ARTICOLO

Versioning delle banche dati aperte nelle Digital Humanities con Crossmark: teoria e pratica

Giovanni Salucci e Michael Bassi

Il contributo prende in esame l'attuale stato dell'arte nella gestione delle banche dati aperte per la provenienza e la citabilità: vengono proposti il DOI e il servizio Crossmark come possibili chiavi di volta ad un corretto approccio al problema del *versioning* (o storicizzazione dei dati); ovvero la corretta gestione, conservazione, tracciabilità, disseminazione dei prodotti della ricerca scientifica, come pure il riconoscimento di tutte le autorialità intervenute nel tempo. Dopo una rassegna delle differenze tra le edizioni digitali, tendenzialmente statiche, e le banche dati, ontologicamente dinamiche, viene proposto un innovativo modo di ripensare il DOI: non solo identificatore persistente di un oggetto fisso, bensì nodo di una fitta rete di relazioni. Seguono alcuni casi di studio e alcune linee guida volte a stabilire un primo standard nella gestione del *versioning* all'interno del mondo delle *Digital Humanities*.

This contribution examines current approaches to the management of open databases with respect to provenance and citability. It proposes DOIs and the Crossmark service as key tools for addressing versioning challenges, including the management, preservation, traceability, and dissemination of research outputs. By contrasting relatively static digital editions with inherently dynamic databases, the paper reconceptualizes the DOI not merely as a persistent identifier, but as a relational node. The contribution concludes with selected case studies and preliminary guidelines for versioning management in the field of Digital Humanities.

Parole chiave: Umanistica digitale, banche dati digitali, Crossref, Crossmark
Keywords: Digital Humanities, Data Versioning, DOI, Crossref, Crossmark

Peer review

Submitted	15/12/2025
Accepted	13/03/2026
Published	21/03/2026

DOI 10.35948/DILEF/2026.4405

1. Introduzione

Negli ultimi dieci anni le *Digital Humanities* (DH) hanno vissuto un aumento considerevole nella domanda e nella creazione di banche dati digitali, utilizzate sempre più non solo come supporto alla ricerca, ma come veri e propri oggetti scientifici. I corpora testuali, le banche dati biografiche e bibliografiche, i *database* lessicali, i *repository* di documenti storici e le collezioni di metadati culturali rappresentano ora le strutture di base per l'analisi, l'interpretazione e la condivisione del sapere umanistico. Tuttavia, nonostante questa centralità, le sfide metodologiche più significative connesse al carattere dinamico di queste risorse sono state poco affrontate.

Uno dei problemi più interessanti, oltreché nevralgici, ma meno sistematizzati, riguarda il *versioning* delle banche dati, cioè la storicizzazione dei dati. Infatti, le banche dati nelle DH sono soggette a un processo continuo di aggiornamento, revisione e riorganizzazione, in contrasto con le pubblicazioni tradizionali: nuove fonti vengono inserite, errori corretti, criteri interpretativi ridefiniti, schemi di dati modificati, e tali operazioni possono essere eseguite anche da ricercatori diversi, che intervengono dopo le azioni degli autori originali. Questi cambiamenti, sebbene spesso cruciali per far progredire la ricerca, sono raramente documentati in modo formale e standardizzato, rendendo difficile stabilire quale versione di una banca dati sia stata utilizzata in un determinato studio, o come i risultati possano essere verificati e riprodotti nel tempo, o a chi appartenga la responsabilità scientifica, a quale titolo e in che misura.

Il problema del *versioning* non è esclusivamente tecnico, ma coinvolge questioni più profonde di natura sia conoscitiva che editoriale. Nelle DH, i dati non sono meri contenitori di informazioni: incorporano scelte interpretative, modelli concettuali e pratiche disciplinari specifiche. Di conseguenza, ogni versione di una banca dati rappresenta uno stadio particolare di un processo di conoscenza in evoluzione e si fa testimone di scelte progettuali ben precise. L'assenza di pratiche condivise di *versioning* e di citazione delle versioni rischia quindi di compromettere la trasparenza, la riproducibilità e la validità scientifica della ricerca che utilizza quei dati digitali.

In questo contesto, risulta particolarmente interessante il confronto con le pratiche consolidate della comunicazione scientifica, dove il problema delle versioni è stato affrontato, almeno in parte, attraverso politiche e scelte editoriali specifiche. Tra queste, *Crossmark*, un servizio sviluppato da *Crossref*, si propone di rendere visibile lo stato delle pubblicazioni digitali, segnalando aggiornamenti, correzioni o ritrattazioni e distinguendo la cosiddetta *version of record* (VoR, versione di riferimento) dalle sue revisioni successive. Crossmark non è un sistema di *versioning* dei dati, in quanto è stato concepito per gli articoli scientifici; tuttavia, può rappresentare un modello

funzionale di trasparenza e tracciabilità e può offrire spunti utili per riflettere sulle esigenze delle banche dati nelle Digital Humanities.
L'obiettivo di questo contributo è analizzare il problema del *versioning* delle banche dati nelle DH da una prospettiva teorico-metodologica, confrontando le pratiche emergenti nel campo dei dati umanistici-digitali con il modello di *versioning* editoriale proposto da Crossmark. Attraverso una rassegna critica della letteratura e un'analisi concettuale delle diverse tipologie di *versioning*, si intende evidenziare le principali criticità e proporre un quadro interpretativo che possa contribuire allo sviluppo di pratiche più trasparenti, sostenibili e condivise per la gestione delle versioni delle banche dati DH.

