⊗ ARCHIVIO GRAFTON9

WEB ARCHIVING

Il web non è un archivio.

Copiare, conservare e rimettere in rete le pagine web.

Una panoramica breve: concetti, strumenti, formati.

XX SIMposio di storia della conflittualità sociale
Storie in Movimento / Zapruder
Il Poggiolo — Rifugio Re_Esistente, Marzabotto (BO) · 23–26 luglio 2026
https://storieinmovimento.org

WEB ARCHIVING

DEFINIZIONE

Archiviare **sul** web ≠ archiviare **il** web.

Non inventari, digitalizzazioni o cataloghi da pubblicare online.

Fare copie di pagine web che già esistono — con il loro aspetto e
il loro funzionamento — e renderle accessibili quando l'originale non c'è più.

Scomparsa del web

Per come è fatto, il web è instabile: la vita media di una pagina è di pochi anni.

Il 38% delle pagine web esistenti nel 2013 non è più raggiungibile;
un quarto di tutte le pagine apparse tra il 2013 e il 2023 è già scomparso.[1]

Le cause sono tante: server che si rompono, domini abbandonati o rubati,
siti sequestrati, piattaforme che chiudono, contenuti cancellati.

La memoria del web è fragile. **Il web non è un archivio.**

[1] Pew Research Center, *When Online Content Disappears*, 2024 — https://pewresearch.org · https://blog.archive.org/2026/04/23/gone-but-not-forgotten-recovering-the-dead-web

Chi si occupa di conservazione del web oggi

1. Conservazione dei beni culturali (archivi, biblioteche nazionali, Internet Archive)
2. Informatica forense e ambito legale
3. Giornalismo d'inchiesta e investigativo (OSINT)
4. Comunità informali e attivismo digitale (Archive Team, archivi di movimento)

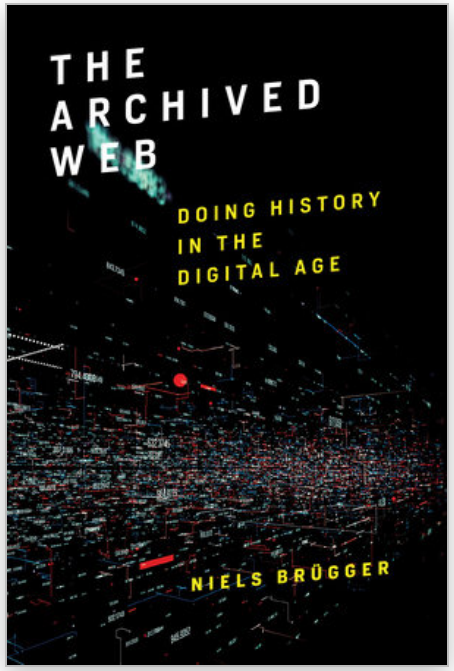
L'interesse commerciale è limitato a nicchie, quasi nullo (compliance legale).

I processi sono studiati in ambito accademico;
la quasi totalità dei software è libera, sviluppata da organizzazioni no profit.

Digitalità (Digitality)

1. Digitalizzato (Digitized) — nato analogico, poi scansionato
2. Nativo digitale (Born digital) — nato direttamente nel web
3. Ricostruito digitale (Reborn digital) — archiviato e ripubblicato

L'oggetto che consultiamo in un archivio web non è mai identico all'originale: è una **ricostruzione**, con una sua storia.



Niels Brügger, *The Archived Web: Doing History in the Digital Age*, MIT Press, 2018. https://doi.org/10.7551/mitpress/10726.001.0001

Pratiche informali, sbagliate

1. Copia-incolla: selezionare il testo di una pagina e incollarlo altrove
2. Fare uno screenshot della pagina
3. Stampare la pagina in PDF dal browser

Preservano in qualche modo il contenuto, ma perdono il contesto:
l'URL di pubblicazione, la data di cattura, l'aspetto grafico, i collegamenti.

E non offrono alcuna garanzia di integrità: sono facili da alterare,
difficili da verificare.

Pratiche corrette, obiettivo

Creare una copia che mantenga l'unità informativa e visiva del contenuto,
insieme alle proprietà tipiche di una pagina web:
URL originale, **data** di cattura, **aspetto** originale, **navigabilità** da browser diversi.

Due qualità da preservare:

- **Fedeltà** (fidelity): quanto la copia somiglia e si comporta come l'originale
- **Provenienza** (provenance): dove, quando e come è stata fatta la cattura

Esistono formati standard e processi consolidati, frutto di anni di
esperienza e collaborazione tra le istituzioni.

Passaggi

1. **Selezione**: cosa archiviare? `nomination`
2. **Acquisizione**: `record`
3. **Descrizione** con metadati `description`
4. **Accesso**: `replay`

Passaggi

Cattura (**Record**) → **Accesso** (**Replay**)

Prima si cattura la pagina o l'intero sito (record),
poi la si rende di nuovo navigabile (replay).

