AIDAinformazioni
Rivista semestrale di Scienze dell'Informazione
Anno 42
N. 3-4 – luglio-dicembre 2024

Contributi

Rubriche

mundaneum

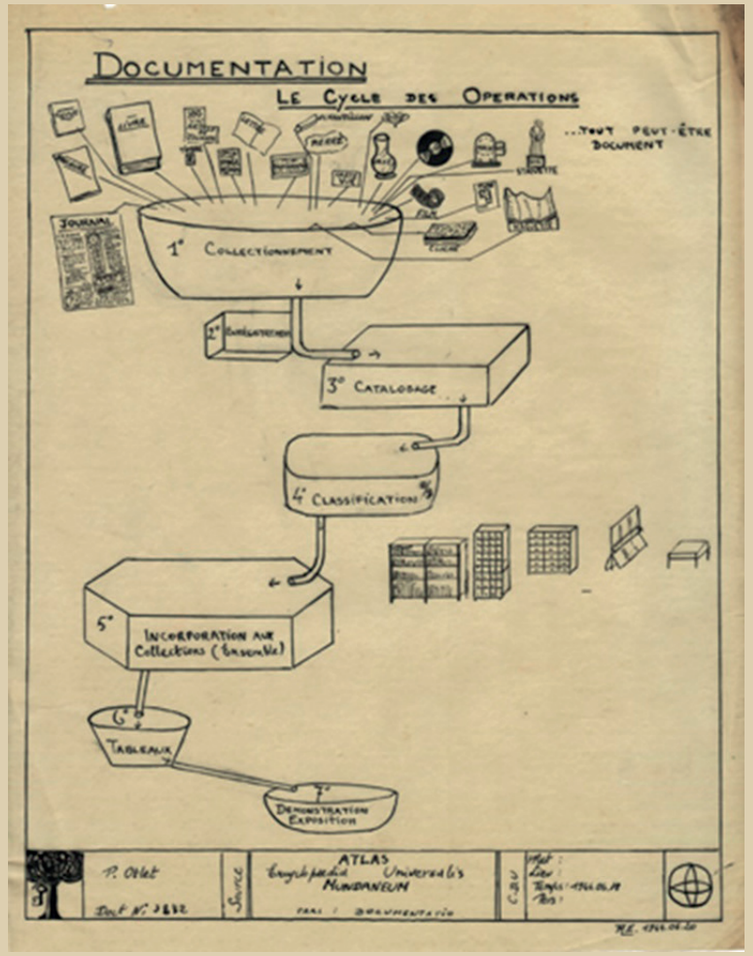
In copertina
Disegno di Paul Otlet, Collections Mundaneum, centre d'Archives, Mons (Belgique).

ISBN 979-12-5965-456-4
ISSN 1121-0095
9 791259 654564
9 770112 100950

AIDAinformazioni Anno 42 – N. 3-4 – luglio-dicembre 2024

AIDAinformazioni
RIVISTA SEMESTRALE DI SCIENZE DELL'INFORMAZIONE

NUMERO 3-4
ANNO 42
LUGLIO-DICEMBRE 2024



CACUCCI EDITORE
BARI

AIDAinformazioni

Rivista semestrale di Scienze dell'Informazione

Fondata nel 1983 da Paolo Bisogno

Proprietario della rivista:
Università della Calabria

Editrice: Cacucci Editore S.a.s.
Via D. Nicolai, 39 – 70122 Bari (BA)
www.cacucciedi tore.it
e-mail: riviste@cacucciedi tore.it
Telefono 080/5214220

AIDAinformazioni

Rivista semestrale di Scienze dell'Informazione

«AIDAinformazioni» è una rivista scientifica che pubblica articoli inerenti alle Scienze dell'Informazione, alla Documentazione, all'Archivistica, alla Gestione Documentale e all'Organizzazione della Conoscenza ma amplia i suoi confini in ulteriori campi di ricerca affini quali la Terminologia, la Linguistica Computazionale, la Statistica Testuale, ecc. È stata fondata nel 1983 quale rivista ufficiale dell'Associazione Italiana di Documentazione Avanzata e nel febbraio 2014 è stata acquisita dal Laboratorio di Documentazione dell'Università della Calabria. La rivista si propone di promuovere studi interdisciplinari oltre che la cooperazione e il dialogo tra profili professionali aventi competenze diverse, ma interdipendenti. I contributi pubblicati affrontano questioni teoriche, metodologie adottate e risultati ottenuti in attività di ricerca o progettuali, definizione di approcci metodologici originali e innovativi, analisi dello stato dell'arte, ecc.

«AIDAinformazioni» è riconosciuta dall'ANVUR come rivista di Classe A per l'Area 11 – Gruppo Scientifico Disciplinare 11/HIST-04 – Scienze del libro, del documento e storico-religiose e come rivista scientifica per le Aree 10 – Scienze dell'antichità, filologico-letterarie e storico-artistiche; 11 – Scienze storiche, filosofiche, pedagogiche e psicologiche; 12 – Scienze giuridiche; 14 – Scienze politiche e sociali. È anche annoverata dall'ARES (Agence d'évaluation de la recherche et de l'enseignement supérieur) tra le riviste scientifiche dell'ambito delle Scienze dell'Informazione e della Comunicazione. La rivista è, inoltre, indicizzata in: ACNP – Catalogo Italiano dei Periodici; BASE –Bielefeld Academic Search Engine; ERIH PLUS – European Reference Index for the Humanities and Social Sciences – EZB – Elektronische Zeitschriftenbibliothek – Universitätsbibliothek Regensburg; Gateway Bayern; KVK – Karlsruhe Virtual Catalog; The Library Catalog of Georgetown University; SBN – Italian union catalogue; Ulrich's; Union Catalog of Canada; LIBRIS – Union Catalogue of Swedish Libraries; Worldcat.

I contributi sono valutati seguendo il sistema del *double blind peer review*: gli articoli ricevuti sono inviati in forma anonima a due referee, selezionati sulla base della loro comprovata esperienza nei topics specifici del contributo in valutazione.

AIDAinformazioni

Anno 42

N. 3-4 – luglio-dicembre 2024

CACUCCI EDITORE

BARI

© 2024 Cacucci Editore – Bari
Via Nicolai, 39 – 70122 Bari – Tel. 080/5214220
http://www.cacucciedit ore.it e-mail: info@cacucci.it

Sommario

Contributi

Rubriche

Contributi

AIDAinformazioni
ISSN 1121–0095
ISBN 979-12-5965-456-4
DOI 10.57574/596545641
pag. 9-28 (luglio-dicembre 2024)

Il nuovo regolamento eIDAS e alcune "quisquilie" archivistiche

Alessandro Alfier*

Abstract: The essay analyses the rules relating to electronic archiving services, as regulated by the new European regulation called eIDAS 2. The analysis, starting from the use of the ambiguous term archiving, tries to clarify the relationship between the new electronic archiving services, on the one hand, and the contexts and systems of records management, digital preservation and digital custody, on the other. It therefore seeks to contextualize the new archiving services within the life cycle of electronic records, as defined by archival theory and methodology. The analysis finally includes a comparison between the paradigm of the electronic archiving services and the Italian model of digital preservation, highlighting some of its functional anomalies.

Keywords: Digital preservation, eIDAS, Electronic archiving service, Electronic record, Records management.

1. Introduzione

Negli ultimi mesi si è molto dibattuto del regolamento europeo eIDAS, che nella sua nuova versione innova la precedente risalente al 2014[1]. Le esigenze che hanno indotto a modificare il regolamento sono molteplici, come anche emerge dai considerando che anticipano gli articoli del testo normativo vero e proprio. Da un punto di vista generale, si può però ritenere che eIDAS 2 – come è stata ribattezzata la nuova versione del regolamento – miri soprattut-

* Ministero dell'Economia e delle Finanze – Direzione dei sistemi informativi e dell'innovazione, Roma, Italia. alessandro.alfier@mef.gov.it.

[1] Regolamento (UE) n. 2024/1183 del Parlamento europeo e del Consiglio dell'11 aprile 2024, entrato in vigore il 20 maggio 2024 e che modifica il Regolamento (UE) n. 910/2014 del Parlamento europeo e del Consiglio del 23 luglio 2014 in materia di identificazione elettronica e dei servizi fiduciari per le transazioni elettroniche nel mercato interno. Il testo consolidato del Regolamento n. 910/2014, a seguito delle modifiche introdotte nel 2024, è consultabile all'indirizzo https://eur-lex.europa.eu/legal-content/IT/TXT/?uri=CELEX%3A02014R0910-20240520.

to a migliorare l'efficacia del precedente strumento normativo[2], per realizzare una maggiore armonizzazione contro i rischi e i costi dell'odierna frammentazione, dovuta all'uso di soluzioni nazionali divergenti in tema di identità digitale e di servizi fiduciari. L'orizzonte a cui si guarda come a un obiettivo è, dunque, quello di un'Unione europea in cui si possa avere, finalmente, un coeso ecosistema digitale a dispetto delle barriere transfrontaliere[3]. Tanto è vero che il nuovo regolamento si preoccupa, in particolare, di garantire un livello uniforme per la qualità, l'affidabilità e la sicurezza dei processi di identificazione elettronica e dei servizi fiduciari, indipendentemente dal luogo in cui essi prendono forma[4]. Proprio questa finalità di carattere generale spiega poi il perché del ricorso alla normazione sovranazionale. Lo sviluppo di un effettivo quadro europeo per l'identità digitale e per i servizi fiduciari non può essere conseguito, in misura sufficiente, dalle azioni dei singoli stati membri, potendo invece essere più efficacemente supportato da un intervento dell'Unione europea, chiamata allora a legiferare in base al principio di sussidiarietà e a quello di proporzionalità[5] e ricorrendo allo specifico strumento del regolamento in luogo della direttiva[6].

Tra le molte innovazioni realizzate, in eIDAS 2 si ritrova anche una novità che, in prima battuta, sembra più direttamente riconducibile alle dinamiche del mercato e agli sviluppi tecnologici in essere: quella dell'espansione dell'elenco dei servizi fiduciari presi in considerazione dallo stesso regolamento. Rispetto al testo del 2014, oggi si aggiungono tre nuovi servizi fiduciari la cui prestazione, anche qualificata, è normata da eIDAS: quelli legati all'archiviazione elettronica, ai registri elettronici e ai dispositivi per la generazione di firme e sigilli elettronici a distanza. Con riferimento ai primi, che probabilmente più interessano coloro che si occupano di scienze e tecniche documentali, non sembra che la loro inclusione nel regolamento sia stata motivata da particolari riflessioni, al di là delle esigenze emerse dal mercato delle tecnologie per

---

[2] (Parlamento europeo e Consiglio dell'Unione europea. 2024, considerando n. 1).

[3] (Parlamento europeo e Consiglio dell'Unione europea. 2024, considerando n. 7).

[4] (Parlamento europeo e Consiglio dell'Unione europea. 2024, considerando n. 71).

[5] (Parlamento europeo e Consiglio dell'Unione europea. 2024, considerando n. 76).

[6] Il caso eIDAS sembra rientrare in un preciso orientamento adottato di recente dagli organismi legislativi europei: «sotto il profilo della tecnica della normazione, dopo una prima fase più rispettosa delle prerogative dei singoli stati membri, è prevalsa la tendenza all'adozione di atti direttamente vincolanti. A partire dai primi anni duemila, si è passati dalle direttive di armonizzazione alla moltiplicazione di strumenti regolamentari. Il superamento del modello della direttiva è stato giustificato con la necessità di eliminare le barriere al funzionamento del mercato interno, riducendo la frammentazione normativa – storicamente un ostacolo alle dinamiche concorrenziali – e contribuendo a raggiungere una maggiore certezza giuridica, attraverso un insieme armonizzato di regole fondamentali che fanno ricorso a standard tecnologici comuni» (Belisario 2024, 34).

l'informazione e la comunicazione. A tal proposito, in un considerando che precede gli articoli veri e propri si legge:

> molti stati membri hanno introdotto requisiti nazionali per i servizi che forniscono un'archiviazione elettronica sicura e affidabile al fine di consentire la conservazione a lungo termine di dati elettronici e documenti elettronici, nonché per i servizi fiduciari associati. Al fine di garantire la certezza giuridica, la fiducia e l'armonizzazione in tutti gli stati membri, è opportuno istituire un quadro giuridico per i servizi di archiviazione elettronica qualificati, ispirato al quadro per gli altri servizi fiduciari di cui al presente regolamento. Il quadro giuridico per i servizi di archiviazione elettronica qualificati dovrebbe offrire ai prestatori di servizi fiduciari e agli utenti un pacchetto di strumenti efficienti che comprenda requisiti funzionali per il servizio di archiviazione elettronica, nonché chiari effetti giuridici in caso di utilizzo di un servizio di archiviazione elettronica qualificato. Tali disposizioni dovrebbero applicarsi ai dati elettronici e ai documenti elettronici creati in forma elettronica e ai documenti cartacei che sono stati scannerizzati e digitalizzati[7].

Al di là di quali siano state le reali motivazioni del legislatore europeo, si può però osservare che tale inclusione va a colmare una lacuna importante della precedente versione di eIDAS. Se si vuole costruire uno spazio giuridico all'interno dell'Unione europea, in cui il ricorso alle tecnologie digitali permetta ai cittadini, alle imprese e ai soggetti pubblici di interagire in piena fiducia e senza incertezze, risulta allora inevitabile affrontare anche il tema della credibilità dei documenti prodotti elettronicamente[8]. Ciò in ragione del fatto che da sempre lo strumento documentale ha una natura performativa che gli permette di plasmare, con modalità efficaci e collaudate, i rapporti tra i membri del consesso sociale (Alfier 2023; Yeo 2017; Yeo 2010). La previsione di servizi per l'archiviazione elettronica, che permettano di riproporre su scala digitale la forza performativa del documento già apprezzata nel tradizionale contesto analogico, costituisce di fatto una risposta a quella esigenza, al di là del grado di consapevolezza presente negli organismi europei che hanno emanato eIDAS 2. Forse questa novità è anche la cartina di tornasole di una maggiore maturità sui temi più strettamente documentali da parte dei decisori europei. Alcuni osservatori, infatti, si sono spinti ad affermare che dall'esigenza

---

[7] (Parlamento europeo e Consiglio dell'Unione europea. 2024, considerando n. 66).

[8] Osserva a questo proposito Patrizia Sormani: «uno dei principali scopi del [nuovo] regolamento eIDAS è quello di facilitare le transazioni elettroniche […], permettendo di effettuare operazioni in modo sicuro. Ogni transazione, di qualunque genere sia, di fatto genera dati che aggregati nelle diverse forme divengono documento informatico, anch'esso da tutelare. Con il nuovo regolamento si assiste ad un'evoluzione […] Quanto previsto fino ad ora non basta più, non è più sufficiente a garantire le transazioni elettroniche e le persone poste al centro della transazione, in quanto soggetti attivi protagonisti. Va tutelata la persona, i suoi dati, favorite le transazioni sicure e *tutelati i documenti e dati informatici generati dalle stesse*» (Sormani 2024, 129. Il corsivo è di chi scrive [*N.d.A.*]).

di regolare i servizi per l'archiviazione elettronica traspare un cambiamento di approccio: finalmente «si riconosce l'insufficienza della firma [e delle connesse tecniche criptografiche] per garantire l'integrità e l'autenticità dei documenti digitali a lungo termine» (Belisario 2024, 40)[9]. Tra l'altro, nel considerando sopra citato, si insiste su una nozione piuttosto ampia di documento elettronico: inclusiva non solo di quei documenti elettronici testuali che si presentano con forme facilmente intelligibili dagli umani, in quanto richiamano da vicino quelle della documentazione cartacea, ma anche dei sempre più diffusi flussi di dati, strutturati in linguaggi *machine-readable*. Nozione questa che, fortunatamente, appare coerente con quanto già previsto dal nostro legislatore nazionale, che nel *Codice dell'amministrazione digitale* definisce il documento informatico come quella particolare fattispecie di documento elettronico che veicola non solo la rappresentazione informatica di atti o fatti giuridicamente rilevanti, ma anche la rappresentazione di dati significativi sul piano giuridico (Decreto Legislativo 7 marzo 2005, n. 82)[10].

Sembra così emergere uno scenario piuttosto rassicurante per coloro che hanno a cuore i temi documentali. La più recente versione di eIDAS pare, infatti, dischiudere una nuova fase: quella in cui, nel dominio dei servizi fiduciari e non sottoposti alla regolamentazione europea, il documento digitale e il correlato archivio elettronico trovano la loro giusta collocazione, aprendosi al contempo degli spazi di azione per le tecniche di conservazione documentale e per quelle teorie archivistiche che hanno saputo – più di altre – aggiornarsi sotto la spinta dell'attuale scenario tecnologico. Tutto bene, dunque, sotto il cielo di eIDAS 2?

2. I servizi di archiviazione elettronica: un'incerta collocazione tra gestione documentale e conservazione?

Nel dibattito italiano sono molti coloro che, nel commentare la nuova versione del regolamento europeo, hanno evidenziato come la previsione dei servizi per l'archiviazione elettronica costituisca una grande opportunità per il

[9] Andrea Lisi indica addirittura un cambio di passo rivoluzionario: «in eIDAS 2 finalmente ci si occupa (e ci si preoccupa) di "e-archiving", [dunque] di registrazioni affidabili e di documenti in grado di salvaguardare la nostra memoria digitale, preservando la fonte di provenienza. E, in un mondo ormai tristemente caratterizzato dalla diffusione sistematica di deep fake e fake news […] questa attenzione alle regole dell'archiviazione elettronica finisce per essere un'attività rivoluzionaria». Lo stesso Lisi sottolinea poi come tale cambiamento di approccio sia anche dovuto al riconoscimento della maturità dell'elaborazione normativa e delle pratiche nazionali di alcuni stati membri, tra cui il nostro: «possiamo senz'altro sottolineare che l'Italia abbia fatto da apripista per favorire questa evoluzione del senso stesso e della necessità (oltre che dei modi e mezzi) di documentare affidabilmente nel mondo digitale» (Lisi 2024, 44, 50).

[10] Art. 1, comma 1, lettera p.

nostro paese. L'Italia, infatti, sotto la spinta di una normativa consolidata nel campo dei servizi per la conservazione digitale, può vantare una maturità non solo in termini di tecnologia, ma anche di organizzazione e di processi destinati all'archiviazione elettronica. Ricorda, a questo proposito, una studiosa come Maria Guercio che «l'esperienza italiana è di grande importanza per l'Europa, dato che nessun altro paese ha alle spalle un'applicazione ventennale di norme specifiche sulla conservazione e dieci anni di attività nel campo dell'accreditamento e, quindi, di analisi critica dei sistemi conservativi digitali esistenti» (Guercio 2023)[11]. Ancor più ottimistiche appaiono poi le previsioni espresse da un soggetto direttamente chiamato in causa, l'associazione che riunisce le imprese che operano nel settore dei servizi per la conservazione digitale: «è evidente come l'esperienza fino ad oggi maturata in ambito italiano rappresenterà un valore aggiunto e giocherà un ruolo importante [rispetto allo sviluppo dei servizi per l'archiviazione elettronica previsti dal regolamento europeo] […] Dunque, non ci aspettiamo impatti considerevoli e, anzi, è naturale pensare alla possibilità concreta che il modello italiano possa essere esportato all'interno di un contesto europeo» (Pomarico 2024). Nelle retrovie di tanto ottimismo, giustificato per diverse ragioni se retrospettivamente si guarda alla prima versione del regolamento europeo che non affrontava il dominio documentale, mi sembra però di intravvedere alcune criticità piuttosto dense. Queste si fanno più visibili se valutiamo i servizi per l'archiviazione elettronica, previsti da eIDAS 2, alla luce della sostenibilità del ciclo di vita del documento digitale e degli archivi elettronici.

Il groviglio si evidenzia a partire dalla scelta terminologica compiuta dagli estensori del regolamento europeo: il ricorso, piuttosto infelice, al termine di archiviazione[12]. Infelice perché foriero di ambiguità. Lo segnala, in qualche modo, anche il white paper *eIDAS Trust Electronic Archiving Services supported*

[11] «Nonostante alcuni limiti inziali, il modello seguito fino al 2020 in Italia – dopo un lungo e tortuoso percorso iniziato addirittura nel 1994 – ha avuto il merito di responsabilizzare il settore pubblico sulle criticità e difficoltà del mantenimento nel tempo delle memorie digitali e costretto le aziende di settore a confrontarsi con misure di qualità e audit basate su standard internazionali […] L'approvazione con determinazione AgID del 2020 delle *Linee guida per la formazione, gestione e conservazione di documenti informatici* ha consolidato e ulteriormente qualificato il percorso adottato in Italia, anche se si è reso necessario – in quell'occasione proprio a seguito di un intervento della Commissione europea chiamata a esprimersi preventivamente sulla normativa italiana – abolire il regime nazionale di accreditamento dei depositi di conservazione e stabilire un più limitato sistema di controlli definito da un regolamento sui criteri per la fornitura dei servizi di conservazione dei documenti informatici del dicembre 2021. Non è escluso che proprio il passaggio obbligato a Bruxelles, per l'approvazione della nostra regolamentazione nazionale […] abbia richiamato l'attenzione delle istituzioni europee sulla necessità di un intervento normativo sovra-nazionale di allineamento e normalizzazione, quello appunto che […] si è concretizzato con la modifica del regolamento eIDAS» (Guercio 2023).

[12] *Archiving* nella versione inglese di eIDAS 2 e *archivo* in quella spagnola.

by the eArchiving Initiative[13]. Questa fonte riconosce che «in some contexts, digital preservation could be used as a synonym for electronic archiving» (eArchiving Initiative 2024, 4): come a dire che in altri contesti l'espressione archiviazione elettronica potrebbe, invece, essere intesa con un significato non avvicinabile a quello di preservazione digitale o – secondo gli usi terminologici prevalenti nel nostro paese – di conservazione digitale. Il termine archiviazione, infatti, nel contesto della disciplina, delle metodologie e delle pratiche archivistiche può essere dilatato per identificare delle sedimentazioni documentarie che si collocano in fasi assai diverse del ciclo di vita documentale: tanto in quella della gestione documentale, quanto in quella della conservazione digitale vera e propria (Giunta e Marti 2024, 163)[14]. Se poi si guarda alla

[13] Nella pagina web di presentazione del white paper sono illustrate le ragioni che hanno motivato la sua redazione: «the eArchiving Initiative of the European Commission has produced a White Paper setting out how we can support the implementation of new electronic archiving services established under the amended Regulation (EU) No 910/2014, known as eIDAS 2 […] In essence, the eArchiving Initiative represents a concerted endeavour to fortify the digital infrastructure of the European Union […] As a result of the outcomes of the Archiving Initiative and preceding projects, a set of tools, specifications and procedures can be offered to support the implementation of the electronic archiving trust services […] The rationale behind the eIDAS 2 electronic archiving trust service is largely consistent with the purpose and objectives of the eArchiving Initiative» (European Union 2024). Quanto poi alla eArchiving Initiative, si tratta di un progetto finanziato dalla Commissione europea e gestito per conto di quest'ultima dall'E-ARK Consortium. Tanto è vero che l'eArchiving Initiative si può considerare come l'ultima di una serie di iniziative, sponsorizzate sempre dalla Commissione europea, che nel tempo hanno coinvolto l'E-ARK Consortium (il primo E-ARK project, seguito poi dall'E-ARK4ALL e dall'E-ARK3 project). Questo consorzio ha la finalità di sviluppare gli standard, gli strumenti e le buone pratiche con cui affrontare le sfide correlate alla preservazione non solo dei documenti elettronici, ma di una pluralità di forme con cui si rappresentano oggi i contenuti digitali, diffondendo poi il patrimonio di conoscenze elaborato. Nella sua funzione di supporto all'implementazione dei nuovi servizi di archiviazione elettronica previsti da eIDAS 2, l'eArchiving Initiative partecipa al CEN/TC 468 Technical Committee on preservation of digital information, che a sua volta ha avviato la definizione di requisiti funzionali e tecnologici inerenti proprio all'archiviazione elettronica. Dall'attività di questo organismo potrebbero derivare quelle indicazioni tecniche di maggior dettaglio previste dalla nuova versione di eIDAS: «entro il 21 maggio 2025 la Commissione […] stabilisce un elenco di norme di riferimento [*reference standards* nella versione in lingua inglese del regolamento] e, se necessario, stabilisce specifiche e procedure applicabili ai servizi di archiviazione elettronica qualificati. Si presume che i requisiti dei servizi di archiviazione elettronica qualificati siano rispettati ove un servizio di archiviazione elettronica qualificato sia conforme a tali norme, specifiche e procedure» (Testo consolidato del Regolamento (UE) n. 910/2014, art. 45 undecies, comma 2).

[14] A titolo d'esempio nella normativa italiana in vigore il termine di archiviazione compare: all'art. 50, comma 4 del D.P.R. 28 dicembre 2000, n. 445, riferito al contesto della gestione documentale; all'art. 68, comma 3 dello stesso D.P.R., riferito però all'ambito della conservazione permanente; in più punti della *Linee guida sulla formazione, gestione e conservazione dei documenti informatici* (Agenzia per l'Italia Digitale 2021), riferito sia alla gestione documentale che alla conservazione digitale.

definizione che il regolamento europeo fornisce per i servizi di archiviazione elettronica, l'ambiguità non sembra alleggerirsi più di tanto. Nella versione inglese di eIDAS l'*electronic archiving* è definito come «a service ensuring the receipt, storage, retrieval and deletion of electronic data and electronic documents in order to ensure their durability and legibility as well as to preserve their integrity, confidentiality and proof of origin throughout the preservation period»[15]. A una prima lettura, sembrerebbe che si stia facendo riferimento a un servizio che non possa funzionalmente essere collocato nel contesto di un sistema di gestione documentale. Quel servizio, infatti, appare destinato a ricevere dei dati e dei documenti elettronici originariamente generati altrove e su cui poi, in una fase successiva a quella di produzione, lo stesso è chiamato a svolgere una serie di funzioni. Questa impressione iniziale muta però se ci soffermiamo a raffrontare tale definizione normativa con la nozione di *records management* che si ritrova nel principale standard di gestione documentale, l'ISO 15489-1:2016: «field of management responsible for the efficient and systematic control of the creation, receipt, maintenance, use and disposition of records» (International organization for standardization 2016, 3). In questo confronto colpiscono alcuni parallelismi. In primo luogo il ricorso comune al termine *receipt*. Coloro che sono più esperti nel campo del *records management* sanno bene che un sistema di gestione documentale non è propriamente un contesto in cui il documento elettronico possa essere generato, ma in termini più precisi esso si identifica con uno scenario in cui l'oggetto documentale è ricevuto o meglio catturato: *capture* è, non a caso, il termine inglese che ricorre in questo standard ISO[16] e che con riferimento alla terminologia italiana di settore potremmo tradurre con registrazione. Cattura con cui sottomettere il documento alle policy, alle funzioni e alle responsabilità che compongono la gestione documentale, così da farlo sedimentare, con modalità coerenti e

[15] Testo consolidato del Regolamento (UE) n. 910/2014, art. 3, comma 48. Al comma 49 dello stesso articolo si definisce poi il servizio qualificato di archiviazione elettronica come un servizio «fornito da un prestatore di servizi fiduciari qualificato e che soddisfa i requisiti di cui all'articolo 45 undecies». Quest'ultimo, a sua volta, stabilisce che «i servizi di archiviazione elettronica qualificati soddisfano i requisiti seguenti: a) sono forniti da prestatori di servizi fiduciari qualificati; b) utilizzano procedure e tecnologie in grado di garantire la durabilità e la leggibilità dei dati elettronici e dei documenti elettronici oltre il periodo di validità tecnologica e almeno per tutto il periodo di conservazione legale o contrattuale, preservandone nel contempo l'integrità e l'esattezza dell'origine; c) assicurano che tali dati elettronici e tali documenti elettronici siano conservati in modo tale da essere protetti dal rischio di perdita e alterazione, ad eccezione delle modifiche riguardanti il loro supporto o il loro formato elettronico; d) consentono alle parti autorizzate, facenti affidamento sulla certificazione, di ricevere una relazione in un modo automatizzato, in cui si conferma che i dati elettronici e i documenti elettronici consultati da un archivio elettronico qualificato godono della presunzione di integrità dei dati, dall'inizio del periodo di conservazione fino al momento della consultazione».

[16] Lo stesso standard prevede uno specifico processo denominato *capturing records* (International organization for standardization 2016, 16-17).

controllate, nell'archivio digitale dell'organizzazione e a prescindere da dove originariamente lo stesso documento sia stato generato. Tanto è vero che lo stesso standard ISO, quando si addentra più nel dettaglio del concetto di *records system*, tralascia del tutto il termine *creation*: «information system which captures, manages and provides access […] to records over time […] A records system can consist of technical elements, such as software […] and non-technical elements including policy, procedures, people and other agents, and assigned responsabilities» (International organization for standardization 2016, 3). Se consideriamo, infatti, il dominio di una certa organizzazione le fonti di produzione documentale sono diverse, ma in senso stretto tutte esterne al sistema di gestione documentale propriamente detto: lo sono certamente quelle da cui derivano i documenti che provengono dal fronte esterno dell'organizzazione, ma poi anche quelle da cui si originano i flussi documentali ad essa interni. Tra quest'ultimi si devono annoverare non solo i documenti generati per mezzo di software di *document management system* o di *content management system* e finalizzati alla creazione collaborativa di contenuti – software più o meno integrati con il sistema di gestione documentale – ma anche i documenti prodotti dai cosiddetti *applicativi verticali*, finalizzati non a un uso trasversale da parte dell'organizzazione, ma rispondenti alle esigenze legate ad ambiti settoriali di attività dell'ente[17]. Ebbene questa pluralità di fonti documentali riesce ad alimentare un archivio digitale – unitario e completo – solo se il sistema di gestione documentale si dimostra efficace nei processi di cattura[18]. Tanto che potremmo dire che quella frammentazione di cui è portatore lo scenario digitale, in particolare sul fronte documentale, trova un possibile baluardo solo in sistemi di *records management* pervasisi nelle loro capacità di cattura e

[17] Ad esempio gli applicativi per la gestione del personale, per la gestione del budget o per il controllo di gestione.

[18] Non a caso sul processo di cattura o registrazione insistono, come su un aspetto davvero rilevante, i *Model Requirements for the Management of Electronic Records*, al punto da prevedere proprio per tale processo una serie di requisiti ad hoc*:* «documents made or received in the course of business become records when they are set aside, that is, "captured" into the ERMS [Electronic records management system] […] In many cases, documents that are set aside, or captured, become records by being bound to a business process, for example as happens in a workflow […] In other cases there may be a policy that every document relating to a business matter must become a record, even if it does not formally participate in a business process. In yet other circumstances however, the process of capture will be initiated selectively by a user. Determination of which documents should be captured into a records system should be based on an analysis of the regulatory environment, business and accountability requirements and the risk of not capturing the records […] Electronic documents that are generated or received in the course of business processes originate from both internal and external sources […] They may arrive through different communication channels e.g. local area network, wide area network, electronic mail, facsimile, letter post (to be scanned) […] A flexible input system is required to capture documents with good management controls so that these diverse requirements are addressed» (European Commission 2001, 12, 39).

quindi massimamente ricettivi rispetto alla miriade di fonti documentali che investono, esternamente e internamente, una certa organizzazione.

Le similarità che ci inducono a ritenere, almeno in teoria, il servizio di archiviazione elettronica previsto da eIDAS 2 più ubiquo rispetto a quanto appaia a prima vista – tanto da poterlo avvicinare anche alla fase di gestione documentale e non solo a quella di conservazione digitale – sono però anche altre. Insistendo ancora nel raffronto con l'ISO 15489-1:2016, ci si accorge di una definizione di documento che ricalca quella prevista dal regolamento europeo, per quanto attiene al superamento del tradizionale dualismo tra dato e documento. Lo standard, infatti, definisce quest'ultimo proprio nei termini di informazione: un'informazione da usarsi come evidenza[19], per provare l'esecuzione di una transazione e utilizzabile pertanto anche a fini giuridici (International organization for standardization 2016, 2). In questo modo l'ISO sembra rifarsi a una nozione di documento la cui cattura, in un idoneo sistema di gestione documentale, lo rende un *surrogato* rappresentativo. Capace quindi, con piena validità ed efficacia, di stare in luogo di quanto rappresentato e dunque di una porzione di realtà[20] che, dal qui ed ora in cui si è concretizzata, deve essere proiettata e mantenuta in vita – appunto in forma di evidenza o surrogato – per un tempo e uno spazio potenzialmente indefiniti. Una nozione che quindi inevitabilmente riporta in auge le capacità performative del documento nel plasmare i rapporti tra i consociati. Questo rimando alla dimensione di evidenza è interessante perché, a sua volta, richiama da vicino alcune previsioni di eIDAS proprio in tema di servizi per l'archiviazione elettronica:

> ai dati elettronici e ai documenti elettronici conservati mediante un servizio di archiviazione elettronica non vengono negati gli *effetti giuridici* né l'ammissibilità come prova in procedimenti giudiziari per il solo motivo della loro forma elettronica o perché non sono conservati mediante un servizio di archiviazione elettronica qualificato […] I dati elettronici e i documenti elettronici conservati mediante un servizio di archiviazione elettronica qualificato godono della *presunzione* della loro integrità e della correttezza della loro origine per la durata del periodo di conservazione da parte del prestatore di servizi fiduciari qualificato[21].

---

[19] *Evidence* nell'originale testo inglese dello standard, che addirittura ricorre all'espressione *authoritative evidence of business* per delineare l'essenza di un documento in quanto catturato e mantenuto all'interno di un sistema di gestione documentale (International organization for standardization 2016, 3-4).

[20] O per meglio dire di una *transaction*, tanto per usare la terminologia propria dello standard: «transaction: smallest unit of a work process […] consisting of an exchange between two or more participants or systems […] work process is one or more sequences of actions required to produce an outcome that complies with governing rules» (International organization for standardization 2016, 3).

[21] Testo consolidato del Regolamento (UE) n. 910/2014, art. 45 undecies. I corsivi sono di chi scrive [*N.d.A.*].

Anche in questo passaggio del regolamento europeo – come nello standard ISO dedicato alla gestione documentale – si sta, dunque, trattando del documento elettronico come evidenza, con conseguenti effetti sul piano giuridico, un'evidenza che in alcuni casi può assumere una particolare forza: quella della presunzione di genuinità, tale per cui la falsità del documento non può essere semplicemente dichiarata, ma deve essere puntualmente dimostrata. Tra l'altro questo stesso documento può operare performativamente, come evidenza o surrogato, solo nella misura in cui esso possegga delle precise qualità, che gli sono indistintamente garantite sia dall'assoggettamento ai servizi di archiviazione elettronica previsti dalla normativa europea, sia dalla cattura e mantenimento da parte di un sistema di gestione documentale conforme allo standard ISO. In eIDAS 2, infatti, il ricorso ai nuovi servizi ha il fine di garantire ai dati e ai documenti elettronici la «durability and legibility as well as […] their integrity […] and proof of origin»[22]. Qualità queste che, sostanzialmente, trovano un loro esatto corrispettivo nell'ISO 15489-1:2016, allorquando si afferma che la gestione documentale si basa su documenti intesi come informazioni che «regardless of form or structure, are authoritative evidence of business when they possess the characteristics of authenticity, reliability, integrity and usability» (International organization for standardization 2016, 3)[23].

Alle considerazioni finora svolte va aggiunto un ulteriore elemento che, dal punto di vista teorico e al di là delle intenzioni del legislatore europeo, fanno oscillare, come un pendolo, i nuovi servizi per l'archiviazione elettronica tra la fase della gestione documentale e quella della conservazione digitale. Nella versione italiana e in quella inglese del testo consolidato del regolamento eIDAS si legge che un servizio di archiviazione elettronica consente

> la ricezione, la *conservazione*, la consultazione e la cancellazione di dati elettronici e documenti elettronici al fine di garantirne la durabilità e leggibilità nonché di *preservarne* l'integrità, la riservatezza e la prova dell'origine per tutto

[22] Testo consolidato del Regolamento (UE) n. 910/2014, art. 3, comma 48.

[23] L'*authenticity*, così come intesa nel dominio ISO, non è altro se non quella provenienza del documento che può essere provata e indicata anche nel regolamento europeo; la *usability* citata nello standard ricomprende, come un suo aspetto importante, la *legibility* individuata da eIDAS. L'unica caratteristica documentale che non compare, espressamente, nella norma europea è quella della *reliability*, che nella terminologia adottata da ISO fa riferimento alla capacità rappresentativa del documento, dunque alla sua veridicità rispetto alla porzione di realtà rappresentata. È anche vero però che la qualità della *reliability* è implicita in eIDAS 2, considerato che lo stesso regolamento europeo stabilisce che ai dati e documenti elettronici, in ragione del fatto di essere conservati da un servizio di archiviazione elettronica, non possono essere negati «gli effetti giuridici né l'ammissibilità come prova in procedimenti giudiziari» (Testo consolidato del Regolamento (UE) n. 910/2014, art. 45 undecies, comma 1). Sarebbe, infatti, una vera contraddizione in termini che il documento, riconosciuto come dotato di validità sul piano giuridico e di efficacia sul piano probatorio, non fosse al contempo veridico, seppure relazionato a una nozione di veridicità non assoluta, ma relativa, in quanto conforme alle convezioni del sistema di diritto e del sistema sociale.

il *periodo di conservazione*

the receipt, *storage*, retrieval and deletion of electronic data and electronic documents in order to ensure their durability and legibility as well as to *preserve* their integrity, confidentiality and proof of origin throughout the *preservation period*[24].