2. Principi FAIR, *Versioning* e Digital Humanities

La condivisione dei dati è essenziale per garantire la trasparenza, la validazione e la riproducibilità del processo di ricerca, nonché per favorire ulteriori scoperte basate sul riutilizzo dei dati; è quindi possibile affermare che «il valore dei dati risiede nel loro utilizzo» [Hauf et al, 2021]. Per consentirne il riuso, i dati devono essere condivisi secondo i principi FAIR (*Findable*, *Accessible*, *Interoperable*, *Reusable*), che ne assicurano la reperibilità, l'accessibilità, l'interoperabilità e il riutilizzo [Wilkinson et al, 2016]. I principi FAIR forniscono linee guida per la pubblicazione delle risorse digitali e descrivono un insieme di caratteristiche, attributi e comportamenti che avvicinano progressivamente le risorse digitali al raggiungimento di tale obiettivo. E il *versioning* dei dati è rilevante per la corretta applicazione dei principi FAIR, perché versioni ben identificate facilitano interoperabilità, riproducibilità, provenienza e riuso dei dati. [Klump et al, 2021]
Tuttavia, specialmente nel campo delle banche dati realizzate in progetti di DH, la riflessione sul *versioning* si è sviluppata in modo frammentario e spesso indiretto, prevalentemente all'interno degli studi sulle edizioni digitali e sui progetti di *digital scholarly editing* ([Bürgermeister 2020] per una rassegna). In questo ambito, il concetto di "versione" è tradizionalmente legato all'idea di edizione, intesa come specifica rappresentazione critica e interpretativa di un testo o di un insieme di fonti. Tuttavia, sebbene il tema delle versioni sia centrale nelle edizioni digitali, esso viene raramente formalizzato in termini di politiche di *versioning* esplicite. Le modifiche apportate a un progetto – correzioni, aggiunte di materiali, cambiamenti di interfaccia o di codifica – a volte sono registrate nei metadati (ad esempio, in caso di edizioni XML TEI, all'interno della sezione <editionStmt> contenuta nelle intestazioni dei file XML, [Broyles, 2020]), ma spesso sono documentate in modo non ufficiale o affidate a note di progetto, senza una chiara distinzione tra versioni né attraverso criteri condivisi per la loro identificazione e citabilità. Questo approccio risulta particolarmente problematico quando le edizioni digitali assumono la forma di

banche dati aperte, utilizzate come dei “semilavorati” – potenzialmente o effettivamente in continuo aggiornamento – e fonti primarie per studi e approfondimenti successivi.
Nel complesso, se da un lato la letteratura DH evidenzia una crescente consapevolezza della natura dinamica delle risorse digitali, dall’altro non offre ancora modelli consolidati per il *versioning* delle banche dati come oggetti scientifici autonomi, distinti sia dalle edizioni testuali sia dai semplici strumenti di supporto alla ricerca. C’è dunque una presa di coscienza della situazione attuale e di quella in divenire ma non sono ancora stati definiti conseguenti standard di approccio condivisi.

2.2 *Versioning*, provenienza e gestione dei dati

Al di fuori delle DH, il problema del *versioning* è stato affrontato in modo più sistematico nell’ambito della *data curation* e della *data science* (ad es: [Chavan et al, 2015]). In questi contesti, il *versioning* dei dati è strettamente connesso ai concetti di provenienza, tracciabilità e riproducibilità della ricerca. Modelli concettuali come la *PAVOntology* (*Provenance*, *Authoring and Versioning* [Ciccarese et al, 2013]) e lo standard W3C PROV[1] mirano a descrivere in modo formale la storia di un oggetto digitale, includendo informazioni su chi ha prodotto una risorsa, quando e attraverso quali processi; analogamente, iniziative quali *RDA Data Versioning Research Group* mirano a orientare i *data providers* e i *data collectors* verso una pratica coerente di *versioning* dei dati nella definizione dei propri protocolli e delle proprie procedure di gestione dei dati. [Klump et al, 2021]
Questi approcci enfatizzano la necessità di distinguere tra diverse versioni di uno stesso *dataset*, documentando le trasformazioni subite nel tempo e consentendo la ricostruzione dei passaggi intermedi. Nel contesto delle scienze computazionali e dei dati, tali pratiche sono spesso supportate da sistemi di controllo versione e da infrastrutture progettate per gestire grandi volumi di dati strutturati.
Nonostante la loro rilevanza, questi modelli sono stati solo parzialmente adottati nelle DH. Anche recenti rassegne impegnate a esplicitare la provenienza e il tracciamento delle modifiche nelle DH [Massari et al, 2025] confermano le specificità dei dati umanistici fortemente interpretativi, eterogenei e spesso non normalizzati, che rendono complessa l’applicazione diretta di soluzioni sviluppate per altri ambiti disciplinari. Tutto questo ha contribuito ad aumentare il divario tra le pratiche avanzate di *versioning* dei dati e le esigenze concrete dei progetti DH.

2.3 Le banche dati come oggetti scientifici non statici

Un elemento ricorrente nella letteratura più recente è il riconoscimento delle banche dati digitali come oggetti scientifici dinamici, il cui significato evolve nel tempo. A differenza delle fonti tradizionali, le banche dati non si limitano a raccogliere informazioni in forma sostanzialmente statica, ma strutturano attivamente il sapere attraverso modelli, categorie, relazioni e criteri di inclusione ed esclusione. Ogni modifica a questi elementi incide direttamente sull'interpretazione dei dati e, di conseguenza, sui risultati della ricerca che li utilizza.
Questa dimensione dinamica solleva interrogativi cruciali sul piano metodologico: come documentare l'evoluzione di una banca dati? Come rendere esplicite le scelte interpretative incorporate nelle diverse versioni? E ancora, come garantire che un'analisi basata su una versione precedente resti riproducibile e verificabile nel tempo?
In molti casi, le banche dati continuano a essere citate come entità monolitiche, senza riferimento a una versione specifica, nonostante il loro contenuto e la loro struttura possano cambiare in modo significativo nel corso degli anni. Una prassi del genere evidenzia la necessità di sviluppare concetti e pratiche condivise, per gestire questi strumenti come realtà collocate nel tempo, ognuna con il suo percorso di varianti. Per una corretta gestione delle banche dati, anche in vista del loro riconoscimento accademico come entità citabili e tracciabili, diventa quindi strettamente necessaria l'applicazione di una certa modellizzazione nella struttura dei dati che si intende rappresentare e diffondere.
In altre parole, quello che deve cambiare è proprio l'approccio nei confronti delle banche dati aperte nell'ambito delle Digital Humanities, in virtù del *medium* nel quale queste nascono, vivono e proliferano. Si ha che fare con entità ontologicamente dinamiche e in continua mutazione; basti pensare che le pagine web che costituiscono una banca dati aperta vengono *caricate* – in termini di editoria canonica potremmo dire *stilate* – ogniqualvolta un visitatore vi accede, in maniera pressoché istantanea. Basti questo a delineare la considerevole distanza che separa quanto appena descritto da un prodotto che si cristallizza nel momento stesso della sua uscita pubblica, come uno nato e pensato per la carta stampata o comunque redatto in una forma che alla stampa si rifà per vesti editoriali e fissità, come ad esempio il PDF.