Processi che avvengono in momenti distinti, e possono essere operati da agenti diversi.

Cattura (Record)

aka harvesting, crawling, ...

Software dedicati, che impersonano l'attività di una persona che naviga il sito con un browser e salvano il contenuto man mano che lo navigano.

Alcuni sono automatici (crawler, browser headless), in altri la navigazione avviene manualmente: dipende dalla natura tecnica del sito da archiviare.

Siti difficili da catturare: contenuti caricati via JavaScript, scroll infinito, login e paywall, video in streaming, piattaforme social chiuse.

Formati: WARC e WACZ

WARC (Web ARChive) — standard ISO 28500.[1]
In parole semplici è un contenitore, come un file ZIP:
racchiude richieste, risposte e metadati della cattura.

WACZ (Web Archive Collection Zipped) — formato portabile di Webrecorder:[2]
impacchetta WARC + indici + metadati in un singolo file,
consultabile direttamente nel browser e firmabile per verificarne l'autenticità.

[1] https://iipc.github.io/warc-specifications/ · https://en.wikipedia.org/wiki/WARC_(file_format)
[2] https://specs.webrecorder.net/wacz/latest/

Cattura (Record)

SOFTWARE

☞ ArchiveWeb.page — https://archiveweb.page/
☞ SingleFile — https://github.com/gildas-lormeau/SingleFile
Browsertrix Crawler — https://github.com/webrecorder/browsertrix-crawler
Heritrix — https://github.com/internetarchive/heritrix3
Wget — https://www.gnu.org/software/wget/
AutoArchiver — https://bellingcat.gitbook.io/toolkit/more/all-tools/auto-archiver

SERVIZI

☞ SPN (Save Page Now) — https://web.archive.org/save/
Browsertrix — https://webrecorder.net/browsertrix/

☞ = adatti anche a chi inizia: estensioni per browser o servizi web, nessuna riga di comando.

Accesso (Replay)

Software che rendono nuovamente disponibile alla navigazione il contenuto precedentemente archiviato.

È una copia statica e immutabile: alcune funzioni di interattività del sito di origine non possono essere riprodotte (form, ricerche).

SOFTWARE

☞ ReplayWeb.page — https://replayweb.page/ (funziona interamente nel browser, senza server)
PyWB — https://github.com/webrecorder/pywb
OpenWayback (deprecato) — https://github.com/iipc/openwayback

Dove conservare le copie

Una copia archiviata va poi custodita nel tempo:

- **Internet Archive** — deposito pubblico, gratuito, ma centralizzato
- **Deposito legale** — BNCF Firenze (L. 106/2004)
- **Self-hosting** — pieno controllo, ma richiede cura e ridondanza

Nessuna soluzione basta da sola: la ridondanza è parte della conservazione.

Prevenzione

Necessità di formare e sensibilizzare chi produce contenuti a renderli archiviabili.
Scelte corrette di tecnologie e piattaforme: HTML statico, URL stabili,
niente dipendenze superflue da servizi terzi.

Creating Preservable Websites
https://www.loc.gov/programs/web-archiving/for-site-owners/creating-preservable-websites/

Discussione aperta

Effimero vs permanente: molti contenuti social sono pensati per una visibilità a tempo determinato.

Consenso ed etica: chi decide cosa archiviare? Diritto all'oblio,
sicurezza delle persone, contenuti critici.

Infrastrutture di archiviazione: chi le mantiene, chi le paga, chi le controlla.

Condivisione delle conoscenze e delle competenze.

Cosa è un "documento"? Un meme, una chat, un vocale su Telegram sono fonti storiche?
Dove finisce la comunicazione e inizia la documentazione?

Risorse

DIY Web Archiving: For the Web (& World) You Care About
https://zinebakery.com/homemade-zines/bakeshop-2-diywebarchiving

What Is Web Archiving?
https://webrecorder.net/resources/what-is-web-archiving/

IIPC Training materials
https://netpreserve.org/web-archiving/training-materials/

What is a web archive?
https://www.youtube.com/watch?v=ubDHY-ynWi0

Risorse

Una nota polemica sul web archiving in Italia, Federico Mazzini
https://doi.org/10.25430/pupb-9788869384394-10

404: file not found. Il caso Indymedia Italia, Alice Corte, Zapruder 45
https://italy.indymedia.org/library/Zap45_8-Schegge1.pdf

S. Allegrezza — Web e social media come nuove fonti per la storia
https://doi.org/10.6092/issn.2532-8816/15665

Slides e PDF di questa presentazione:

https://grafton9.net/talks/simposio2026

https://grafton9.net/talks/simposio2026/slides.pdf

void@grafton9.net

https://grafton9.net