Così nel passaggio da una lingua all'altra della stessa norma si ritrova, sorprendentemente, che al termine inglese di *storage* è subentrato quello italiano di conservazione[25]. Sorprendentemente perché essi non sono traducibili l'uno nell'altro. Nel linguaggio tecnico essi, infatti, afferiscono a processi ben distinti: mi verrebbe da dire abissalmente distinti[26]. Altra differenza che emerge, analizzando questo articolo di eIDAS nelle sue due diverse formulazioni linguistiche, è legata all'espressione usata per indicare la durata temporale rispetto a cui devono essere assicurate, dal servizio di archiviazione elettronica, le qualità documentali essenziali: nella versione italiana si fa riferimento al *periodo di conservazione*, mentre in quella inglese si rimanda al *preservation period*. Rileggendo il testo italiano – in cui compaiono ben due occorrenze di conservazione e un'unica occorrenza di preservazione – si ha quasi l'impressione che per il suo redattore la conservazione sia null'altro che un sinonimo di preservazione, come se si trattasse di due termini esattamente intercambiali, senza alcuna sfumatura di significato. Ma è davvero così o siamo forse dinnanzi a una forzatura semantica e concettuale?

3. I servizi di archiviazione elettronica: un'incerta collocazione tra conservazione e preservazione?

Se pure volessimo riconoscere che nelle intenzioni del legislatore europeo i nuovi servizi di archiviazione elettronica devono essere ascritti non alla fase della gestione documentale, ma a quella della conservazione digitale – a dispetto di una serie di criticità concettuali – così da evitare, tra l'altro, di dover smorzare l'entusiasmo di quegli operatori economici che nel nostro paese hanno visto in eIDAS 2 la riproposizione del modello nostrano di conservazione digitale, non tutti i dubbi sarebbe comunque tacitati. Sullo sfondo rimarrebbe, infatti, un interrogativo: ammesso e non concesso che i nuovi servizi di archiviazione elettronica non siano pertinenti alla gestione documentale, come essi si collocano invece rispetto alla differenza funzionale tra conservazione

[24] Testo consolidato del Regolamento (UE) n. 910/2014, art. 3, comma 48. I corsivi sono di chi scrive [*N.d.A.*].

[25] Più correttamente nella versione spagnola del regolamento europeo l'inglese *storage* è reso con il termine castigliano di *almacenamiento* e non con il termine, pure esistente, di *conservación*.

[26] Sulle differenze sostanziali tra i processi di storage e quelli di conservazione si rinvia ad alcune osservazioni di Federico Valacchi (Valacchi 2006, 78-79).

e preservazione digitale, considerata tra l'altro l'incertezza terminologica che traspare dalla versione italiana del regolamento europeo osservata in chiusura del capitolo precedente?

In ambito italiano questo quesito probabilmente appare oggi come singolare, in ragione del fatto che nell'attuale contesto si è persa, per lo più, la percezione della distinzione tra quei due concetti. In realtà il termine preservazione, da sempre utilizzato anche se in modo marginale, ha conosciuto a partire da una certa fase un nuovo e più vasto uso: in concomitanza con la diffusione, su larga scala, dei documenti e degli archivi elettronici e sull'onda dell'equivalente termine inglese *preservation* impiegato, in particolare, dalle comunità archivistiche anglosassoni che, per prime, si sono trovate a confrontarsi con le sfide poste dalla nuova documentazione digitale. L'enfasi che allora, a partire da quel momento, si è inteso comunicare con il ricorso al termine di preservazione, è ben spiegata da Maria Guercio, nel suo commento a come il nuovo scenario digitale abbia comportato profonde innovazioni per le metodologie e per le prassi archivistiche: «a cominciare dalla constatazione che la conservazione non sia neppure concepibile se non si avviano le attività che la rendono possibile al momento stesso della formazione del sistema documentario informatico e se del personale professionalmente preparato non sia presente nel disegno stesso del sistema» (Guercio 2019, 148). Come a dire che la conservazione della documentazione elettronica richiede di essere largamente anticipata, non solo alla fase di formazione dei documenti, ma addirittura a quella dedicata alla progettazione di quel sistema documentario destinato alla loro cattura. Su questo principio, che impone alla conservazione di farsi *pre-servazione*, si trova a convergere sostanzialmente tutta la letteratura archivistica. Tanto che tale principio può considerarsi pacifico già dalla fine degli anni Novanta del secolo scorso. Non a caso nella *Guide for managing electronic records from an archival perspective*, predisposta nel 1997 dal Committee on electronic records, costituito all'interno dell'International Council on Archives, si afferma:

> the preservation should be addressed as early as possible in the life-cycle of records, at the conception stage, and appropriate follow-on actions should be taken in the creation and maintenance stages […] The best practice is to articulate preservation requirements for archival records at the conception stage, when a record keeping system is being designed […] A preservation plan should be formulated around these requirements. The plan should delineate how the records should be preserved across time and technology (International Council on Archives – Committee on electronic records 1997, cap. 1.4)[27].

[27] Sempre nella *Guide* si delinea come la preservazione debba operare, senza soluzione di continuità, ben al di là della fase di design del sistema documentario, dunque anche nelle successive fasi di creazione e mantenimento dei documenti elettronici: «when archival records are identified and a preservation plan established at the conception stage, preservation activity in the creation stage involves monitoring records creation practices to ensure that the records

Su questa necessità che la conservazione dei documenti elettronici si riconfiguri, funzionalmente, come preservazione si insiste ancora oggi nel già citato white paper *eIDAS Trust Electronic Archiving Services supported by the eArchiving Initiative*:

> a common characteristic of all types of electronic data and electronic documents is their rapid obsolescence, which is primarily due to their technical/technological characteristics. Risks to the integrity, authenticity, and usability of electronic data and electronic documents increase over time […] [So this scenario] has forced the simultaneous development of approaches, methods, and services to mitigate the risks on the electronic data and electronic documents […] For this reason, electronic archiving is strongly linked to digital preservation measures. The simple storage or replication of electronic data and electronic documents without considering preservation needs has become the main source of loss of information usability. *Acting in the early stages of electronic archiving is the best option to guarantee the integrity, authenticity and usability of the electronic data and electronic documents* as long as they are needed (European Union 2024, 3-4)[28].

Dall'insieme di queste considerazioni emerge, allora, come la preservazione si componga di una serie di funzioni – configurate secondo precisi requisiti – che non devono essere concentrate su una specifico sistema ad hoc, dovendo al contrario essere ubiquamente distribuite su tutti i sistemi che fanno uso di documenti elettronici: a partire da quelli di gestione documentale, che nelle organizzazioni tanto pubbliche quanto private sono chiamati a catturare e a mantenere – e dunque a questo punto anche a preservare – flussi ingenti di documenti elettronici che si sedimentano in archivi digitali. La preservazione, nello scenario digitale, è in altri termini l'altra faccia dell'uso: in qualsiasi contesto in cui si fa un utilizzo, non estemporaneo, di documenti elettronici lì deve operare anche la preservazione, senza sosta e per tutto il periodo di tempo

---

are created as expected and that they can be available, retrievable and understandable over long periods of time. When there is no archival involvement in the conception stage, the archives should seek to be involved at the earliest possible phase in the life of a system. Solving problems after the fact is likely to be more difficult than preventing such problems ahead of time. Preservation problems are apt to get worse, and more difficult to solve over time. If preservation was not addressed at the conception stage, archivists should analyse records creation practices to determine if it will be possible to preserve the archival records, and to identify any changes that would improve or facilitate preservation […] During the maintenance stage, monitoring and corrective follow-up actions are necessary to ensure that decisions taken in the conception and creation stages continue to be respected. In the maintenance phase, the record keeping system should be monitored to identify when changes occur, or are likely to occur, that might impact the availability, retrievability or understandability of the records over time. Such changes could occur in the records life cycle, the record keeping system, the enabling technology, or the custody or control of the records» (International Council on Archives – Committee on electronic records 1997, cap. 1.4).

[28] Il corsivo è di chi scrive [*N.d.A.*].

in cui quello stesso uso si dispiega. Non è un caso che lo standard ISO 15489-1:2016 preveda una serie di processi che sono tipici della *preservation*: lo *storing records*, lo *use and reuse* e il *migration and converting records* (International organization for standardization 2016, 17-18). La preservazione così definita deve essere allora distinta dalla conservazione, che più precisamente andrebbe sempre indicata come conservazione permanente. Questa si esercita sulla documentazione elettronica che ha oramai esaurito le ragioni d'uso per le quali è stata originariamente catturata e utilizzata da una certa organizzazione e per un definito periodo di tempo e che nonostante questa perdita di valore d'uso è stata comunque valutata come meritevole di custodia a tempo indefinito, in quanto nel frattempo essa ha acquisito una sopraggiunta utilità d'uso: *in primis* per la ricerca storica e per la ricerca scientifica. Così possiamo tracciare la seguente distinzione: la preservazione deve considerarsi come un complesso di funzioni trasversalmente distribuite nei diversi contesti e sistemi che catturano e fanno uso di documenti elettronici, per il compimento delle attività che sono proprie delle organizzazioni; la conservazione, invece, deve essere riconosciuta come un complesso di funzioni che sono relegate e si concentrano in contesti e sistemi specifici, progettati per il fine esclusivo di custodire il patrimonio documentale in essi trasferito e per metterlo a disposizione di una comunità di utenti terza rispetto alle organizzazioni che, originariamente, si sono servite di quella stessa documentazione elettronica[29]. Si potrebbe anche dire che mentre la preservazione ha una natura sincrona in rapporto agli usi che presiedono agli originari processi di produzione e cattura dei documenti elettronici, non così la conservazione che in relazione ad essi si pone, invece, su un piano diacronico, tanto da qualificarsi come la fase finale del ciclo di vita documentale. Di queste distinzioni sembra in qualche modo avvedersi lo stesso legislatore europeo, che in uno dei considerando del regolamento che modifica eIDAS riconosce che «le attività degli archivi nazionali e delle istituzioni della memoria, in qualità di organizzazioni preposte alla conservazione del patrimonio documentario nell'interesse pubblico, sono generalmente disciplinate dal diritto nazionale e non forniscono necessariamente servizi fiduciari ai sensi del presente regolamento»[30].

Così dinnanzi alla divaricazione funzionale che sussiste tra la dimensione della preservazione e quella della conservazione, si dovrebbe ritenere che i servizi di archiviazione elettronica previsti dal nuovo regolamento europeo

---

[29] In termini archivistici propri i soggetti produttori.

[30] (Parlamento europeo e Consiglio dell'Unione europea 2024, considerando n. 67). Va in ogni caso tenuto in conto che, come ha stabilito la sezione V della Corte di cassazione con l'ordinanza 7 marzo 2022, n. 7280, in tema di interpretazione delle fonti del diritto unionale i considerando riportati in un regolamento europeo hanno il compito di illustrare le ragioni dell'intervento normativo e ne integrano la motivazione, ma non contengono enunciati di carattere normativo.

siano propriamente di natura preservativa. Interpretazione questa che appare confermata dalla versione in lingua inglese di eIDAS 2, che rispetto alla traduzione italiana si caratterizza – come si è visto – per una maggiore coerenza terminologica. Richiamandoci, dunque, al significato più tecnico del termine *preservation* e contestualizzandolo nel ciclo di vita della documentazione elettronica, possiamo concludere quanto segue: i nuovi servizi di archiviazione, voluti dal legislatore europeo, definiscono degli ambiti funzionali di natura preservativa associati ai diversi contesti e sistemi utilizzati per gli usi correnti – a scopo di business o per finalità giuridiche – dei documenti elettronici e dei relativi archivi digitali. In questa prospettiva la stessa previsione dei servizi di archiviazione elettronica potrebbe essere ricompresa, eventualmente, nella sfera della gestione documentale, considerato che anch'essa necessita di idonee funzioni di preservazione, né più né meno di qualsiasi altro contesto e sistema che abiliti all'utilizzo dei documenti elettronici. In un'ulteriore conferma, pertanto, di quanto già osservavo nel precedente capitolo di questo contributo.

4. I servizi di archiviazione elettronica e il modello italiano

Se si concorda su questa interpretazione, su di essa si può far leva per affrontare un ulteriore interrogativo: quanto di questo paradigma europeo dei servizi di archiviazione elettronica può ritrovare un suo equivalente, già realizzato, nel modello italiano? Nella ricerca di una possibile risposta credo si debba partire dal constatare come, tra i principi cardine su cui negli ultimi anni sono stati sviluppati in Italia i servizi di conservazione digitale, vi siano le regole tecniche contenute, da ultimo, nelle *Linee guida sulla formazione, gestione e conservazione dei documenti informatici*. In particolare, quella che prevede che «il sistema di conservazione sia almeno logicamente distinto dal sistema di gestione informatica dei documenti» (Agenzia per l'Italia Digitale 2021, 31)[31]. Si tratta, in realtà, di un'indicazione è che da tempo presente nel nostro ordinamento giuridico. Essa, infatti, compariva già nelle abrogate *Regole tecniche in materia di sistema di conservazione*, con un'articolazione leggermente differente*:* «il sistema di conservazione opera secondo modelli organizzativi esplicitamente definiti che garantiscono la sua distinzione logica dal sistema di gestione documentale, se esistente» (Decreto del Presidente del Consiglio dei Ministri 3 dicembre 2013)[32]. Nella letteratura di settore questa regola tecnica non ha suscitato particolari perplessità. Essa però, per come è stata interpretata nei concreti processi di sviluppo dei servizi di conservazione digitale, ha avuto un preciso effetto: quello di espungere dai sistemi di gestione documentale le funzioni

[31] Approvate dall'Agenzia per l'Italia Digitale con determinazioni del direttore generale 9 settembre 2020, n. 407 e 17 maggio 2021, n. 371.

[32] Art. 5, comma 1.

di preservazione, riaggregate allora in sistemi ad hoc nettamente distinti dai primi. Potremmo, pertanto, riferirci al modello italiano come a un paradigma preservativo anomalo: da un lato, infatti, è indubbio che quelli che nel nostro paese indichiamo come servizi di conservazione digitale siano da considerare, più correttamente, come dei servizi di preservazione, in quanto destinati a documenti elettronici d'uso corrente da parte delle organizzazioni[33]; dall'altro lato però le funzioni di preservazione non operano – come ci si aspetterebbe in base alla nozione di *preservation* in senso stretto – all'interno dei sistemi di *records management*, essendo autonomamente articolate in contesti e sistemi propri. Le ragioni che hanno imposto questa particolare configurazione credo dipendano dal punto di origine della normativa di settore nel nostro paese[34]: le *Regole tecniche per la riproduzione e conservazione di documenti su supporto ottico idoneo a garantire la conformità dei documenti agli originali* (Centro nazionale per l'informativa nella pubblica amministrazione 2004). Quelle prime regole tecniche definivano, non a caso, la conservazione *sostitutiva*. Questa era chiamata a surrogare quella parte della gestione documentale che, all'epoca, si riteneva tecnologicamente e organizzativamente non in grado di garantire le funzioni di preservazione della documentazione elettronica. E tale sottovalutazione delle performance preservative dei sistemi di gestione documentale si è poi mantenuta, sottotraccia, in tutta la normativa posteriore.

In ogni caso, a prescindere dalle cause che hanno contribuito a strutturare in questo modo peculiare il modello italiano, esso risulta per alcuni suoi aspetti disfunzionale. In primo luogo in ragione del fatto che i documenti elettronici non sono concretamente trasferiti dalle applicazioni di gestione documentale a quelle di conservazione digitale, ma sono più semplicemente versati. La documentazione permane, infatti, presso gli originari sistemi di *records management* e al contempo dei *secondi esemplari* sono consegnati ai sistemi di conservazione. L'effetto finale è quello di duplicare su due fronti l'archivio digitale corrente della stessa organizzazione, con l'onere aggiuntivo di doverli costantemente tenere allineati, considerato che una sedimentazione documentale di natura corrente non è mai, per definizione, stabilizzata in senso assoluto. La disfunzione è, in secondo luogo, legata a una generale incertezza che abbraccia l'in-

[33] Non è un caso che l'art. 44, comma 1-bis del D. Lgs. 7 marzo 2005, n. 82 (*Codice dell'amministrazione digitale*) solleciti le amministrazioni pubbliche a versare quanto prima possibile – quasi si trattasse di un'urgenza – i propri documenti elettronici in forma aggregata (fascicoli e serie informatiche) dagli originari sistemi di gestione documentale ai sistemi di conservazione digitale: «il sistema di gestione dei documenti informatici delle pubbliche amministrazioni è gestito da un responsabile […] Almeno una volta all'anno [questo] responsabile della gestione dei documenti informatici provvede a trasmettere al sistema di conservazione i fascicoli e le serie documentarie anche relative a procedimenti non conclusi».

[34] Per una disamina critica di come si è sviluppata in Italia la normativa relativa alla conservazione dei documenti elettronici si rinvia ad una serie di considerazioni di Federico Valacchi (Valacchi 2006, 80-96).

tero scenario: l'organizzazione – in particolare un'amministrazioni pubblica – svolge i propri compiti servendosi degli esemplari dei documenti elettronici che persistono nella propria applicazione di gestione documentale, in quanto quest'ultima dispone oggi di sofisticate integrazioni con strumenti irrinunciabili: *in primis* quelli per la gestione dei flussi documentali e per la gestione dei processi di lavoro. In questo modo però l'organizzazione stessa, basando la propria operatività sugli esemplari catturati e persistenti nel proprio sistema di gestione documentale, riconosce ad essi una piena credibilità: dunque una piena validità ed efficacia o, per usare il linguaggio dell'ISO 15489-1:2016, un effettivo valore di evidenza. Stando così le cose, non si intende allora quale sia, rispetto ad esso, il *surplus di credibilità* conferito ai secondi esemplari custoditi dai sistemi di conservazione. I termini del problema non sono certo risolti se si invoca l'assai episodica necessità di esibire il documento elettronico come mezzo probatorio davanti a un giudice. Non vi è, infatti, una norma che obblighi il giudice a privilegiare, come prova documentale, l'esemplare del documento elettronico presente nell'applicazione di gestione documentale rispetto all'esemplare dello stesso documento custodito nel sistema di conservazione. Così ciò che sembra emergere è un interrogativo basilare su quale sia la concreta e irrinunciabile utilità di duplicare gli archivi digitali attraverso dei servizi di preservazione esterni ai sistemi di gestione documentale.

Il modello italiano potrà, dunque, costituire un riferimento utile per gli approfondimenti necessari a sviluppare il paradigma dei servizi di archiviazione elettronica previsti da eIDAS 2: certamente per quanto riguarda gli aspetti tecnologici, gli assetti organizzativi e le policy di vigilanza adottate dalle autorità pubbliche. Ad esso però si dovrebbe anche guardare come a un caso che ci impartisce, ahimè, una lezione amara, ma basilare: l'efficacia e l'efficienza dei futuri servizi di archiviazione non si giocherà esclusivamente sul terreno delle innovazioni tecnologiche, per quanto spinte, ma anche su quello di una loro idonea collocazione funzionale nel contesto del ciclo di vita del documento elettronico – per come si è andato definendo a livello di teoria e metodo verificati in questi decenni di esperienza sul campo – così da costruire un ecosistema documentario digitale davvero sostenibile.

Riferimenti bibliografici

Agenzia per l'Italia Digitale. 2021. *Linee guida sulla formazione, gestione e conservazione dei documenti informatici*. https://www.agid.gov.it/sites/agid/files/2024-05/linee_guida_sul_documento_informatico.pdf.

Alfier, Alessandro. 2023. "Per una rigenerazione teorica dell'archivistica in Italia, a partire dal concetto di documento." *AIDAinformazioni*, anno 41, no. 3-4 (luglio-dicembre): 9-26.

Belisario, Ernesto. 2024. "Il regolamento eIDAS 2.0 e l'impatto sulla gestione documentale delle PA: quali prospettive?" *Rivista elettronica di diritto, economia e management* 15 (4): 32-41. https://www.clioedu.it/documenti/eventi-live-ondemand/rivista-elettronica/Rivista-elettronica-4_2024.pdf.

Centro nazionale per l'informatica nella pubblica amministrazione – CNIPA. 2004. "Deliberazione 19 febbraio 2004, n. 11 Regole tecniche per la riproduzione e conservazione di documenti su supporto ottico idoneo a garantire la conformità dei documenti agli originali - Art. 6, commi 1 e 2, del testo unico delle disposizioni legislative e regolamentari in materia di documentazione amministrativa, di cui al decreto del Presidente della Repubblica 28 dicembre 2000, n. 445." *Gazzetta Ufficiale* no. 57, 9 marzo 2004.

Decreto del Presidente del Consiglio dei Ministri 3 dicembre 2013. "Regole tecniche in materia di sistema di conservazione ai sensi degli articoli 20, commi 3 e 5 -bis, 23 -ter, comma 4, 43, commi 1 e 3, 44, 44 -bis e 71, comma 1, del Codice dell'amministrazione digitale di cui al decreto legislativo n. 82 del 2005." *Gazzetta Ufficiale* no. 59, 12 marzo 2014, Suppl. Ordinario no. 20.

Decreto Legislativo 7 marzo 2005, n. 82. "Codice dell'Amministrazione Digitale." *Gazzetta Ufficiale* no. 112, 16 maggio 2005 - Suppl. Ordinario no. 93.

eArchiving Initiative. 2024. *eIDAS Trust Electronic Archiving Services supported by the eArchiving Initiative.* White Paper, 28 maggio. https://ec.europa.eu/newsroom/repository/document/2024-22/eIDAS_Trust_Electronic_Archiving_Services_supported_by_the_eArchiving_Initiative_to4u8jdCPacvkjEhLY4ncfFRA_105792.pdf.

European Commission. 2001. *Model requirements for the management of electronic records: MoReq specification*. Office for official publications of the European Communities. https://web.archive.org/web/20110720155236/http://dlmforum.eu/index.php?option=com_jotloader&view=categories&cid=23_75067adade55e2da39ea036bc400a33f&Itemid=100&lang=en.

European Union. 2024. "eArchiving White Paper on eIDAS 2." Last update 4 June. https://digital-strategy.ec.europa.eu/en/library/earchiving-white-paper-eidas2.

Giunta, Enrico, e Federica Marti. 2024. "Le potenzialità di eIDAS 2 sui servizi di conservazione digitale: stato dell'arte, impatti tecnologici e prospettive." *Rivista elettronica di diritto, economia e management* 15 (4): 145-63. https://www.clioedu.it/documenti/eventi-live-ondemand/rivista-elettronica/Rivista-elettronica-4_2024.pdf.

Guercio, Maria. 2019. *Archivistica informatica. I documenti in ambiente digitale*. Carocci.

Guercio, Maria. 2023. "Nuovo eIDAS, le proposte per archiviazione e conservazione: verso più controllo." *Agenda digitale*, 19 settembre. https://www.agendadigitale.eu/documenti/nuovo-eidas-le-proposte-per-archiviazione-e-conservazione-verso-piu-controllo/.

International Council on Archives – Committee on electronic records. 1997. *Guide for managing electronic records from an archival perspective*. ICA. https://www.ica.org/resource/ica-study-n8-guide-for-managing-electronic-records-from-an-archival-perspective/.

International Organization for Standardization. 2016. *International Standard ISO 15489-1. Information and documentation - Records management. Part. 1: Concepts and principles*. 2nd ed. ISO.

Lisi, Andrea. 2024. "L'evoluzione del documento informatico nel nuovo quadro giuridico dell'eIDAS 2." *Rivista elettronica di diritto, economia e management* 15 (4): 42-53. https://www.clioedu.it/documenti/eventi-live-on-demand/rivista-elettronica/Rivista-elettronica-4_2024.pdf.

Parlamento europeo e Consiglio dell'Unione europea. 2014. *Regolamento (UE) n. 910/2014 del 23 luglio 2014 relativo all'identificazione elettronica e ai servizi fiduciari per le transazioni elettroniche nel mercato interno e che abroga la direttiva 1999/93/CE*. Gazzetta ufficiale dell'Unione europea L 257, 28 agosto 2014.

Parlamento europeo e Consiglio dell'Unione europea. 2024. *Regolamento (UE) 2024/1183 dell'11 aprile 2024 che modifica il regolamento (UE) n. 910/2014 per quanto riguarda l'istituzione del quadro europeo relativo a un'identità digitale*. Gazzetta ufficiale dell'Unione europea L 1183, 30 aprile 2024.

Pomarico, Raffaele. 2024. "La bozza di eIDAS 2 introduce il concetto di e-archiving: vediamo di che cosa si tratta e cosa cambia per il settore dei servizi fiduciari." *Agenda digitale*, 26 marzo. https://www.agendadigitale.eu/documenti/e-archiving-ecco-come-funziona-il-servizio-eidas-2/.

Sormani, Patrizia. 2024. "Dalla conservazione digitale all'e-archiving. I requisiti che un qualified trust service provider deve possedere per erogare il servizio di e-archiving: prospettive e scenari per il mercato." *Rivista elettronica di diritto, economia e management* 15 (4): 126-44. https://www.clioedu.it/documenti/eventi-live-ondemand/rivista-elettronica/Rivista-elettronica-4_2024.pdf.

Valacchi, Federico. 2006. *La memoria integrata nell'era digitale. Continuità archivistica e innovazione tecnologica*. Titivillus.

Yeo, Geoffrey. 2010. "Representing the Act: Records and Speech Act Theory." *Journal of the Society of Archivists* 31 (2): 95-117.

Yeo, Geoffrey. 2017. "Information, Records, and the Philosophy of Speech Acts." In *Archives in Liquid Times*, edited by Frans Smit, Arnoud Glaude-mans, and Rienk Jonker. Stichting Archiefpublicaties.

AIDAinformazioni
ISSN 1121–0095
ISBN 979-12-5965-456-4
DOI 10.57574/596545642
pag. 29-50 (luglio-dicembre 2024)

Exploration du réseau numérique YouTube autour de la santé des militaires : quelles sont les thématiques des discours, les sources d'informations et les acteurs de la communication ?

Fetta Belgacem, Marc Tanti*

Abstract: The YouTube platform presents a great number of information published by multiples sources discussing the theme of military health. This information takes diverse forms; video/ text/ comment, and provides us a precious space to apprehend military health issues. Our article presents an analyse of the multiple YouTube publications about the health of military persons and their families from the year of 2021. The aim is to identify the different medias communicating about this subject and the major military health themes spread within YouTube social media.

Keywords: Military health, YouTube, Comments, Mixed methodology, Digital practices.

1. Introduction

L'humanité produit chaque année un volume d'informations numériques de l'ordre du zettaoctet, soit presque autant que d'étoiles dans l'Univers (Demarthon 2012). La population militaire, comme la population civile, s'est appropriée les médias sociaux et est devenue partie prenante d'un monde de plus en plus digitalisé. Notamment, cette population s'exprime, dans cet espace public démocratisé sur ses préoccupations, son état de santé… Les données fournies par cette population sur le Web social[1], peuvent constituer des informations à très forte valeur ajoutée pour les institutions militaires, une fois exploitées. Ainsi, Twitter est aujourd'hui utilisé par de nombreux organismes de santé publique pour éduquer, informer et surveiller l'état de santé de la population civile, en particulier en cas de catastrophe (González-Padilla et Tortolero-Blanco 2020, Tanti 2023). Cet article présente les résultats originaux d'une recherche qui consiste en la collecte et l'analyse des données massives en lien avec la santé des militaires français. Ces données, provenant du Web social,

* Centre d'épidémiologie et de santé publique des armées – CESPA ; Institut méditerranéen des sciences de l'information et de la communication, Aix-Marseille Université, France. fetta020@gmail.com, mtanti@gmx.fr.

[1] Réseaux sociaux, Forums, Blogs.

ont pour objectif d'examiner les discours sanitaires exprimées par les militaires et par leurs familles sur ces espaces numériques. Ces masses d'informations numériques produites par ces acteurs présentent un intérêt stratégique pour les armées et « forment une "matière première" particulièrement profitable » (Boullier 2015). Elles permettent, une fois traitées et organisées, d'anticiper les problèmes de santé de leurs employés, à condition de les inscrire dans un processus plus global d'analyse et de classification. Dans son article, Dominique Boullier explore comment les Big Data transforment les sciences humaines et sociales en introduisant les « traces numériques » comme un nouveau matériau pour observer le social et l'humain. Ces traces permettent de détecter des comportements et des interactions en temps réel, redéfinissant les cadres d'analyse traditionnels vers une approche plus dynamique et réactive.

Ainsi, en s'appuyant sur cette approche des traces numériques, ce travail applique les méthodes du Big Data pour explorer et catégoriser les thématiques de santé les plus récurrentes sur YouTube. En revanche, cette étude se concentre sur l'observation et l'analyse des comportements en ligne, sans toutefois aller jusqu'à formuler des recommandations de prévention ou des actions ciblées pour anticiper les risques identifiés.

L'objectif principal consiste en l'analyse des pratiques info-communicationnelles autour de la santé des militaires diffusés sur la plateforme YouTube. L'expression « pratiques info-communicationnelles » englobe l'ensemble des activités liées à l'information et à la communication, telles que la recherche d'informations, la production de contenu, l'analyse des informations et le partage de celles-ci, entre autres (Boumhaouad 2017). Nous examinons ici deux éléments clés : les vidéos partagées ainsi que les commentaires associés au sujet de leur état de santé.

Pour commencer, nous mettons en œuvre un processus de catégorisation des thèmes de santé les plus discutés, ce qui nous permet d'identifier les diverses sources d'information traitant de ces questions et d'identifier les profils des personnes impliquées aussi bien dans le processus de création/ production que dans la réception/ consommation de ce contenu. Enfin, nous chercherons à comprendre, expliquer et caractériser les pratiques info-communicationnelles en ligne des utilisateurs lorsqu'ils abordent le sujet de la santé des militaires.

2. État de l'art

Bien que « la hiérarchie, la discipline et la centralisation soient les pierres angulaires de l'organisation militaire, et représentent par conséquent des concepts clés de l'idéologie corporative » (Doorn 1970, tel que cité dans Schweisguth 1978, 382), l'arrivée d'Internet a permis à certains militaires de s'exprimer en faisant sauter le verrou de la hiérarchie (Resteigne et al. 2012). Cela accroît considérablement leur visibilité et leur offre de nouvelles formes d'expression

diversifiées et horizontales. Le partage virtuel de données entre les militaires et la publicisation des retours d'expérience sur Internet leur apportent du réconfort et un soutien psychologique leur permettant de se constituer en « communautés virtuelles » (Rheingold 1994) pour discuter de leurs préoccupations individuelles en collectif. C'est d'ailleurs ce que confirme le rapport publié par le Ministère de la Défense sur Internet où il présente les grandes lignes de sa stratégie sur les réseaux sociaux numériques. Les réseaux sociaux permettent finalement aux soldats d'avoir un espace de discussion « les blogs, les groupes virtuels de mobilisation ont des vertus cathartiques qui permettent d'évacuer les insatisfactions du moment » (Centre interarmées 2013). Ces modes d'interaction en ligne principalement sur les blogs, les forums, les réseaux sociaux ou sur les plateformes-vidéos renforcent les liens entre les militaires ainsi qu'entre les militaires et la société civile. Il ne s'agit pas d'espaces non hiérarchisés puisque certains comptes numériques sont plus visibles que d'autres. Mais cette visibilité n'est pas conditionnée par le statut professionnel ou l'affiliation des internautes, mais par « l'organisation sociale des jugements portés par les internautes qui produit une hiérarchie de la visibilité de ces pages » (Cardon 2012, 68).

Cette situation peut cependant s'avérer préjudiciable pour l'armée, car elle expose les informations personnelles, y compris les données de localisation, à une divulgation potentielle, ce qui pourrait mettre en danger les opérations. De plus, elle favorise la propagation de fausses informations, de rumeurs et de propagande, ce qui peut saper la confiance du public dans l'institution militaire. Pour Laurence Ifrah, « l'isolement et l'éloignement les [les militaires] incitent à communiquer à leurs proches leurs états d'âme : bien souvent, ils se laissent aller à raconter des anecdotes vécues sur les théâtres d'opérations, à publier des photos d'eux et de leurs camarades et par conséquent à divulguer des informations pouvant être utilisées par les forces ennemies » (Ifrah 2010, 60).

Toutefois, ces plates-formes peuvent être utiles aux forces armées, sous réserve qu'elles soient capables d'exploiter leur potentiel. En effet, diverses organisations ont recours, par exemple, à des médias sociaux tels que X (anciennement Twitter) dans un but éducatif ou préventif, en suivant l'état de santé de leur population afin de minimiser les risques.

Sur le plan académique, le « foyer virtuel[2] » (Lévi-Strauss 1977) que représentent les médias numériques pour de nombreux militaires – en raison des pratiques interactives qu'ils ont engendrées et qui ont favorisé la création des

[2] L'expression « foyer virtuel » est utilisée chez Lévi-Strauss pour faire référence à la notion d'identité. L'identité est, selon lui, une sorte de « foyer virtuel auquel il nous est indispensable de référer pour expliquer un certain nombre de choses, mais sans qu'il n'ait jamais d'existence réelle » (Lévi-Strauss 1977, 332). L'identité est donc aussi un ensemble de représentations qu'une personne assigne aux réalités qui l'entourent. Nous pensons que les médias numériques sont aussi un foyer virtuel pour les communautés en ligne. Ces derniers se forgent une identité à travers les interactions continuelles entre les membres.

communautés en ligne – a éveillé l'intérêt des chercheurs. Ces derniers ont tiré parti de cette source de données pour les analyser et les comprendre dans l'objectif d'accélérer la détection des maladies ou des épidémies. Prenons, par exemple, l'étude menée par Aude Berger, dont la thèse en médecine portait sur les usages numériques des patients militaires dans le cadre de la gestion de leur santé. Son travail a mis en évidence les sujets de santé les plus fréquemment cherchés sur Internet par cette population. Les soldats ont recours à des plates-formes numériques pour se renseigner sur une variété de problèmes de santé, allant des maladies graves aux maladies ponctuelles, et pour s'informer sur les traitements et la santé maternelles infantiles (Berger 2015).

Au-delà de cette étude, il faut mentionner que la recherche sur le triptyque « militaire, réseaux sociaux et santé » est principalement américaine. Pour Svitlana Volkova et al., la santé des militaires n'est pas sans enjeux et peut impacter la sécurité de toute une nation, raison pour laquelle la prévention est, selon elle, primordiale en milieu militaire (Volkova et al. 2017).

Glen Coppersmith (2014) explique dans son étude autour de l'identification de la dépression sur Twitter que les réseaux sociaux numériques offrent un potentiel considérable en matière de prévention des troubles mentaux chez les militaires américains. En effet, la fréquence des publications sur Twitter, ainsi que les tonalités et les changements langagiers associées aux messages, peut révéler les signes d'une maladie mentale.

Alors que Coppersmith se concentre sur le syndrome de stress post-traumatique visible dans les publications numériques des militaires, Craig J. Bryan et al. (2018) s'intéresse plutôt à la prévention du risque suicidaire via les réseaux sociaux. Ses résultats indiquent que certains contenus numériques peuvent prédire la cause du décès et fournir une estimation sur les profils en ligne susceptibles de mourir de suicide. À partir du moment où une personne décide de confier ses souffrances et ses envies de suicide plutôt à sa communauté virtuelle qu'à son médecin, ces espaces numériques deviennent un terrain pertinent pour prévenir les tentatives de suicide.

Contrairement aux études américaines citées, qui se concentrent principalement sur des domaines spécifiques de la santé tels que le risque suicidaire pour Bryan, le stress post-traumatique pour Coppersmith ou le syndrome pseudo-grippal pour Volkov, cette étude a pour objectif d'englober l'ensemble des thématiques de santé afin de les identifier et de les catégoriser. En adoptant cette approche plus globale, l'étude fournit une vue d'ensemble des problématiques de santé et permet de mettre en lumière des catégories parfois sous-représentées dans les études précédentes et d'inclure santé mentale et santé physique. De plus, bien que quelques recherches aient exploré le thème global de la santé militaire sur les médias numériques (Pavalanathan et al. 2016), la majorité se concentre sur la plateforme X (anciennement Twitter). YouTube,

en revanche, reste peu étudiée dans le contexte militaire, malgré son rôle croissant dans la diffusion d'informations sanitaires.

Selon une enquête Ipsos réalisée en avril 2024, 99% des jeunes déclarent avoir visité la plateforme YouTube au cours des trois derniers mois (Ipsos 2024). La population militaire étant avant tout jeune, il nous a donc paru légitime de nous interroger sur les types de discours véhiculés en lien avec leur santé en premier lieu sur cette plateforme. Notamment nous nous sommes interrogés sur ce que sont les préoccupations sanitaires émises par cette population sur elle-même et par elle-même, les sources d'informations utilisées[3] et les autres acteurs communiquant sur leur santé ?

3. Méthodologie de travail

La collecte et l'analyse des données numériques diffusées par les militaires et autour des militaires sur la plateforme YouTube, permettent d'appréhender les comportements discursifs sur le sujet de la santé des militaires dont l'objectif est d'identifier les signaux faibles et de repérer les préoccupations de santé. Pour rappel, ces données sont comme des traces informationnelles qui permettent d'accéder au social dans ses formes les plus intimes. Non seulement, elles renseignent sur le réseau et le collectif, mais l'individu et ses préoccupations y prennent une place centrale.