2.4 DOI e *versioning*: identificazione persistente stratificata e gestione del cambiamento

Il rapporto tra DOI (*Digital Object Identifier*) e *versioning* costituisce un elemento cruciale per comprendere le modalità attraverso cui la comunità scientifica affronta il

problema del cambiamento nel tempo delle risorse digitali. Il DOI nasce come identificatore persistente volto a consentire l'accesso stabile a una risorsa digitale. Essendo stato adottato principalmente nell'ambito dell'editoria scientifica accademica, il DOI è prevalentemente associato all'idea di stabilità, citabilità e permanenza del record scientifico, indipendentemente dalla sua collocazione o dalla piattaforma di distribuzione. Circa l'utilizzo del DOI nei progetti di DH, e in particolare nelle banche dati letterarie e nei progetti lessicografici si rimanda ad altri articoli pubblicati su questa stessa Rivista ([Salucci 2022], [Salucci 2023])

Tuttavia, alla luce delle nuove esigenze derivanti dalle specifiche caratteristiche dei prodotti digitali (che spingono a rivedere e riadattare l'applicazione e la funzione del DOI), la crescente diffusione di oggetti digitali dinamici – articoli aggiornabili, *preprint* soggetti a modifiche, *dataset* in banche dati aperte – ha messo in discussione l'associazione implicita tra DOI e staticità dell'oggetto identificato: risulta evidente un certo scarto tra DOI inteso come identificatore del corrispettivo digitale dei prodotti dell'editoria accademica tradizionale (articoli e monografie) e DOI quale identificatore di risorse native digitali dinamiche, svincolate da qualsivoglia fissità (record di database). Il problema non riguarda il DOI in sé, ma il modo in cui esso viene interpretato e utilizzato all'interno delle pratiche editoriali e scientifiche. In particolare, emerge una tensione tra due esigenze apparentemente contrapposte: da un lato, la necessità di identificatori persistenti che garantiscano la stabilità delle citazioni; dall'altro, il riconoscimento del carattere dinamico e aggiornabile delle risorse digitali.

Nel sistema accademico tradizionale, questa tensione viene risolta attraverso una distinzione relativamente netta tra l'oggetto identificato dal DOI — la cosiddetta VoR — e le sue eventuali modifiche successive. In molti casi, le correzioni e gli aggiornamenti non comportano l'assegnazione di un nuovo DOI, ma vengono gestiti come documenti separati (*errata*, *corrigenda*, *addenda*) collegati all'articolo originale. Crossmark si inserisce precisamente in questo spazio, offrendo un meccanismo utile per segnalare lo stato dell'oggetto identificato dal DOI senza compromettere la sua identità persistente.

Crossmark, infatti, fornisce un sistema online e multipiattaforma per individuare rapidamente lo stato di un prodotto della ricerca, insieme a metadati aggiuntivi riferiti al processo editoriale [Hendricks et al, 2020]. In aggiunta, il bottone Crossmark può essere incorporato anche nei file PDF, consentendo così di essere informati su eventuali modifiche anche a distanza di mesi o persino di anni dal download: un chiaro esempio della distanza e interconnessione tra documento PDF stampabile (statico) e la sua controparte pagina web HTML (dinamica). Crossmark, in quanto servizio offerto da Crossref associato al DOI, non introduce nuove versioni identificabili, ma fornisce metadati di stato che qualificano l'oggetto nel tempo. In questo senso, esso rappresenta una soluzione che privilegia la continuità dell'identificatore rispetto alla frammentazione in molteplici DOI per ogni modifica.

Tale approccio è funzionale alla logica delle pubblicazioni scientifiche, ma presenta limiti evidenti quando viene considerato in relazione a oggetti digitali più complessi, come le banche dati delle DH.
Nel caso dei *dataset* DH, infatti, il cambiamento non è marginale né accessorio, ma spesso strutturale e interpretativo: talvolta non mancano l'intervento di nuovi autori o modifiche sostanziali ai contenuti già pubblicati. L'uso di un singolo DOI per rappresentare un contenuto in evoluzione rischia di occultare differenze significative tra versioni, rendendo problematica la ricostruzione del contesto nel quale determinati risultati di ricerca sono stati prodotti. Al tempo stesso, l'assegnazione di un nuovo DOI per ogni aggiornamento può risultare impraticabile e concettualmente riduttiva, soprattutto quando le modifiche sono incrementali o frequenti.
Questo scenario suggerisce la necessità di un uso più flessibile e consapevole del DOI nel contesto delle Digital Humanities. Una possibile soluzione, in corso di realizzazione nel progetto *OIM*[2] (*Osservatorio degli Italianismi nel Mondo*), consiste nell'utilizzo di porzioni di metadati riservati a Crossmark per registrare le variazioni del contenuto e le differenti versioni, mantenute accessibili per la consultazione diacronica. Si riporta in appendice un esempio, relativo al tracciamento di *versioning* occorso per una voce del dizionario degli italianismi del portoghese. In questo modello, il DOI non coincide più con un oggetto statico, ma diventa parte di un sistema di identificazione stratificato, capace di riflettere la natura temporale e le differenti versioni dei dati.
In tale prospettiva, Crossmark può essere considerato non come una soluzione completa al problema del *versioning*, ma come strumento al servizio di una più ampia trasformazione del ruolo del DOI: da semplice identificatore di un oggetto fisso a nodo di una rete di relazioni che documentano lo stato, la storia e le trasformazioni di una risorsa digitale. Trasporre questa logica alle banche dati delle DH implica ripensare il DOI non solo come strumento di citazione, ma come componente di un'infrastruttura capace di implementare pratiche di provenienza, autorialità e responsabilità, grazie a metadati di versione e politiche di conservazione a lungo termine.

