Cassandra - Mastodon - Buttondown.com
Videorubrica "Quattro chiacchiere con Cassandra"
Lo Slog (Static Blog) di Cassandra
L'archivio di Cassandra: scuola, formazione e pensiero

**Cassandra per i posteri: l'archivio su Internet Archive**

*Licenza d'utilizzo: i contenuti di questo articolo, dove non diversamente indicato, sono sotto licenza Creative Commons Attribuzione—Condividi allo stesso modo 4.0 Internazionale (CC BY-SA 4.0),* tutte le informazioni di utilizzo del materiale sono disponibili a questo link.

# *Archivismi: La scomparsa dei media fisici*

*(685) - La sparizione di dischi, dischetti, cartucce, CD e DVD dai nostri cassetti potrebbe essere vista come una benedizione. E' invece l'ascensore, in discesa, che conduce al girone infernale degli "Archivisti Caduti". Vediamo come non finire anche noi laggiù…*

**11 agosto 2026** - in passato, aldifuori della saga di Archivismi, Cassandra aveva ripetutamente esternato sul tema della conservazione del digitale su supporti fisici, questo perché la preservazione della cultura moderna è uno dei valori fondanti della sua attività.

Il concetto di base è rendere ridondante la conservazione dell'informazione, quindi non linkare, ma copiare, perché l'evanescente cloud non diventi l'unico (ed indegno!) "custode della cultura".

Ma copiare sui supporti fisici, pur essenziale, non basta.

Infatti i bit memorizzati sugli archivi informatici, dal 1970 ad oggi, si sono rivelati non il mezzo di conservazione più longevo della storia, ma anzi quello più cagionevole. Quantità infinite di informazioni e software non rimpiazzabili sono andati perduti in questi ultimi decenni, e queste perdite continuano. Persino l'evoluzione crescente delle unità di memorizzazione contribuisce a rendere obsoleti ed inaccessibili i supporti che già non si fossero degradati e cancellati da soli.

Il primo articolo sopra citato termina con queste parole:

*“Non siamo eterni, e tra anni o decenni soltanto pochi bit in rete ricorderanno chi e che cosa eravamo. Saranno bit contenenti testi, immagini, video ma anche programmi. E ripetiamolo, i programmi non sono solo informazioni, sono la legge e la struttura del ciberspazio, che solo di software è costituito. E’ davvero bello pensare che si stia costruendo una strada che, forse, riuscirà a traghettare verso il futuro la parte più importante del “noi” digitale.”*

Il “noi digitale” è, appunto, un tipo particolare di dati, i dati eseguibili, il software in quanto tale. Ed il software, per funzionare, ha bisogno di due cose, un supporto di memorizzazione ed un sistema di esecuzione.

Detto in termini un po’ datati ma assai più semplici, un disco ed un computer.

Questi due “oggetti” non sono necessariamente “materiali”; al giorno d’oggi posso avere un “disco” memorizzato nel “cloud” ed una macchina virtuale, anche questa nel “cloud”. Ma i dati che costituiscono il disco virtuale e la macchina virtuale devono anche loro essere conservati da qualche parte.

Per quanto riguarda il software scritto fino a pochi anni fa, quello che si installava “in locale”, un mezzo fisico di installazione era la norma.

Cassette audio, poi floppy disk a capacità crescente, poi cd-rom, anche questi di capacità crescente (Cd-rom 700 MB, 800 MB, DVD 4,3 GB, 9GB, Blu-Ray 25 GB, 50 GB) hanno supportato la crescita delle capacità dei computer, del software e delle reti.

Ad un certo punto il cloud, affiancato dalle unità esterne di memorizzazione riscrivibili, come smartphone, chiavette e dischi USB3, è diventato mainstream, ed ha fatto scomparire, una dopo l’altra, le unità di memorizzazione su supporti fisici “non riscrivibili”, rendendo quindi inaccessibili le informazioni contenute sui supporti, se non gestite in maniera “attiva” ricopiandole su supporti più moderni.

Fino qui abbiamo raccontato una storia di progresso, che tuttavia, come già detto, dobbiamo riconsiderare in termini di conservazione dell’informazione.

Perché se è vero che mai come oggi le persone, anche quelle digiune di informatica, hanno a disposizione “spazi digitali” cosi grandi da essere inimmaginabili solo pochi anni orsono, è anche (ahimè) vero che questi spazi sono ancora più cagionevoli delle memorie fisiche degli anni scorsi.

Uno spazio cloud è infatti quanto di più evanescente si possa concepire per conservare informazioni; senza uno sforzo di gestione, sia economico che di amministrazione, tutto quanto memorizzato nel cloud scompare in pochi anni, spesso dimenticato dal proprietario che nemmeno ne sentirà la mancanza.

Ed invece, **per avere reali garanzie di conservazione**, dobbiamo copiare le informazioni su supporti durevoli.

Se gli assiri sono riusciti a farci arrivare le loro contabilità su tavolette di argilla cotte, che hanno attraversato indenni più di 4 millenni, noi dobbiamo trovare metodi altrettanto buoni per conservare i bit (di per sé eterni, come la termodinamica ci insegna) per periodi anche più lunghi.

E cosa si può fare a livello personale se qualche nostro dato ci pare “degno” di conservazione? I nostri cari e sudati scritti, oppure anche solo le foto delle vacanze?

Tre cose, comuni alle strategie di backup industriale di dati.

1) **almeno** tre copie dei dati;

2) su **almeno** due tipi di supporto;

3) in **almeno** due locazioni geografiche diverse.

Uno di questi supporti/locazioni può tranquillamente essere il cloud, un altro, virtualmente eterno, può essere l’Arctic World Archive, alle isole Svalbard nel circolo polare artico. E tutti gli altri metodi di archiviazione che ci sono nel mezzo, a vostra scelta, **incluso almeno un supporto fisico che stia a casa vostra, saldamente nelle vostre mani**.

**La cosa più difficile e faticosa, in tutto questo, è una gestione “attiva” dei dati**. Dovete sapere dove conservate cosa, altrimenti finirete nella massa di coloro che perdono tutto anche le foto ed i contatti, ogni volta che traslocano o cambiano cellulare.

E se non volete assumervi nemmeno questo onere ma sentite di avere qualcosa che davvero vorreste far durare nel tempo e diffondere, **potete affidarlo, con garbo e praticamente in modo gratuito, ad Archive.org**.

**Tutti i dettagli** su come fare per far approdare in modo sicuro i vostri dati ed i vostri programmi in queste due durature locazioni **li trovate spiegati, anzi raccontati, da Cassandra nella saga di “Archivismi”**, raccolta in un comodo volumetto.

Enjoy, e buon Ferragosto.

---

Scrivere a Cassandra - Mastodon - Buttondown.com
Videorubrica “Quattro chiacchiere con Cassandra”
Lo Slog (Static Blog) di Cassandra
L’archivio di Cassandra: scuola, formazione e pensiero

**Cassandra per i posteri: l’archivio su Internet Archive**

*Licenza d’utilizzo: i contenuti di questo articolo, dove non diversamente indicato, sono sotto licenza Creative Commons Attribuzione—Condividi allo stesso modo 4.0 Internazionale (CC BY-SA 4.0),* tutte le informazioni di utilizzo del materiale sono disponibili a questo link.