Le réseau YouTube est devenu le média vidéo le plus populaire, et permet de faire circuler du contenu audiovisuel pouvant atteindre près de 1 milliard d'individus dans plus de 88 pays. La parole jusqu'ici unidimensionnelle, devient dès lors horizontale et se déterritorialise pour inclure des communautés virtuelles élargies. La frontière entre émetteurs et récepteurs, producteurs et consommateurs, amateurs et professionnels semble s'estomper devant les possibilités offertes par ce réseau. C'est le lieu où se diffuse et se propage la parole individuelle en dehors des espaces d'expression officiels, élargissant ainsi l'expression vers des arènes alternatives en ligne. YouTube est ici étudiée comme une plateforme centrée non seulement sur l'individu militaire, mais également sur les différentes communautés militaires, telles que les pages dédiées aux membres de l'armée de terre, de l'armée de l'air, ou des marins-pompiers. Les individus amateurs créateurs de chaînes YouTube et leurs abonnés s'unissent autour de la production d'information sur un sujet donné et forment une communauté. C'est ainsi que ce service de partage de vidéos interpersonnel est devenu la première communauté vidéo au monde sur le réseau Internet (Snickars et Vonderau 2009).

La plateforme YouTube regorge de nombreuses données sur la santé des militaires. Avant d'entamer les premières collectes, nous avons effectué une

[3] Sur la plateforme YouTube, diverses sources d'information – médias/institutions/utilisateurs ordinaires – disposent d'une chaîne YouTube et publient du contenu.

première exploration manuelle où l'objectif est d'identifier les sources d'où proviennent ces données diffusées sur la plateforme. Il est à noter que toutes les sources nous intéressent dans le cadre de notre analyse, qu'elles émanent des sites officiels du ministère des Armées, des sites des médias traditionnels ou des influenceurs ayant pour contenu l'univers militaire. Après avoir identifié les différentes sources d'information, nous remarquons un nombre important de commentaires associés aux vidéos publiées. Ce sont ces commentaires qui nous intéressent le plus pour l'analyse étant donné que ces données proviennent essentiellement de la communauté militaire elle-même et de leur entourage[4]. Les commentaires sur YouTube, qu'ils soient des récits sur la thématique de santé ou des hommages pour les militaires décédés, recèlent une mine d'informations qui ont été collectées de manière méthodique dans notre étude. Ces informations ont ensuite été classées et analysées pour en extraire des connaissances.

Nous avons tout d'abord sélectionné un corpus de 43 vidéos à partir desquelles nous avons extrait 6 318 commentaires analysés avec le logiciel Alceste pour procéder à une analyse thématique. Cette démarche analytique permet de repérer des thématiques dominantes et d'offrir une compréhension approfondie des discours émergeants dans un corpus vidéo.

4. Collecte de données : extraction et formatage du corpus

Pour accéder aux divers discours sur la santé des militaires français qui circulent sur YouTube, il nous a fallu réfléchir aux différentes étapes de collecte et d'analyse. Ces données sont massives, non structurées et protéiformes : elles incluent des images, des vidéos et du texte sous forme de commentaires. Elles se distinguent par le triptyque « volume, variété, vélocité » (Boullier 2015), ce qui soulève des questions quant aux outils et aux méthodes permettant leur extraction.

Pour mener à bien notre étude, nous avons choisi une méthodologie de recherche mixte, alliant des techniques de collecte provenant des sciences humaines et sociales, telles que l'observation approfondie des réseaux sociaux numériques et des forums, à celles des sciences informatiques, comme la collecte par API (Application Programming Interface), ainsi que des analyses textuelles qualitatives et quantitatives. Cette approche a exigé la mise en place d'une méthode de collecte automatisée à l'aide d'API, une fonction offerte par le réseau YouTube aux développeurs. Ce qui rend possible la collecte automatisée de 50 vidéos par mot clé. Nous nous sommes basées sur un script pour faire appel à cet API. Afin de simplifier son utilisation, nous avons également conçu une

[4] Ces commentaires collectés à des fins de l'analyse sont anonymisés dans le cadre de cet article.

interface utilisateur ou ce qu'on appelle une fenêtre d'interaction. Grâce à celle-ci, l'utilisateur n'a pas besoin de modifier le code ; il doit simplement fournir les informations nécessaires au bon fonctionnement de l'outil et à sa recherche.

Cette interface (Fig. 1) offre plusieurs options notamment :

- Sélectionner le dossier où seront enregistrées les informations collectées ;
- Choisir le fichier Excel qui contiendra la liste des vidéos YouTube ;
- Entrer les termes de recherche (mots-clés) ;
- Décider d'extraire les données et, le cas échéant, d'indiquer le nombre maximal de vidéos à inclure ;
- Régler le nombre maximal de commentaires à récupérer (un chiffre spécifique ou l'ensemble) ;
- Opter pour la mise à jour des commentaires existants ;
- Choisir d'assembler tous les fichiers texte au formatage spécifique pour chaque mot-clé en un seul ;
- Choisir d'assembler tous les fichiers texte au formatage spécifique pour tous les mots-clés en un seul ;
- Modifier la clé API.

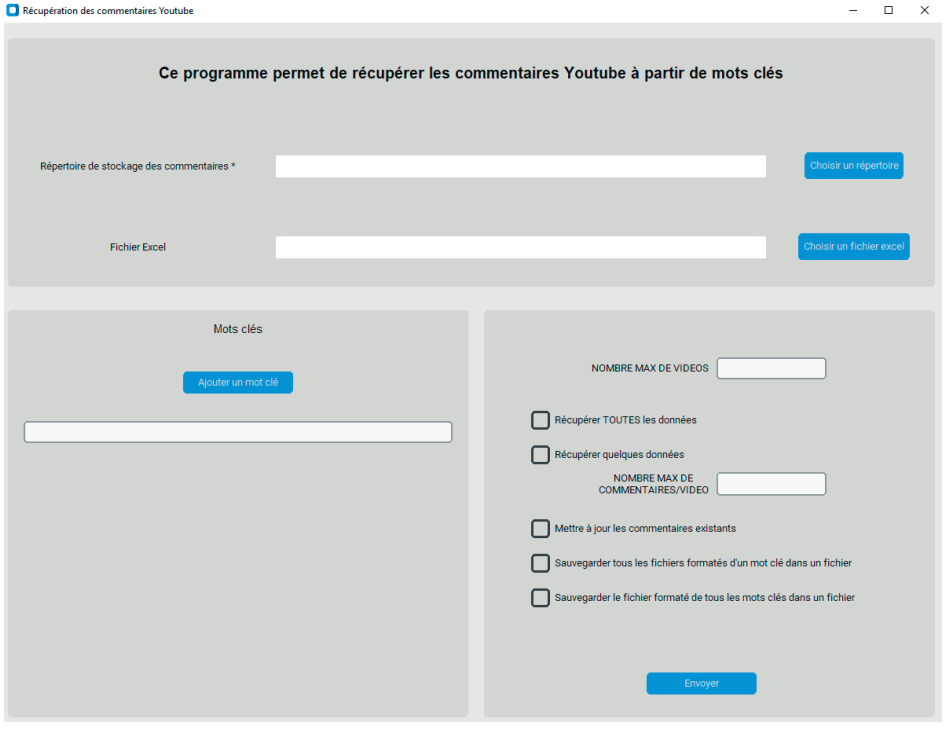


Figure 1: Interface de collecte de données YouTube.

Pour concevoir cette interface, nous avons utilisé la bibliothèque standard de Python *Tkinter* et une de ses extensions *Customtkinter*. Cette dernière offre une personnalisation avancée de l'outil, ce qui contribue grandement à son aspect visuel attrayant.

Nous y avons inséré tous les mots clés en lien avec les conditions sanitaires des militaires français. Ces données collectées sont stockées sous forme de fichiers textes, prêts à être utilisés par le logiciel d'analyse ALCESTE®. Il était crucial que ces fichiers soient conformes aux exigences spécifiques d'ALCESTE®, car leur formatage influence directement les résultats de l'analyse. Le format choisi pour chaque corpus impacte l'analyse. Par conséquent, nous avons établi une hiérarchie au sein du fichier et mis en évidence uniquement certaines informations. Chaque fichier formaté suit la structure suivante (Fig. 2) :

```
**** *titre_de_la_video *ann_numero_de_lannee_de_publication
-*mot_cle
Descriptif de la vidéo

@nom_utilisateur_1
Commentaire

@nom_utilisateur_2
Commentaire
```

Figure 2: Format des fichiers formatés pour le logiciel d'analyse Alceste®[5].

Nous avons collecté 43 vidéos traitant de la santé des militaires français à partir de l'année 2021. Après avoir construit notre corpus en incluant les différents mots-clés, nous l'avons ensuite soumis à une analyse textuelle multiforme à l'aide du logiciel ALCESTE®.

5. Analyses menées

Les données récoltées ont été abondantes (6 318 commentaires) et ont nécessité l'adoption d'une méthode d'analyse rigoureuse permettant de les qualifier, les classer et les catégoriser afin de les traiter et de les décrire en détail.

5.1. Analyses descriptives

Les analyses automatiques ont relevé l'existence de 43 vidéos au total. Les vidéos ont été sélectionnées en fonction de plusieurs critères : leur pertinence

[5] Les informations en couleur « orange » sont remplacées et adaptées pour chaque vidéo YouTube.

par rapport au mot clé « santé des militaires français », la variété des perspectives qu'elles offrent ainsi que la diversité des sources d'information. Par exemple, nous avons inclus des vidéos issues à la fois des sources officielles et non- officielles pour obtenir un panorama complet du sujet.

Il a été ensuite décidé d'examiner attentivement la provenance de ces contenus ayant suscité un intérêt pour la santé des militaires :

- Informations officielles en ligne fournies par le gouvernement et l'armée française (gouvernement-l'armée française) ;
- Informations médiatiques en ligne provenant de sources traditionnelles telles que France 2, France 5, Public Sénat, RTL[6], « Le Point », « Le Parisien », RMC[7], « Le HuffPost », Euronews en français et « Télérama » ;
- Informations médiatiques en ligne issues de magazines de santé comme « allo Docteurs » ;
- Informations en ligne issue des médias indépendants (LEGEND, Defense Zone) ;
- Informations médiatiques en ligne diffusées par des influenceurs tels que Poisson Fécond, Les Retex d'Hika, Vincent Firelife et Philteam.

Cette étude des sources d'information abordant le sujet de la santé des militaires français est suivie par l'étude des commentaires afin d'analyser les thématiques de santé qui suscitent l'interaction des militaires et de leur famille.

Nous avons analysé un total de 6 318 commentaires, le plus court se limitant à un seul mot, tandis que le plus long en comptait 1 021 issus des vidéos sélectionnées. Au cours de cette analyse, le logiciel ALCESTE ®, nous a référencé 70 termes les plus significatifs et récurrents dans notre corpus, regroupés en trois classes. Ces termes renvoyant aux commentaires complets d'utilisateurs nous ont permis d'identifier des tendances dans les problématiques sanitaires les plus évoquées par les militaires et leurs familles sur la plateforme YouTube.

5.2. Analyses thématiques

Nous avons utilisé des méthodes d'analyse statistique lexicale pour examiner le corpus de données collectées, dans le but de mettre en évidence les principaux sujets de santé associés à la santé des militaires. La première étape de l'analyse consiste à préparer un corpus textuel formaté, dans lequel les unités de sens sont organisées pour faciliter le traitement statistique. ALCESTE ®, crée ensuite un dictionnaire des termes et de leurs racines, tout en prenant en compte leur fréquence d'apparition.

---

[6] RTL correspond au média Radio Télévision Luxembourg.

[7] RMC est aujourd'hui un groupe de médias français, principalement connu pour sa station de radio qui propose des émissions d'actualité, de sport et de divertissement.

Cette classification initiale permet non seulement de dénombrer les occurrences de chaque terme, mais aussi de révéler des associations sémantiques en fonction de leur répartition au sein du corpus. Le texte est ensuite divisé en sections homogènes, chacune contenant un nombre de mots optimal pour assurer la cohérence sémantique. Ces segments sont ainsi analysés et regroupés selon leurs similitudes et oppositions lexicales, contribuant à une première ébauche des thématiques principales.

L'étape suivante, dite de « classification hiérarchique descendante » ou « méthode Reinert », est au cœur du fonctionnement d'ALCESTE ®, et est fréquemment utilisée dans les sciences humaines et sociales pour sa capacité à regrouper les énoncés significatifs de façon logique et structurée. Cette méthode consiste à subdiviser le corpus selon les cooccurrences des mots, afin d'obtenir des regroupements qui révèlent les grandes oppositions thématiques. Le processus identifie les unités de sens dominantes et organise les termes en classes sémantiques, permettant ainsi d'isoler des concepts centraux à partir des données textuelles. Dans les sciences sociales, ce procédé est particulièrement pertinent pour saisir les préoccupations collectives ou les discours dominants dans un groupe, comme ici dans le contexte de la santé militaire. Pour renforcer la compréhension des thèmes extraits, nous nous appuyons sur le concept de thématisation, un principe central en analyse textuelle qui permet de structurer le contenu en thèmes représentatifs du sujet étudié. Alex Mucchielli et Pierre Paillé décrivent cette approche en expliquant que « la thématisation constitue l'opération centrale [..] à savoir la transposition d'un corpus donné en un certain nombre de thèmes représentatifs du contenu analysé et ce, en rapport avec l'orientation de recherche » (Paillé et Mucchielli 2021, 231). Cette étape est essentielle pour organiser les données en fonction de la problématique de recherche, créant ainsi un cadre logique et structuré qui donne une vue d'ensemble des enjeux de santé dans le corpus militaire étudié.

À la suite de cette classification, nous avons atteint un taux de classification élevé : environ 95 % des unités textuelles ont été intégrées dans les classes identifiées, tandis que seuls 5 % des éléments ont été écartés en raison d'une pertinence moindre ou d'un manque de cohérence avec les autres segments. Ce niveau de pertinence est significatif, car il témoigne de la rigueur et de l'efficacité du logiciel ALCESTE®, dans le traitement de grandes quantités de données complexes. Les unités classées ont été regroupées en trois catégories principales appelées « classes d'énoncés significatifs » ou simplement « classes », chacune d'elles représentant un ensemble cohérent de préoccupations de santé militaire. Ce processus de dépouillement des données à travers la construction de grilles thématiques a non seulement permis d'organiser les informations de manière structurée, mais aussi de mieux saisir l'univers sémantique du corpus. Vous trouverez ici un exemple de graphe issu du logiciel

ALCESTE®, et qui nous visualise les termes les plus caractéristiques associés à la modalité[8] « année 2023 ».

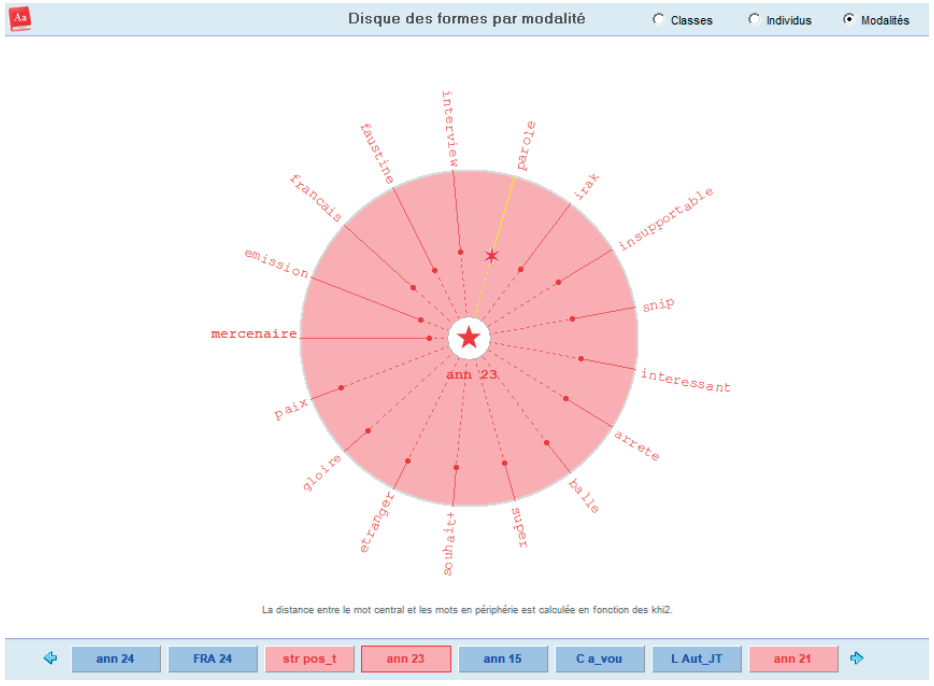


Figure 3: Disque de formes par modalité « année 2023 ».

6. Résultats

6.1. Sources d'informations

Nos résultats permettent d'identifier 13 sources d'information. Nous entendons ici par source d'information les chaînes YouTube qui diffusent des contenus vidéos autour de l'univers de la santé militaire. Nous observons un pic de vidéos diffusées par la chaîne audiovisuelle France 2 sur l'état de santé des militaires. L'émission, intitulée « ça commence aujourd'hui », en produit 11, centrée sur les témoignages de militaires et de leurs conjoints. On dénombre également quatre vidéos publiées par des influenceurs, deux vidéos provenant de médias indépendants gérés par des journalistes professionnels (Legend, Défense, Zone) et deux autres issues de sites institutionnels de l'armée française et du gouvernement. En termes de nombre de publications vidéo, ce sont les médias traditionnels qui dominent le traitement de ce sujet (Fig. 4).

[8] Nous entendons par modalité, un critère de segmentation de notre corpus par année.

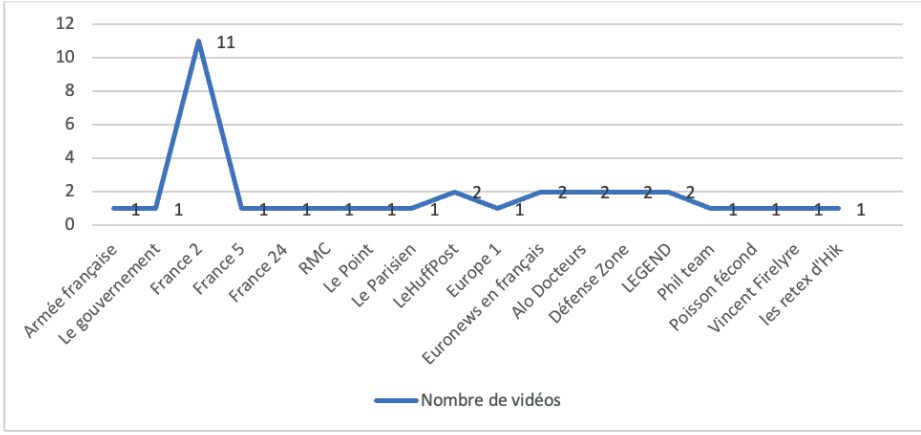


Figure 4: Nombre total de contenu vidéos collectées sur la plateforme YouTube.

Il est intéressant de noter que ce sont principalement les médias indépendants et les chaînes YouTube des influenceurs qui réussissent à fidéliser leur audience et à susciter une forte interaction avec celle-ci. En effet, quatre de ces médias indépendants ont recueilli plus de 500 commentaires, même si la chaîne du média « Le Parisien » a atteint le record de commentaires (Fig. 5)[9].

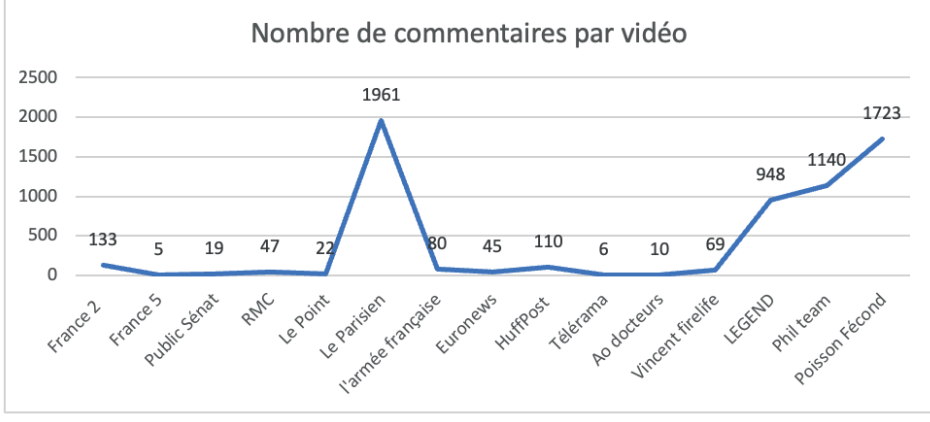


Figure 5: Courbe de croisement de la modalité « source d'information » et « nombre de commentaires ».

La vidéo du Parisien est suivie d'un contenu publié par l'influenceur Poisson Fécond. Cette vidéo intitulée *Pourquoi Ces Soldats Sont Devenus Fous ?!* a enregistré 1,1 million de vues et 1 723 commentaires entre à la fois des témoignages de personnes ordinaires, des militaires mais aussi des familles de mili-

[9] Voir infra en thématique 2.

taires. Ce ne sont donc plus exclusivement les journalistes professionnels, qui sont les seuls aptes à filtrer l'information essentielle en direction de l'opinion publique, mais ce sont aussi les personnes ordinaires ou les influenceurs qui sélectionnent l'information à transmettre à l'opinion publique. Cette logique de publication concerne également l'environnement militaire et procure autant d'adresses où peuvent venir s'agréger de nombreux témoignages de soutien entre communauté militaire et une volonté de libération de la parole. L'examen de tous les commentaires sur chaque vidéo de la figure 3 nous a permis d'identifier certaines questions de santé les plus fréquemment abordées dans ces différentes sources.

6.2. Thématiques retrouvées

L'analyse retrouve trois grandes thématiques :
- Les décès de soldats français ;
- Les préoccupations physiques ;
- La santé psychologique.

6.1.1 Décès de militaires français

Les décès liés au service, par exposition au combat, sont intégrés dans les analyses de santé militaire et sont étudiés en termes de prévention. Une analyse de 2012 des pertes au combat au cours de la première décennie des conflits qui ont suivi le 11 septembre 2001 en Irak et en Afghanistan a révélé que, sur 87,3 % des décès, environ un sur quatre aurait pu être évité au point de vue médical, ce qui signifie que les soins préhospitaliers et l'évacuation ont influencé ces décès au combat (Eastridge et al. 2012). Même si la mort de soldats français ne peut pas être considérée comme une «information de santé» au sens strict, elle est associée à des problématiques de santé publique à partir du moment où elle suscite des inquiétudes quant au bien-être physique ou mental des soldats ou des futurs soldats.

Dans l'émission LEGEND diffusée sur le réseau YouTube, produite et animée par Guillaume Pley, Pascal B, « sniper » de l'armée française, raconte ses souffrances psychologiques. Nous constatons de nombreux commentaires de jeunes qui souhaitent rejoindre l'armée, et qui ressentent déjà un grand stress face aux décès des militaires qui peut survenir lors des déplacements en opération :

> Très touchante cette interview surtout quand il parle des funérailles de ses collègues je suis pas encore militaire, mais bientôt j'espère et clairement quand j'entends qu'un Gendarme est décédé ou qu'un militaire s'est fait tuer au combat ça me met les larmes aux yeux alors que j'ai 19 ans tête brûlée normalement faut y aller pour me décrocher une larme, mais dès que ça parle de l'armée

> c'est pas pareil pour moi c'est des héros qui se font tuer des gens qui sont là pour nous protéger et quand un d'entre eux part ça me rend assez triste et je n'imagine même pas la sensation quand tu sais qu'un de tes potes est mort au combat ça doit être 1 x pire.

Cette thématique rassemble tous les commentaires de toutes les vidéos qui font état du décès de soldats français dans le cadre des opérations menées au Mali entre 2020 et 2022, ainsi que celui de deux autres soldats tués en Irak en 2023. Cette catégorie représente 23 % du corpus global.

Dans un premier temps, nous relevons plusieurs messages écrits concernant le décès des militaires français au Mali. En particulier, dans ces messages, il est demandé le retrait des troupes françaises du Sahel, par exemple dans une vidéo publiée par la chaîne YouTube du média Euronews (2021), *Sahel : les décès de cinq militaires relancent le débat sur l'engagement français*.

Dans l'actualité médiatique du corpus traité, en 2023, le chef de l'État français, Emmanuel Macron, a annoncé le décès d'un militaire en Irak, soit le deuxième en deux jours consécutifs. Cette nouvelle a fait l'objet d'une large diffusion, notamment grâce à une vidéo diffusée par le magazine « Le Point » le 21 août. Au côté des vidéos issues des médias traditionnels, les médias officiels de l'institution militaire et du gouvernement diffusent également du contenu sur YouTube en hommage aux militaires décédés en opération. La chaîne de l'armée française a publié une vidéo intitulée *À la mémoire de nos militaires français morts au combat*. Cette vidéo a été visionnée 41 0000 fois, suivi de 80 commentaires[10], lesquels envoient des messages de condoléances à ces soldats et leurs familles et témoignent du respect envers les combattants : « Force et honneur à nos soldats. A eux le pouvoir, car eux seuls sont capables de donner leurs vies pour la Patrie et le sens de servir, protéger le peuple c'est eux seuls qui l'ont aujourd'hui. Reposez en paix près de vos frères d'armes. Un grand soutien aux familles ».

La figure 6 ci-dessous présente la courbe des différents médias qui publient sur la thématique du décès des militaires ou encore des commentaires de militaires eux-mêmes qui évoquent leur douleur face à ces morts aux combats : « Ancien 2éme REP, même formaté, on restera des hommes. Mes camarades mes frères la douleur restera toujours dans nos mémoires. Que dieu vous grade Patria Nostra ».

---

[10] Voir infra (courbe nombre de commentaires par média).

Figure 6: Courbe de croisement de la modalité « décès des militaires » et le nombre de vidéos par année.

C'est donc l'année 2021 et l'année 2023 qui présentent le plus de vidéos autour de la thématique du décès des militaires français. Les sources d'information publiant sur cette thématique sont : Euronews, « Le Parisien » et les chaînes YouTube officielles du gouvernement et du ministère des armées français.

6.1.2 Préoccupations physiques des militaires

Les informations sur l'état de santé physique des militaires représentent environ un quart de l'ensemble des données YouTube. Ces données sont principalement issues des sources traditionnelles, avec une prédominance pour le journal « Le Parisien », qui a d'ailleurs généré le plus grand nombre de réactions dans notre corpus global. Dans une vidéo du corpus, nous assistons à une séquence où deux soldates de la Marine nationale semblent avoir eu un malaise pendant les vœux du Président Macron aux Forces armées. Cette scène, captée en direct, a été relayée par divers médias officiels et a engendré de multiples commentaires (Fig. 7).

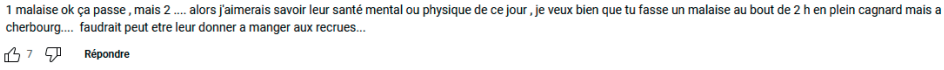


Figure 7: Commentaire autour du malaise de deux soldates de la Marine nationale pendant les vœux aux Forces armées.

En examinant le nuage de mots associés à cette catégorie, on note non seulement les mots tels que « malaise », « vagal », « tomber », mais aussi des

termes clés comme « dose », « vaccin » et « injecter », qui sont pertinents dans ce contexte (Fig. 8).

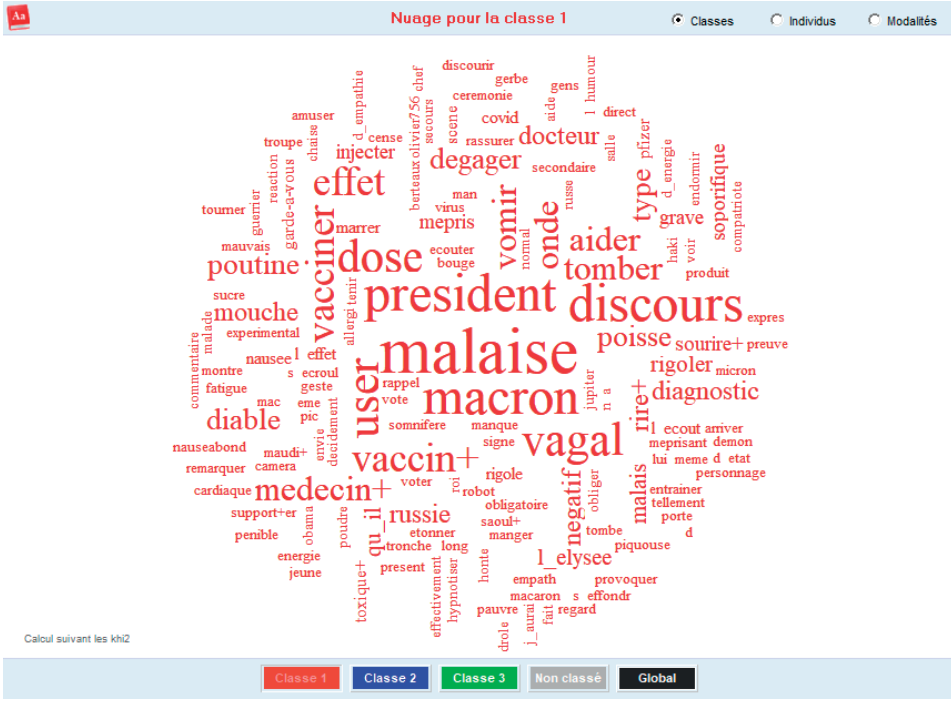


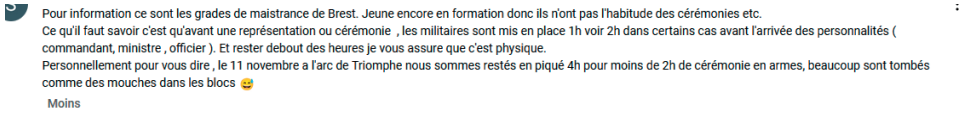
Figure 8: Nuage des mots de la thématique 2 issu du logiciel ALCESTE®.

En effet, de nombreux commentateurs YouTube associent ce malaise des femmes militaires à des réactions indésirables liées aux vaccins contre la Covid-19. Voici quelques commentaires relevés :

- « Ben en même temps quand tu t'es vacciné avec 3 doses ne faut pas s'étonner de terminer avec des symptômes du moyen âge » ;
- « Va donc savoir si ce n'est pas la conséquence de tous les vaccins qu'on leur injecte » ;
- « Policiers, militaires etc.… sont-ils contraints à continuer de se faire vacciner contre le Covid-19 ??? Deux personnes qui font un malaise. Étrange » ;
- « Ces malaises sont dus aux effets secondaires des vaccins. Et non à un malaise vagal comme il le prétend » *;*
- *«* effets secondaires des injections géniques expérimentales sous les yeux du principal ordonnateur de ce massacre passe présent et à venir ».

Les militaires et leurs familles réagissent et affirment que ces malaises sont monnaie courante au sein de l'armée. Voici un exemple de commentaire : « Pour avoir eu un fils militaire, je sais que ceux-ci attendent plusieurs heures immobiles avant que les "huiles" débutent les divers discours. J'ai vu lors de la fin des "classes" et de la remise de galons des 1ere classes, un marin musicien qui tenait la grosse caisse s'écrouler de fatigue. C'était impressionnant ».

Ou encore le témoignage de ce militaire, qui met en perspective l'évènement, en s'appuyant sur sa propre expérience :


Pour information ce sont les grades de maistrance de Brest. Jeune encore en formation donc ils n'ont pas l'habitude des cérémonies etc.
Ce qu'il faut savoir c'est qu'avant une représentation ou cérémonie , les militaires sont mis en place 1h voir 2h dans certains cas avant l'arrivée des personnalités (commandant, ministre , officier). Et rester debout des heures je vous assure que c'est physique.
Personnellement pour vous dire , le 11 novembre a l'arc de Triomphe nous sommes restés en piqué 4h pour moins de 2h de cérémonie en armes, beaucoup sont tombés comme des mouches dans les blocs 😅
Moins

La participation en ligne des militaires et de leur famille apporte ici une forme de connaissance de terrain et cadre les pratiques communicationnelles des personnes ordinaires, dans une logique de confrontation des points de vue, et de transfert des informations expérientielles (Sedda et al. 2022) à des personnes connectées. Cette confrontation des idées accouplée à une logique de mise en scène de soi et de son expérience positionne le militaire au croisement d'un récit sur l'intime et pour le collectif, sur l'institutionnel pour l'ordinaire. Expliquer les valeurs de l'institution militaire et communiquer à des fins de recadrage de sens est une des pratiques info-communicationnelles les plus fréquentes des militaires relevées sur la plateforme YouTube à propos des sujets de santé.

6.1.3 Santé psychologique des militaires

Selon l'Organisation mondiale de la santé (OMS), la qualité de vie est « associée de manière complexe à différents facteurs, incluant : la santé physique, l'état psychologique, le niveau d'indépendance, les relations sociales, la relation avec l'environnement, la culture et la politique » (Tap et Roudès 2008). La qualité de vie des militaires dépend donc de leurs conditions à la fois physiques et psychologiques. L'étude des commentaires YouTube retrouvés dans ce corpus nous permet d'identifier les femmes de militaires comme principales actrices communiquant sur cette thématique, aux côtés des voix de militaires. Qu'elles prennent la parole comme invitées d'émissions télévisées diffusées ensuite sur la plateforme YouTube, ou qu'elles réagissent avec des commentaires suite à la publication de vidéos sur ce sujet, elles sont au cœur des pratiques info-communicationnelles entourant la santé psychologique des militaires. La figure 9 ci-dessous présente la courbe des différents médias qui publient sur la thématique de la santé psychologique des militaires (Fig. 9).

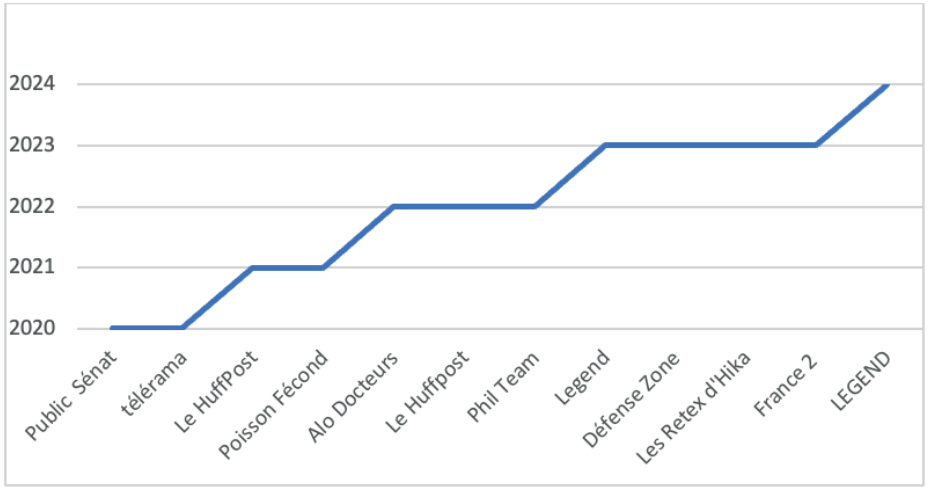


Figure 9: Croisement modalité « source d'information » et « année » concernant la thématique santé psychologique des militaires.

Les médias du service public ont diffusé 9 vidéos sur la plateforme YouTube autour de l'état psycho-sociale des militaires et de leurs conjointes. L'émission *Ça commence aujourd'hui* de France 2 publie, elle seule, 5 vidéos. Il s'agit, dans cette émission, de présenter les récits de militaires et de femmes de militaires qui racontent l'angoisse des départs en opération.

La douleur de l'absence, le manque, la peur au ventre sont autant de descriptifs employés par ces femmes dans ces émissions pour qualifier le poids du métier de leurs conjoints sur leurs vies conjugales. La vidéo, en conjuguant des images poignantes de femmes de militaires en détresse et de leurs témoignages sincères, invite à réfléchir sur les réalités souvent ignorées de leurs époux soldats en mission, tout en créant un lien empathique fort avec l'audience.

Ces récits de femmes de militaires en vidéo ont d'ailleurs suscité de nombreux commentaires sur le réseau YouTube. En plus des commentaires de courage adressés à ces femmes, d'autres femmes prennent la parole et racontent leurs angoisses en commentaires, dont voici un exemple : « Femme de marin, cette émission est très poignante. Être femme de marin militaire, c'est beaucoup de sacrifices : l'éloignement, l'absence d'un mari d'un père est parfois très dure à supporter, mais ce qui est véridique, toutes femmes de militaire le diront, nous sommes très soudés nous ne montrons rien de nos émotions on passe nos journées à les attendre, car on aime nos conjoints on remonte le moral on s'occupe de tout, mais le plus beau cadeau c'est de les retrouver à leur retour de mission l'adrénaline est toujours présente ».

Cette forme de communication de ces femmes consistant à publiciser une partie de leur vie sur Internet est une pratique d'extimité (Tisseron 2011), qui

révèle un besoin qu'ont les femmes de militaires d'extérioriser publiquement des situations personnelles traditionnellement réservées à la sphère privée. Prendre la parole est ici une tentative de soigner la blessure des non-dits et un désir de sortir de l'entre soi.