2.5 Crossmark e il tracciamento delle versioni editoriali

Nel contesto dell'editoria accademica digitale, Crossmark rappresenta uno dei pochi tentativi istituzionalizzati di affrontare il problema del *versioning* in modo standardizzato e interoperabile. Sviluppato da Crossref come servizio online, applicabile a documenti pubblicati sia in formato HTML che PDF, Crossmark non nasce come sistema di gestione delle versioni in senso tecnico, bensì come strumento per garantire la trasparenza e l'affidabilità del record scientifico, consentendo ai lettori di verificare lo stato corrente di una pubblicazione digitale. [Meyer 2011], [Lamney 2014]

Il principio alla base di Crossmark è relativamente semplice: ogni oggetto dotato di DOI che aderisce al servizio incorpora un indicatore visivo (il cosiddetto *Crossmark button*) attraverso il quale è possibile accedere a informazioni aggiornate sullo stato del documento. Tali informazioni possono includere la presenza di correzioni, errata, addenda, aggiornamenti editoriali o, nei casi più elaborati, la ritrattazione dell'articolo. In questo senso, Crossmark introduce una distinzione esplicita tra la VoR e le sue eventuali revisioni successive, rendendo visibile al lettore la dimensione temporale e processuale della pubblicazione scientifica.
Dal punto di vista concettuale, Crossmark può essere interpretato come una forma di *versioning* editoriale dichiarativo. A differenza dei sistemi di controllo versione utilizzati nello sviluppo software o nella gestione dei dati strutturati, Crossmark non registra le singole modifiche né consente il confronto puntuale tra versioni. Piuttosto, segnala l'esistenza di uno stato aggiornato e rinvia a documenti separati che descrivono la natura dell'intervento editoriale. La granularità del *versioning* è dunque limitata e orientata alla stabilità del record bibliografico, più che alla tracciabilità analitica delle trasformazioni del contenuto.
Questa impostazione risponde a esigenze specifiche del sistema editoriale accademico, in cui la citabilità, l'autorità e la persistenza delle pubblicazioni hanno un ruolo centrale. Tuttavia, proprio per queste caratteristiche, Crossmark offre uno spunto rilevante anche per la riflessione sul *versioning* delle banche dati nelle Digital Humanities; come visto in precedenza, non può essere adottato direttamente come soluzione per il *versioning* delle banche dati DH, ma può essere considerato un modello concettuale di trasparenza.
Nella porzione dei metadati di Crossmark, che quindi segnaleranno la "storia editoriale" del contenuto registrato sotto quel DOI, sarà possibile documentare la successione degli interventi, come pure rendere disponibile l'accesso alle versioni precedenti del contenuto; in questo modo si renderà immediatamente visibile lo "stato" di una risorsa digitale – aggiornata, superata, corretta – per poter essere adattata alle infrastrutture di ricerca DH, affiancando o integrando sistemi più sofisticati di provenienza e *data versioning*.

2.6 *Versioning* delle pubblicazioni *versus versioning* delle banche dati

Il concetto di *versioning* assume significati differenti a seconda dell'oggetto cui viene applicato. Nel caso delle pubblicazioni scientifiche tradizionali, il *versioning* è storicamente legato a pratiche editoriali orientate alla stabilità del testo e alla persistenza del record bibliografico. Anche nella sua declinazione digitale, la pubblicazione mantiene una struttura relativamente chiusa: l'articolo o il volume vengono identificati dalla VoR, che costituisce il punto di riferimento per la citazione e

l'attribuzione dell'autorialità scientifica. Eventuali modifiche successive sono rare, trattate come interventi marginali e vengono segnalate attraverso strumenti editoriali specifici, tra cui Crossmark, senza compromettere l'identità dell'opera originaria.
Le banche dati digitali, al contrario, sono caratterizzate da una natura intrinsecamente aperta e dinamica. Esse non vengono generalmente pubblicate come prodotti conclusi, ma si sviluppano attraverso cicli di aggiornamento che coinvolgono nuovi autori, nuovi contenuti, modifiche alle strutture e ai modelli interpretativi. In questo contesto, il *versioning* non riguarda soltanto la correzione di errori o l'aggiunta di nuovi dati, ma investe l'evoluzione complessiva dell'oggetto, inclusi i criteri di selezione, le ontologie di riferimento e le relazioni tra le entità rappresentate.
Questa differenza strutturale ha implicazioni rilevanti sul piano metodologico. Mentre il *versioning* delle pubblicazioni mira a preservare la stabilità e l'affidabilità del testo citato, il *versioning* delle banche dati deve confrontarsi con l'esigenza di documentare il cambiamento come parte integrante del processo conoscitivo. Nel caso dei *dataset* delle DH, ogni versione può incorporare scelte interpretative diverse, rendendo inefficace identificare una singola "versione di riferimento" nell'ottica di non ridurre la complessità del lavoro scientifico sottostante. Ecco che quanto di più digitale e dinamico le DH hanno oggi da offrire sembra ricongiungersi metodologicamente a quella critica delle varianti di Gianfranco Contini, dove anche le carte preparatorie sono da considerarsi importanti tanto quanto la versione pubblicata dall'autore, meritorie dell'attenzione dello studioso e – diremmo oggi – di costituire un nodo che testimonia lo stato di avanzamento.
Crossmark opera all'interno di un contesto in cui la versione corrente (l'ultima, o VoR) è centrale, ma allo stesso tempo possono essere ugualmente rilevanti versioni precedenti, magari già citate in altre pubblicazioni scientifiche. Le banche dati DH, quindi, richiedono strumenti capaci di rendere esplicita la storia delle trasformazioni, permettendo di comprendere non solo se una risorsa è stata aggiornata, ma come e perché tali aggiornamenti abbiano modificato il significato dei dati, e chi ne abbia avuto la responsabilità e con quale ruolo.
In questo senso, il *versioning* delle pubblicazioni e quello delle banche dati rispondono a logiche diverse ma complementari. Se il primo privilegia la stabilità necessaria alla comunicazione scientifica, il secondo deve valorizzare la dimensione processuale e interpretativa della ricerca. La sfida per le Digital Humanities consiste nel trovare un equilibrio tra questi due modelli, integrando la trasparenza editoriale ispirata a strumenti come Crossmark con pratiche di *versioning* dei dati più granulari e orientate alla tracciabilità del cambiamento.