L'analyse de notre corpus nous a également permis d'identifier quatre influenceurs qui diffusent des vidéos sur cette thématique : Poisson Fécond, Vincent Firelife, Les RETEX d'Hika et Phil Team. Ces influenceurs créent un espace d'interaction où nous pouvons entendre la voix du militaire en dehors des officines institutionnelles.

La vidéo publiée sur la chaîne YouTube Les RETEX d'Hika en 2023 retrace le témoignage d'Aurélien, blessé psychique du Mali, et de Sabrina, sa femme.

Sabrina parle de solitude des femmes face à ces troubles de leurs compagnons militaires. Voici un de ses commentaires : « On est l'appui administratif parce que nos hommes, il arrive un moment où ils ne peuvent pas remplir un document parce que leur syndrome ne leur permet plus, donc moi je suis obligée de prendre ça en charge, mais je n'ai pas d'appui pour m'aiguiller vers les directions à prendre ».

On note que la demande de soutien et la quête de reconnaissance (Honneth 2013) sont très présentes dans les discours des femmes de militaires. En voici un exemple : « Je suis femme de militaire. Les départs en Opex toujours compliqué... La peur, la boule au ventre. Le travail. Les 4 enfants à gérer et à rassurer. Oui un peu de reconnaissance ne serait pas un mal. J'ai celle de mon homme c'est quand même le plus important ».

7. Discussions- Conclusions

Notre étude a collecté 43 vidéos traitant de la santé des militaires français à partir de l'année 2021 sur la plateforme YouTube. Après avoir réalisé une analyse textuelle multiforme de notre corpus de 6 318 commentaires provenant de 43 vidéos YouTube à l'aide du logiciel ALCESTE®, nous avons identifié trois grandes thématiques : les décès de soldats français, les préoccupations sur la santé physique et psychologique avec différents commentaires, témoignages et influenceurs présentés dans cet article.

En discussion, la plateforme YouTube retient l'attention des militaires, anciens militaires et de leurs familles qui y voient matière à faire circuler leurs représentations de l'institution militaire et leurs témoignages sur leur état de santé. Ces pratiques engendrent une forme de sympathie entre les militaires et le grand public, et bouleversent le schéma traditionnel où la rigidité et la discrétion sont les maîtres-mots. Cet article met en évidence trois points essentiels. L'état de santé militaires sur les médias numériques n'est pas un sujet tabou. Nous assistons au contraire à une déterritorialisation de la parole pour inclure des communautés virtuelles. De plus, la plateforme numérique You-

Tube est une chambre d'écho des médias traditionnels qui trouvent un lieu pour diffuser leurs contenus et donc ainsi susciter une large participation de la part des utilisateurs YouTube autour de la santé militaire. Enfin, nous observons une interaction marquée des internautes, générée tant par les médias traditionnels que par les chaînes YouTube des influenceurs et des journalistes indépendants, raison pour laquelle il nous est difficile d'opposer médias traditionnels et médias numériques dans la cadre de ce sujet.

En conclusion générale, notre travail met en évidence que les pratiques info-communicationnelles des militaires et de leurs familles se traduisent par un discours numérique centré à la fois sur le récit informatif et affectif. L'information numérique concernant l'état de santé des militaires, tant physique que psychologique, permettent de partager des expériences vécues liées à des problèmes de santé. Cela contourne le devoir de réserve et l'obligation de silence en utilisant les médias sociaux, traduisant ainsi un désir, direct ou indirect, de réparation ou de reconnaissance. Ce phénomène se manifeste principalement à travers les témoignages de femmes de militaires. D'autres pratiques numériques de militaires viennent faire circuler leur expérience de terrain afin de consolider le point de vue militaire et de poser un cadre à la communication des personnes ordinaires sur l'univers militaire. En plus de ces pratiques info-communicationnelles en ligne à visée informative, nous relevons des témoignages et des commentaires YouTube avec une forte dimension affective, permettant une construction d'une communauté numérique basée sur la sympathie et l'expression de la solidarité à l'égard du sujet militaire.

Références bibliographiques

Berger, Aude. 2015. "Le Patient Militaire Internaute en tant que consommateur d'information santé en ligne. Étude qualitative basée sur le ressenti de 16 patients du 3ème Régiment d'Hélicoptères de Combat d'Étain." Thèse de Doctorat, Université de Lorraine. http://docnum.univ-lorraine.fr/public/BUMED_T_2015_BERGER_AUDE.pdf.

Boullier, Dominique. 2015. "Les sciences sociales face aux traces du big data: Société, opinion ou vibrations ?" *Revue française de science politique* 65 (5): 805-28. https://doi.org/10.3917/rfsp.655.0805.

Boumhaouad, Hatim. 2017. "Pratiques info-communicationnelles des usagers des dispositifs numériques. Théorie de l'acteur-réseau." *Les Cahiers du numérique* 13(3) : 137-66. https://shs.cairn.info/revue-les-cahiers-du-numerique-2017-3-page-137?lang=fr.

Cardon, Dominique, et Marie-Carmen Smyrnelis. 2012. "La démocratie Internet : Entretien avec Dominique Cardon." *Transversalités* 123 (3) : 65-73. https://doi.org/10.3917/trans.123.0065.

Centre interarmées de concepts, de doctrines et d'expérimentations. 2013. "Réseaux sociaux. Nature et conséquences pour les forces armées." *Réflexion doctrinale interarmées.* RDIA-2013/001_RS. https://www.defense.gouv.fr/sites/default/files/cicde/20130419-NP-CICDE-RDIA-3.10.1-RESEAUX-SOCIAUX-2013_ex-RDIA-2013-001.pdf.

Coppersmith, Glen, Craig Harman, and Mark Dredze. 2014. "Measuring Post Traumatic Stress Disorder in Twitter." *Proceedings of the International AAAI Conference on Web and Social Media* 8 (1): 579-82. https://doi.org/10.1609/icwsm.v8i1.14574.

Craig, Bryan J., Jonathan E. Butner, Sungchoon Sinclair, Anna Belle O. Bryan, et al. 2018. "Predictors of Emerging Suicide Death Among Military Personnel on Social Media Networks." *Suicide and Life-Threatening Behavior* 48(4): 413-30. https://doi.org/10.1111/sltb.12370.

Demarthon, Fabrice. 2012. "Le Big Data : un enjeu économique et scientifique." *CNRS Le journal*. https://lejournal.cnrs.fr/articles/le-big-data-un-enjeu-economique-et-scientifique.

Eastridge, Brian J., Robert L. Mabry, Peter Seguin, et al. 2012. "Death on the Battlefield (2001–2011): Implications for the Future of Combat Casualty Care." *Journal of Trauma and Acute Care Surgery* 73 (6): S43137. https://doi.org/10.1097/TA.0b013e3182755dcc.

González-Padilla, Daniel A., and Leonardo Tortolero-Blanco. 2020. "Social Media Influence in the COVID-19 Pandemic." *International Braz j Urol : Official Journal of the Brazilian Society of Urology* 46 (suppl.1): 120-24. https://doi.org/10.1590/S1677-5538.IBJU.2020.S121.

Honneth, Axel. 2013. *La lutte pour la reconnaissance*. Gallimard.

Ifrah, Laurence. 2010. "Le Web 2.0 et les réseaux sociaux, l'envers du décor." *Que sais-je ?*, no. 3881 (juin) : 60-72.

Ipsos. 2024. "Le monde en chiffres - Mythes et réalités des réseaux sociaux." 12 avril. https://www.ipsos.com/fr-fr/le-monde-en-chiffres-mythes-et-realites-des-reseaux-sociaux.

Lévi-Strauss, Claude, Jean Petitot, et Jean Marie Benoist. 1977. *L'identité, Séminaire interdisciplinaire (1974-1975)*. Bernard Grasset.

Paillé, Pierre, et Alex Mucchielli. 2021. "L'analyse thématique." In *L'analyse qualitative en sciences humaines et sociales*, 231-314. Armand Colin. https://shs.cairn.info/l-analyse-qualitative-en-sciences-humaines--9782200624019-page-269?lang=fr.

Pavalanathan, Umashanthi, Vivek Datla, Svitlana Volkova, et al. 2016. "Discourse, Health and Well-being of Military Populations Through the Social Media Lens." *AAAI Workshop: WWW and Population Health Intelligence*, https://pdfs.semanticscholar.org/330a/a7bd1612370206956743efc739229563f745.pdf.

Resteigne, Delphine, et John Tavernier. 2012. "Internet, un moteur de transformation des organisations militaires ?" *Pyramides. Revue du Centre d'études et de recherches en administration publique* (24) : 63-75. https://journals.openedition.org/pyramides/927?lang=en.

Rheingold, Howard. 1994. *The Virtual Community. Homesteading on the Electronic Frontier*. Harper Perennial Editions.

Schweisguth, Etienne. 1978. "L'institution militaire et son système de valeurs." *Revue française de sociologie* 19 (3): 373-90. https://doi.org/10.2307/3321050.

Sedda, Paola, Nataly Botero, et Myriam Hernández Orellana. 2022. "Influenceurs et influenceuses santé : les récits et les savoirs du corps sur les réseaux sociaux." *Études de communication. langages, information, médiations*, no. 58 (septembre) : 7-24. https://doi.org/10.4000/edc.14155.

Snickars, Pelle, et Paul Vonderau. 2009. *The YouTube reader*. National Library of Sweden.

Tanti, Marc. 2023. "Quelles sont les fakes news, théories du complot et controverses qui ont circulé à travers les frontières numériques lors de la première vague de Covid-19 ?" *AIDAinformazioni*, année 41, no. 1-2 (janvier-juin) : 133-52.

Tap, Pierre, et Rolande Roudès. 2008. "Qualité de vie, souffrances et identité(s)." *Le Journal des psychologues* 260 (7) : 41-47. https://doi.org/10.3917/jdp.260.0041.

Tisseron, Serge. 2011. "Intimité et extimité." *Communications* (88) : 83-91. https://doi.org/10.3917/commu.088.0083.

Volkova, Svitlana, Lauren E. Charles, Josh Harrison, and Courtney D. Corley. 2017. "Uncovering the Relationships between Military Community Health and Affects Expressed in Social Media." *EPJ Data Science* 6 (1): 1-23. https://doi.org/10.1140/epjds/s13688-017-0102-z.

AIDAinformazioni
ISSN 1121–0095
ISBN 979-12-5965-456-4
DOI 10.57574/596545643
pag. 51-78 (luglio-dicembre 2024)

Assisted morbidity coding: the SISCO.web use case for identifying the main diagnosis in Hospital Discharge Records

Elena Cardillo[*], Lucilla Frattura[**]

Abstract: Coding morbidity data using international standard diagnostic classifications is increasingly important and still challenging. Clinical coders and physicians assign codes to patient episodes based on their interpretation of case notes or electronic patient records. Therefore, accurate coding relies on the legibility of case notes and the coders' understanding of medical terminology. During the last ten years, many studies have shown poor reproducibility of clinical coding, even recently, with the application of Artificial Intelligence-based models. Given this context, the paper aims to present the SISCO.web approach designed to support physicians in filling in Hospital Discharge Records with proper diagnoses and procedures codes using the International Classification of Diseases (9th and 10th revisions), and, above all, in identifying the main pathological condition. The web service leverages NLP algorithms, specific coding rules, as well as ad hoc decision trees to identify the main condition, showing promising results in providing accurate ICD coding suggestions.

Keywords: Coding Support Systems, Hospital Discharge Records, ICD, Morbidity coding, Coding Rules.

1. Introduction

The proper use of standard classifications, such as the International Classification of Diseases (ICD) and coding of morbidity data has always been fundamental for all general epidemiological and many health-management purposes (WHO 2016). One example is the use of the information flow of the Hospital Discharge Records (SDO) collected in national databases for monitoring hospitalization episodes provided in public and private hospitals and thus the provision of hospital assistance. This has become an indispensable tool for both administrative analyses (i.e., for accurate billing) and clinical

[*] Institute of Informatics and Telematics, National Research Council (IIT-CNR), Rende, Italy. elena.cardillo@iit.cnr.it.
[**] Azienda Sanitaria Universitaria Giuliano Isontina (ASUGI), Udine, Italy. lucilla.frattura@asugi.sanita.fvg.it.

elaborations (e.g., health quality assessment), which can bring to the planning of new measures to support healthcare and welfare activities or to more strictly clinical-epidemiological and outcome analyses.

In this frame, although approaches to coding vary across institutions, clinical coding specialists frequently perform coding retrospectively. The assignment of codes to each patient episode of care during hospitalization is determined by different factors, among others by the coder's interpretation of the available case notes or the completeness of the electronic health records. As a result, accurate coding is dependent on both the intelligibility of the case notes and the coders' knowledge of medical terminology (Sundararajan et al. 2015).

Several studies have indicated poor reproducibility of clinical coding (Tatham 2008) and poor accuracy which seems not dependent on the version of the standard coding system used, which in the case of SDO is ICD (Quan et al. 2014).

In recent years, even if the application of artificial intelligence (AI) has begun to attract and, in some cases, assist clinicians in the practice of medical coding, the performances achieved by AI models do not meet expectations. Many studies have proven this, especially concerning inadequate levels of data coding accuracy (less than 50%) and high computational costs (Falis et al. 2024; Soroush et al. 2024). This means that more reliable and trustworthy systems are required to support physicians or coders in speeding up the coding process while retaining the necessary precision.

Given this context, the paper aims to describe the results of the "SISCO.web" project[1], whose scope was to design and implement a Coding Support System (CSS), in the form of a web service, to improve accuracy in coding health conditions in Italian Hospital Discharge Records (SDO). The main objective of the service is to support Italian physicians (coders) in morbidity coding, and more specifically in the coding of diagnoses and procedures/interventions using ICD-9th revision, Clinical Modifications (ICD-9-CM), mandatory in Italy, and, more notably, in identifying the "main condition" to be filled in SDOs.

The paper is structured as follows: Section 2 provides background information on using and coding SDO, and describes the applied methodology. Section 3 showcases the results and includes a preliminary evaluation. Section 4 presents some related works, and finally, Section 5 offers conclusions and future directions.

---

[1] The "SISCO.web" project, funded by the Friuli Venezia Giulia (FVG) Region and coordinated by the Italian Collaborating center of the World Health Organization Family of International Classifications (WHO-FIC) in Udine through the Azienda Sanitaria Bassa Friulana Isontina n. 2 (incorporated now into the "Azienda sanitaria universitaria Giuliano Isontina" - ASUGI) was executed from 2017 to 2021 and led to the development of a prototype (SISCO.web service) which can assist clinicians in coding SDO data using ICD-9-CM, but it is also set up to support ICD-10 coding.

2. Materials and Methods

2.1. Hospital Discharge Records

The Hospital Discharge Record Database was established, in Italy, with the Decree of the Ministry of Health on 28 December 1991. It serves as a tool for collecting information about each patient discharged from public and private hospitalization institutions across the country. The information gathered in each SDO includes, beyond the patient's characteristics (e.g., age, sex, etc.), the peculiarities of the hospitalization (e.g., institution and discharge discipline, method of discharge, etc.) and, above all, clinical features (e.g., the main diagnosis, concomitant diagnoses, diagnostic or therapeutic procedures, and interventions), excluding information relating to drugs administered during hospitalization[2].

Subsequently, other decrees introduced new regulations for the information flow transmission to the Ministry of Health, expanded the information content of the SDO, and adopted the international classification ICD-9-CM version 1997 (Italian Ministry of Health 2000) for the coding of diagnoses and diagnostic and therapeutic procedures, then updating this regulation with the adoption of the 2007 Italian version and introducing the adoption of the Diagnosis Related Group classification (DRG), version 24 for hospital admissions (Italian Ministry of Labor, Health and Social Affairs 2008a).

In 2011, the "It.DRG Project", coordinated by the Ministry of Health, was launched to develop a new classification and assessment method for inpatient care, specific and representative to the Italian context (Sforza et al. 2021). The objective of this project was: the development and testing of an updated version of the ICD-10 classification (International Classification of Diseases and Health Related Problems-10th Revision) that incorporates WHO-approved updates and makes minor changes, finalizing the so-called Italian modification of ICD-10 (ICD-10-IM); the development and testing of the Italian classification of Interventions and Procedures (CIPI), a version of the section on procedures and interventions of ICD-9-CM modified and supplemented, to adapt it to specific Italian needs and to provide for integration with codes that allow for the detection of information on: (i) Procedures/treatments provided (also) in ambulatory care; (ii) Medical-surgical devices; (iii) High-cost drugs; and iii) finally, a new version of the DRG system (Nonis et al. 2018).

Despite the significant outcomes of the "It.DRG project" for innovating and improving SDO data management, there is a need to create a roadmap for implementing the new classifications, especially ICD-10, in a more simplified manner. This involves using crosswalking tables to ICD-10-IM and

[2] Hospital Discharge Records database (HDR/SDO), see for details European Health Information Portal (2023).

confirming the planned current version of DRG classification. The attention in this paper is paid primarily to a tool for coding diagnoses and intervention using ICD-9-CM, with the understanding that the mentioned crosswalking tables for coding diagnoses in ICD-10-IM can be easily implemented in the tool's architecture.

2.1.1. The International Classification of Disease

The International Classification of Disease is the most known and widely used standardized WHO classification system, which was originally intended to facilitate the statistical analysis of health data (Moriyama et al. 2011). Each successive revision to the ICD, typically spanning 10-20 years, has sought to address new use cases while adapting to advances in medicine and healthcare and has continued to grow in number of total codes (Williamson et al. 2024). The tenth version has approximately 14,000 codes for health conditions, signs, symptoms, and reasons to encounter health services. This revision has then been renewed with the implementation of the eleventh revision of the classification, ICD-11 (World Health Organization 2019/2021), developed thanks to an unprecedented collaboration between WHO working groups, knowledge engineers and informaticians from Stanford University (USA), and professionals all over the world to become a global standard for health data, clinical documentation and statistical aggregation. It presents a new coding structure compared to previous revisions and is fully digital for the first time. The basic component is an underlying ontology database containing all ICD entities (over 55,000 unique entities)[3]. The new structure, its digital nature, and the tools provided to support the use of the classification enhanced its application flexibility. Moreover, it is interoperable with health information systems and other coding systems.

As mentioned above, in Italy, ICD-9-CM is used for morbidity coding, containing over 15,000 diagnosis codes. Its use is also recommended in primary care prescription documents and for diagnoses and problems encoding in the Italian Patient Summary (Italian Permanent working table for Digital health in Regions and Autonomous Provinces 2010) each entity within the ICD-9-CM is encoded by a unique identification string consisting of three to five digits and an optional single letter prefix corresponding to a supplementary category. Practical applications of the ICD in healthcare have expanded and now have come to include the indexing of health record data in hospitals, the

[3] These entities include diseases, injuries, external causes, signs and symptoms, substances, drugs, anatomy, etc., pointing to about 17,000 categories, for over 120,000 clinical terms covered, allowing the description of health conditions at any level of detail by combining codes.

coding of medical billing claims (Moriyama et al. 2011), and the assessment of quality of patient care (O'Malley et al. 2005).

2.1.2. The coding of the main condition

A coded health data record can have a varying number of diagnostic codes. Some authors, considering that there is no uniform definition of "main condition", noted that one of these diagnoses must be coded as the main condition, known also as "main diagnosis", "primary diagnosis", "principal diagnosis or "discharge diagnosis" (Sukanya 2017).

Two definitions have been used for the main condition in ICD-coded health data: a "resource use" definition and a "reason for admission" definition. In Italy, the first definition is implemented, as said above, in detecting and coding the discharge diagnosis using ICD-9-CM, 2007 version (Italian Ministry of Labor, Health and Social Affairs 2008b). In the Italian SDO, it is necessary to code the main diagnosis, and several other diagnoses related to the hospital episode of care. The mentioned national database on SDO contains more than 290 million records (7,957,647 only in 2023). Annual reports are available for download from the website of the Italian Ministry of Health (Italian Ministry of Health 2024). Coding of these records is made directly by clinicians and health professionals, with some levels of accuracy monitoring at the hospital and regional level before the data are sent to the Ministry of Health periodically. This richness of data must face with its accuracy. Several Italian studies are available showing low accuracy in coding. Hospital discharge data were found to be specific but insensitive in many fields. For example, the reporting of acute ischemic stroke and thrombolysis provides misleading indications about both the quantity and quality of acute ischemic stroke hospital care in many studies (Rinaldi et al. 2003; Spolaore et al. 2005). Other studies show that Hospital discharge records appear to poorly reflect the incidence of amyotrophic lateral sclerosis and can be used only after clinical verification of the diagnosis (Chiò et al. 2002). Moreover, looking at (Amodio et al. 2014), the diagnosis of influenza seems to be overcoded. Nevertheless, based on the retrieved evidence, administrative databases can be employed to identify primary breast cancer. The best algorithm suggested is ICD-9 or ICD-10 codes located in the primary position (Abraha et al. 2018). At an international level, many studies confirmed that physicians do not code the disease in SDOs according to the main diagnosis principles (Wang et al. 2021). It is observed that in many cases, the main diagnosis is mistaken for an outpatient diagnosis, making it more difficult to identify when multiple diseases occur simultaneously or in cases of complications. These studies reveal that physicians still require support to collect, classify, analyze, and use medical record information according to disease classification criteria.

2.2. The SISCO.web approach

The main scope of the SISCO.web service, as mentioned above, is to support the coding of SDOs, guiding the physicians to identify and code the main condition, allowing the most appropriate ICD-9-CM codes, and in the future ICD-10 codes. This means that its function is to guide the user before the compilation of the SDOs, to choose and assign appropriate ICD codes to the diagnostic formulations available in medical record documentation collected during patient hospitalization, and, further, to identify among different diagnoses, the main one (Cardillo et al. 2019). Peculiarities of this support system are:

- A knowledge base containing clinical concepts, related terms, and mappings to ICD-9-CM for managing the transition from the usual scientific language to the coding language. This means, the integration of such resources with the ICD-9-CM systematic index, the ICD-9-CM alphabetical index, and other additional terms (synonyms, acronyms, linguistic variants, common medical terms, etc.);
- Standardized coding rules (e.g., "diagnostic and procedure codes are to be used at their highest level of specificity"; "three-digit codes are to be assigned only if there are no four-digit codes within that code category"; etc.);
- A rule engine for managing these rules, represented by the Business Rules Management System (BRMS) "Drools".

As shown in Fig. 1, the SISCO.web architecture includes three main layers:

1. Presentation layer: handling the interactions that users have with the software. Here the web component, has a multi-tier architecture, deployed on a Tomcat web server, offering two web user interfaces (WUIs) to support the compilation of SDOs. The WUIs make JSON calls to the Web Services of the underlying levels, which access the data resources built by the batch component. The two WUIs allow for two specific tasks: i) the text encoding WUI (TEM module), which serves as a coding tool, since it allows for searching clinical terms (diagnoses and procedures) and suggests the most appropriate ICD-9-CM codes based on search algorithms and related terms derived from the knowledge base; ii) the identification of the main diagnosis WUI (IMDM module), based on a rule engine that implements a specific decision tree for choosing and coding the main condition among the multiple diagnoses selected in the previous step. These two modules will be described in detail in the following paragraphs;
2. Application layer: handling the main code definitions and the most basic functions of the developed application. In SISCO.web this layer

includes five main functions which will be detailed later (e.g., search, autocomplete, code Details, use of related Terms for improving search, coding rules application through the Drools engine);
3. Data layer: which is mainly devoted to data storage. In fact, it houses not only data but indexes and tables. Here the batch component is aimed to build the data resources, i.e., the SISCO.web knowledge base, which is stored on the Apache Lucene Index.

The Apache Lucene Index[4], was chosen because it is a valid open-source tool for retrieving data and information. It provides straightforward Java APIs for creating text indexes and full-text search with options such as proximity search, fuzzy search, and score-based sorting, weighted filter search.

To implement the RESTful layer of web services within the system architecture, we chose Jersey[5], an open-source framework based on the JAX-RS API using annotation-based programming, which simplifies the creation of RESTful web services. It also facilitates the representation of data in standard formats such as JSON, XML, and HTML.

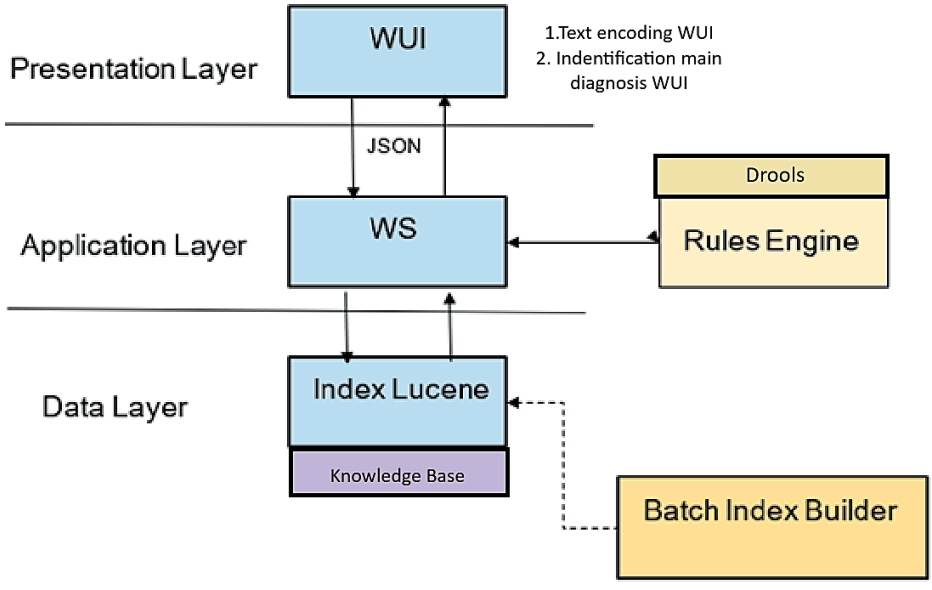


Figure 1: SISCO.web architecture.

[4] Apache Lucene is available for download at (Apache Lucene n.d.).
[5] Eclipse Jersey is available for download at (Eclipse Foundation n.d.).

The main process to reach the supported coding of morbidities and procedures and the identification of the main condition is shown in Fig. 2 and can be briefly described as follows:

1. Using the first module, i.e., TEM, the user starts searching for a diagnosis (one at a time) using the ones reported in the discharge letter (LDO) of the patient, to look for its ICD-9-CM code;
2. The system applies classic Natural Language Processing (NLP) algorithms such as Tokenization, text similarity algorithms to assign the most appropriate code to the diagnosis plus Decision Trees, and Symbolic NLP algorithms, i.e., rule-based and knowledge-based algorithms, relying on predefined linguistic rules and knowledge representations. For this reason, dictionaries, grammars, and ontologies are used to process language;
3. Every time the user searches for a diagnosis and selects one of the results suggested by the system, a list of coded diagnoses is generated to allow the user to identify, among these diagnoses the main condition;
4. The same procedure is used to search for procedures and interventions if reported in the discharge letters, and a second list of coded procedures/interventions will be generated by the system to be used as well by the IMDM module;
5. These two lists of codes represent the input data for the decision tree algorithm, which, as described in Subsection 2.2.3., will guide the user to identify the main pathological condition based on the defined coding rules.

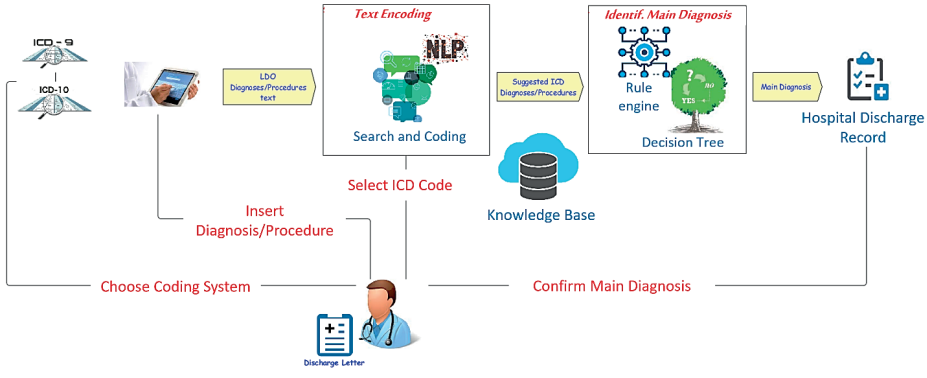


Figure 2: The SISCO.web main process.

To better understand the above-mentioned process, the next subsections will give details on the knowledge base, the algorithms, the decision tree and the coding rules used in the two modules to suggest the most appropriate ICD-9-CM codes.

2.2.1. The SISCO.web Knowledge Base

The knowledge base (KB) built for the project and used in the TEM integrates a series of terminological resources related to diagnoses and interventions/procedures in EHRs. The main data sources, as shown in Fig. 3, are represented by the Italian versions of:

- ICD-9-CM (v. 2007), Systematic index of diagnoses and procedures, considering the codes at the maximum level of specification;
- ICD-9-CM (v. 2007), Alphabetic index of diagnoses, and Alphabetic index of procedures.

For this project, an ontological version of ICD-9-CM has been created starting from the available ministerial tables of the classification, bringing to the development of the ICD-9-CM Ontology in OWL.

The lists of terms present in ICD-9-CM, in some cases inappropriate or outdated jargon, were supplemented with terms taken from other sources such as:

- Ad hoc created glossaries of diagnoses derived from physicians' scientific language, developed during a previous project (Cardillo et al. 2018);
- A glossary of diagnoses coded in ICD-9-CM extracted from the FVG Emergency Department (ED) EHRs database;
- Rare Diseases terms (Prime Minister's Decree 2017);
- Italian MeSH diagnoses and procedures terms (Istituto Superiore di Sanità n.d.).

All the terms derived from these sources were in most cases already mapped to the corresponding ICD-9-CM codes and were qualified as exact or approximate mapping.

Regarding the resource extracted from FVG Emergency Department "SEI Database", in the beginning, a list of 425 common pathological conditions in the ED was proposed by the ED FVG regional working group. On this list, a further analysis was performed to verify the use of technical/scientific terms and the correctness of the ICD-9-CM coding associated to these pathological conditions, bringing in the end to a glossary of 696 diagnoses (2,530 words) which enriched the SISCO.web KB.

Resources	Version	N. of Terms
ICD-9-CM systematic index	IT- 2007	16,294
ICD-9-CM alphabetical index	IT- 2007	289,834
Physicians' Glossary of diagnoses	v. 2017	1,421
Rare Diseases terms	v. 2017	683
Emergency physicians' diagnoses and pathological conditions (SEI database)	v. 2018	696
MeSH synonyms for diagnoses and procedures	v. 2017	641
Neoplasms related terms	v. 2017	13,290
Total		**322,859**

Table 1: Knowledge Base SISCO.web (Cardillo et. al 2019).

As observable, the total number of terms in the KB, considering the whole Italian ICD-9-CM resource and the above-mentioned additional resources, is about 323,000. It's important to note that the entire SISCO.web KB, particularly the data extracted from the SEI dataset, is not publicly accessible.

Regarding the ICD-9-CM Ontology, as said above, we created a processable version of the Ministerial file published online, since the original .xls file missed important details about each ICD-9-CM code. This information includes descriptions, inclusion and exclusion criteria, and notes, which are crucial for giving coding support based on ICD. To solve this problem, we developed a script that builds a lightweight ontology in OWL which can also be used to search for inconsistencies in the ICD-9-CM hierarchy or in the attribute's association. The ontology classes are based on the structure of the ICD-9-CM systematic index. At the top level, there are two main classes representing the ICD-9-CM main sections: *Diseases and Injuries*, and *Procedures and Interventions*. Within the *Diseases and Injuries* section, there are 17 classes that correspond to the ICD-9-CM "chapters" in this category, along with two additional classes for supplementary classifications: one for *external causes of injury and poisoning* and another for *factors influencing health status and contact with health services*.

Each chapter has its own class hierarchy, following the index structure that includes *blocks*, *categories*, *subcategories*, and *subclassifications*. To help with navigation, we labelled chapter classes with chapter numbers (e.g., Chapter I, Chapter II) and use E and V for the above mentioned additional classes. Similarly, in the *Procedures and Interventions* section, each category is organized under ranges such as the *Nervous System Intervention*, which covers codes 01-05. Each class/subclass in the ontology connects to the relevant data type annotations and, when needed, to Object Properties (i.e., relationships betwe-

en classes) and axioms. Access to the ICD-9-CM Ontology is currently restricted. However, we are planning to make it available on public repositories or GitHub shortly.

2.2.2. The Text encoding module

The first module is designed for searching the appropriate code for one or more diagnoses and procedures/interventions mentioned in the patient's discharge letter. The user enters a diagnosis in the search box using free text, which can be a single word or a multi-word term (T1). As the user starts typing, the system provides suggestions for autocompletion based on the knowledge base (KB), drawing from systematic or alphabetical indexes, MeSH synonyms, glossaries of general practitioners or emergency physicians, rare diseases, etc. These suggestions are the ones that have the entered text as their prefixes. Subsequently, the system conducts a syntactic search on the description of each attribute associated with ICD-9-CM classes in all types of resources in the KB. Different weights are assigned to each attribute based on its source and position. The search yields a list of ICD classes (diagnoses/procedures) that meet the search criteria, i.e., one or more attributes containing T1. The results are displayed in descending order based on their score.

To enhance the search function for the coder, the system permits filtering of the results in the list. This is achieved by incorporating the terms used in the query with related terms suggested by the system. These suggestions are based on their co-occurrence with the searched term within the ICD descriptors. To be more specific, the descriptions of the resulting ICD classes are tokenized to extract the most significant words (stop words are not considered). Moreover, to facilitate the tokenization and subsequent counting of term occurrences, the following ICD attributes are to be considered:

- The main description of the ICD class, along with any supplementary descriptions and inclusion terms in the systematic index;
- The description of the entry terms in the alphabetical index.

The system counts the number of times each token/term appears in the list of ICD classes resulting from the search. It then arranges the terms in descending order based on the number of occurrences and presents them to the user as related terms in a separate box. The user can choose one of the related terms or continue entering other free text in the search box. The system provides suggestions for autocompletion as the user enters more terms (T1, T2, etc.). The result list of ICD codes (diagnoses or procedures, depending on the user's initial selection) is updated to consider the search criteria, ensuring that one or more attributes contain all the input terms (T1, T2, etc.), and co-occurrences, making the search more precise. A similar approach is used in

the ICD-11 Coding Tool[6], which, unlike SISCO.web, allows also to use ICD chapters and ranges as research filters. From here on the algorithm performs the same steps, until the user selects a specific diagnosis/procedure among the ICD classes displayed in the search results which is always a leaf code. Once the diagnosis/procedure is selected, the system adds it to the list of candidate diagnoses/procedures used by the decision tree algorithm for identifying the main condition.

Is worth mentioning that the search and coding algorithm for procedures follows the same steps as that for diagnoses, but the Knowledge base which supports the process is smaller. In fact, in the case of procedures, the NLP algorithm examines only terminological resources related to interventions and procedures, therefore fewer terms are indexed. Specifically, the search is conducted almost entirely on the classes contained in the systematic index of ICD-9-CM section procedures, as well as on the procedure terms present in the ICD-9-CM alphabetical index, and the external resource MeSH.

2.2.3. The Identification of the main diagnosis module

To support physicians in the identification of the main condition, a decision tree was created to adhere to the WHO guidelines for morbidity coding in ICD-10 (Zavaroni et al. 2018). This includes following, on one hand, the WHO ICD-10 rules and guidelines for morbidity coding (WHO 2016)[7], which are up-to-date compared to ICD-9-CM 2007 rules, and on the other hand the WHO definition of the main condition, i.e., « the condition, diagnosed at the end of the episode of health care, primarily responsible for the patient's need for treatment or investigation» (WHO 2016, 147).

Furthermore, interventions and procedures were also considered in the decision-making process. To manage the extensive array of ICD codes (about 5,000), they were grouped into three sets:

1. "relevant surgery": encompassing interventions or procedures typically requiring an operating room, or those with resource consumption comparable to operating room costs;
2. "selected non-relevant surgical interventions": encompassing interventions or procedures, other than relevant surgery, that require significant resources, mostly higher than a non-surgical treatment of a condition;

[6] ICD-11 Coding tool is used to find the correct ICD-11 code for a specific diagnosis and it is connected to the ICD-11 browser to allow user to see further details for a searched diagnosis. It is available at (WHO 2024).

[7] This guideline has been updated during the publication of the sixth edition of ICD-10 in 2019 and later with the publication of ICD-11 release.

3. "residual non-relevant surgical interventions": encompassing interventions or procedures that necessitate fewer resources than non-surgical treatments.

Conditions were categorized into "conditions" (including diseases and clinical manifestations or normal physiological changes) and "pathological conditions" (abnormal anatomy or functioning constituting diseases).

The decision tree hierarchy includes: i) specific hospital settings which are highly specialized by age and changes of particular conditions, such as "neonatology" and "pregnancy, delivery, and puerperium", foreseen specific or partially specific paths; ii) paths for the other hospital settings, according to the general rules and, iii) the interventions/procedures set. Notably, the third group of interventions/procedures mentioned above is excluded as a viable option for identifying the main condition.

In Summary, the coding of certain health conditions is driven by the condition itself (pregnancy and related conditions, neonatal health), whereas for others, resource consumption due to procedures is the primary determinant. Thus, when a relevant intervention/surgery is identified, it influences the choice of the targeted condition. The decision tree rules are integrated into the rule engine module of the SISCO.web service.