2.7 Esempi di *versioning* in banche dati, corpora e dizionari nelle Digital Humanities

Il concetto di *versioning* nelle Digital Humanities ha trovato applicazioni concrete in diversi tipi di risorse, dai corpora testuali ai dizionari digitali, fino ai *dataset* linguistici e storici. Nonostante la pratica non sia sempre formalizzata come nei sistemi di controllo versione software, numerosi progetti dimostrano empiricamente strategie efficaci per gestire l'evoluzione dei dati.

Un primo caso di studio è *Drama Corpora*[3], un'infrastruttura *open* per lo studio del teatro europeo, che utilizza *Git* e *repository GitHub* per tracciare modifiche ai corpora, inclusi testi, annotazioni e metadati.

Un secondo esempio è *Papyri.info*[4], un'infrastruttura *open* di ricerca per la conservazione, lo studio e la consultazione dei papiri. Il sistema si articola principalmente in due componenti: il *Papyrological Navigator* (PN) e il *Papyrological Editor* (PE). Il PN consente agli utenti di esplorare e interrogare l'insieme delle risorse attraverso strumenti di navigazione e ricerca avanzata, mentre il PE, accessibile a ricercatori e studiosi autorizzati, permette di creare nuove schede testuali o modificare quelle esistenti. L'editing e la marcatura dei testi avvengono secondo lo standard EpiDoc[5] (TEI XML); tutte le modifiche effettuate sono sottoposte a un processo di revisione tra pari, che assicura il controllo scientifico dei contenuti prima della loro accettazione definitiva e della loro pubblicazione. Un elemento centrale del funzionamento di Papyri.info è il controllo di versione, realizzato tramite sistemi di *versioning* distribuito (Git), che garantisce la tracciabilità, la trasparenza e la riproducibilità di ogni intervento editoriale. Solo le versioni approvate attraverso il processo di revisione vengono integrate nel repository ufficiale e rese pubblicamente disponibili.

Analogamente, progetti di edizioni digitali di testi letterari o manoscritti storici adottano numerazioni di versione e date di aggiornamento nei file TEI. Per esempio, il progetto *Codex Sinaiticus*[6] integra numeri di versione e <revisionDesc> nei suoi testi XML per documentare le modifiche apportate alle trascrizioni e alle annotazioni. Questo approccio consente di tracciare la storia evolutiva del testo digitale, rendendo ciascuna versione citabile e garantendo la trasparenza dei processi editoriali.

E ancora, il progetto di *digital library Perseus*[7] così come *Digital Latin Library*[8] e anche *Open Greek and Latin*[9]: tutti si appoggiano a repository Git per preservare e tenere traccia delle modifiche ai contenuti.

Se passiamo da progetti autonomi a piattaforme progettate per le DH, quale ad esempio la piattaforma europea CLARIN[10] si nota che il tema del *versioning* emerge soprattutto nel contesto dei repository di dati di ricerca: è previsto l'uso degli

identificatori persistenti al fine della riproducibilità scientifica. Ad esempio, nella piattaforma CLARIN:EL, costituita per raccogliere, documentare, curare e distribuire risorse linguistiche digitali in Grecia, sono documentate e implementate politiche di *versioning*[11]; in questo caso viene assegnato un identificatore persistente (*handle*) a ciascuna risorsa, al momento del caricamento nel repository. Nel caso di nuove versioni viene sempre generato un nuovo *handle*, ma attraverso una procedura è possibile registrare il collegamento fra le versioni (con un collegamento tra i due *handle*). Questo comportamento è del tutto analogo a quello che accade in altri repository, ad esempio Zenodo[12], il repository ufficiale dell'*Open Science* gestito dalla Commissione Europea: il *versioning* si applica sempre a livello dell'intero *dataset* o della risorsa caricata, ma non può raggiungere il livello di descrizione granulare dei singoli record (o schede) contenute nella collezione. Ancora una volta, quindi, il *dataset* è considerato come una unità autonoma, ma si perde il dettaglio della evoluzione e varietà delle trasformazioni delle parti che lo costituiscono (i singoli *record*).

3. Come utilizzare Crossmark per il *versioning* delle banche dati DH

Alla luce delle criticità emerse nelle sezioni precedenti, appare evidente l'opportunità di stabilire delle politiche concettuali insieme a linee guida operative che, utilizzando DOI e Crossmark, consentano di affrontare in modo sistematico il *versioning* delle banche dati nelle Digital Humanities, sia a livello del *dataset* che dei *record* che lo costituiscono. Un tale *framework* non deve limitarsi a fornire considerazioni teoriche, editoriali e infrastrutturali, ma deve soprattutto individuare soluzioni tecniche, per favorirne l'applicabilità in un'ampia varietà di progetti di umanistica digitale.
La sezione seguente è dedicata all'illustrazione della proposta, mediante una guida operativa che ne chiarisce i principali passaggi.
Nel loro insieme, questi passaggi delineano un contesto che non concepisce il *versioning* come un semplice problema tecnico, ma come una pratica scientifica essenziale per la trasparenza, la riproducibilità e l'affidabilità della ricerca nelle DH, così come la corretta attribuzione delle autorialità.

#1 – Utilizzo dei DOI con Crossref in previsione dell'utilizzo di Crossmark
Il primo requisito è quello dell'adozione sistematica degli identificativi permanenti, in particolare i DOI da assegnare a ciascun livello di descrizione (*database*, *collection*, *record*) all'interno della banca dati. Per fare ciò, occorrerà scegliere Crossref come agenzia di registrazione, per poter accedere al servizio Crossmark come specificato in seguito.
Ovviamente, per poter beneficiare dell'intero ventaglio di opportunità – a tutto vantaggio anche di valorizzazione e diffusione – risulta fondamentale prestare

particolare attenzione nella redazione dei metadati, curandone completezza e qualità.

#2 – Scelta e criteri di adozione delle politiche di *Versioning*

Oltre all'adozione degli identificatori persistenti, occorre stabilire a priori quando una modifica giustifichi o meno la creazione di una nuova versione citabile, e di conseguenza in che caso assegnare DOI distinti alle versioni rilevanti.

Nel flusso di pubblicazione delle banche dati, in analogia a quanto accade, per esempio, per gli articoli scientifici pubblicati nelle Riviste, occorre definire le politiche di *versioning*: fatto salvo per interventi minimi (quali la correzione di meri refusi o errori materiali), in tutti gli altri casi le modifiche apportate a un contenuto dopo la sua pubblicazione richiederanno di produrre una nuova versione e di modificare conseguentemente i metadati del DOI, o di registrarne di nuovi.

Nel caso di modifiche sostanziali (quale ad esempio l'intervento di un nuovo autore), sarà necessario conservare la versione precedente e pubblicarne una nuova, registrando un nuovo DOI; nei metadati di registrazione si farà attenzione a collegarlo al precedente, per tenere traccia delle modifiche ma anche della provenienza.

In generale, quando si rilasciano delle modifiche, occorre prevedere che il sistema crei in automatico una nuova versione della scheda e archivi la precedente, lasciando al responsabile la decisione di mantenere il DOI già assegnato o registrarne uno nuovo.

In ogni caso, la versione precedente della scheda resta consultabile online: il collegamento tra le due versioni della scheda, registrato nei metadati e nella sezione dello schema riservato a Crossmark, viene reso consultabile all'interno delle schede stesse (o del PDF generato) grazie al bottone "*Check for updates*" di Crossmark.

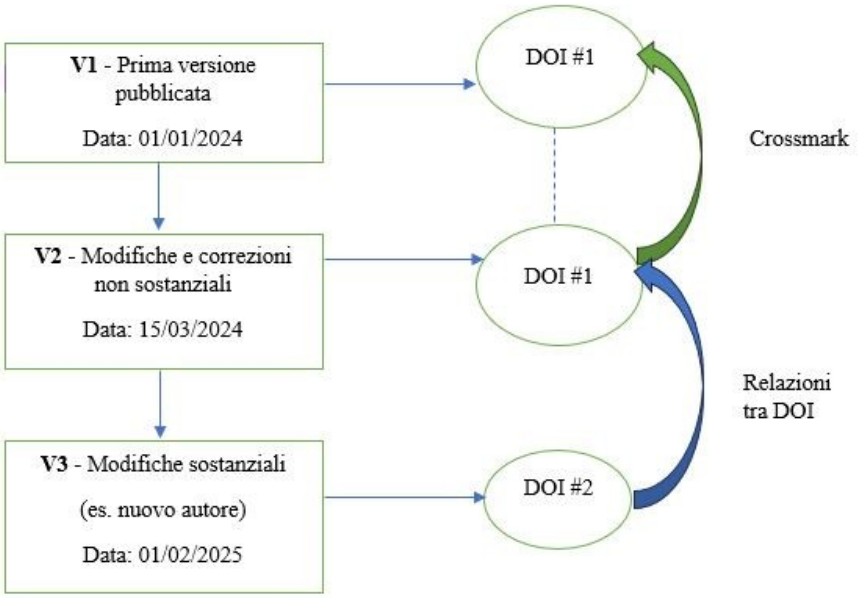

Fig. 1: Schematizzazione della rispondenza tra le diverse versioni di una stessa scheda e i relativi DOI.

In questo modo saranno proprio i metadati a garantire la tracciabilità e la rispondenza tra il DOI e la versione di un certo contenuto scientifico, così come era stato pubblicato in una specifica data.

#3 – Scelta dei metadati per il *versioning*
Un altro elemento chiave è rappresentato dai metadati per il *versioning*. Ogni versione di qualunque *record* della banca dati deve essere accompagnata da un insieme completo di informazioni standardizzate che ne documentino la genesi, l'autorialità, la versione. Tali metadati devono includere, oltre all'identificatore persistente e agli altri metadati per l'identificazione e la descrizione della risorsa digitale cui si riferisce, anche informazioni sulla data di rilascio, indicazioni sugli autori o responsabili delle modifiche, sulla natura del cambiamento e sulla relazione con le versioni precedenti.
Può essere utile introdurre anche la distinzione esplicita tra versioni. A differenza delle pubblicazioni testuali, le banche dati DH richiedono una classificazione delle versioni in base alla tipologia di cambiamento intervenuto. È pertanto consigliato distinguere almeno tra versioni che comportano modifiche strutturali (ad esempio allo schema o al modello concettuale), versioni che introducono cambiamenti nei contenuti o nella autorialità, e versioni che riflettono revisioni interpretative o metodologiche. Questa distinzione consentirà di attribuire un significato scientifico alle diverse versioni, evitando di ridurle a semplici aggiornamenti incrementali.
In tutti i casi, si utilizzerà l'elemento contenitore <crossmark> per registrare le informazioni circa le varie versioni, le modifiche apportate, e inserire i link per consultare le versioni precedenti; in questo modo l'utente potrà essere informato delle modifiche, ed eventualmente accedere alla versione (anche precedente) di interesse. Le informazioni registrate in questo contenitore saranno poi esposte online quando il lettore accederà al servizio di Crossmark.

#4 – Archiviazione e accessibilità delle versioni
Prima ancora della citabilità delle versioni, occorre garantire la storicizzazione dei dati (e metadati) e la disponibilità online delle differenti versioni, pubblicate possibilmente ciascuna con un proprio indirizzo univoco (URL). Affinché le banche dati possano essere utilizzate come fonti scientifiche verificabili, è essenziale che le singole versioni siano consultabili e citabili, in modo tale da ricostruire il contesto dei risultati prodotti.

#5 – Pubblicazione della *Policy* di Crossmark
Per attivare il servizio Crossmark, occorre pubblicare online la Policy. Non viene eseguito da parte di Crossref nessun controllo circa il contenuto o le scelte effettuate;

il renderle pubbliche è una operazione di trasparenza nei confronti degli autori e degli utilizzatori.

#6 – Registrazione del DOI alla policy di Crossmark
Una volta pubblicata, occorre assegnare e registrare un DOI per la pagina della Policy; per la registrazione del DOI, in questo caso, si consiglia di utilizzare il formato *Posted-content* che Crossref suggerisce appositamente per contenuti anche non scientifici pubblicati su siti internet e blog.

#7 – Registrazione dei DOI per tutti gli oggetti della banca dati
Nelle banche dati, si può in generale prevedere uno scenario a tre livelli, così rappresentato gerarchicamente:

1. il livello del database, che corrisponde all'intera banca dati
2. il livello della collezione (e quindi il livello della raccolta di schede omogenee);
3. il livello del record, cioè della singola scheda o contenuto scientifico.