The algorithm which determines the main condition, uses a Drools-based rule engine. Drools is an open-source Business rule management system (BRMS), released under the Apache License 2.0., that can easily be embedded in any Java application, which include an inference engine based on forward and backward chaining (Proctor 2012). The primary function of the Drools rule engine is to match incoming data, (i.e., facts), to the conditions outlined in the rules. It then determines whether and how to execute these rules. Key components in Drools are the following: *rules*; *facts* that are matched against the conditions of the rules to execute the applicable ones; a *production memory* (i.e., where the rules are kept); a *working memory* (i.e., location for the facts)[8].

In our implementation, the system consists of four components, developed in Java, and utilizes the RabbitMQ message broker (see Fig. 4). The primary component is the SISCO Drools Engine, serving as a wrapper for the Drools engine. It takes input data that triggers the execution of one or more rules down to a node, corresponding to a decision (leaf node), or the generation of a request for other parameters. The modules exchange messages in JSON format. On the web server side, the SISCO Rules Web Service component implements a servlet for dynamically creating content based on the interaction with the engine invoked by the main page of the SISCO.web system. The

[8] More details on the Drools key components can be found at Red Hat, Inc., Drools rule engine. Full documentation section (Drools n.d.).

SISCO Rules Data Receiver and the SISCO Rule Data Sender components, finally, act as interfaces with the message broker, transforming the asynchronous communication with the broker into the classic synchronous request/response client/web server communication.

The decision tree represents knowledge in the form of "if P then Q" rules. In the decision tree diagram, non-leaf nodes have two outgoing arcs: YES and NO. The rules defined for each node determine the selection of the outgoing arc and, consequently, the next computed node, based on terminological codes and user responses to the engine. The rules defined on two arcs from the same node are mutually exclusive to ensure the path's clarity. The decision algorithm takes two ICD-9-CM code lists as input: Pathological conditions (PC) and Procedures and Interventions (PI). The selection of the outgoing arc can be determined in two ways: automatically, based on the KB terminology codes feeding the engine, or decided by the user if no knowledge is available in the KB. If the rule engine is unable to ascertain the fulfilment of a rule based on incoming terminological codes, or when a decision necessitates the clinician judgment (e.g., *Are the pathological conditions related to each other?*), the engine will prompt user intervention by formulating a question within the web interface. This question may seek a binary YES/NO response (e.g., *Has it caused complications?*) or the selection of one or more terminological codes (e.g., *Identify the most complex event*). Subsequently, the engine will generate a JSON message encompassing all requisite details for presenting the question, including the query text, answer type (binary or selection of codes), and permissible response values (e.g., YES/NO, TRUE/FALSE, or specific codes).

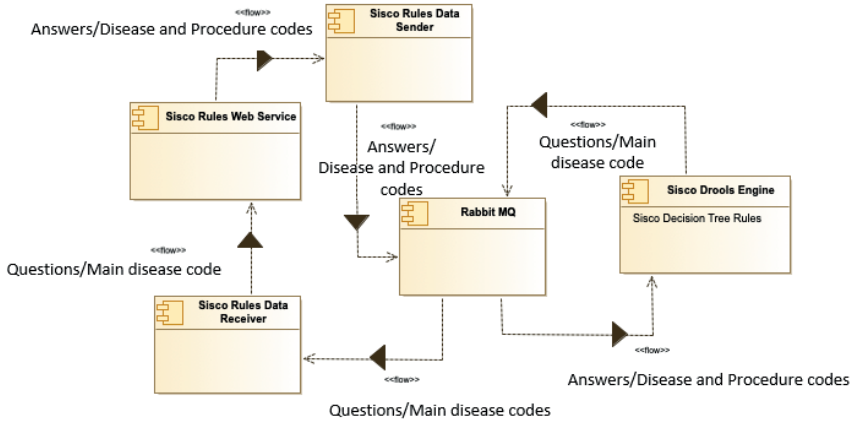


Figure 4: The Rules Engine Component Diagram.

In this way, the WUI content is automatically created by the browser, generating fields based on the answer type. For example, radio buttons are used for exclusive choices and check buttons for multiple choices. Fig. 4 shows two Drools rules for states 18 and 19 in the decision tree diagram. The "S18_ask" rule prompts the user to indicate one or more pathologies not related to the intervention. The "S19_true" rule manages the arrival of the response and determines the next transition from state 19 ("is it a single pathological condition?") based on whether the user has selected one or more codes among the relevant conditions. The result of the rule execution is reaching a leaf node associated with one or more codes suggested for the main pathological condition, which is then displayed in the SISCO.web interface.

```
rule S18_ask
  salience 120
  when
     c:BinaryParameters(i:id, s:state)   from entry-point "entry"
     p:Parameters(ip:id,dd:diagnosis) from entry-point "entry";
     eval(s==18);
     eval(i==ip);
  then
    System.out.println(i +" S18 Identificare una o più condizioni patologiche non correlate all'intervento ");
    JSONObject json = new JSONObject();
    json.put("id", i);
    json.put("state", s);
    json.put("message", "Identificare una o più condizioni patologiche non correlate all'intervento ");
    json.put("type", "ask_multicode");
    CEventsNotification.notify("out.siscoweb.notify_message",json.toString());
    modify(c){state=19}
 end

rule S19_false
  when
     c:BinaryParameters(i:id, s:state)   from entry-point "entry"
     ba:StateParameter(sb:state, t:type, cds:codes, ib:id) from entry-point "entry";
     eval(s==19);
     eval(t=="ask_multicode");
     eval(s==sb);
     eval(i==ib);
     eval(cds.size>1);
  then
    modify(c){state=20}
end
```

Figure 5: Drools S18-S19 rules example.

3. Evaluation

After an internal test conducted by the project's informaticians and terminologists, a more detailed usability test was performed by three physicians: This evaluation employed a subset of pathological conditions extracted from the SEI database mentioned in Section 2, along with diseases and interventions drawn from several anonymized patient discharge letters. These LDO contained multiple diagnosis and interventions/procedures, particularly focusing on complex cases characterized by comorbidities and intricate diag-

nostic definitions. The aim was to assess the tool's effectiveness in suggesting appropriate codes, required for completing the SDO. At this stage, the evaluation was more qualitative than quantitative, as the physicians were unable to access an LDO/SDO database for the project. Nonetheless, initial results indicate that the system performed well, successfully suggesting the most appropriate ICD-9-CM diagnosis even in instances where the input text in the search box of the TEM module was complex or included comorbidities. On average, SISCO.web provided precise ICD-9-CM code suggestions for 80% of 30 use cases tested by physicians, with improved accuracy when using the related terms feature. An example of diagnosis coding (in this case "diabetes") is given in Fig. 6. Here, when a user types "diabete" (diabetes) into the search box, the system auto-completes with suggestions like "diabete-nanismo-obesità" (diabetes-nanism-obesity) and "pre-diabete" (prediabetes). After selecting "diabete", the system displays matching classes in the search results section (considering all the attributes associated to the class, such as title, other description, inclusions, exclusions, alphabetic index terms, etc.), ordered by score. It also suggests related terms (on the left of the page) that co-occur with "diabete" in the ICD-9-CM descriptors. The user can then select a related term like "mellito" (mellitus), prompting the system to refine results based on both selected terms.

At each iteration, the system displays matching classes and related co-occurring terms based on user input. The search progressively narrows down until the user identifies and selects the correct ICD-9-CM class, which is then added to the "Selected Diagnoses" section at the bottom left of the page. Before selecting the proper code, for each code in the results list, the user can view code details (displayed if present on the right of the page and represented by symbols), including:

- *Leaf nodes*: Indicates to select a leaf code from the list presented, being the selected code not a leaf code;
- *Exclusion criteria*: Lists conditions excluded by that ICD-9-CM class;
- *Basic diseases* attribute: Advises coding a basic disease before using the selected code;
- *Use additional codes*: Recommends additional codes relevant to the selected class.

These features resulted helpful for avoiding inconsistencies, providing alerts to key ICD-9-CM coding rules, such as the necessity of coding a basic disease first or using leaf codes instead of general three-digit diagnosis codes (unknown rule by professionals or, in some cases, taken for granted).

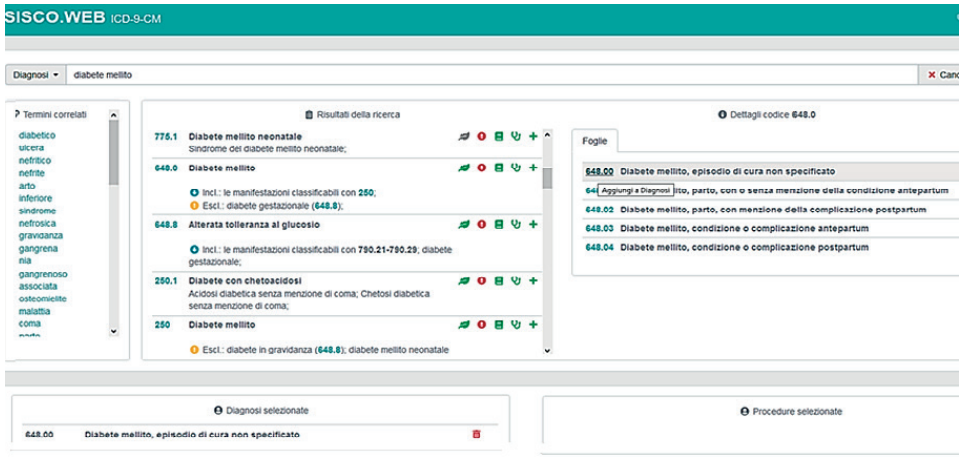


Figure 6: SISCO.web Interface: An example of coding for "diabetes mellitus" diagnosis.

Not completely known is also the need for the combined use of ICD-9-CM alphabetical and systematic indexes (both part of the KB) to extend knowledge about a code, providing references to additional codes related to the selected one, etc. Another useful feature of the TEM module was considered the possibility to show, starting from an ICD-9-CM class in the search results, the hierarchy of the classes, derived from the ICD-9-CM ontology, including all the details for each code.

Regarding the second module focused on the identification of the main condition (IMCM) the WUI, illustrated in Fig. 7, consists of three main sections: the upper section displays the two lists of codes (for diagnoses and procedures) selected by the user in the TEM module; the central section interacts with the user during the decision tree process, and the lower section reveals the main diagnosis once it has been identified.

When the user opens the module page, he will see two lists of codes at the top and a progress bar further down. At this point, the backend navigates the decision tree until it hits the first node that requires user input. At this stage, the rule engine requests the user input the necessary parameters to continue the navigation of the tree. These may include, for instance, the "most resource-consuming pathological condition during hospitalization" among the coded diagnoses (in case of multiple diagnoses). The user then selects one from a combo box, thereby entering the required parameter into the module. Subsequently, the rule engine resumes the path of the tree until the final node is reached, i.e., the identification of the main condition, which is finally displayed to the user for confirmation via a dedicated button. The central part of the page displays only a partial representation of the decision tree structure, including nodes requiring manual input, and the final three stages. This

should help the user understand the operations performed by the rule engine to determine the main diagnosis.

Nevertheless, the system can autonomously perform certain steps in the decision tree, utilizing previously provided information, the formalized coding rules, and inferences derived from the KB.

Furthermore, the WUI provides a button that cancels the rule engine operations and returns to the text encoding module WUI.

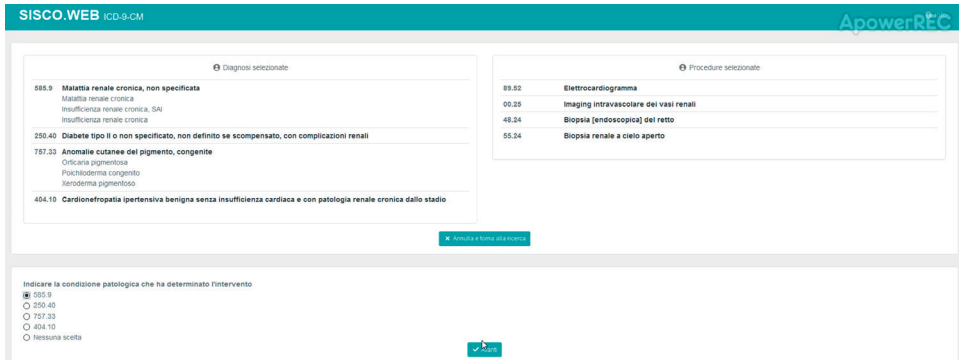


Figure 7: SISCO.web Interface: Rule engine support to identify the main condition.

The SISCO.web system was tested both in terms of the usability and efficiency of the search algorithms, by the doctors involved in the project, and in terms of functionality and performance, by the team of computer experts and terminologists who developed the service. The test highlighted that the search results for ICD-9-CM diagnoses obtained using the mentioned algorithms are substantially superimposable. However, it is noted that:

- The weights assigned to the various ICD-9-CM attributes associated with each ICD class in the search results appear inconsistent concerning the relationship between the importance of the various resources present in the KB and the recurrence of the terms (roots);
- Hierarchical algorithm guarantees greater appropriateness in the selection of ICD-9-CM categories since it maintains the relationship of importance between the resources present in the KB even in the event of their enrichment.

Hence, it was necessary to refine the weights assigned to the various attributes[9] to guarantee appropriateness in the selection of ICD-9-CM categories

[9] In particular, weights ranges from 0 to 10: the main description of the ICD class in the systematic index was still considered the most important with weight 10, the additional terms of the ICD class title have weight 7,5; inclusion terms have weight 2,5; alphabetical index "entry term" has weight 2,5, while its indentations (from the first to the sixth one) were assigned weight 0,1; neoplasm entry term in the alphabetic were assigned weight 2,5, and

even in the event of moving KB resources from one step to another. The steps of the algorithm implemented for the coding activity of a diagnosis were confirmed.

The test revealed issues in the identification of the diagnosis module, which is almost related to the formalization and computerization of the decision tree, particularly for some steps of the tree where the physician's input is necessary. This is especially true when the physician selects multiple interventions, as it's crucial, at a certain point of the process to indicate the relevance of each one. The decision tree is not fully computerized in terms of additional resources for automating certain steps (as it can be for example a list of relevant interventions aligned to anatomical sites or mapped to diagnosis categories, which although available in pdf, is still under elaboration for the integration into the rule engine) and allowing the physician to select multiple options. Currently, the computerized decision tree enables the physician to identify the main pathological condition by answering a series of YES/NO questions.

4. Related works

Different coding support systems have been developed in the last two decades. Some of them aimed to support the coding of causes of death, generally coded using ICD-10. Examples of these tools are MICAR-ACME, of the US National Center for Health Statistics (Israel 1990), and the IRIS system developed by a European consortium (Pavillon et al. 2007). The main issue encountered in these systems is the processing of natural language, which, in the last twenty years has been faced with developing automated coding tools based on NLP algorithms (Friedman et al. 2004). Only a few systems were based on properly defined coding rules, as done by (Farkas and Szarvas 2008) and (Cardillo et al. 2018), both focused on the ICD-9-CM coding. In recent years, challenges have been encountered, from the perspective of Artificial Intelligence (AI) and NLP, based on the literature. Many researchers and companies started applying more sophisticated methods such as Neural Networks or Large Language Models (LLM) to enable EHR data coding (Rios and Kavuluru 2018). This trend is confirmed also by the results of the CLEF ICD10 task[10], held in 2020, focused on ICD-10 coding for clinical textual data in Spanish and including, in particular, two subtasks for evaluating systems that predict ICD-10-CM (diagnostic) and ICD-10-PCS (procedural) codes using the Spanish CodiEsp corpus. Here most of the participants used Machine learning approaches and deep learning language models (prefer-

indentations had 0,1, which is also the weight assigned to the main description of diagnoses / procedures derived from the other glossaries in the KB.

[10] CLEF eHealth 2020 – Task 1: Multilingual Information Extraction (CLEF eHealth Lab Series n.d.).

ring fine-tuned Multilingual BERT), but the highest mean average precision (MAP) for the prediction of ICD-10 diagnostic codes (0.593) resulted by the combination of a XGBoost classifier and a Jaro Winkler string matching system (Miranda-Escalada et al. 2020). Other studies focused on the application of general-purpose LLMs (e.g., ChatGPT 3.5/4, LLAMA, etc.) to test their performances in the task of automated coding of diagnoses extracted from Discharge summaries by using ICD-10. Nevertheless, gaps between the current deep learning-based approach applied to clinical coding and the need for explainability and consistency in real-world practice were reported (Dong et al. 2022). Some studies indicate alternative methods or frameworks specifically designed for automatic ICD coding. For example (Chao-Wei Huang 2022) used a pre-trained language model for ICD coding, sharing a similar idea with BERT-XML, an extension of BERT designed for ICD coding. This model was pre-trained on a large collection of EHR clinical notes using an EHR-specific vocabulary (Zhang et al. 2020). Additionally, (Kim and Ganapathi 2021) introduced the Read, Attend, and Code (RAC) framework for accurate ICD code prediction. Another approach involved the use of off-the-shelf pre-trained generative LLMs to perform ICD coding, without labelled training examples and leveraging the hierarchical nature of the ICD ontology, thus relying on dynamic searches for clinical entities within the ontology (Boyle et al. 2023).

It is worth observing in this context the lack of available datasets for ICD coding to train AI-based models, especially in some languages, such as Italian. Few approaches show how to mitigate this issue. In (Almagro et al. 2019) a cross-lingual approach based on Machine Translation methods is proposed to code death certificates with ICD-10 through supervised learning. In brief, they tried to code Italian death certificates using certificates from another language (French), so combining collections of different languages to increase the availability of coded documents. Improvements in the system performance here were observed for codes assigned to labels with few occurrences. Silvestri et al. (2020) conducted a study on cross-lingual XLM fine-tuning aimed at predicting and classifying ICD-10 codes. A preliminary evaluation of a model fine-tuned on short medical notes written in English using an Italian test set was provided, but results indicated the need for further experiments to increase the number of samples in the test set, to better assess the model's ability to generalize.

A more recent overview on the topic is provided by the study conducted by the Icahn School of Medicine at Mount Sinai in New York revealed significant shortcomings in the performance of LLMs in clinical coding. The analysis showed that the existing models, including the highest-performing GPT-4, achieved less than 50% accuracy in matching medical codes to clinical texts. Such inaccuracies can result in serious billing errors and compliance issues

within healthcare systems. The study also highlighted varying performance levels among different LLMs, posing challenges in clinical environments where precise coding is essential for billing and ensuring accurate patient care (Soroush et al. 2024).

These results emphasize the need for refinement and validation of these technologies before considering clinical implementation, thus providing customized AI tools specifically designed for medical coding, instead of using general-purpose LLMs.

Given this overview, we can state that SISCO.web performances are comparable with most of the mentioned systems. Unlike existing systems and the most recent AI-based coding support, SISCO.web offers dual support. Firstly, it helps in finding the appropriate ICD-9-CM (or in the future ICD-10) code for a diagnosis or procedure by utilizing NLP techniques combined with the application of trustworthy coding rules, which are necessary to know when dealing with the selected classification system. Secondly, it assists in identifying the main diagnosis (the most serious and/or resource-intensive during hospitalization or the inpatient encounter) among multiple diagnoses, which is often a challenging and underestimated task. The advantages of this approach also stem from the integration of decision tree algorithms, which expand the system's functionalities.

5. Conclusions and future directions

This paper shows the approach used to develop a web service aimed at supporting physicians in the compilation of the SDO, while coding the main condition, secondary pathologies, procedures and interventions in ICD-9-CM and, where necessary, in ICD-10. The system also proposes a module based on a series of formal rules that represent a decision tree specifically designed for identifying the main pathological condition, which needs to be indicated and coded in a separate field in SDO. The evaluation of the TEM module, allowing for the search and suggestion of ICD-9-CM coding for diseases and procedures, has reached good performances in terms of the accuracy of the coding suggestions, the efficiency of the system, and regarding the usability of the system. Differently, some limitations are highlighted concerning the rule engine module, which allows, through a series of steps and interactions with the user, the identification of the main diagnosis. In this case, the initial formalization of the rules provided by the decision tree did not yield the expected results. It has therefore become necessary to update the rules and, above all, to make available ad hoc terminological resources to be submitted to the rule engine to automate some steps of the decision tree, thus ensuring the required performance compared to other support systems available in the literature. Considering that ICD-9-CM is currently mandatory in Italy for

coding diagnosis into SDO, the prototype and tests of SISCO.web uses this ICD version to be used in hospital coding. Nevertheless, the system has been designed to work using also ICD-10, including a decision tree specifically set for ICD-10 for identifying the main diagnosis. This possibility, recently, resulted advantageously since, as mentioned in Section 1, the Italian Ministry of Health, to be aligned to European guidelines on cross-boarding care, is working on a roadmap to shift from ICD-9-CM to ICD-10-IM for the coding of morbidities in SDO, leveraging the results of the It.DRG project. For this reason, future work will be the extension of the system, in terms of integration of the KB with the Italian version of ICD-10 (the mentioned ICD-10-IM) and the necessary crosswalking tables as well as the implementation of the already defined ICD-10-based decision tree in the rule engine. At the same time, it will be possible to set up versions of this support system able to manage classifications of interventions other than those used in Italy. Another possible future work is the development of a JavaScript library to distribute the service to interested parties and test it on a large scale (i.e., some hospital wards). As observed in Section 4, automated clinical coding holds promise for AI despite the technical and organizational challenges, but coders need to be involved in the development process, as done in the present work. Given this understanding, it can be argued that SISCO.web could serve as a good compromise, particularly if focusing on a new research direction that could be pursued over the next five years. This would involve improving the approach using LLMs + Retrieval-Augmented Generation (RAG) to enhance both the text encoding module and the implementation of the decision tree in the rule engine. Another possible future work could be to use a complementary approach for the analysis, through NLP/DL, of the diagnostic sections of hospital discharge letters (LDOs in Italy), which are very detailed reports. In our use case, a sample of these documents was used to test the performances of SISCO.web in terms of capacity to support coding for complex search records (e.g., comorbidities, very detailed diagnoses, etc.). In the future, it would be valuable to explore the possibility of providing coding support while registering LDOs' data, particularly in the diagnostic section.

References

Abraha, Iosief, Alessandro Montedori, Diego Serraino, et al. 2018. "Accuracy of administrative databases in detecting primary breast cancer diagnoses: a systematic review." *BMJ Open* 8:e019264. https://doi.org/10.1136/bmjopen-2017-019264.

Almagro, Mario, Raquel Martínez, Soto Montalvo, and Victor Fresno. 2019. "A cross-lingual approach to automatic ICD-10 coding of death certificates by exploring machine translation." *Journal of biomedical informatics* 94 (2019): 103207. https://doi.org/10.1016/j.jbi.2019.103207.

Amodio, Emanuele, Fabio Tramuto, Claudio Costantino, et al. 2014. "Diagnosis of influenza: only a problem of coding?". *Med Princ Pract* 23:568-73. https://doi.org/10.1159/000364780.

Apache Lucene. n.d. "Welcome to Apache Lucene." Last accessed November 10, 2024. https://lucene.apache.org/.

Boyle, Joseph S., Antanas Kascenas, Pat Lok, Maria Liakata, and Alison Q. O'Neil. 2023. "Automated clinical coding using off-the-shelf large language models." Accepted to the *NeurIPS 2023 workshop Deep Generative Models For Health (DGM4H)*, arXiv preprint arXiv:2310.06552 (2023).

Cardillo, Elena, Claudio Eccher, Anna Perri, Vincenzo Della Mea, and Francesco Talin. 2018. "A rule-based Support System for the Validation of Diagnoses coding in the Patient Summary." In *Proceedings of the International Conference on Medical Informatics Europe 2018 (MIE2018), Gothenburg, Sweden, April 24-26, 2018*.

Cardillo, Elena, Lucilla Frattura, Salvatore Ciambrini, Claudio Eccher, Elia Nardo, and Carlo Zavaroni. 2019. "Towards the Development of a Web Support System for Improving Accuracy in Coding Discharge Diagnosis." In *Proceedings of the 2019 IEEE Symposium on Computers and Communications (ISCC), Barcelona, Spain*, 1147-52. https://doi.org/10.1109/ISCC47284.2019.8969649.

Chao-Wei, Huang, Shang-Chi Tsai, and Yun-Nung Chen. 2022. "PLM-ICD: Automatic ICD coding with pre-trained language models." In *Proceedings of the 4th Clinical Natural Language Processing Workshop*, 10–20, Seattle, WA: Association for Computational Linguistics.

Chiò, Adriano, Giovannino Ciccone, Andrea Calvo, et al. 2002. "Validity of hospital morbidity records for amyotrophic lateral sclerosis. A population-based study." *J Clin Epidemiol* 55(7): 723-27. https://doi.org/10.1016/s0895-4356(02)00409-2.

CLEF eHealth Lab Series. n.d. "CLEF eHealth 2020 – Task 1: Multilingual Information Extraction." Last Accessed November 10, 2024. http://clefehealth.imag.fr/clefehealth.imag.fr/index135c.html?page_id=187%20%3E.

Dong, Hang, Matúš Falis, William Whiteley, et al. 2022. "Automated clinical coding: what, why, and where we are?" *NPJ Digit. Med*. 5(159). https://doi.org/10.1038/s41746-022-00705-7.

Drools. n.d. Last Accessed November 10, 2024. https://www.drools.org.

Eclipse Foundation. n.d. "About." Last Accessed November 10, 2024. https://eclipse-ee4j.github.io/jersey.

European Health Information Portal. 2023. "Hospital Discharge Records database." Last Updated January 10, 2023. https://www.healthinformation-portal.eu/health-information-sources/hospital-discharge-database-2.

Falis, Matúš, Gema Aryo Pradipta, Dong Hang, et al. 2024. "Can GPT-3.5 generate and code discharge summaries?" *Journal of the American Medical Informatics Association* 31(10): 2284-93. https://doi.org/10.1093/jamia/ocae132.

Farkas, Richárd, and György Szarvas. 2008. "Automatic construction of rule-based ICD-9-CM coding systems." *BMC Bioinformatics* 9 (3): S10. https://doi.org/10.1186/1471-2105-9-S3-S10.

Friedman, Carol, Lyudmila Shagina, Yves Lussier, and George Hripcsak. 2004. "Automated Encoding of Clinical Documents Based on Natural Language Processing." *JAMIA* 11(11): 392-402. https://doi.org/10.1197/jamia.M1552.

Israel, Robert A. 1990. "Automation of mortality data coding and processing in the United States of America." *World Health Stat Q.* 43(4): 259-62. https://pubmed.ncbi.nlm.nih.gov/2293494/.

Italian Ministry of Health. 2000. "Ministerial Decree October 27, 2000, no. 380 - Regolamento recante norme concernenti l'aggiornamento della disciplina del flusso informativo sui dimessi dagli istituti di ricovero pubblici e private." *Gazzetta Ufficiale*, 19 dicembre 2000, n. 295.

Italian Ministry of Health. 2024. *Rapporto sull'attività di ricovero ospedaliero. Dati SDO Anno 2022*. https://www.salute.gov.it/portale/documentazione/p6_2_2_1.jsp?lingua=italiano&id=3441.

Italian Ministry of Labor, Health and Social Affairs. 2008a. "Ministerial Decree December 18, 2008. "Aggiornamento dei sistemi di classificazione adottati per la codifica delle informazioni cliniche contenute nella scheda di dimissione ospedaliera e per la remunerazione delle prestazioni ospedaliere." *Gazzetta Ufficiale*, 9 marzo 2009, n. 56.

Italian Ministry of Labor, Health and Social Affairs. 2008b. *Classificazione delle malattie, dei traumatismi, degli interventi chirurgici e delle procedure diagnostiche e terapeutiche. Versione italiana della ICD-9-CM, 2007*. Roma: Istituto Poligrafico e Zecca dello Stato.

Italian Permanent working table for Digital health in Regions and Autonomous Provinces. 2010. *Specifiche tecniche per la creazione del "profilo sanitario sintetico" secondo lo standard HL7-CDA rel. 2*. Department for the Digitization of Public Administration and Technological Innovation.

Istituto Superiore di Sanità. n.d. "Medical Subject Headings 2019." Last Accessed November 10, 2024. https://old.iss.it/site/Mesh/.

Kim, Byung-Hak, and Ganapathi Varun. 2021. "Read, attend, and code: Pushing the limits of medical codes prediction from clinical notes by machines." In *Machine Learning for Healthcare Conference*, 196-208. PMLR.

Miranda-Escalada, Antonio, Aitor Gonzalez-Agir, Jordi Armengol-Estapé, and Martin Krallinger. 2020. "Overview of Automatic Clinical Coding: Annotations, Guidelines, and Solutions for non-English Clinical Cases at CodiEsp Track of CLEF eHealth 2020." In *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, CEUR-WS 2696*.

Moriyama, Iwao Milton, Ruth M. Loy, Alastair Hamish, Tearloch Robb-Smith, Harry Michael Rosenberg, and Donna L. Hoyert. 2011. *History of the statistical classification of diseases and causes of death*, edited and updated by H. M. Rosenberg, D. L. Hoyert. DHHS publication, no. (PHS) 2011-1125.

Nonis, Marino, Luigi Bertinato, Laura Arcangeli, et al. 2018. "The evolution of DRG system in Italy: the It-DRG project." *European Journal of Public Health* 28, no. 4 (November), cky218.095. https://doi.org/10.1093/eurpub/cky218.095.

O'Malley, Kimberly J., Karon F. Cook, Matt D. Price, Kimberly Raiford Wildes, John F. Hurdle, and Carol M. Ashton. 2005. "Measuring diagnoses: ICD code accuracy." *Health services research* 40(5p2): 1620-39. https://doi.org/10.1111/j.1475-6773.2005.00444.x.

Pavillon, Gérard, Lars A. Johansson, D. Glenn, S. Weber, B. Witting, and S. Notzon. 2007. "Iris: A Language Independent Coding System For Mortality Data." In *WHO – Family of International Classifications Network – FIC. Annual Meeting. Trieste, Italy, 28 October - 3 November 2007*.

Prime Minister's Decree 12 January 2017. "Definizione e aggiornamento dei livelli essenziali di assistenza, di cui all'articolo 1, comma 7, del decreto legislativo 30 dicembre 1992, n. 502." *Gazzetta Ufficiale*, 18 marzo 2017, no. 65, Allegato 7.

Proctor, Mark. 2012. "Drools: A Rule Engine for Complex Event Processing." In *Applications of Graph Transformations with Industrial Relevance*. AGTIVE 2011, edited by A. Schürr, D. Varró, G. Varró. *Lecture Notes in Computer Science* 7233. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-34176-2_2.

Quan, Hude, Łukasz Moskal, Alan J. Forster, et al. 2014. "International variation in the definition of 'main condition' in ICD-coded health data." *Int J Qual Health Care* 26(5): 511-15. https://doi.org/10.1093/intqhc/mzu064. Epub 2014 Jul 2.

Rinaldi, Rita, Luca Vignatelli, Massimo Galeotti, Giuseppe Azzimondi G., and Piero De Carolis. 2003. "Accuracy of ICD-9 codes in identifying ischemic stroke in the General Hospital of Lugo di Romagna (Italy)." *Neurol Sci* 24: 65-69. https://doi.org/10.1007/s100720300074.

Rios, Anthony, and Ramakanth Kavuluru. 2018. "EMR Coding with Semi-Parametric Multi-Head Matching Networks." In *Proceedings of the conference. Association for Computational Linguistics. North American Chapter Meeting 2018*: 2081-91. https://doi.org/10.18653/v1/N18-1189.

Sforza, Vincenzo, Duilio Carusi, Luigi Bertinato, Marino Nonis, and Silvia Surricchio. 2021 "L'approccio del PROGETTO IT.DRG per la rilevazione dei costi standard delle prestazioni ospedaliere. Il modello IT:COST." *Bilancio Comunità Persona*, n. 2: 82-116. https://dirittoeconti.it/articolo-rivista/lapproccio-del-progetto-it-drg-per-la-rilevazione-dei-costi-standard-delle-prestazioni-ospedaliere-il-modello-itcost/.

Silvestri, Stefano, Francesco Gargiulo, Mario Ciampi, and Giuseppe De Pietro. 2020. "Exploit multilingual language model at scale for icd-10 clinical text classification." In *2020 IEEE Symposium on Computers and Communications (ISCC*), Rennes, France, 2020, 1-7. https://doi.org/10.1109/ISCC50000.2020.9219640.

Soroush, Ali, Benjamin S. Glicksberg, Eyal Zimlichman, et al. 2024. "Large Language Models Are Poor Medical Coders - Benchmarking of Medical Code Querying." *NEJM AI* 1(5) (April 19, 2024). https://doi.org/10.1056/AIdbp2300040.

Spolaore, Paolo, Stefano Brocco, Ugo Fedeli, et al. 2005. "Measuring accuracy of discharge diagnoses for a region-wide surveillance of hospitalized strokes." *Stroke* 36, no. 5 (May): 1031-34. https://doi.org/10.1161/01.STR.0000160755.94884.4a.

Sukanya Chongthawonsatid. 2017. "Validity of Principal Diagnoses in Discharge Summaries and ICD-10 Coding assessments based on national health data of Thailand." *Health Inform Res* 23, no. 4 (October): 293-303. https://doi.org/10.4258/hir.2017.23.4.293.

Sundararajan, Vijaya, Patricia S. Romano, Hude Quan, et al. 2015. "Capturing diagnosis-timing in ICD-coded hospital data: recommendations from the WHO ICD-11 topic advisory group on quality and safety." *Int J Qual Health Care* 27(4): 328-33. https://doi.org/10.1093/intqhc/mzv037.

Tatham, Andrew J. 2008. "The increasing importance of clinical coding." *British Journal of Hospital Medicine* 69(7): 372-3.

Wang, Cheng, Chenlong Yao, Pengfei Chen, Jiamin Shi, Zhe Gu, and Zheying Zhou. 2021. "Artificial Intelligence Algorithm with ICD Coding Technology Guided by the Embedded Electronic Medical Record System in Medical Record Information Management." *J Healthc Eng* 30;2021:3293457. https://doi.org/10.1155/2021/3293457.

Williamson, Ashton, David de Hilster, Amnon Meyers, Nina Hubig, and Amy Apon. 2024. "Low-resource ICD Coding of Hospital Discharge Summaries." In *Proceedings of the 23rd Workshop on Biomedical Language Processing, August 16, 2024*, 548-58. Association for Computational Linguistics. https://aclanthology.org/2024.bionlp-1.45.pdf.

World Health Organization (WHO). 2016. "International Statistical Classification of Diseases and Related Health Problems 10th Revision. Volume 2." Geneva: World Health Organization.

World Health Organization (WHO). 2019/2021. *International Classification of Diseases, Eleventh Revision (ICD-11).* https://icd.who.int/browse11.

World Health Organization (WHO). 2024. "ICD-11 Coding Tool." https://icd.who.int/ct/icd11_mms/en/release.

Zavaroni, Carlo, Antonia Fanzutto, Elia Nardo, Vincenzo Della Mea, and Lucilla Frattura. 2018. "Morbidity coding in ICD-11 (and ICHI): a decision tree to identify the main condition." In *WHO-FIC Annual Meeting Booklet*. Seoul, 22-27 October 2018. WHO. #307.

Zhang, Zachariah, Liu Jingshu, and Razavian Narges. 2020. "BERT-XML: Large scale automated ICD coding using BERT pretraining." In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, 24-34, Online. Association for Computational Linguistics. https://doi.org/10.48550/arXiv.2006.03685.

AIDAinformazioni
ISSN 1121–0095
ISBN 979-12-5965-456-4
DOI 10.57574/596545644
pag. 79-100 (luglio-dicembre 2024)

A humanistic approach to *datafication*

Two case studies: digital and digitized

Valeria Federici*

Abstract: The term *datafication* has the ability to embrace a series of aspects that span from the field of computer science to social and cultural studies. While the process of *datafication* (taking aspects of life and turning them into data) is surrounded by a lure of abstraction and neutrality; similarly to other computational processes, *datafication* reflects cultural biases, flaws, and implications that affect knowledge and knowledge production. This article explores datafication as related to the semantic web, web ontologies, and other systems of classification as both method and structure of art historical analysis. By analyzing two digital repositories that run on MediaWiki, the goal of this investigation is to incentivize a model that, under the umbrella of digital art history, unifies aspects pertaining to digital curatorship and digital preservation. The two case studies are: *The History of Early American Landscape Design (HEALD)* and *The Educational Encyclopedia of Digital Arts (EduEDA)*.

Keywords: Datafication, Semantic Web, Web Ontologies, Digital, Digitized.

1. Introduction

In her book *The Age of Surveillance Capitalism*, Shoshana Zuboff defines *datafication* as «the application of software that allows computers and algorithms to process and analyze [data]» (Zuboff 2019, 187-188). *Datafication* is «a technological process that turns several aspects of the life of an individual, a group, or a society into data. *Data is then turned into information* that acquires new values, including economic value» (Treccani 2020, Emphasis added)[1]. The term *datafication* thus has span from the field of computer science to social and cultural studies (Zuboff 2019). This article explores *datafication* in relation to the use of the semantic web, web ontologies, and other classification systems as both methods and structures of art historical analysis. In

* Center for Advanced Study in the Visual Arts, National Gallery of Art, Washington, D.C., USA. v-federici@nga.gov.

[1] Also, the term *Datafication* appeared in the Italian newspaper "La Repubblica" as early as 1986 (Translated by the author).

particular, it investigates two case studies: the *HEALD*, which is a project by the Center of Advanced Study in the Visual Arts of the National Gallery of Art in Washington, D.C.; and *The Educational Encyclopedia of Digital Arts* (*EduEDA*), a collective effort with numerous media partners, supported by both the Academy of Fine Arts of Carrara and of Florence, Italy. Since these two projects are both built using MediaWiki, the open access software than runs Wikipedia, they are analyzed in conversation with one another, in order to investigate the potential and the limits of data-driven analysis as offered by an open access platform. Specifically, the article delves into the implications of using the semantic web and standardized vocabularies to apply meaning to data, making it readable first by machines and then by humans.

In the first case study, I analyze how the latest upgrade to the digital repository *HEALD* enhanced the use of semantic web to foster investigation of the material available on its platform and to support its preservation. *HEALD* can be considered as a repository of digitized items, i.e. physical objects that underwent a process of digitization to be made available digitally. Since developments in digital technology are rapidly evolving, the upgrade helped address technological obsolescence as an endemic issue in digital formats and frameworks. Such developments often bring changes and mandatory updates that impact the way we can or cannot use a platform that originally seemed to serve our digital goals well, and for the longest time. Despite our best intentions, at an early stage of a digital project's life, we might be already looking for alternative digital formats, new databases, or an entirely new *host* in order to give our project a new virtual life. On the one hand, a solution can be to create a dataset that uses standardized parameters and ontologies in order for content to remain available in the future, and/or to be safely migrated to a new platform. On the other hand, standardization would carry over some implications as well as the question of how to preserve the work's original context, i.e. the digital environment in which the project, or the artwork, was first created. Overall, this analysis offers a way to deal with these implications.

In the second case study, I explore *EduEDA*, *The Educational Encyclopedia of Digital Arts*. Originally called WikiARTpedia – a project that received an Honorary Mention at the Ars Electronica Festival of Linz, Austria in 2009 – in 2012, WikiARTpedia became *EduEDA*, an encyclopedia of new media arts and a research platform for networks dedicated to information technology cultures. As expressed on the project website, the main goal of *EduEDA* is «to create a national and international network of people and institutions in order to collaboratively promote and disseminate digital arts» (EduEDA 2022). The idea of an interconnected repository of new media art is certainly in line with the vision of early Internet communities as expressed initially by the creator of the World Wide Web Berners-Lee, since it fosters collaborations as well as horizontal and collective forms of knowledge production. A vision then car-

ried out by the project Linked Open Data, which had been joined by several cultural institutions and museums over the course of the years (Berners-Lee 2006). As illustrated by the initiator of *EduEDA* Tommaso Tozzi – who is an artist and a pioneer of new media art – in order to maintain the collective character of the project, it was necessary to adopt an open source software that allows everyone to contribute. *EduEDA* does not contain artworks, but links to websites that store the artworks, built by artists themselves or by institutions. The platform also includes a link to the artwork's profile page, which at times features still images. *EduEDA* is here considered mainly for its crossed research tool, as well as for being an example of early digital curatorship. It is only partially a repository of digital items, or so-called digital-born objects. As we will see over the course of this article, *EduEDA* hosts links to either *reproduced* or *duplicable* items (Ippolito 2008, 118)[2]. This distinction is of particular interest for digital curatorship.

Others have discussed digital curatorship at length, giving many possible solutions to display and investigate digital work. Christiane Paul's curatorial approach for digital art is certainly still a beacon for this discipline (Paul 2008). In addition, I consider Ippolito's characterization of digital work in relation to the possibility of reproducing or simulating obsolete technology. In general, I endorse a case-by-case approach to the artefact, where a strategy for display and conservation is developed, whenever possible, in collaboration with artists and makers. When this is not possible, such as in the case of *HEALD*, it is still necessary to clarify that the work has been digitized and to provide details of the process. I will return to the issue of digital vs digitized later in this article, and in the conclusions. Broadly speaking, since operating within the digital realm implies similar challenges for artists, curators, digital humanists, and researchers, this article attempts to offer a roadmap for the investigative fruition of digital content.

This digital art historical study draws from a lineage of scholarship rooted in media studies, which can provide an insightful analysis of the practices and methodologies employed in the field of digital art history and, most importantly, their ramifications. Stemming from this approach, *datafication* is here intended both as a structural aspect of information technology, as well as a cultural one. By considering the epistemological umbrella under which data acquires value, this approach invites a reflection on how data is collected and made available in the field of the digital humanities and digital art history, and advocates for an active role of the humanists in shaping digital methodological practices. This translates into a non-hierarchical and dialogical relationship between information technologies and the humanities, meaning that objects,

[2] Ippolito writes: «We chose the term "reproduced" for any medium that loses quality when copied, including analog, prints, photographs, film, audio, and video […] In contrast, we reserved the word "duplicable" for media that can be cloned».

object production, and consumption through information technology, are considered in material and historically contextualized terms. Drawing from existing tools and experiences, the goal of the article is to find a path – through a humanistic lens on *datafication* – to reconcile practices and methodologies that regard digital and digitized artworks as distinct, even though they share the same data-driven mediums and are influenced by the same lucrative *technological solutionism* (Morozov 2013). This approach is not conclusive; rather, it attempts to demonstrate how theories rooted in media studies can enhance a humanistic approach to digital tools and help explain the premises and outcomes of digital art history projects.

The field of media studies has eloquently illustrated how the digital realm has been characterized by a series of catchy words and phrases that rarely have clear meanings. Datafication, network, media, digital, and algorithms are terms that have multiple connotations but remain elusive, often adopted interchangeably. For instance, Wendy Chun (2011) noted as the term *network* is used as a placeholder for interconnectivity, sociality, or simply the Internet. Chun explains how through a series of linguistic metaphors all media become more transparent. By becoming more transparent, they tend to blend with the environment and to become invisible, and their significance becomes even more occult. Transparency is here intended not only as the ability of media to fit within our surroundings (for instance by being portable), but also as their ambition to predict our actions or reactions, in order to be smoothly assimilated into our life (Schäfer and van Es 2017). This alleged transparency allows media to run without their process being fully explained or questioned. Rather, the process is often regarded as *magic* or outside of human control (Morozov 2013), while the working of machines has been mythologized, and locating agency within digital tools remains an open controversy (Bucher 2018, 52, 60). As I argued elsewhere, a similar process applies to data (Federici 2022).

As Artificial Intelligence moves into the realm of the Digital Humanities, these aspects become more and more relevant. For instance, while speaking about Artificial Intelligence, Jonnie Penn traced this metaphorical trend back to 1976, when it was noted that «words […] served as "incantations" for a desired result, rather than sober descriptions of a mechanism or function» (Penn 2021, 338). Data, network, and algorithms are often surrounded by a lure of abstraction and neutrality. Abstraction is here intended as a process through which it seems possible, or enticing, to conduct scientific analysis divested of human subjectivity. Similarly, neutrality refers to the alleged ability of data to represent evidence through indexicality – a direct connection between the object represented and its record – as if there were no interpretation in the process of displaying content and visualizing data. However, as shown in this study and elsewhere, the opposite mechanism takes place when working with data, and in particular with the semantic web. In fact, notwithstanding the

misleading narrative around information technology and its processes, the mechanisms behind those afore-mentioned terms are achieved through complex preparation, selection, and elaboration. In other words, they are highly mediated. For instance, standardized vocabularies, or ontologies, which are used to apply semantic meaning to data for the machine to read and interpret a given information, are achieved through a linguistic selection that is, above all, cultural, and entails compromise, and standardization. The same linguistic selection can, at times, obliterate historical presence or exclude underrepresented individuals or groups.

Standardization in computational process derives from the nineteenth century pursue of mechanical objectivity (Daston and Gallison 1992; Porter 1995) and it requires the same scrutiny as any other operations in the digital realm. Such scrutiny is possible through an understanding of how these processes work. As it has been noted, in order to successfully combine quantitative research within the humanities, digital art historians have been relying on old models of art historical investigation (mostly revitalizing Panofsky's concept of iconography or Warburg's approach to image association) that have been surpassed by new theories in art history. These old methods are linked to a determinist approach to computing and its use for quantitative analysis, that precludes new paths of investigation (Näslund and Wasielewski 2021). In addition, such approach prevents a thorough analysis of how digital tools operate or can operate. This reflection on the deterministic outcomes of standardization should impinge the mythological aura that surrounds media in general, and digital tools in particular. Dialogically, it should also help to question methodologies in the humanities, in order to flag biases and assumptions.

Along with media theories and art historical methodologies, the premises of this article are indebted to the many who have poignantly analyzed the manifold aspects and interconnections between the Digital and the Humanities, the so-called digital turn (or computational turn), under a methodological and ontological lens. Fundamental is certainly the distinction that Joanna Drucker drew between *data* and *capta* (Drucker 2011), the former indicating the information given (potentially available), and the latter the information taken (collected and elaborated in order to be made available). Drucker invites us to consider data as something constructed, extrapolated, originated by choices, compromises, and therefore *prepared*. Data cannot be considered as an absolute value, and the term *capta* serves to clarify its actual forms and uses. *Capta* is therefore the data that has been turned into information, it is the data we work with.

The contributors to the volume *Raw data is an oxymoron* have exposed that there is no data divested of meaning (Gitelman et al. 2013). A concept carried on further by Taina Bucher's analysis of the algorithm (which, as she argues, should be rather considered in the plural form *algorithms*) as well as by Evgeny

Morozov's observations on the afore-mentioned *technological solutionism*, or the belief that computing has a solution to everything (Bucher 2018; Morozov 2013). On a similar note, as observed by Sven Spieker, the «archive does not record experience so much as its absence» (Spieker 2008, 3). Therefore, as part of a critical approach to data, it is mandatory to consider not only what data shows, but also what it does not show. A claim that has also been made by Stephanie Porras when speaking about the network visualization of archives (Porras 2017). Manovich's approach to data as a medium, along with his concept of *Cultural Analytics,* emphasizes the implications of computing as a technology of culture (Manovich 2020) while Theodor Porter had previously defined quantification as a *social technology* and clarified that it emerged much earlier than the digital turn took place, revealing a longstanding tradition of quantitative analysis that spans over three centuries (Porter 1995, 50). It has been extensively observed how the computational turn in the humanities forced us to rethink how to utilize digital tools and methodologies by attempting to incentivize interdisciplinarity and push for human intervention. The edited volume *Research Methods for the Digital Humanities* has already introduced a compelling scope to «expand the field [...] rather than establish definitive boundaries» (Levenberg et al. 2018, 2). Finally, the «decolonial turn in data and technology» as highlighted by Nick Couldry and Ulises Ali Mejais (2021), is another stepping stone for conducting research in the realm of the digital and digital knowledge production. This leads us to reflect on and rethink standardization as it is currently possible through *datafication*.

This article and its outcomes stand on the shoulders of those analyses and approaches, with a particular focus on MediaWiki for its employment of the semantic web and its characteristic of being an open access platform based on the possibilities of sharing information, creating communities for scholarship, and working collectively. I explore these aspects further in the sections that follow. The two case studies under consideration serve to discuss, and eventually to attempt to come to terms with aspects of the digital realm that pertain to both digital-born (digital) and non-digital-born (digitized) artefacts, in order to contribute critically to the making and usage of digital tools by embracing complexity rather than standardization, by emphasizing processes, and by operating openly within the limitations of the tools used. This investigation thus suggests the possibility of intertwining *digitized* art history, *digital* art history, digital curatorship, and digital preservation. While the first two concepts, borrowed by Johanna Drucker, have been extensively analyzed, all these fields of investigation remain separate from one another (Drucker 2013; Brown 2020). As mentioned, this article ultimately ponders the benefits of a cross-pollination among them to potentially become one expanded field that draws from the experiences and implications of working within the digital realm.

2. HEALD – History of American Landscape Design

The digital resource *HEALD* pertains to «the language of early American landscape aesthetics and garden design in the colonial and national periods» (HEALD 2021a)[3]. *HEALD* combines thousands of texts with more than 1700 images from collections across the United States. The goal of the project is to «trace the development of landscape and garden terminology from British colonial America to the mid-19th century». As mentioned, *HEALD* runs on MediaWiki, an open access software based on JavaScript (and its derivatives). *HEALD* main structural features (database, editor, interface) were upgraded in order to adopt standardized semantic ontologies to ensure the usability, interoperability, and longevity of data (HEALD 2021b)[4]. *HEALD* online content is organized into three main categories: *keywords*, *places*, and *people*. Content was enriched with metadata (by using Semantic MediaWiki or SMW) in order to represent the complex relationships between these three categories, while the MediaWiki software was customized through extensions. Extensions are parts of the MediaWiki software, often coded or edited by computer scientists and a community of software engineers that keeps MediaWiki up to date and functional. For the most part, in line with MediaWiki open access policy, extensions are shared openly and widely (MediaWiki 2024a)[5].

In *HEALD*, a specific term (*keyword*) is described through its usage in common texts (letter, inventory, surveys, diaries) or by citations in dictionaries, treaties, and published material; as well as through its relationship to visual sources, which are categorized into *inscribed*, *associated*, or *attributed*[6]. The relationship between keywords and historic visual documents was first established in the book on which the repository is based (O'Malley 2010), while additional relationships were formulated following an analysis of the specificity of the content in a digital environment. At the same time, similarly to a dictionary or an encyclopedia, the repository includes descriptive pages pertaining to *keywords*, *places*, and *people*. A descriptive page for a keyword helps define how and when the term emerged and how it changed overtime. A descriptive page of a place and/ or of a person, traces and contextualizes their history. Descriptions have been written by multiple contributors over

[3] This investigation stems from my experience working on *HEALD* in collaboration with the Director of the project and former Associate Dean Therese O'Malley and the Digital Research Officer Matthew J. Westerby. *HEALD* digital repository is based on the publication by Therese O'Malley (2010).

[4] A full description of this upgrade is available on the *HEALD* website.

[5] According to one's familiarity with MediaWiki, it might be necessary to consult with a software developer in order to use, install, update an extension.

[6] An *inscribed* image incorporates the word; An *associated* image is related to the term less directly, by a contemporaneous description of the feature; *attributed* images, are those for which there are no *inscribed* terms or *associated* texts.

time[7]. A person (under the category *People*) is featured for using a keyword in writing or for their relevance to the overall topic of the project. A location (under the category *Places*) is featured in reference to a person or to a keyword (for instance, *The National Monument*). *Keywords*, *Places*, and *People* are interconnected throughout the repository not only via descriptions, but also via internal hyperlinks, an indexical way to establish connections among items included in a MediaWiki page. This feature is native to the software markup language (MediaWiki 2024b). Although the content in *HEALD* is organized so to prioritize *Keywords*, the written texts intertwine and are in conversation with images of paintings, drawings, architectural plans, ceramics, photographs, and more.

Regarding the semantic values added to *HEALD* following the upgrade, standard vocabularies (Getty AAT, TGN, ULAN[8]; Library of Congress (LOC) Name Authority) were used, when applicable, to label people's and locations' name, dates, coordinates, etc. Other novel attributes interlace an image to a person (through the value *[has person]*) or to a location (through the value *[has place]*) or both. Inserting values within square brackets is also a native aspect to the WikiMedia markup language. In order to record whether a keyword is *inscribed*, *associated*, or *attributed* to an image, such detail was added as a semantic value, which is exportable (see RDF[9] string below). Both standard and customized semantics are applied with the MediaWiki extension Page Form[10]. Last, cited publications are gathered in a dedicated Zotero library (Zotero 2021). When applicable, a Zotero ID appears in the image page so to allow a direct link from a visual source to a publication[11]. The layering of attributes makes the content accessible by multiple points: via its descriptive texts; via the image collection; via the relationship between images and *keywords*, *places*, or *people*; via its extended bibliography.

As mentioned, MediaWiki is set to interact with new software thanks to a community of worldwide developers invested in the tenet of open access. In line with this principle, and with a recent trend in museum openness, the customized code used for *HEALD* is available on GitHub (GitHub 2021). MediaWiki, as utilized by *HEALD*, integrates a clean interface with SQL, a widely used database language. All these characteristics make *HEALD* a digital product easy to maintain, to implement, and to possibly migrate. The semantic web is used to turn data into information and to preserve content

---

[7] At the time of writing, there are approximately more than 220 content pages.

[8] Getty Art and Architecture Thesaurus, Thesaurus of Geographic Names, Union List of Artist Names.

[9] Resource Description Framework.

[10] Page Forms has been developed by Wikiworks.

[11] See "HEALD: Anonymous, Two Ornamental Ice Houses Above Ground, 1846" (HEALD n.d.a).

since the descriptive metadata can be exported in RDF and retrieved. The use of standardized vocabularies (AAT, TGN, ULAN, LOC) makes possible to interlace the history of places and people as uniquely featured in *HEALD* with potentially other datasets that use the same sets of attributes. Nonetheless, the standardization posed a limit to the relationships expressed within the project, in particular as it relates to keywords and images. This limit was overcome by adding a string of property to be exportable as RDF:

<property:**Keyword** *rdf:resource*="http://heald.nga.gov/mediawiki/index.php/Special:URIResolver/Icehouse"/>
<property:**Keyword_relationship** *rdf:datatype*="http://www.w3.org/2001/XMLSchema#string">Inscribed</property:Keyword_relationship>[12]

Even though the two strings are unique to *HEALD*, they exist within a set of parameters (in this case, *rdf:resource* and *rdf:datatype*) that makes them recognizable and reusable. In this case, in particular, the goal is to preserve the relationship between the image and the keyword (*inscribed*) and to make this information retrievable. To extend upon this example, a more articulated section of the RDF export shows how information pertaining to the aforementioned relationships, object details, as well as bibliographic references intertwine:

<swivt:masterPage
rdf:resource="http://heald.nga.gov/mediawiki/index.php/Special:URIResolver/**File-3A0999.jpg**"/>
<swivt:wikiNamespace rdf:datatype="http://www.w3.org/2001/XMLSchema#integer">6</swivt:wikiNamespace>
<property:Keyword rdf:resource="http://heald.nga.gov/mediawiki/index.php/Special:URIResolver/**Picturesque**"/>
<property:Keyword rdf:resource="http://heald.nga.gov/mediawiki/index.php/Special:URIResolver/**Icehouse**"/>
<property:Keyword_relationship rdf:datatype="http://www.w3.org/2001/XMLSchema#string">**Inscribed**</property:Keyword_relationship>
<property:Keyword_relationship rdf:datatype="http://www.w3.org/2001/XMLSchema#string">**Associated**</property:Keyword_relationship>

File-3A0999.jpg is the object name; "Picturesque" is the term *associated* with the architectural style depicted; "Icehouse" is the *inscribed* term con-

[12] This string is taken from the RDF export of an image in the *HEALD* online repository: *File-3A0999.jpg*.

tained in the historic publication. Then, a second part of the same extract contains information about the publication:

<swivt:Subject
rdf:about="http://heald.nga.gov/mediawiki/index.php/Special:URIResolver/**File-3A0999.jpg-23Publication**">
<property:Date rdf:datatype="http://www.w3.org/2001/XMLSchema#gYearMonth">1846-12</property:Date>
<property:Date-23aux rdf:datatype="http://www.w3.org/2001/XMLSchema#double">2395631.5</property:Date-23aux>
</swivt:Subject>
<owl:DatatypeProperty rdf:about="http://heald.nga.gov/mediawiki/index.php/Special:URIResolver/**Property-3AReference_ID**"/>

The RDF above has been edited to reflect only the file name and the publication date (File-3A0999.jpg-23Publication), along with the publication ID (Property-3AReference_ID), which refers to its record in Zotero.

One of the main concerns of implementing *HEALD* semantically, was to avoid divesting *HEALD* content of its context, both digital (the current platform used) and historical (the elements described in the essays that explain *keywords*, *people*, and *places*). It is clear that a process of reduction must occur in order to create the strings of code necessary to capture the relationship between these elements. Adding semantic values to textual descriptions implied a reduction of the content to essential details, such as relationship, which are expressed with the value *Keyword_relationship*. The semantic value has been utilized as additional content to be read, analyzed, and considered in conjunction with existing descriptions, and with the art historical research at the core of the project. In other words, semantic values were taken and used for what they could offer, i.e., retrievable and archivable data, and were elaborated through the lens of the art historical research central to the project.

This way of recording content cannot be considered as a way of preserving in its true sense, for it will not recreate the digital environment in which the data was originally featured. However, it allows for both data and metadata to be reloaded in a new digital environment, in order to be further utilized for data visualizations or data analysis or else. By populating the semantic data with additional information that speaks to the context within which the history of American landscape design unfolds, users can export content for further research or adopt or expand on the customized relational model for their digital art history research projects. Along with data enrichment based on the semantic web, the project will maintain fundamental aspects of *HEALD*'s digital functionalities. The most relevant outcome pertaining to this analysis is then to consider the semantics as additional values and not as a reduction of content and context.

Based on this project, I continue to delve into aspects of digital obsolescence and digital curation by moving onto my second case study.

3. EduEDA, The Educational Encyclopedia of Digital Arts

In 2004, Tommaso Tozzi, a Florence-based net artist and activist, initiated an online project called WikiARTpedia that received an Honorary Mention at the Ars Electronica Festival of Linz, Austria in 2009 (Ars Electronica 2022). In 2012, WikiARTpedia became *EduEDA, The Educational Encyclopedia of Digital Arts*: an open research platform for networks about information technology culture. The main goal of *EduEDA*, as expressed on its website, is «to create a national and international network of people and institutions in order to collectively promote and disseminate digital arts» (EduEDA 2022). *EduEDA* is a collective effort with numerous media partners and supported by both the Academy of Fine Art of Carrara and of Florence, among other institutions. The platform is an incredibly vast repository of network art and art practices that use information technology, and it includes instances of precursors to such practices, such as conceptual art and Fluxus. *EduEDA* also includes artistic experiences and practices that remained marginalized in the overall art historical narratives of digital art. Unfortunately, among these artistic experiences, those pertaining to the Italian context, continue to be underrepresented in the global histories of digital art[13]. The recently launched *Net Art Anthology* by Rhizome at the New Museum has almost entirely bypassed the experience of Italian Net Art, with the exception of *Life Sharing* (2000-2003) by Eva and Franco Mattes, aka 01.ORG (Net Art Anthology 2017).

EduEDA, which like *HEALD,* runs on MediaWiki, only partially includes image files and – for the most part – compiles links to artists' or institutions' websites. It resembles Wikipedia in the way it is organized, with a menu that includes items such as *artists*, *artworks*, *genres* or *artistic movements*, *cross search*, *space*, *time*, *macro categories*, and related descriptive pages. Equally to *HEALD*, *EduEDA* is a collection of works that often reside in other locations, at times preserved by different institutions, online or offline, and it features internal hyperlinks. As mentioned, this is a core characteristic of MediaWiki markup language, made possible by inserting the name of a page in the back-end editor within double square brackets, as in the example that follows: *[[name of the page]]* (MediaWiki 2024d).

In order to maintain the collective character of *EduEDA*, Tommaso Tozzi adopted an open source software that allows everyone to contribute. However,

[13] For instance, the BBS called *The Thing*, founded by Wolfgang Sthaele around the same time than Tozzi made *Hacker Art BBS* is often cited as part of the history of Net Art along with the netstrikes conducted by Ricardo Dominguez since 1998; while *Hacker Art BBS* is absent from art historical accounts on the subject.

unlike *HEALD*, *EduEDA* does not use Semantic MediaWiki (SMW). Let's consider whether the project could benefit from semantic values and how. On *EduEDA*, the works are listed in alphabetical order (which is a default setting of MediaWiki that creates page lists by using a label called *Category*). As noted, the platform does not use standard vocabularies, which means that it is not currently possible to interlace, compare or contrast its content with similar information available online, and it is not possible to export any of it via RDF or to share it through Linked Open Data.

EduEDA features a *crossed search* section that allows users to find elements in the repository that are intertwined. If one clicks on the *crossed search* tab, they are taken to a page where it is possible to launch queries into a database in which information is organized into categories: dates, locations, topics, publications, and so forth. This section of *EduEDA* has been implemented via Prototype, whose last version was released in 2015 (Prototype 2015). This is a significant tool within *EduEDA,* especially since search capabilities are often problematic, given the complexity of data relationships and interconnections among this kind of objects. The categories are ordered in a JSON file and are retrievable. However, in order to be found, they require a less intuitive path than an RDF export, as offered by SMW. Here below is an extract from a JSON file pertaining to the *crossed search* within the category *Net Art* (EduEDA n.d.a):

```
if(cat == "arte_delle_reti"){
elementi = ["Arte digitale", "Arte elettronica", "Arte in rete", "Arte telematica", "Ascii art", "Browser art", "Software art", "Web art", "Conservazione dell'arte digitale", "Cracker art", "Cyberfemminismo", "Database art", "Flood net", "Form art", "Game art", "Hacker art"];
```

Two important observations can be drawn from analyzing this portion of the JSON file. First, this categorization shows that extensive manual labor would be necessary to potentially retrieve this content for preservation purposes or to utilize it on a different platform. Second, from a strictly digital art historical standpoint, this categorization reveals the difficulty of classification. It is compelling to see that, in order to strategically include as many categories as possible under the label of Net Art, a macro-label was created and then subdivided into "elementi" (elements). Unfortunately, the *crossed search* section of *EduEDA* does not seem to have been created for purposes other than browsing its content online, and does not offer other insights into the digital methodological approach to information visualization. This section of *EduEDA* therefore allows for exploration of how its content interlaces, but it does not currently enable exploration of those relationships computationally.

A series of changes could be applied to the current structure of *EduEDA* or similar repositories. Whether or not these changes take place, they could

be helpful in suggesting a roadmap for digital repositories that can be used for several inter-sectorial purposes within disciplines that share a common digital ground of investigation. In addition, since MediaWiki is already set up for the semantic web, it might be possible to upgrade *EduEDA* without necessarily rebuilding it. I believe it is necessary for a current version of *EduEDA* to be kept as an additional tile in the technological history of network culture. As noted by Christiane Paul, «writing an history of new media and preserving the art itself will require new models and criteria for documenting and preserving process and instability» (Paul 2008, 6). The scenario of new media art is fragmented and these new models and criteria seem to be inevitably unstable, transitory, and multifarious. In order to formulate a proposal to use the material made available by Tozzi for research and exhibition purposes, I take into consideration two factors: first, such a material can be divided into *reproduced* and *duplicable* following the model laid out by Ippolito (2008). With regard to the artistic practices similar to Tozzi's, *reproduced* work consists of digitized tapes of live performances, happenings, or events that were transferred to a digital format. In contrast, *duplicable* work consists of those projects that were born digital. Additionally, an element of the work is unreproducible, namely the context in which these artistic experiences took place, specifically those produced during the two decades prior to the 2000s.

That being said, the areas to edit or enhance would be the following:

1. Description of the work (adopting standard vocabularies);
2. Format of the available files;
3. Search feature and tagging (adding metadata; using Semantic MediaWiki);
4. Usability of the content (RDF, CSV, XML exports).

A platform including more uniform descriptions, with details on file formats, duration of the work, code structure, and enriched metadata, should also enable the use of its content by exhibition makers. I will not go into details regarding points 1 to 3, which are rather self-explanatory. Rather, in order to further investigate digital curatorship, I will focus on point *4: Usability of the content*.

This feature can be particularly useful for exhibition purposes. While putting together a new media art exhibition, it should be possible to borrow extant projects from an online repository with proper acknowledgement of all the stakeholders involved. Enhancing this feature could lead to the creation of a shared new media art platform from which any institution could borrow the work (by downloading it or by presenting it in a browser) for the duration of an exhibition. Such a platform would then serve the much larger goal of preserving, presenting, and researching new media art experiences as occurred

prior to the 2000s. The latter is historically illustrated in *EduEDA* by descriptive pages whose content is not marked semantically.

This way, *EduEDA* will have more than just links to existing resources, as it will include the work, or a simulation of it, in whichever available format. It will still serve as the aggregator it is now – extensive and complex – and it will build on its already well-established practice of working collaboratively with multiple institutions. Another difference between this platform and existing ones will be the inclusion of the neglected history of Italian new media art. To serve this purpose, the platform should, however, be available in multiple languages (EduEDA n.d.b)[14].

Similarly to *HEALD*, the repository could have a point of entry to its content through image files rather than exclusively via text. *HEALD* image collection allows users to start exploring its vast content by directly clicking on one of the images contained in the collection (HEALD n.d.b). Images reveal their relationship to other elements of the projects through icons, and through a system of filters that guides user's exploration. Last, I contend that what was once considered the disturbing office-like aspect of early new media art exhibitions, will no longer concern exhibition makers nor viewers (Federici 2019). It is plausible to consider that, for the most part, viewers are now familiar with information technology, and will be more inclined to explore simulations of early new media artwork. The technological and temporal distance between current devices and those employed prior to the 2000s may facilitate an exploration of the latter in a gallery or museum setting due to the visual cues that would allow the viewer to see the old devices as ancestors of more recent apparatuses.

By combining knowledge, methods, and tools already available across disciplines and projects that deal with the digital realm, the investigation of digital objects can be further enhanced. All these disciplines already share the same computational and datafied language. The purpose of this exploration is to find a common ground from which they can all benefit. I will summarize a possible common path able to serve projects as diverse as *HEALD* and *EduEDA* in the conclusions.

4. Conclusions

This article focused on the open access software MediaWiki, utilized by both *HEALD* and *EduEDA*, which «is used by tens of thousands of websites and thousands of companies and organizations. [MediaWiki is] powerful, multilingual, free and open, extensible, customizable, reliable, and free of charge» (MediaWiki 2023a). Nonetheless, it is not intuitive and arguably it is

[14] At the time of writing, only 99 items in *EduEDA* have been translated into English.

not user friendly. However, the general consensus is that MediaWiki is a *good* software insofar as it is free and open, well maintained, and widely used. It is designed to allow users to edit, update, and delete content[15]. All textual content of MediaWiki.org is licensed under the Creative Commons Attribution/Share-Alike License (CC BY-SA) and the GNU Free Documentation License (GFDL) – software can be copied and modified – except for pages that explicitly state that their contents are in the public domain (MediaWiki 2023b). A Creative Commons license entails that the software can be shared – copied and redistributed in any medium or format – and adapted – remixed, transformed, and built upon (Creative Commons n.d.).

Although it was never conceived as a digital art history tool, MediaWiki embeds significant components and possibilities that can serve the discipline well. One of these components is Semantic MediaWiki; while its extensions can make the software quite versatile since their development is ongoing and responds to an ever-changing digital environment. One example is the latest implementation of an extension that works with IIIF (International Image Interoperability Framework), a tool widely used in image-based research (MediaWiki 2024c). The combined analysis of projects as diverse as *HEALD* and *EduEDA* demonstrates how adaptable MediaWiki can be and how it can support a digital art historical investigation, since it allows for the handling of complex object relationships between textual and visual content.

MediaWiki records a project's development by retaining the history of how and when each page of a digital repository is edited. This functionality can be significantly helpful. Process history is generally accepted, or even in demand, in database-driven tools for digital asset management (particularly in museums, in order to avoid losing an object's provenance or exhibition history, for instance). Nonetheless, it is often overlooked in the field of digital humanities and digital art history, whose tools tend to replicate the transparency (obliteration) model mentioned in the introduction. This is particularly true in case of data visualizations, which are highly mediated. Along with clarifying the process through which a digital project is conceptualized, and eventually displayed, the history of process itself is a rather critical aspect to consider. For instance, it helps to track users' interventions, in particular when a large team is working on the same project over an extended period of time. In addition, it testifies to the various steps, successes, and pitfalls that might characterize the creation of large repositories, and perhaps it may lead to a specific change of direction or compromise.

The overall goal is not to present MediaWiki as a panacea for the many aspects discussed throughout this article that pertain to the larger field of digital art history and are highlighted in the cited literature, but rather to contem-

[15] This functionality can be turned off, and in general users' accessibility and editing capabilities can be diversified.

plate the complexity of MediaWiki as an advantage. Given the structure and potential of the software, one wonders whether it can be used to combine those aspects. Complexity and contextualization are necessary to unravel research in the humanities. Along with standard ontologies, MediaWiki allows for the use of additional unscripted semantic categories. As a result, the digital repositories analyzed here are an example of how to integrate elements of computing with the multifarious landscape of research in the humanities – not necessarily innovatively, but coordinately.

Another goal of this investigation has been to reflect on how data is turned into information by looking at the interstices of this transformation rather than taking it for granted. Accordingly, I explored possible ways to intervene on how digital content is made and displayed. As a critical reflection on the term *datafication* and the implications of using the semantic web and web ontologies in digital art history, this article investigated the affordance given to data, in general, and its impact on the humanities, in particular. In addition, it touched upon aspects of digital curatorship and preservation, encouraging an active use of existing digital tools and practices with the intent to shape the digital realm through humanistic methodologies and approaches, rather than yielding to computing unconditionally. The turning of data into information, which is one of the main conundrums for humanists dealing with computational methods, can be done by shifting attention to the process, by highlighting and discussing the choices made, and most importantly, by engaging with the complexity of both computing and humanistic research. This approach, which is not novel, is not dismissive of quantitative research, but it does not prioritize it. Rather, it considers its potential – such as the exportability of data – without denying its limits.

By analyzing the recent upgrade to *HEALD* (HEALD 2021b), I highlighted that web vocabularies are only one way to express meaning and do not necessarily entail one way of interpreting it. Although the semantic web now allows for *HEALD*'s online content to be retrieved, exported, migrated, and further analyzed, it will not prevent the loss of the original digital environment in which the project took form. This aspect speaks to digital preservation as well as to digital curatorship. The latter, in particular, is another element that tends to go unnoticed, even though digital projects are always delivered via a web-based interface or a website available to the public, and they are often developed with an ideal user and/or a specific mode of navigation in mind. The fact that users browse platforms differently and have varying needs in terms of inclusivity and accessibility – which is a fundamental aspect that interests digital curatorship greatly – is not always considered in database-driven tools. As highlighted in the discussion on the *HEALD* upgrade and further suggested by the investigation of *EduEDA*, allowing users to enter or explore content

from multiple points of access can unleash research potential, enhance content exploration, and improve navigation experience.

It should also be noted that, for the most part, the case studies presented here pertain to archives of reproductions. That means that much of the work on display underwent a process of digitization. While all image files in *HEALD* pertained to digitized objects, the work featured in *EduEDA* is either digital or digitized. In the case of *EduEDA*, digitization was used as needed in order to align the work under the same computational language. This includes, for instance, early performance work, and other artistic interventions that were recorded on tape and then digitized, as well as works that used to run on or were made with software that is no longer available. These works can re-live as a simulation by using a software that acts like the *old* version of the one originally used. Other have explored this path. For instance, the project ArtBase by Rhizome offers emulations of "expired" software to reproduce an artwork on a current framework (Rhizome 2021)[16]. The question yet to be answered is not whether good tools for digital art history exist, rather whether we can create models, structures, and roadmaps to avoid redundancy while embracing interdisciplinary methods.

Therefore, I have juxtaposed aspects of seemingly distinct disciplines – namely digital art history, digitized art history, digital curatorship, and digital preservation – in order to glean insights that these fields can share as we navigate the digital realm. *HEALD* and *EduEDA* shed light on, and eventually helped to come to terms with, relevant aspects of the digital that pertain to both digital-born and non-digital-born artefacts, namely the difference between *reproduced* and *duplicable* items. This difference can be stretched further to be considered close to the distinction that Drucker (2013) draws between digitized art history and digital art history. Perhaps this distinction is more useful at a granular level of object analysis rather than at macro disciplinary level.

Since the digital realm and tools allow for information to be reproduced, duplicated, and replicated, we are no longer dealing with unique artefacts or objects, but items that are ubiquitous and whose *originality* as a concept and methodological approach has run into the sand. This is not to say that a clear distinction should not exist. The information pertaining to the origin of content is still precious and should be included within the metadata as well as indicated by labels. At the same time, specific scholarly competencies within the different fields should be regarded. However, once the object enters the digital sphere, it becomes ubiquitous, reproducible, and transferable, and our approach to it should embrace these inherent aspects. For this reason, and because of the ways in which the digital realm has changed the work that humanists do and how it is done, aspects of digital curatorship, digital conservation,

[16] Rhizome has recently relaunched their project ArtBase in an attempt to continue preserving digital born artefacts.

and digitized art history could exist under the umbrella of digital art history, adopting those distinctions as they pertain to the object at a micro level. These disciplines all operate on and with overlapping methodologies, and could all benefit from a crosspollination rather than compartmentalization.

Finally, by highlighting *HEALD* upgrade and suggesting similar interventions for *EduEDA*, I hope to have demonstrated how to practically move step by step toward an extended and extensive, yet obviously not comprehensive digital art history, characterized by *datafication*. The first step of this analysis illustrated the use of freely available software (MediaWiki), particularly for its ability to record process-oriented projects. Second, I showed how the relationships characterizing the material (*reproduced* or *duplicable*) have been conceptualized, defined, and described (semantic web). By doing so, I observed how information has been reduced to fit the process of *datafication*; Third, I considered how to best represent these relationships for preservation purposes (RDF export). Fourth, I addressed user interaction and project accessibility (points of access); Fifth, I also considered the opportunity to share data via existing open data initiatives (Linked Open Data; GitHub). Last, I discussed the possibility of customizing the software to adapt its functionality to the content under investigation (extensions). This path can offer different layers of content fruition and different layers of object analysis. It can help maintain a clear focus on how to approach the *digital* to serve different objectives and incentivize the creation of additional shared models (similarly to the shared vocabularies) for others to adopt.

References

Ars Electronica. 2022. "Archive: Prix." http://archive.aec.at/prix/.