Questo scenario viene rappresentato nella Figura sottostante (Fig. 2)

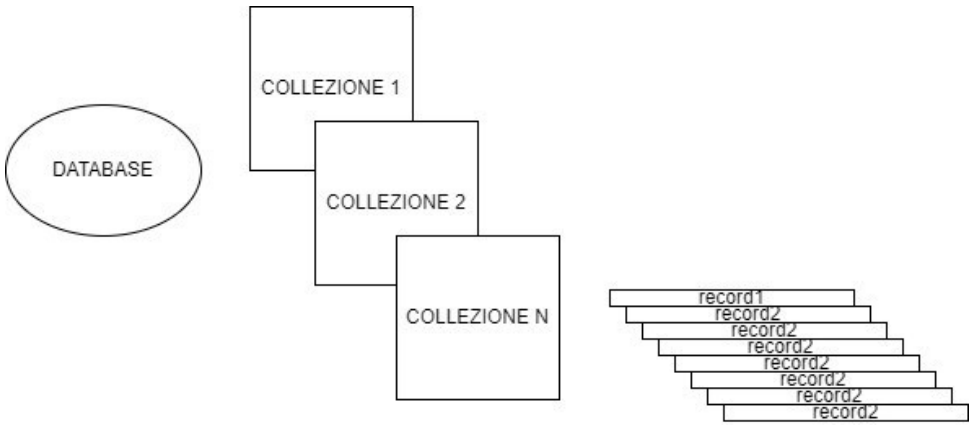

Fig. 2: La struttura tipica di un progetto DH, a tre livelli di descrizione: il livello globale del database, il livello intermedio rappresentato dalle collezioni; il livello di dettaglio costituito dalla singola scheda (*record*).

Qualora siano presenti tutti i tre livelli in esame, è evidente che per l'identificazione e l'assegnazione dei DOI si dovrebbe prevedere:

- un primo DOI (e relativi metadati) per il database nella sua interezza;
- un numero limitato di DOI (e relativi metadati) per le varie collezioni, uno per ciascuna collezione (che corrisponde di solito alla raccolta di schede prodotte da ciascuna unità o gruppo di lavoro o secondo altri criteri di omogeneità, ad esempio tematico o cronologico);

- un numero elevato di DOI (e relativi metadati), uno per ciascuna scheda (record).

I metadati da assegnare e compilare per ciascun DOI possono quindi essere di volta in volta riempiti con informazioni (metadati) specifiche per quel particolare DOI, oppure provenire da scelte effettuate in precedenza, cioè essere "ereditati" a cascata dai livelli superiori; questo perché, come visto in precedenza, esiste una struttura gerarchica tra database > collezione > record.
Nell'insieme dei metadati compilati è indispensabile che siano correttamente gestite le informazioni per garantire il collegamento ad altre entità (le cosiddette relazioni) e il tracciamento delle versioni. Anche per questo motivo, predisporre metadati ricchi di qualità è una attività dispendiosa, sia che venga effettuata manualmente, sia che preveda l'utilizzo di sistemi informatici che riducano i tempi necessari al *data-entry*, grazie a una gestione automatizzata dei valori ereditati.

#8 – Pubblicazione delle schede, con bottone di Crossmark e metadati
Al momento della pubblicazione, sarà inserito nella scheda (e nella versione PDF della stessa) l'apposito bottone di Crossmark per accedere alla storia delle versioni e verificare l'attualità di quella consultata (Fig. 3).

Fig. 3 – Bottone di Crossmark[13] che permette di verificare lo storico delle versioni per il contenuto visualizzato.

4. Conclusioni

L'analisi dello stato dell'arte ha mostrato come la riflessione sul *versioning* nelle Digital Humanities si sia sviluppata prevalentemente nell'ambito delle edizioni digitali, ma non si sia ancora estesa sistematicamente alle banche dati intese come infrastrutture di ricerca dinamiche. Il confronto tra *versioning* delle pubblicazioni e *versioning* delle banche dati ha permesso di chiarire come questi due modelli rispondano a esigenze e conseguenti metodologie d'approccio differenti.
In questo contributo si è proposta, in maniera innovativa, l'adozione sistematica del DOI insieme al servizio Crossmark per il *versioning* nelle banche dati umanistiche, con lo scopo di offrire informazioni attendibili di provenienza e tracciabilità, supportate da un insieme articolato di metadati. Tale approccio intende fornire un contributo significativo per colmare il divario tra la dinamicità intrinseca delle banche dati e le esigenze della comunicazione scientifica, offrendo così strumenti più chiari e trasparenti per il loro utilizzo e riuso nella ricerca.

Accanto alle motivazioni di carattere teorico, il presente contributo ha delineato i principali passaggi operativi necessari alla realizzazione dell'infrastruttura proposta, offrendo un inquadramento metodologico volto a favorirne la comprensione e l'adozione in contesti di ricerca umanistica. L'analisi approfondita dei tracciati dei metadati, così come la discussione delle soluzioni di natura più strettamente tecnica e implementativa, non rientra tuttavia negli obiettivi di questo lavoro e sarà sviluppata in un contributo successivo, nel quale tali aspetti verranno affrontati in modo sistematico e dettagliato.
Infine, per abbattere le barriere di ingresso (economiche e tecnologiche) e favorire un utilizzo più ampio nei vari progetti di banche dati nelle Digital Humanities, l'adozione di Crossmark dovrebbe essere integrata in una prospettiva infrastrutturale e istituzionale più larga. Il *versioning* delle banche dati non può essere demandato esclusivamente ai singoli progetti, ma richiede il coinvolgimento di istituzioni, archivi digitali e infrastrutture di ricerca, in grado di garantire la conservazione a lungo termine delle versioni e la loro accessibilità. In conclusione, l'allineamento con i principi FAIR e con le pratiche di *data curation* rappresenta un passaggio fondamentale per rendere il *versioning* una componente stabile e sostenibile delle Digital Humanities.

Ringraziamenti

Si ringrazia la dott.ssa Benedetta Fontanella per il prezioso contributo alla discussione nella definizione del tracciato dei metadati, nonché per il supporto nella conduzione degli esperimenti con Crossmark. Si ringraziano i revisori per il loro contributo propositivo durante la fase di referaggio.