Brown, Kathryn. 2020. *The Routledge Companion to Digital Humanities and Art History*. Milton: Taylor & Francis Group.

Berners-Lee, Tim. 2006. "Linked Data." Archived June 6, 2022. https://web.archive.org/web/20220606143535/http:/www.w3.org/DesignIssues/LinkedData.

Bucher, Taina. 2018. *If … Then: Algorithmic Power and Politics*. New York, N.Y.: Oxford University Press.

Chun, Wendy Hui Kyong. 2011. *Programmed Visions: Software and Memory*. Cambridge, Mass.: The MIT Press.

Couldry, Nick, and Ulises Ali Mejias. 2021. "The Decolonial Turn in Data and Technology Research: What is at Stake and where is it Heading?" *Information, Communication & Society* 26 (4): 786-802. https://doi:10.1080/1369118X.2021.1986102.

Creative Commons. n.d. "Licenses." Accessed October 7, 2024. https://creativecommons.org/licenses/by-sa/3.0/.

Daston, Lorraine, and Peter Galison. 1992. "The Image of Objectivity." *Representations*, no. 40: 81–128. https://doi.org/10.2307/2928741.

Drucker, Johanna. 2011. "Humanities Approaches to Graphical Display." *Digital Humanities Quarterly* 5 (1).

Drucker, Johanna. 2013. "Is There a 'Digital' Art History?" *Visual Resources* 29 (1-2): 5-13. https://doi.org/10.1080/01973762.2013.761106.

EduEDA. 2022. "EduEDA - The EDUcational Encyclopedia of Digital Arts." Archived August 17, 2022. https://web.archive.org/web/20220817082401/http://www.edueda.net/index.php?title=EduEDA_-_The_EDUcational_Encyclopedia_of_Digital_Arts.

EduEDA. n.d.a "Config:Ricerca:JS:Form." Accessed October 7, 2024. https://www.edueda.net/config/ricerca/js/form.js.

EduEDA. n.d.b "Index." Archived June 30, 2020. https://web.archive.org/web/20200630032503/http://www.edueda.net/index.php?title=Categoria:EN.

Federici, Valeria. 2019. "Curating New Media Art in Italy in the 1980s: The Uneasiness of Medium Contingency." *Interdisciplinaryitaly.org.* https://interdisciplinaryitaly.org/curating-new-media-art-in-italy-in-the-1980s-the-uneasiness-of-medium-contingency/.

Federici, Valeria. 2022. "The Age of Datum or Data as a Methodological Paradigm." *The Italianist* 42 (3): 323–43. https://doi.org/10.1080/02614340.2023.2223877.

Gitelman, Lisa, ed. 2013. *"Raw Data" Is an Oxymoron*. Cambridge, Mass.: The MIT Press.

GitHub. 2021. "NationalGalleryOfArt/heald-packages. https://github.com/NationalGalleryOfArt/heald-packages.

HEALD. 2021a. "Home." https://heald.nga.gov/mediawiki/index.php/Home.

HEALD. 2021b. "Recent Upgrade." https://heald.nga.gov/mediawiki/index.php/Recent_Upgrade.

HEALD, n.d.a. "File:0999.jpg: Anonymous, Two Ornamental Ice Houses Above Ground, 1846." Accessed October 7, 2024. https://heald.nga.gov/mediawiki/index.php/File:0999.jpg.

HEALD, n.d.b. "Image Collection." Accessed October 7, 2024. https://heald.nga.gov/mediawiki/index.php/History_of_Early_American_Landscape_Design:Image_Collection.

Ippolito, Jon. 2008. "6 Death by Wall Label." In *New Media in the White Cube and beyond: Curatorial Models for Digital Art*, 106–32. https://doi.org/10.1525/9780520942349-008.

Levenberg, Lewis, Tai Neilson, and David Rheams, eds. 2018. *Research Methods for the Digital Humanities*. Cham, Switzerland: Palgrave Macmillan.

Manovich, Lev. 2020. *Cultural Analytics*. Cambridge, Mass.: The MIT Press.

MediaWiki. 2023a. "MediaWiki." Last modified December 29, 2023 at 18:14 (UTC). https://www.mediawiki.org/wiki/MediaWiki.

MediaWiki. 2023b. "Project:Copyrights." Last modified October 1, 2023 at 12:06 (UTC). https://www.mediawiki.org/wiki/Project:Copyrights.

MediaWiki. 2024a. "Manual: Extensions." Last modified August 24, 2024 at 09:11 (UTC). https://www.mediawiki.org/wiki/Manual:Extensions.

MediaWiki. 2024b. "Markup_spec." Last modified July 20, 2024 at 19:24 (UTC). https://www.mediawiki.org/wiki/Markup_spec.

MediaWiki. 2024c. "Extension:IIIF." Last modified September 9, 2024, at 7:51 (UTC). https://m.mediawiki.org/wiki/Extension:IIIF.

MediaWiki. 2024d. "Internal_links." Last modified October 5, 2024 at 01:16 (UTC). https://www.mediawiki.org/wiki/Help:Links/en#Internal_links.

Morozov, Evgeny. 2013. *To Save Everything, Click Here: The Folly of Technological Solutionism*. New York, N.Y.: PublicAffairs.

Näslund, Anna, and Amanda Wasielewski. 2021. "The Digital U-Turn in Art History." *Konsthistorisk Tidskrift/Journal of Art History* 90 (4): 249–66. https://doi.org/10.1080/00233609.2021.2006774.

Net Art Anthology. 2017. "Life Sharing." https://anthology.rhizome.org/life-sharing.

O'Malley, Therese. 2010. *Keywords in American Landscape Design*. New Haven, C.T.: Yale University Press.

Paul, Christiane, ed. 2008. *New Media in the White Cube and beyond: Curatorial Models for Digital Art*. Berkeley: University of California Press.

Penn, Jonnie. 2021. "Algorithmic Silence: A Call to Decomputerize." *Journal of Social Computing* 2 (4): 337–56. https://doi.org/10.23919/JSC.2021.0023.

Porras, Stephanie. 2017. "Keeping Our Eyes Open: Visualizing Networks and Art History." *Artl@S Bulletin* 6 (3).

Porter, Theodore M. 1995. *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*. Princeton, N.J.: Princeton University Press.

Prototype. 2015. "Home." http://prototypejs.org/.

Rhizome. 2021. "Artbase: Main Page." https://artbase.rhizome.org/wiki/Main_Page.

Schäfer, Tobias Mirko, and Karin van Es, ed. 2017. *Datafied Society: Studying Culture through Data.* Amsterdam University Press.

Spieker, Sven. 2008. *The Big Archive: Art from Bureaucracy*. Cambridge, Mass.: The MIT Press.

Treccani. 2020. "Datificazione." Archived February 18, 2020. https://web.archive.org/web/20200218181114/http://www.treccani.it/vocabolario/datificazione_(Neologismi)/.

Zotero. 2021. "keywords_in_early_american_landscape_design/library." https://www.zotero.org/groups/54737/keywords_in_early_american_landscape_design/library.

Zuboff, Shoshana. 2019. *The Age of Surveillance Capitalism*. New York, N.Y.: PublicAffairs.

AIDAinformazioni
ISSN 1121–0095
ISBN 979-12-5965-456-4
DOI 10.57574/596545645
pag. 101-118 (luglio-dicembre 2024)

Testimonianze di un impegno culturale per l'Università di Salerno.

Le carte di Alfonso Menna

Rosa Parlavecchia*

Abstract: The contribution aims to explore the life and activities of Alfonso Menna - a central figure in the administrative and cultural history of Salerno - through his personal archive. Menna began his administrative career in Sarno and later in Salerno, where he held numerous important positions, including mayor from 1955 to 1970. During his tenure, he led significant urban development projects, and his cultural commitment is highlighted by the promotion of events, conferences, and the commemoration of illustrious local personalities through his writings. Alfonso Menna's personal archive, kept at the Salerno State Archive, bears witness to his lively activity through a collection of documents personally ordered by Menna before his donation.
The detailed analysis of his papers offers a valuable overview of the administrative and cultural dynamics of Salerno, providing a fundamental tool for scholars and researchers interested in local history and the evolution of educational institutions, in particular the University of Salerno.

Keywords: Alfonso Menna, Personal archive, Salerno State Archive, University of Salerno, Libraries.

1. Introduzione

Ci sono «uomini che hanno ben meritato della civica riconoscenza» (Menna 1993, 1996), impegnati nel servizio delle istituzioni pubbliche, che si distinguono per la loro straordinaria abilità nel canalizzare l'energia, la passione politica e civile, l'integrità morale e la competenza nell'organizzazione amministrativa. Tra questi spicca indubbiamente Alfonso Menna, esempio paradigmatico di amministratore pubblico: visionario, moralmente irreprensibile, competente ed efficiente. Una figura di rilievo istituzionale per la storia della città di Salerno di cui fu sindaco dal 10 luglio del 1955 al 17 ottobre 1970.

* Dipartimento di Scienze del patrimonio culturale, Università degli studi di Salerno, Fisciano (SA), Italia. rparlavecchia@unisa.it.

Nato nel 1890 a Domicella in provincia di Avellino, Menna si distinse sin da giovane per la sua determinazione nell'ambito degli studi, conseguendo nel 1911 il Diploma di Segretario comunale. La sua carriera prese avvio come Vicesegretario ragioniere presso il comune di Sarno. Successivamente, grazie al superamento di concorsi ed esami, si trasferì al Comune di Salerno, dove assunse il ruolo di Vicesegretario di sezione, operando con impegno all'interno della Segreteria generale.

Con il trascorrere degli anni e la costante dedizione al lavoro, Menna scalò gradualmente i gradini della carriera amministrativa. Dopo un percorso professionale di oltre trent'anni, durante il quale si fece notare per le sue competenze riconosciute a livello nazionale, raggiunse il ruolo di Vicesegretario generale e successivamente fu promosso con la qualifica di Segretario generale di prima classe (Sbrescia 2022, 793).

La sua lunga carriera, infatti, lo ha visto ricoprire ruoli di rilievo quali appunto Segretario comunale, Commissario prefettizio del Comune di Battipaglia[1], Segretario generale della città di Salerno, Commissario dell'Istituto nazionale per la gestione delle imposte di consumo, sindaco di Salerno per circa tre lustri e successivamente Presidente dell'Istituto per lo Sviluppo Economico dell'Italia Meridionale (ISVEIMER), costituito con regio decreto-legge del 31 giugno 1938, n. 883[2].

La dedizione al servizio di Menna emerse con forza anche in momenti di crisi, come l'alluvione del 1954, un tragico evento inciso nella memoria collettiva che segnò profondamente la città di Salerno. Tale dedizione non passò inosservata, tanto che nel 1958 fu insignito del titolo di Grande Ufficiale dell'Ordine al Merito della Repubblica Italiana, in riconoscimento dei suoi evidenti sforzi[3].

---

[1] Commissario per l'Amministrazione provvisoria dalla costituzione del Comune di Battipaglia dal maggio del 1929 al novembre del 1931. «Battipaglia, elevata in comune, inizierà la sua nuova vita, e, sotto l'Egida del Littorio, contribuirà, con quella tenacia che è caratteristica dei rurali, alla costruzione dell'edificio al cui vertice è la redenzione economica della Patria» (Menna 1930, 1931, 1932).

[2] L'ISVEIMER, con sede a Napoli, esercitava il credito a medio termine in favore delle medie e piccole imprese industriali al fine di mettere in valore risorse economiche e possibilità di lavoro nel Meridione. Al fondo di dotazione dell'Istituto parteciparono, oltre alla Cassa del Mezzogiorno, il Banco di Napoli nella misura del 40% del fondo stesso, e, nella misura del 20% complessivamente, le Casse di risparmio ed altri istituti di credito nei territori interessati (Istituto per lo Sviluppo Economico dell'Italia Meridionale 1959).

[3] Inoltre, nel 1963 gli fu conferita l'onorificenza dell'Ordine al Merito della Repubblica Italiana e nel 1970 il titolo di Cavaliere dell'Ordine di Vittorio Veneto. Nel 1968 ricevette la Medaglia d'oro al merito civile con queste motivazioni: «Poneva le sue eccezionali capacità di amministratore e di organizzatore al servizio della collettività, promuovendo e potenziando, con feconda continuità, ogni iniziativa volta al processo di rinnovamento del Mezzogiorno ed allo sviluppo sociale ed economico della città di Salerno. Uguale contributo di capacità e di dedizione egli offriva in occasione di due drammatiche inondazioni della città che lo vedeva-

Nel 1955, Menna si ritrovò quasi catapultato nella vita politica locale, risultando eletto sindaco di Salerno con la Democrazia Cristiana. Questo risultato sorprendente, pervenuto nonostante un grave problema di salute che lo aveva tenuto lontano dalla città e dalla campagna elettorale, è testimone sia di quella che era la sua autorevolezza personale, sia dell'apprezzamento e della fiducia che la cittadinanza nutriva nei suoi confronti.

Questa breve introduzione risulta soltanto un frammento della lunga vicenda biografica di Menna, il quale, scomparso nel 1998 all'età di 108 anni, ha lasciato un'impronta significativa nel tessuto sociale e culturale di Salerno.

Testimonianza della sua vivace attività sono le carte donate a più riprese dallo stesso all'Archivio di Stato di Salerno poco prima della sua morte (Menna 1993, 1996). Ancor prima di effettuare la donazione, l'ex sindaco aveva riordinato i documenti per cui il materiale si è presentato al personale addetto già ben organizzato e classificato. Il Fondo Alfonso Menna è costituito da due serie: *Università* (buste 1-4) e *Profili* (buste 5-11), per un totale di undici buste[4].

Per quanto riguarda la serie *Profili*, la documentazione copre un arco cronologico che va dal 1851 al 1996 e raccoglie materiale eterogeneo relativo a molte personalità di spicco per la città di Salerno. Ogni fascicolo presenta il nominativo dell'autorità di cui Menna si stava occupando – artisti, pittori, poeti, politici, ecclesiastici – e al loro interno sono presenti diversi documenti classificati per tipologie: articoli di giornale, corrispondenza, *curriculum vitae*, pubblicazioni.

I fascicoli ospitano, inoltre, opuscoli di mostre artistiche, immagini di lapidi commemorative, preghiere, libri, cartoline, pubblicazioni di vario genere, nonché inviti a eventi di diversa natura. Ognuno di questi fascicoli contribuisce a tessere il vasto mosaico della storia salernitana dagli anni Quaranta agli anni Novanta del secolo scorso, secondo quella che era la visione 'archivistica' di Menna.

L'importanza di queste personalità emerge in modo tangibile dai *Profili* pubblicati dallo stesso Menna per il "Corriere del Mezzogiorno" e nei due volumi editi per i tipi di De Luca nel 1993, *Come li ricordo. Sono uomini che*

---

no impegnato, con gravissimi rischi personali, in una illuminata, pronta ed umanitaria opera di soccorso ed assistenza. L'imponenza delle realizzazioni da lui conseguite nel corso di una pluridecennale attività ha trovato rispondenza nella commossa riconoscenza delle popolazioni interessate e nella larga ammirazione ovunque suscitata. 1943-1968» (Presidenza della Repubblica italiana 2021).

[4] Fondo Alfonso Menna, Archivio di Stato, Salerno. La bibliografia sugli archivi di persona è amplissima. Ci si limita a indicare (D'Addario 1972, 45; Barrera 2006, 617-657; Di Domenico e Sabba 2020). Per una sintetica definizione si rimanda alle *Linee guida sul trattamento dei fondi personali* a cura della Commissione nazionale biblioteche speciali, archivi e biblioteche d'autore (2019) dell'Associazione italiana biblioteche, versione 15.1.

hanno ben meritato della civica riconoscenza. Infatti, come riportato nella *Prefazione*

> nello stendere le ultime pagine del libro *Il banco e la cattedra* (Menna 1991) mi sono accorto che devo ancora adempiere il dovere di ricordare, ai presenti e ai venturi, coloro che con me, e certamente più di me, diedero le proprie risorse per aprire a Salerno la strada che doveva condurla al suo divenire. […] Esortato anche da comuni amici, mi propongo di mettere sulla carta poche cose. Sono profili di persone colte nelle loro tendenze, in atteggiamenti e posizioni caratteristiche (Menna 1993, 7).

Il lavoro svolto dall'ex sindaco si concentra principalmente sul lato umano delle personalità esaminate, mirando a comprendere il loro modo di agire, vivere e pensare. Si tratta di biografie di uomini illustri che rivelano un'attenzione particolare ai dettagli e una volontà marcata di tramandare il ricordo di personaggi rilevanti per le generazioni future. Molte delle figure citate sono amici dello stesso Menna, la cui memoria è elogiata e commemorata attraverso bozze di discorsi e lettere ai parenti attraverso le quali l'autore riesce a ricostruire vicende personali e professionali con dovizia di particolari.

2. Le carte dell'Università di Salerno nel Fondo Menna

Lo studio universitario nella città di Salerno vanta una storia che affonda le radici nell'antichità e si snoda attraverso periodi di rinascita culturale e trasformazioni socio-politiche. Le sue origini possono essere fatte risalire alla Scuola Medica Salernitana, rinomata istituzione medica dell'Alto Medioevo, che fiorì tra il IX e il XIII secolo. Questa scuola, considerata da alcuni la prima università medievale al mondo, attrasse studenti e studiosi da ogni dove con il suo programma di studi innovativo e l'attenzione alla ricerca scientifica[5].

Tuttavia, l'Università di Salerno nella sua forma moderna ha radici ben più recenti[6]. È stata fondata nel 1944 poco dopo il passaggio del governo Badoglio a Salerno e con la nomina di un salernitano, Giovanni Cuomo[7], alla guida

[5] Per approfondimenti sulla storia della Scuola Medica Salernitana si rimanda a (De Renzi 1857; Kristeller 1986; Gallo 1994; Galasso 1995; Leone e Sangermano 2003; Jacquart e Paravicini Bagliani 2007; Vitolo 2007; Galdi 2021).

[6] Per una storia dell'Università di Salerno si vedano almeno (Musi, Oldoni e Placanica 2001-2004; Zaccaria e Amato 2018, 31-103; Andria 2020, 309-314).

[7] La figura di Giovanni Cuomo fu di grande rilievo nel panorama politico cittadino salernitano per circa vent'anni prima dell'ascesa del regime fascista. Dopo l'armistizio e lo Sbarco di Salerno, ritornò alla vita politica come principale esponente locale e in seguito come membro della Costituente. La sua presenza e influenza nel contesto politico locale e nazionale sono stati significativi, specialmente durante i periodi di transizione e cambiamento come quello seguito alla fine della Seconda Guerra Mondiale. Si vedano almeno (D'Auria 1985; D'Aniello 1994; Bonani 2007, 2008a, 2008b).

del Ministero dell'Educazione Nazionale[8]. L'istituzione fu creata inizialmente come una Facoltà di Magistero, sebbene l'idea fosse stata proposta già nel 1942, come testimonia un documento dell'Amministrazione podestarile di Salerno indirizzato al Ministro dell'Educazione Nazionale, Giuseppe Bottai[9]. La proposta, infatti, non fu accolta inizialmente a causa di varie complicazioni burocratiche e finanziarie. Con il sostegno del governo e il desiderio di fornire opportunità educative ai giovani delle regioni meridionali, l'Università di Salerno iniziò a crescere e ad espandersi nel corso degli anni.

Nei primi anni Cinquanta, grazie al supporto di una convenzione finanziaria tra Comune, Provincia e Camera di Commercio vi fu un definitivo riconoscimento dell'Istituto di Magistero Universitario di Salerno grazie a un secondo atto costitutivo (D.P.R. del 9 ottobre 1951, n.1300) a cui fece seguito uno statuto in cui, all'articolo 1 comma 3, si consentiva l'accesso alla sola popolazione maschile in virtù della vicinanza con il "Suor Orsola Benincasa" di Napoli destinato alla frequentazione di sole donne[10].

Tuttavia, nel giro di pochi anni apparve evidente che la frequenza degli studenti fosse qualitativamente e quantitativamente scarsa per cui si pensò di porre rimedio con una modifica dello statuto affinché fossero ammesse a frequentare anche le donne. Infatti, grazie all'intervento del costituzionalista Vincenzo Sica, membro di un comitato tecnico di nomina ministeriale, venne dichiarata ammissibile la proposta di modifica statutaria e pertanto, con il D.P.R. del 22 marzo 1961, le donne furono ammesse a frequentare l'Istituto di Magistero di Salerno.

Tra le carte di Menna conservate all'Archivio di Stato di Salerno, sono numerosi i documenti che riguardano questa annosa questione che ha visto l'allora sindaco esporsi in prima linea. Sono presenti, infatti, verbali di deliberazione commissariali e del Consiglio Comunale aventi ad oggetto l'ammissione delle donne ai corsi di studio[11]; e poi un documento, datato 5 giugno 1954, in cui il Commissario prefettizio richiede al Ministro della Pubblica Istruzione di

[8] Deliberazione commissariale del 16 febbraio 1944 n. 44. Registro delle Deliberazioni del Commissario Prefettizio, 1943, dall'1 al 59; 1944 dall'1 al 735, Sezione Storica, Archivio Comunale, Salerno.

[9] Il cuore della questione risiedeva nella notevole distanza delle Regioni dell'Italia meridionale dalle città che ospitavano le Facoltà di Magistero, come Napoli con la sua sezione femminile presso l'Istituto "Suor Orsola Benincasa", Messina e Roma. Questa distanza costituiva un ostacolo significativo per molti laureati provenienti dalla Campania, dall'Irpinia, dalla Puglia, dalla Lucania e dalla Calabria, impedendo loro di continuare gli studi in maniera agevole, Istituto Sup. di Magistero, Funzionamento, Istituto Sup. di Magistero, Istituzione, Categoria IX, classe 3, fasc. 34/2, 1942-1963, Archivio Comunale, Salerno.

[10] Questa disposizione fu fortemente contrastata dalla politica salernitana che chiese fossero ammesse a frequentare almeno le studentesse residenti su tutto il territorio provinciale (Musi 2004, 14; Andria 2020, 310).

[11] Consiglio Comunale di Salerno 1944-1968. Verbali di deliberazione commissariali, busta 2, fascicolo 1, Fondo Alfonso Menna, Archivio di Stato, Salerno.

accogliere il voto unanime della città di Salerno relativo all'accesso all'Istituto di Magistero delle donne nate o residenti a Salerno; alcune lettere di Menna indirizzate al professori Giordano, Tesauro e Quagliariello, quest'ultimo rettore dell'Università di Napoli, sempre relative all'ammissione delle donne, ma anche relazioni e numerosi articoli di giornali dedicati alla questione ("Il setaccio"; "Il Messaggero"; "Il Mattino"; "L'Unità"; "La voce di Salerno"; "Il giornale d'Italia"; "Il quotidiano", "La guida del popolo", "Il Tempo")[12].

Sempre legati all'ammissione delle donne ai corsi di studio, si segnalano, inoltre, una lettera e un telegramma del sindaco indirizzati al Ministro della Pubblica istruzione, Aldo Moro[13]; e poi la dichiarazione di Menna, datata 17 maggio 1959, relativa alla modifica dello statuto dell'Istituto di Magistero e la delibera dell'Istituto Universitario "G. Cuomo", datata 24 ottobre 1960, sull'ammissione delle donne ai corsi di studio[14].

All'interno delle buste della serie *Università* si conservano, dunque, documenti relativi a due questioni di primaria importanza: la prima riguarda l'appena citata ammissione delle donne ai corsi, mentre la seconda riguarda la necessaria individuazione di una adeguata collocazione finalizzata all'espansione della sede universitaria.

Fin dal principio, gli spazi messi a disposizione per l'Istituto di Magistero risultarono piuttosto esigui: appena cinque locali dati in concessione dal Comune e dalla Provincia in quella che era la ex sede della Corte di Assise in via Tasso e in quella che era la sede della Biblioteca provinciale, presso il Palazzo Pinto. Con il passare degli anni, grazie al diretto intervento di Menna, verranno messi a disposizione nuovi locali per lo svolgimento delle lezioni, presso la nuova sede a Piazza Malta, nel pieno centro cittadino.

Un'altra tappa importante nella storia dell'Ateneo salernitano è, infatti, il processo di statalizzazione che portò il Magistero parificato "Giovanni Cuomo" a diventare Istituto universitario di Magistero statale con l'approvazione da parte delle due Camere della Legge n. 199 dell'8 marzo 1968 (Andria 2020, 311). Tra le priorità dell'allora rettore, Gabriele De Rosa, vi furono «la riqualificazione culturale dell'ateneo, la questione del nuovo insediamento collegata al modello di università, il rapporto assai complesso con il potere politico» (Musi 2004, 16).

«L'espansione dell'organismo appena passato alle competenze dello Stato […] risulterà decisiva per il futuro dell'ateneo» (Andria 2020, 312), infatti, nel giro di pochi anni nasceranno nuove facoltà (Lettere e filosofia, Lingue e let-

---

[12] Consiglio Comunale di Salerno 1944-1968. Verbali di deliberazione commissariali, busta 2, fascicoli 11-15, 17-20, 24-17, 29.

[13] Consiglio Comunale di Salerno 1944-1968. Verbali di deliberazione commissariali, busta 2, fascicolo 23.

[14] Consiglio Comunale di Salerno 1944-1968. Verbali di deliberazione commissariali, busta 2, fascicoli 28, 30.

terature straniere, Scienze matematiche, fisiche e naturali) e numerosi corsi di laurea. Le sedi dislocate per la città risulteranno inidonee e per lo svolgimento delle attività didattiche e i laboratori tanto da portare all'inaugurazione, nel 1978, di un nuovo plesso universitario destinato alle facoltà scientifiche presso il Comune di Baronissi, distante circa 10 chilometri da Salerno.

Il tema della dislocazione urbana delle strutture universitarie in rapida espansione caratterizzò il dibattito salernitano politico e pubblico per diversi anni. Sin dagli inizi degli anni Settanta, infatti, ci era resi conto che i locali cittadini erano inadeguati per cui uno studio condotto dall'ingegnere Corrado Beguinot e dall'architetto e urbanista Giulio De Luca aveva individuato nei paesi della Valle dell'Irno il luogo migliore per delocalizzare l'università.

Con uno studio commissionato all'architetto Paolo Portoghesi, invece, il Comune di Salerno rispose cercando di evidenziare le 'carenze' di un modello Campus ormai tramontato, come dichiarato dallo stesso Menna in occasione di un accorato intervento durante un Consiglio comunale del 21 dicembre 1970:

> […] Salerno, come Città e Capoluogo, si presenta con le carte in regola e con tutto il peso delle sue tradizioni, della sua storia, delle legittime aspettative dei suoi figli, che non intendono, nel modo più assoluto, essere defraudati di ciò che un loro sacrosanto diritto ad opera di una politica di disconoscimento di valori e di esigenze che pur si impongono anche all'occhio del più sprovveduto: intendo riferirmi alla localizzazione dell'Università. […] Vorrei che i Signori amministratori ascoltassero un poco la voce dell'opinione pubblica Salernitana; ovunque si giudica il tentativo con incredulità e addirittura con ilarità. […] Si tende ora a cambiare la denominazione dell'Istituto universitario: nella relazione non si parla più di «Università di Salerno» ma di «Università della Campania»; e ciò facendo si reca una grave offesa alla parte più sensibile dell'anima della Città. […] È davvero inconcepibile il tentativo in atto: nel caso in esame non si tratta di togliere qualcosa ad altro Comune per darla a Salerno; si tratta, invece, di togliere a Salerno ciò che essa è riuscita a darsi attraverso anni di lotte e di sacrifici, senza alcun plausibile motivo.
> E ciò la Città potrebbe consentire? Nel lanciare l'idea di una diversa ubicazione, si sono considerate le ripercussioni che si potrebbero avere nella pubblica opinione? (Menna 1971, 13-15).

Nella busta n. 1 della serie *Università* sono conservati documenti cruciali relativi allo studio di fattibilità redatto dall'architetto Paolo Portoghesi. I tratti distintivi del progetto di fattibilità concepito offrono un'ampia panoramica sulle idee di rinnovamento e riqualificazione urbana che avrebbero caratterizzato la città di Salerno. Tra le carte sono custoditi i risultati di uno sforzo collaborativo e multidisciplinare, espresso attraverso una relazione tecnica esaustiva e tavole di lavoro dettagliate. Questi documenti rappresentano una sorta di manifesto delle aspirazioni della comunità locale verso una trasformazione ur-

bana intelligente e sostenibile. Il progetto proposto dall'architetto Portoghesi, infatti, non si limitava a un semplice rinnovamento degli spazi universitari, ma ambiva a integrare l'università nell'ambiente circostante, facendo leva sul recupero di edifici dismessi e sull'espansione dell'area di intervento verso la zona orientale della città e le colline circostanti.

Questa visione olistica rifletteva un desiderio di armonizzare gli interessi della comunità accademica con quelli della città nel suo complesso. Nonostante il progetto non sia mai stato realizzato fisicamente, la sua esistenza e la sua elaborazione dettagliata rappresentano un contributo di notevole rilevanza al dibattito in corso riguardante lo sviluppo urbano di Salerno. Questo dibattito, come testimoniato dagli sforzi accurati e diligenti del sindaco Menna nei documenti rappresenta il suo impegno relativo all'ubicazione dell'Università di Salerno. L'interesse meticoloso nella raccolta di testimonianze sottolinea l'importanza attribuita a questo dibattito da parte delle istituzioni locali e della comunità salernitana nel suo insieme (Andria 2020, 312)[15].

3. Prospettive di ordinamento e sviluppo delle biblioteche dell'Università di Salerno

Un fascicolo appartenente all'ultima busta della serie *Università* testimonia uno spaccato molto interessante relativo alle questioni organizzative, gestionali e scientifiche dell'Università di Salerno. Al suo interno, infatti, si trovano un insieme di documenti datati tra il 1974 e il 1976 relativi alle nuove prospettive di sviluppo e ordinamento delle biblioteche.

Sin dalla nascita dell'Istituto di Magistero, inizialmente ospitato presso la sede della Biblioteca provinciale di Salerno (Parlavecchia 2020, 292)[16], la dotazione libraria contava solo poche decine di volumi; si trattava principalmente di donazioni da parte di docenti o di acquisizioni realizzate con i pochi fondi di cui si disponeva. Tuttavia, già negli anni Quaranta, grazie all'incremento dei finanziamenti destinati agli acquisti, con i volumi offerti dal Ministero della Pubblica Istruzione e con quelli devoluti da parte di privati, una raccolta libraria di tutto rispetto inizia a prendere forma. Nel 1950, infatti, la biblioteca può

[15] Il Consiglio di amministrazione dell'ateneo deliberò il 3 febbraio 1971 per l'ubicazione extraurbana nella circoscrizione territoriale dei comuni di Baronissi, Mercato San Severino e Fisciano. I lavori di quello che diventerà il Campus di Fisciano presero avvio il 29 gennaio 1982, mentre bisognerà aspettare il mese di settembre del 1987 per il trasferimento di persone e cose.

[16] La Biblioteca provinciale di Salerno era ospitata dal 1910 al primo piano di palazzo Pinto, nel centro storico, per concessione da parte del cavaliere Gaetano Pinto alla Provincia di Salerno. Oltre al palazzo venne successivamente donata un'importante raccolta libraria costituita da oltre 1000 volumi, 234 pergamene databili tra il 993 e il 1761 e un cospicuo numero di manoscritti.

vantare circa 1.565 volumi, un numero che si triplica a circa 3.600 pochi anni dopo raggiungendo un valore patrimoniale significativo (Andria 2008a, 271).

Negli anni Sessanta, le sottoscrizioni alle principali collane e gli abbonamenti a periodici cominciano a delineare una collezione dalla fisionomia sistematica e funzionale. Inoltre, nel 1966 il trasferimento della biblioteca dal centro storico al centro cittadino in ambienti di oltre 600 m^2 segna una svolta importante.

La biblioteca delle facoltà di Magistero e di Lettere e filosofia, insieme all'università, peregrineranno in diverse sedi trovando una sistemazione definitiva a via Irno. In questo periodo, si assiste alla formazione di una vasta collezione umanistica, in grado di supportare pienamente le esigenze degli studenti e dei docenti. Nel giro di otto anni dalla sua nuova collocazione, la biblioteca vedrà crescere esponenzialmente il suo patrimonio librario, triplicando il numero di volumi e introducendo circa 800 testate periodiche in abbonamento. Nonostante alcune preoccupazioni riguardo alle condizioni ambientali nella nuova sede, la biblioteca riuscirà a gestire l'aumento costante di nuove acquisizioni, che si aggireranno intorno alle 5000 unità all'anno.

Risale all'anno accademico 1970/1971, invece, l'istituzione della biblioteca delle facoltà di Economia e Commercio e di Giurisprudenza (Andria 2008a, 272). Grazie a una politica di acquisti decisa e a un flusso continuo di donazioni, la collezione si arricchisce rapidamente. Donazioni di pregio e di rilevanza scientifica hanno contribuito ad ampliare il patrimonio librario. Tra queste, si segnalano i lasciti di illustri personalità come Cenzato (con opere di argomento storico-politico), De Cecco (che comprendeva volumi principalmente di Diritto civile, penale, amministrativo, romano, ma anche di Economia, Storia e Medicina), De Crescenzo (con repertori e testi giuridici dei primi decenni del XX secolo) e Ingrosso (principalmente di carattere economistico). Particolarmente significativo, per qualità e quantità dei volumi, il fondo librario che fu devoluto dagli eredi di Giovanni Cuomo (Andria 2008b, 19-29; Manzo 2021, 243)[17].

[17] La raccolta libraria, composta da circa novemila unità bibliografiche, si caratterizza principalmente per i materiali di supporto alle attività professionali del Cuomo. Le edizioni coprono un ampio arco cronologico che va dal Cinquecento alla metà del Novecento. Il fondo è stato accuratamente catalogato e organizzato in un sistema di classificazione che include le seguenti sezioni: giuridica, umanistica, periodici, estratti e antico. Dal 2011, il Centro bibliotecario di Ateneo ha avviato diverse iniziative volte alla catalogazione e alla valorizzazione del Fondo Cuomo. Tra queste attività rientra anche il progetto di digitalizzazione di alcune opere. Infatti, oltre 600 items sono attualmente fruibili tramite la piattaforma "Liberabit," (Università degli Studi di Salerno, n.d.a.), la biblioteca digitale dell'Università di Salerno che offre la possibilità di consultare documenti che fanno parte del patrimonio culturale dell'Ateneo. Realizzata in ambiente open source con DSpace-GLAM, Liberabit assicura l'archiviazione a lungo termine di oggetti e collezioni digitali. Condivisione e interoperabilità dei dati sono garantite dall'uso di standard e formati aperti. Inoltre, l'integrazione fra DSpace-GLAM e l'ecosistema

I documenti del Fondo Menna testimoniano una fase decisiva per la definizione di una nuova politica delle biblioteche dell'Università. È, infatti, in queste carte che si delinea il progetto che porterà a una centralizzazione del Sistema bibliotecario di Ateneo.

Una prima relazione riguardante le prospettive di ordinamento e sviluppo della biblioteca di Economia e Commercio e Giurisprudenza è affidata al direttore scientifico di quest'ultima, Augusto Placanica (Petrucciani 2005)[18], ed è datata 31 gennaio 1974. Nel documento il professor Placanica ricostruisce l'allora critico stato della biblioteca con un'analisi molto pungente e dai toni giustamente polemici.

In particolare, quella che emerge è una condizione alquanto paradossale dove lo stato di disorganizzazione della biblioteca e le problematiche strutturali ne impediscono un funzionamento efficiente:

> 1) il magazzino librario (il cosiddetto capannone) è addirittura allogato in un corpo di fabbrica separato dal resto dell'edificio, per cui - quand'anche i libri fossero posti in ordine l'addetto dovrebbe andare a prelevare e riporre un libro per volta; e ogni volta dovrebbe percorrere due rampe di scale in discesa e in salita, effettuare una passeggiata nell'interrato e una al piano terra, aprire e chiudere per due volte due porte interne e una addirittura esterna; operazioni tutte da ripetere per la rimessa a posto dello stesso volume; 2) pertanto il magazzino librario è vuoto; 3) ergo, i libri non sono collocati, cioè non hanno né etichetta né posto al centro; 4) ergo, la schedatura generale e centralizzata è stata rinviata; 5) ergo, i cataloghi non esistono; 6) la Biblioteca, dunque, è come se non ci fosse[19].

Sebbene la dotazione libraria conti circa 20.000 volumi e 800 riviste, la biblioteca soffre di gravi mancanze in termini di organizzazione, catalogazione e accessibilità. I libri non sono schedati, né correttamente collocati, rendendo difficile il loro utilizzo da parte di studenti e studiosi. La dislocazione dei volumi nei vari istituti aggrava il problema, causando una frammentazione e complicando ulteriormente l'accesso alle risorse.

Per risolvere la situazione, Placanica propone due soluzioni, una a lungo termine per la quale prevede una ristrutturazione completa della biblioteca in

---

IIIF mette a disposizione funzionalità e modalità di navigazione adatti per la ricerca e lo studio. Si veda il "Fondo Cuomo" 2021 (Università degli Studi di Salerno, n.d.b.). Per quanto riguarda i fondi di persona acquisiti nel corso degli anni dalle biblioteche dell'Università di Salerno, si rimanda a (Andria 2017, 9-31).

[18] Placanica, professore di Storia moderna presso l'ateneo salernitano, diresse dal 1970 al 1974 la Biblioteca comunale di Catanzaro "Filippo De Nobili" e fu professore incaricato di Bibliografia presso l'Università di Messina.