Appendice

Porzione di metadati corrispondenti alla sezione Crossmark per tracciare il *versioning* (estratto dal tracciato di registrazione della scheda *Cimbalino*, DOI: 10.35948/OIM/DIZ_12_SI_10939)

```
<crossmark>
  <crossmark_version>1</crossmark_version>
  <crossmark_policy>10.35948/OIM/crossmark-policy</crossmark_policy>
  <custom_metadata>
    <assertion group_label="Versione 2" group_name="versione2" label="Data pubblicazione" name=
    "datazione" order="1">10/07/2025</assertion>
    <assertion group_label="Versione 2" group_name="versione2" label="A cura di" name="autore" order="2">
    Miraglia, Gianluca</assertion>
    <assertion group_label="Versione 2" group_name="versione2" label="Dettagli versione 2" name=
    "n_versione" order="3">Seconda edizione</assertion>
    <assertion group_label="Versione 1" group_name="versione1" label="Data pubblicazione" name=
    "datazione" order="1">21/10/2021</assertion>
    <assertion group_label="Versione 1" group_name="versione1" label="A cura di" name="autore" order="2">
    Miraglia, Gianluca</assertion>
    <assertion group_label="Versione 1" group_name="versione1" label="Dettagli versione 1" name=
    "n_versione" order="3">Prima edizione</assertion>
    <ai:program name="AccessIndicators">
      <ai:free_to_read />
      <ai:license_ref applies_to="vor">https://creativecommons.org/licenses/by/4.0/</ai:license_ref>
    </ai:program>
  </custom_metadata>
</crossmark>
```

Note

1. https://www.w3.org/TR/prov-overview/
2. https://www.italianismi.org/
3. https://dracor.org/
4. https://papyri.info/
5. https://epidoc.stoa.org/
6. https://www.codexsinaiticus.org/en/
7. https://www.perseus.tufts.edu/hopper/
8. https://digitallatin.org/
9. https://www.opengreekandlatin.org/
10. https://www.clarin.eu/
11. https://www.clarin.gr/en/CLARINELDataCollectionPolicy
12. https://zenodo.org/
13. https://www.crossref.org/images/documentation/Crossmark-check-for-updates.png

Bibliografia

- Broyles 2020 = Broyles, Paul A. *Digital Editions and Version Numbering*, Digital Humanities Quartely Volume 14 Number 2. 2020.
- Bürgermeister 2020 = Bürgermeister, Martina, *An Empirical Study of Versioning in Digital Scholarly Editions in* Cristina Marras, Marco Passarotti, Greta Franzini, Eleonora Litta (a cura di), *Atti del IX Convegno Annuale AIUCD. La svolta inevitabile: sfide e prospettive per l'Informatica Umanistica. 2020.* pp. 55-60
- Chavan et al, 2015 = Chavan, Amit and Huang, Silu and Deshpande, Amol and Elmore, Aaron J. and Madden, Sam and Parameswaran, Aditya; *Towards a Unified Query Language for Provenance and Versioning,* Proceedings of the 7th USENIX Conference on Theory and Practice of Provenance; USENIX Association, 2015
- Ciccarese et al, 2013 = Ciccarese, Paolo; Soiland-Reyes, Stian; Belhajjame, Khalid; Gray, Alasdair JG; Goble, Carole; Clark, Tim. *PAV ontology: provenance, authoring and Versioning*; Journal of Biomedical Semantics 2013, 4:37 - http://www.jbiomedsem.com/content/4/1/37
- Hauf et al, 2021 = Hauf, N.; Furholz, A.; Klaas, V.C.; Morger, J.; Šimukovi˘c, E.; Jaekel, M. Data Reuse in the Social Sciences and Humanities: Project Report of the SWITCH Innovation Lab "Repositories & Data Quality"; ZHAW Zurcher Hochschule fur Angewandte Wissenschaften: Winterthur, Switzerland, 2021. DOI: 10.21256/zhaw-2404
- Hendricks et al, 2020 = Ginny Hendricks, Dominika Tkaczyk, Jennifer Lin, Patricia Feeney; *Crossref: The sustainable source of community-owned scholarly metadata.* Quantitative Science Studies 2020; 1 (1): 414–427. DOI: https://doi.org/10.1162/qss_a_00022
- Klump et al, 2021 = Klump, J, Wyborn, L, Wu, M, Martin, J, Downs, RR, and Asmi, A., *Versioning Data Is About More than Revisions: A Conceptual Framework and Proposed Principles.* Data Science Journal, 20: 12, 2021. pp. 1–13. DOI: 10.5334/dsj-2021-012
- Lamney 2014 = Rachael Lammey, *CrossRef developments and initiatives: an update on services for the scholarly publishing community from CrossRef.* Science Editing 2014;1(1):13-18. DOI: https://doi.org/10.6087/kcse.2014.1.13
- Massari et al, 2025 = Massari, Arcangelo; Peroni, Silvio; Tomasi, Francesca; Heibi, Ivan. *Representing provenance and track changes of cultural heritage metadata in RDF: a survey of existing approaches.* Digital Scholarship in the Humanities, 2025 DOI: 10.1093/llc/fqaf076
- Meyer 2011 = Meyer, C.A., *Distinguishing published scholarly content with CrossMark.* Learned Publishing, 24 (2011) pp. 87-93. DOI: 10.1087/20110202
- Salucci 2022 = Salucci, Giovanni *Utilizzo del DOI (Digital Object Identifier) nei progetti di digital humanities* in Rivista DILEF - II, 2022/2 (gennaio-dicembre), pp. 308-319. DOI: 10.35948/DILEF/2023.4307
- Salucci 2023 = Salucci, Giovanni *Utilizzo del DOI (Digital Object Identifier) per la diffusione di progetti lessicografici digitali* in DILEF. Rivista digitale del Dipartimento di Lettere e Filosofia - 3 (2023), pp. 275-292. DOI: 10.35948/DILEF/2024.4327
- Wilkinson et al, 2016 = Wilkinson, M.D.; Dumontier, M.; Aalbersberg, I.J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L.B.; Bourne, P.E.; et al. The FAIR Guiding Principles for Scientific Data Management and Stewardship. Sci. Data 2016, 3, 160018.