[19] Placanica 1974. Relazione del direttore scientifico della biblioteca delle Facoltà di Economia e Commercio e Giurisprudenza dell'Università di Salerno sullo stato attuale e sulle prospettive di ordinamento e sviluppo della biblioteca stessa, busta 4, fascicolo 25, Fondo Alfonso Menna, Archivio di Stato, Salerno. La documentazione non presenta cartulazione.

un arco di tre anni, centralizzando i volumi e collegando il magazzino librario con la sala lettura, fino a creare un sistema di cataloghi ben strutturato; e una a breve termine per la quale prevede sia una schedatura provvisoria sia la creazione di una sala di consultazione per facilitare l'accesso alle opere fondamentali.

In entrambe le soluzioni, è cruciale il potenziamento del personale e una riorganizzazione sistematica per evitare che la situazione peggiori ulteriormente.

Interessante, senza dubbio, è la visione di Attilio Mauro Caproni, allora «incaricato di Bibliografia e Biblioteconomia alla facoltà di Magistero e funzionario direttivo presso la Biblioteca nazionale Centrale di Roma» per un progetto operativo di sistema bibliotecario dell'Università di Salerno.

La relazione – presentata il 28 aprile 1976 in occasione della riunione della Commissione incaricata dal Consiglio di Facoltà congiunto di Lettere e Filosofia e Magistero[20] – si concentra sulla necessità di costruire una struttura adatta al contesto socioculturale dell'istituzione. Si riconosce che, sebbene sia utile ispirarsi a modelli nazionali ed esteri, è importante adattare tali esperienze alle specifiche esigenze locali.

Il sistema proposto da Caproni, infatti, prevede una Biblioteca Universitaria Centrale che operi in sinergia con le Biblioteche di Istituto, assicurando sia un servizio capillare che una centralizzazione delle risorse documentarie. Il progetto affronta le problematiche legate alla gestione e organizzazione dei fondi documentari, agli spazi e agli organici necessari per un servizio efficiente e sottolinea, inoltre, la necessità di una chiara distribuzione delle responsabilità e di una stretta collaborazione tra personale bibliotecario, docenti e studenti.

Prevede, poi, un approccio modulare per la crescita del sistema, per rispondere progressivamente alle esigenze emergenti dell'università, con l'implementazione di tecnologie di comunicazione avanzate e un piano di acquisti documentari ben programmato al fine di evitare la duplicazione di risorse e garantire un accesso efficiente ai materiali e tenendo conto della crescita della documentazione e delle esigenze specifiche di ciascun istituto, con attenzione alla qualità e all'obsolescenza del materiale. È prevista l'implementazione di tecniche di comunicazione avanzate tra la biblioteca e gli istituti, nonché con altre istituzioni culturali e di ricerca, per favorire lo scambio di informazioni e il coordinamento. Si propone una chiara definizione delle responsabilità tra bibliotecari, docenti e amministrazione, promuovendo la corresponsabilizzazione per il corretto funzionamento del sistema. Viene, infine, sottolineata l'importanza della formazione degli utenti, soprattutto degli studenti, sull'uso degli strumenti bibliografici, e la creazione di servizi didattici e informativi integrati.

La sola stesura del progetto prevede per Caproni:

[20] La Commissione era così composta: Pier Fausto Palumbo, ordinario di Storia medievale presso la Facoltà di Magistero; Italo Gallo, incaricato di Papirologia presso la Facoltà di Lettere e Filosofia e Attilio Mauro Caproni.

> non meno di un anno di lavoro ed un impegno personale così distribuito:
> – due esperti bibliotecnici a metà tempo per l'intero anno, incaricati oltre che della corretta definizione del sistema e delle singole procedure, anche del collegamento tra i diversi centri studi;
> – un architetto per tre mesi a metà tempo per la definizione urbanistica e progettuale del sistema, con particolare riguardo alla sua modularità;
> – un ingegnere per tre mesi a metà tempo, esperto nei problemi di comunicazione, per la definizione del sistema informativo globale e delle tecniche di comunicazione;
> – uno esperto in procedure amministrative per tre mesi a metà tempo per la definizione di una prima bozza dell'intera normativa riguardante il sistema;
> – un esecutivo per un anno a metà tempo per compiti di coordinamento materiale e di stesura dattilografica.
> Solo un tale impegno di persone potrà permettere una effettiva stesura di un progetto realmente operativo, tenuto anche conto che al gruppo di lavoro si dovrà affidare anche il collegamento con i responsabili amministrativi dell'iniziativa della nuova struttura universitaria per i dati di opportuna conoscenza, lo studio sulla effettiva realtà socio-culturale intorno alla quale la nuova struttura graviterà, ed infine la stessa indicazione di tutti gli standards minimi ed una minuta analisi dei costi[21].

Un altro documento, sempre a firma di Augusto Placanica, evidenzia i problemi e le prospettive legati all'organizzazione della futura biblioteca.

In questa relazione si esplora il "dilemma" tra una struttura bibliotecaria centralizzata e una decentralizzata, evidenziando i rischi di dispersione e inefficienza associati a quest'ultima. Si propone una biblioteca centralizzata che garantisca omogeneità nei servizi, aggiornamenti tempestivi e una gestione efficace dei fondi. Tuttavia, si suggerisce anche di avere sale di consultazione specializzate nei dipartimenti per facilitare l'accesso alle risorse.

Infine, si invita a un confronto costruttivo tra tutte le parti interessate per affrontare le sfide legate alla creazione di un sistema bibliotecario efficace, evidenziando l'importanza di una progettazione organica e ben strutturata.

Placanica ha sostenuto con fermezza la necessità di una scelta centralizzata, affermando che

> la biblioteca, essendo strumento tecnico richiedente alti livelli di specializzazione nelle strutture e nel personale, va invece centralizzata al fine di conferire omogeneità nella tenuta e collocazione del patrimonio librario, nella scrupolosa vigilanza sul suo uso, nella spedizione degli acquisti, nelle operazioni contabili relative; analogamente omogenei dovranno essere i servizi forniti consultazione, lettura, prestito interno, esterno ed internazionale, omogenei e unitari i sistemi di catalogazione, classificazione, soggettazione, decimalizzazione, ecc.

[21] Caproni 1976. Per un progetto di sistema bibliotecario dell'Università di Salerno, busta 4, fascicolo 25, Fondo Alfonso Menna, Archivio di Stato, Salerno. La documentazione non presenta cartulazione.

> È evidente che una struttura bibliotecaria centralizzata dotata di attrezzature e personale i più nutriti possibile può rispondere ad esigenze molteplici, soprattutto se si prevede una biblioteca di tipo "aperto".

In accordo con la relazione del professor Pier Luigi Spadolini, rappresentante del Ministero della Pubblica Istruzione, sia i consigli di facoltà sia le commissioni di coordinamento hanno accettato l'idea di una biblioteca centralizzata, pur mantenendo attivi i nuclei dipartimentali. Tuttavia, il processo di realizzazione del sistema bibliotecario universitario fu poi sospeso, subendo ritardi significativi.

Per guidare la transizione verso un sistema unificato e stabilire le basi della centralizzazione, è stato istituito nel 1982 un Ufficio Centrale delle Biblioteche (UCB), che si è occupato di riorganizzare l'attività delle tre biblioteche di facoltà sia a livello amministrativo che catalografico. Successivamente, l'UCB ha ceduto il posto a due entità bibliotecarie autonome: una interdipartimentale per la Facoltà di Scienze a Baronissi, ufficialmente istituita con un decreto rettorale nel 1992, e un Centro di Servizio di Ateneo per le Biblioteche, situato nel campus di Fisciano, creato circa cinque anni dopo a causa dei ritardi nella costruzione della nuova biblioteca centrale.

La storia più recente ha visto nascere il Centro Bibliotecario di Ateneo (decreto rettorale n. 3735 del 30 dicembre 2010), con la disattivazione di due organismi precedenti, completando così l'organizzazione del Sistema Bibliotecario di Ateneo, il cui Regolamento è stato approvato dal decreto rettorale n. 3701 del 24 dicembre 2010. Attualmente, il Centro Bibliotecario di Ateneo dell'Università di Salerno è composto da due biblioteche centrali: la Biblioteca Centrale del Polo Umanistico "E. R. Caianiello", dedicata alle scienze umane e la Biblioteca del Polo Scientifico e Tecnologico, entrambe situate nel campus di Fisciano. A queste si aggiunge la biblioteca di Medicina e Chirurgia, localizzata nel campus di Baronissi[22].

4. Conclusioni

I documenti relativi all'Università di Salerno presenti nel Fondo Menna rappresentano un elemento significativo per la comprensione e la valorizzazione della storia dell'ateneo. Queste carte, testimonianze dirette di una fase cruciale dello sviluppo universitario, offrono una finestra unica su decenni di evoluzione accademica e culturale, e mettono in luce la visione e le strategie che hanno guidato l'istituzione verso la sua affermazione nel panorama universitario nazionale.

---

[22] Per una breve ricostruzione della storia si rimanda al sito del Centro bibliotecario di Ateneo dell'Università degli studi di Salerno (n.d.).

Il Fondo Menna, con le sue fonti inedite, permette non solo di ricostruire in modo più dettagliato le vicende legate alla formazione del sistema bibliotecario e delle infrastrutture culturali dell'Università di Salerno, ma anche di approfondire il ruolo svolto da figure chiave nella crescita dell'ateneo. Queste carte rappresentano un tassello fondamentale per lo studio delle dinamiche interne e dei rapporti con il territorio, evidenziando l'impatto dell'università nello sviluppo sociale e culturale della comunità di appartenenza

Gli sforzi di grandi personalità diventano una fonte di ispirazione per la comunità accademica odierna e contribuiscono a rafforzare l'identità storica dell'Università di Salerno. In un momento di celebrazione come questo, recuperare e studiare le tracce del passato assume un valore simbolico che sottolinea l'importanza della memoria collettiva come strumento di crescita e continuità[23].

La storia del Sistema Bibliotecario dell'Università di Salerno rappresenta un capitolo essenziale nello sviluppo del sistema accademico e culturale dell'ateneo. La sua evoluzione, scandita da momenti chiave come l'istituzione del CBA e la centralizzazione delle risorse, ha segnato un passaggio cruciale verso una gestione più efficiente e organizzata del patrimonio documentale. Questi progressi non solo hanno migliorato l'accesso e la fruizione delle risorse per studenti e docenti, ma hanno anche rafforzato il ruolo della biblioteca come centro di diffusione della conoscenza.

L'analisi di documenti storici inediti legati alla nascita e allo sviluppo del Centro Bibliotecario di Ateneo assume un'importanza centrale per comprendere le scelte strategiche fatte nel corso degli anni e per preservare la memoria istituzionale. Si rende possibile così restituire una visione chiara del processo decisionale e delle sfide affrontate nella costruzione di un sistema bibliotecario moderno e funzionale.

Riferimenti bibliografici

Andria, Marcello. 2008a. "Biblioteche a Salerno fra Otto e Novecento: Spunti per un'indagine." In *Storia di Salerno*, vol. 3, *Salerno in età contemporanea*, a cura di Giuseppe Cacciatore e Luigi Rossi. Salerno: Sellino.

[23] Il 16 ottobre 2024 hanno preso avvio le celebrazioni per l'ottantesimo anniversario dalla nascita dell'Ateneo salernitano. Con la mostra "UNISA 80. 80 anni di storia, secoli di futuro", organizzata presso la Biblioteca del Polo tecnico-scientifico di Ateneo, si è voluto dar via a un percorso fotografico e documentale che raccontasse la crescita culturale e strutturale dell'Università degli Studi di Salerno nel corso degli anni. Il Sistema Bibliotecario di Ateneo, inoltre, ha dedicato un percorso bibliografico digitale attraverso cui è possibile consultare i primi annuari, gli statuti, i notiziari dell'epoca, con riferimenti sulle principali attività della vita accademica e istituzionale dell'ateneo, insieme a pubblicazioni che tracciano la storia dell'università salernitana (Università degli studi di Salerno 2024, n.d.c.).

Andria, Marcello. 2008b. "Libri e letture di un intellettuale salernitano: La biblioteca privata di Giovanni Cuomo." In *Giovanni Cuomo: Una vita per Salerno e il Mezzogiorno*, a cura di Vittoria Bonani. Angri: Gaia.

Andria, Marcello. 2017. "Biblioteca di biblioteche: Fondi privati, donazioni e collezioni speciali nella Biblioteca centrale dell'Università." *Rassegna storica salernitana* 68: 9-31.

Andria, Marcello. 2020. "La lunga durata di una eredità culturale: Tratti di storia recente dell'Università di Salerno." In *Opulenta Salernum. Una città tra mito e storia*, a cura di Giovanni Di Domenico, Angela Pontradolfo e Maria Galante. Roma: Gangemi.

Barrera, Giulia. 2006. "Gli archivi di persone." In *Storia d'Italia nel secolo ventesimo: Strumenti e fonti*, a cura di Claudio Pavone. Roma: Ministero per i beni e le attività culturali. Dip.to per i beni archivistici e librari. Direz. gen. per gli archivi.

Bonani, Vittoria. 2007. *Giovanni Cuomo e il suo tempo: 1943-1948*. Angri: Gaia.

Bonani, Vittoria. 2008a. *Giovanni Cuomo: Una vita per Salerno e il Mezzogiorno: Atti del convegno nazionale di studi. Salerno, 12-14 dicembre 2007*. Angri: Gaia.

Bonani, Vittoria. 2008b. *Giovanni Cuomo: La vita, gli affetti e l'impegno politico*. Angri: Gaia.

Centro Bibliotecario di Ateneo dell'Università degli studi di Salerno. n.d. "Il Centro: Storia." Consultato il 20 settembre 2024. https://www.biblioteche.unisa.it/centro/storia.

Commissione nazionale biblioteche speciali, archivi e biblioteche d'autore dell'Associazione italiana biblioteche. 2019. "Linee guida sul trattamento dei fondi personali." Associazione Italiana Biblioteche. https://www.aib.it/documenti/linee-guida-sul-trattamento-dei-fondi-personali/.

D'Addario, Arnaldo. 1972. *Lezioni di archivistica*. Bari: Adriatica.

D'Aniello, Ennio. 1994. *Ricordo di Giovanni Cuomo (nel 50° anniversario di "Salerno Capitale")*. Salerno: Laveglia.

D'Auria, Elio. 1985. "Giovanni Cuomo." In *Dizionario biografico degli italiani*, vol. 31. Roma: Istituto dell'Enciclopedia Italiana.

De Renzi, Salvatore. 1857. *Storia documentata della Scuola medica salernitana*. Napoli: G. Nobile.

Di Domenico, Giovanni, e Fiammetta Sabba, a cura di. 2020. *Il privilegio della parola scritta: Gestione, conservazione e valorizzazione di carte e libri di persona*. Roma: Associazione Italiana Biblioteche.

Galdi, Amalia. 2020. "La Scuola medica salernitana nel Medioevo: Un'istituzione mediterranea tra storia e leggenda." In *Opulenta Salernum. Una città tra mito e storia*, a cura di Giovanni Di Domenico, Angela Pontradolfo e Maria Galante. Roma: Gangemi.

Galasso, Giuseppe. 1995. "Una scuola e un mito: La Scuola medica di Salerno." *Rassegna storica salernitana* 24: 7-30.

Gallo, Italo, a cura di. 1994. *Salerno e la sua Scuola medica*. Salerno: Arti Grafiche Boccia.

Istituto per lo Sviluppo Economico dell'Italia Meridionale. 1959. *Crediti ed agevolazioni per l'industrializzazione del Mezzogiorno continentale*, a cura di Vittorio Cascetta. Napoli: Arti grafiche SAV.

Jacquart, Daniele, e Agostino Paravicini Bagliani, a cura di. 2007. *La scuola medica salernitana: Gli autori e i testi*. *Convegno internazionale, Università degli Studi di Salerno, 3-5 novembre 2004*. Firenze: SISMEL-Edizioni del Galluzzo.

Kristeller, Oskar Paul. 1986. *Studi sulla Scuola medica salernitana*. Napoli: Istituto Italiano per gli Studi Filosofici.

Leone, Alfonso, e Gerardo Sangermano, a cura di. 2003. *La "Schola Salernitana" e le altre: Atti della giornata di studio (Salerno 1 giugno 2002)*. Salerno: Civitas Hippocratica.

Manzo, Pio. 2021. "Periodo costituzionale transitorio: Testimonianze bibliografiche nella raccolta di Giovanni Cuomo." In *Scaffali come segmenti di storia: Studi in onore di Vincenzo Trombetta*, a cura di Rosa Parlavecchia e Paola Zito. Roma: Quasar.

Menna, Alfonso. 1930. *Per la elevazione di Battipaglia a comune*. Salerno: Tip. L. e A. Lauretano.

Menna, Alfonso. 1931. *Intorno alle origini di Battipaglia*. Salerno: Stamp. R. Beraglia.

Menna, Alfonso. 1932. *Il nuovo comune di Battipaglia*. Salerno: Tip. F.lli di Giacomo.

Menna, Alfonso. 1971. *Intervento di Alfonso Menna al Consiglio Comunale del 21 dicembre 1970: (Programma e sede universitaria)*. Salerno: Graficart di Giovanni Di Giacomo.

Menna, Alfonso. 1991. *Il banco e la cattedra: Dalle materne all'Università*. Salerno: De Luca Editore.

Menna, Alfonso. 1993. *Come li ricordo: Sono uomini che hanno ben meritato della civica riconoscenza*, vol. I. Salerno: De Luca Editore.

Menna, Alfonso. 1996. *Come li ricordo: Sono uomini che hanno ben meritato della civica riconoscenza*, vol. II. Salerno: De Luca Editore.

Musi, Aurelio, Massimo Oldoni, e Augusto Placanica, a cura di. 2001-2004. *Storia dell'Università di Salerno*, Voll. I-II. Salerno: Arti Grafiche Boccia.

Musi, Aurelio. 2004. "Introduzione." In *Storia dell'Università di Salerno: L'età contemporanea (1944-2004)*, vol. 2, a cura di Aurelio Musi. Salerno: Arti Grafiche Boccia.

Parlavecchia, Rosa. 2020. "La Biblioteca provinciale di Salerno." In *Opulenta Salernum: Una città tra mito e storia*, a cura di Giovanni Di Domenico, Angela Pontradolfo e Maria Galante. Roma: Gangemi.

Petrucciani, Alberto. 2005. "Placanica, Augusto." In *Dizionario bio-bibliografico dei bibliotecari italiani del XX secolo*, a cura di Simonetta Buttò e Alberto Petrucciani. https://www.aib.it/aib/editoria/dbbi20/placanica.htm.

Presidenza della Repubblica italiana. 2021. "Menna Alfonso: Medaglia d'oro al merito civile." https://www.quirinale.it/onorificenze/insigniti/526.

Sbrescia, Vincenzo Mario. 2022. "Tecnica amministrativa e visione politica al servizio del bene pubblico: Alfonso Menna, Sindaco riformista della Città di Salerno, lungimirante Presidente Isveimer, brillante saggista." *Rivista giuridica del mezzogiorno* 36 (3): 793-97.

Università degli studi di Salerno. n.d.a. "Liberabit." Consultato il 16 settembre 2024. https://www.liberabit.unisa.it/.

Università degli studi di Salerno. n.d.b. "Fondo Cuomo." Liberabit. Consultato il 16 settembre 2024. https://www.liberabit.unisa.it/cris/fonds/fonds09100.

Università degli studi di Salerno. n.d.c. "UNISA 80: 80 anni di storia, secoli di futuro." Liberabit. Consultato il 30 ottobre 2024. https://www.liberabit.unisa.it/cris/path/path10105.

Vitolo, Giovanni. 2007. *La Scuola medica salernitana come metafora della storia del Mezzogiorno*. Firenze: SISMEL Edizioni del Galluzzo.

Zaccaria, Raffaella Maria, e Salvatore Amato. 2018. "Introduzione." In *L'Archivio storico dell'Università degli studi di Salerno: Inventario*, a cura di Raffaella Maria Zaccaria. Soveria Mannelli: Rubbettino.

AIDAinformazioni
ISSN 1121–0095
ISBN 979-12-5965-456-4
DOI 10.57574/596545646
pag. 119-148 (luglio-dicembre 2024)

CompL-it: a Computational Lexicon of Italian

Flavia Sciolette, Andrea Bellandi, Emiliano Giovannetti, Simone Marchi*

Abstract: This paper describes CompL-it, a new open computational lexicon for contemporary Italian. The resource was constructed from three sources: an already available Italian lexicon, a lemmatized list of inflected forms obtained from a morphological analyser, and a set of treebanks. Integrating these resources required a standardisation process in accordance with the standards of the Linguistic Linked Open Data community, which was necessary for the subsequent conversion into the OntoLex-Lemon model. The resulting computational lexicon comprises approximately 100,000 lexical entries, 790,000 forms, 57,000 senses, and 86,000 semantic relations. The lexicon, thanks to its rich and articulated linguistic structure, can be used, as shown, to enhance information retrieval in the context of full-text search tasks.

Keywords: Computational Lexicon, Linguistic Resources, Linguistic Linked Open Data, OntoLex-Lemon, Information Retrieval.

1. Introduction

While a significant number of digital lexical resources are available for many languages (such as various multilingual WordNets) (Princeton University n.d.; MultiWordNet n.d.; Global WordNet Association n.d.), only a few integrate different layers of linguistic information, such as morphology, semantics, and syntax. This is not surprising, as the construction of a computational lexicon[1] that conveys linguistic information across different layers can be an extremely time-consuming task that requires advanced linguistic expertise.

* CNR-Istituto di Linguistica Computazionale (ILC) "A. Zampolli", Pisa, Italy. flavia.sciolette@ilc.cnr.it; andrea.bellandi@ilc.cnr.it; emiliano.giovannetti@ilc.cnr.it; simone.marchi@ilc.cnr.it.

[1] In this context, a computational lexicon can be defined as a resource that contains information about words, their meanings, and linguistic properties, designed to be used by computer systems for tasks like natural language processing (NLP), machine translation, or text analysis. It typically includes details such as word categories (e.g., noun, verb), syntactic information, and semantic relationships.

On the other hand, the need for resources of this kind is long-standing. Italian linguistics, for example, has always shown an interest for lexical data (Sabatini 2006), which has been encouraged by the increasing availability of many corpus-based resources. As documented in Chiari (2012) many projects involving corpora (monolingual, parallel, domain-specific) have flourished and both digitised traditional dictionaries and computational dictionaries have taken advantage of them, for example to calculate the frequency of words or to increase their lexical coverage (for example by adding neologisms). In terms of exploitation, a number of applications are meant to take advantage of lexical resources, such as sentiment analysis (Prakash and Aloysius 2021), and «semantic role labelling, verb sense disambiguation, and ontology mapping» (Brown et al. 2022, 2).

In the context of archival science and document management, the availability of linguistic resources to support the organisation and retrieval of information has been considered crucial for many years (Chen et al. 1995; Smith 1997; Thompson et al. 2011). In these fields, the development of computational lexicons can provide a fundamental contribution to the community, expanding the potential for knowledge analysis and management. It is believed that a resource capable of formalising a language's lexical and semantic structures in a complex way can improve the efficiency of archiving, classification, and information retrieval activities, within a document management paradigm increasingly supported by IT tools (Bamman and Crane 2010; Hmeidi et al. 2016; Passarotti and Mambrini 2021). The creation of increasingly efficient tools for automating archival and document practices can greatly simplify the management of large volumes of unstructured data, enhancing precision in indexing and retrieving information. Furthermore, a computational lexicon can serve as a key linguistic resource for building Knowledge Organization Systems (KOSs), such as ontologies and thesauri, crucial elements for knowledge organisation (Hodge 2000; Shiri 2015).

In this work, we illustrate CompL-it, an Italian computational lexicon built by leveraging existing resources, whose data have been thoroughly analysed, extracted, converted, and interconnected. CompL-it has been made freely available as Linguistic Linked Open Data (LLOD) on the CLARIN repository (CLARIN-IT n.d.a).

2. State of the art

In order to define the state of the art regarding computational lexical resources for the Italian language we first conducted a search on the Virtual Language Observatory (CLARIN VLO n.d.) (VLO) of the European infrastructure CLARIN (Common Language Resources and Technology Infrastructure). This database, which contains hundreds of thousands of references

to language resources and tools, was browsed using the available filters. In particular, a search was carried out by type of resource, selecting *lexicalResource*, and specifying the Italian language. In addition, from the obtained list we excluded automatically produced resources (i.e, not revised by hand), lists of idiomatic expressions, lists of terms with no linguistic information at all, multilingual named entities, sets of embeddings, parallel corpora, metadata, resources that only appear by virtue of some references to the Italian language and, finally, all the resources whose data are not open and freely available. Among this last type of resources, however, it is worth mentioning BabelNet (Navigli and Ponzetto 2012; BabelNet n.d.), Senso Comune (Vetere et al. 2011), and Italian FrameNet (Basili et al. 2017), particularly for the richness of data they offer.

The resources identified on CLARIN VLO that met the above criteria were 14, and can be classified into two categories: 11 lexical resources and 3 terminological resources.

The available lexical resource that, for this work, has been taken as the main reference (and used as one of the sources) is LexicO (Sciolette, Giovannetti, and Marchi 2023), a multi-layered computational lexicon developed at CNR-ILC and built from Parole-Simple-Clips (PSC) (Bel et al. 2000; Ruimy et al. 2002; ILC4CLARIN CNR 2016). More details on the nature of LexicO will be provided later in section 3.1.

Another very rich lexical resource, also developed at CNR-ILC, is ItalWordNet (Roventini et al. 2003), available as an SQL dump on the VLO in its second version (Roventini, Marinelli, and Bertagna 2016). The VLO also mentions MultiWordNet (Pianta, Bentivogli, and Girardi 2002; MultiWordNet n.d.), realised as an extension of Princeton's WordNet (Miller 1995; Princeton University n.d.), and which also includes data for the Italian language. Multilingual language resources include OmegaWiki (Meijssen 2014), an open and collaborative resource whose aim is «to describe all words of all languages with definitions in all languages» and includes lexical, terminological and ontological information (WikiMedia 2022).

In addition to the resources listed so far, which are structured as lexicons, other resources are available for Italian that convey individual layers of linguistic information. Building on the aforementioned PSC and ItalWordNet, a resource that provides semantic data called the Italian Sense Inventory was created. This resource was developed within the ELEXIS project (ELEXIS n.d.) to support Word Sense Disambiguation tasks.

On the VLO, there are also two resources developed by the same author that complement each other: Italian Function Words (Grella 2018a) and Italian Content Words (Grella 2018b). The former, as the name suggests, contains Italian function words and is designed to support tasks such as POS tagging and syntactic parsing. The second constitutes a morphological dictionary of

over 2 million inflected forms that, however, includes hundreds of thousands of forms that, although morphologically correct, are not represented in linguistic usage. The last two lexical resources we include in this review are Universal Derivations (Kyjánek et al. 2021) and Universal Segmentations (Žabokrtský et al. 2022), both multilingual, in which about 10 thousand Italian lemmas are linked to their respective segmentations, derived forms and compounds.

With regard to the three terminological resources identified in the VLO we first mention Geodomain WordNet, a collection of geographical terms linked to the English and Italian WordNets (Frontini, Del Gratta, and Monachini 2016). The other two resources, developed within the Pan-Latin Terminology Network (Realiter n.d.), are the Pan-Latin Lexicon of Collars and Sleeves in Fashion and Costume (Zanola et al. 2023) and the Pan-Latin Textile Fibres Vocabulary (Dankova, Zanola, and Calvi 2022).

Although not available on CLARIN VLO, Morph-it! (Zanchetta and Baroni 2005) is a freely accessible and rich morphological resource for Italian, consisting of 504,906 inflected forms and 34,968 lemmas. However, as the authors note (Morph-it! 2018), because it is derived from an Italian newspaper corpus, the resource has «many gaps in basic, every-day vocabulary».

Another interesting resource worth mentioning is SimpleLEX-IT, as it was built similarly to CompL-it, i.e., by combining together different existing resources (Mazzei 2016; SimpleLEX-IT n.d.). In particular, SIMPLELex-it was developed by integrating morphological data from the previously cited Morph-it!, the *Vocabolario di base della lingua italiana* by Tullio De Mauro (De Mauro 1980; 2016), two entries of the Italian Wikipedia concerning verbs (Wikipedia 2024a; 2024b) and, finally, data from the Italian Universal Dependencies (UD) treebanks (Universal Dependencies n.d.a).

In the context of Linked Open Data – or, more precisely, LLOD, understood as the reference community for the creation and sharing of resources according to LOD principles (Cimiano et al. 2020; LLOD n.d.) – the linguistic resources currently available for Italian include RDF datasets for the previously mentioned PSC (Del Gratta et al. 2015) and IWN (Bartolini 2016) resources. The LLOD landscape, however, offers resources for different languages, both contemporary and historical varieties; as an illustrative and non-exhaustive example in a constantly expanding field, it is worth mentioning Dbnary (Sérasset 2015), the multilingual resource based on Wiktionary, made available according to LLOD principles. For historical varieties, we cite LiLa – Linking Latin, a knowledge base for Latin that now includes several resources (Mambrini and Passarotti 2023), and the DigitAnt project for ancient language varieties in Italy (Mallia et al. 2024). On the terminological front, CHAMUÇA is noted, a resource for Portuguese loanwords in Asian languages (Khan et al. 2024), and initial studies for a resource related to terms in the Babylonian Talmud (Sciolette 2024), with the formalisation of contex-

ts through the OntoLex module FrAC (Frequency, Attestation and Corpus Information), which is currently under development (Chiarcos et al. 2022; Github n.d.a).

As a source of reference data for the construction of the CompL-it lexicon, and as already mentioned, we chose LexicO, one of the freely available lexical resources for the Italian language. The motivation for this choice is twofold, and is partly also evident from the data reported in section 4.3, where LexicO has been quantitatively compared to five other resources. First of all, LexicO (see section 3.1) is a multi-layered linguistic resource, in which information of various kinds (phonological, morphological, syntactic and semantic) is encoded: in this sense, it constitutes a *unicum* of its kind, given that all the others limit themselves to representing, essentially, lexemes linked to each other through semantic relations. Moreover, from a more quantitative point of view, LexicO with its dense network of relations constitutes an extremely rich resource of linguistic data.

However, LexicO also has limitations, both in terms of coverage in the number of lexical entries, in terms of specific content (such as inflected forms or missing lexical senses) and, finally, in terms of the data format in which it is currently represented.

The idea of building CompL-it arose precisely due to these limitations of LexicO: to ensure maximum lexical coverage, the linguistic data from LexicO was integrated with data from two additional sources. Furthermore, standards defined by the LLOD community were adopted for the model and representation format.

3. The sources

As stated in the previous section, LexicO was selected as the foundational resource for constructing CompL-it. Additionally, we considered two other sources: M-GLF (MAGIC-Generated Lemmatized Forms), a list of lemmatized forms with morphological information generated by the MAGIC tool (Battista and Pirrelli 1999; Pirrelli and Battista 2000), and a set of Italian language treebanks available through the UD repository (Universal Dependencies n.d.b). All these resources have been chosen both for the richness of the data they provide and because they have been manually constructed or validated.

These three resources are very different from each other in terms of formats, models and purposes, and therefore their integration required a process of standardisation, as described in Section 4.1. In the following sections, we describe the three resources together with some specific pre-processing interventions carried out prior to the data standardisation and conversion steps necessary to create CompL-it.

3.1. LexicO

LexicO is a computational lexicon of Italian, available on CLARIN as a relational database (CLARIN-IT n.d.b). This resource is derived from the above mentioned PSC, with which it shares the same model based on the theory of Generative Lexicon by James Pustejovsky (Pustejovsky 1995).

LexicO contains four layers of linguistic information: a morphological layer, which describes lemmas, parts of speech (POS), and inflectional rules; a semantic layer, which includes information about senses and their relationships; a syntactic layer, detailing the syntactic behaviour of units and their phrase structure; and finally, a phonological layer, which involves inflected forms generated from the inflectional rules in the morphological layer. Although each layer operates independently, there are connections between different units, such as between syntactic and semantic entries. Since its initial use in tasks such as full-text search (Giovannetti et al. 2022), it has become evident that there is a need to convert all the data into a format compliant with current standards.

Morphological units form the basis of lexical entries in LexicO. Each unit is associated with a POS value and a set of morphological rules used to generate grammatically correct forms. These forms are defined as a type of entry called phonological units.

Each association between a lemma and a form is described with a POS and a certain number of morphological traits, as shown in Table 1.

Lemma	Form	POS	MorphFeat
abbandonare	abbandonai	V	1|singular|indicative|past

Table 1: *abbandonare* (to abandon) with its form *abbandonai*, its POS (verb), and its morphological features (first person singular, indicative, past).

As already mentioned in the previous section, LexicO is directly derived from the PSC computational lexicon. This initial resource, while already quite comprehensive, contained redundant or duplicated data and some entries lacking in information: an emblematic example is the absence of the form *vado* (I go) of the verb *andare* (to go). Although these issues did not diminish the intrinsic value of the source, they required interventions to address the gaps where possible. All interventions are documented in Sciolette, Giovannetti, and Marchi (2023).

3.2. M-GLF

The second lexical source we used to build up CompL-it was M-GLF, a list of lemmatised forms generated by MAGIC, a morphological analyser for

Italian (Battista and Pirrelli 1999). The tool includes three modules: a lexicon compiler, the morphological analyser itself, and a morphological generator. We used this latter to generate the M-GLF list of forms (CLARIN-IT n.d.c) by starting from a list of morphological rules for lemmas, endings and idiosyncratic entries, contained in a morphological database.

An example of a M-GLF entry follows which is relative to a form of the Italian verb *abbaiare* (to bark):

```
[1]          MACRO:word[l_abbaiare,abbaiera',v_fin,3,!,s,-
fut,ind,!,!]
```

In this example, *l_abbaiare* is the lemma of the form *abbaierà*, which is the third person singular of the finite verb *abbaiare* in the indicative mood and future tense. These morphological traits are indicated in the line, separated by commas. Exclamation points represent *null* values for unspecified features, such as degree, which is only relevant for adjectives.

The MAGIC generation tool is based on rules that constitute an extremely rigorous model. First of all, the tool was unable to generate certain forms, such as the absolute superlative. Moreover, the generation of entries produced some inconsistencies. In particular, we found entries having multiple POS, such as *noun* and *adjective* (e.g., *svedese* can indicate both the noun for a resident of Sweden and the adjective for denoting the quality of being Swedish). In these cases, we decided to intervene by splitting the entries with double POS into distinct entries, each of which having its own POS with the correct morphological traits.

3.3. Treebanks

To further enrich the morphological layer of CompL-it, we also decided to consider lemmas, forms, and morphological information obtained from the available treebanks for Italian. The treebanks are collected in the UD repository, according to a common annotation scheme (Universal Dependencies n.d.c), used for resources in different languages.

We only included treebanks that have been manually revised: three based on balanced corpora of general-purpose texts (such as newspapers, legal documents, etc.) and one from a specific domain. We considered the following treebanks:

- ISDT (Italian Stanford Dependency Treebank) (Universal Dependencies n.d.d): this resource was obtained through a semi-automatic conversion process starting from MIDT (the Merged Italian Dependency Treebank). It is the result of merging pre-existing dependency-based resources, aimed at improving the interoperability of available data.

The schema was partially adapted to account for the specific features of the Italian language (Simi, Bosco, and Montemagni 2014);

- VIT (Universal Dependencies n.d.e): it is a conversion of VIT (Venice Italian Treebank), developed at the Laboratory of Computational Linguistics at Università Ca' Foscari in Venice. Originally a constituency-based treebank, VIT includes linguistic materials of various types, extracted from five text typologies and spoken dialogues. The data underwent conversion to the CoNLL-U format (Universal Dependencies n.d.f), along with several stages of data harmonisation;
- ParTUT (ParallelTut) (Universal Dependencies n.d.g): it is a conversion of a multilingual parallel treebank developed at the University of Turin consisting of a variety of text genres, including talks, legal texts and Wikipedia articles (Sanguinetti and Bosco 2015);
- ParlaMint-It (Universal Dependencies n.d.h): it is a collection of transcriptions of parliamentary sessions of the Italian Senate, annotated in Universal Dependencies. The corpus is part of a larger multilingual collection of parliamentary transcripts built during the ParlaMint project (CLARIN n.d.).

Although the selected treebanks were chosen precisely because they underwent manual revision, they are not entirely free of errors, including gaps and inconsistencies, particularly in morphological features, which can introduce noise.

For example, there are cases where a word appearing in the treebanks is annotated with fewer morphological features than the same word appearing in the other two resources. This is the case for the form *abilitati* (enabled, as in "enabled users"), described in the treebanks only through lemma and POS, while in LexicO and M-GLF, this word is also provided with number (plural) and gender (masculine) features. In all these cases, in CompL-it, the words from the resource richer in linguistic information have been added.

4. The nature of CompL-it

This section illustrates the resource CompL-it, by starting with the necessary standardisation process that had to be carried out, described in Section 4.1. The conversion in RDF format is briefly described in Section 4.2, while a quantitative analysis of the resource is carried out in Section 4.3. To ensure data interoperability we chose Ontolex-Lemon as the backbone model of Compl-It, as it is the *de facto* standard for representing lexical resources in the Linked Open Data community.

In order to proceed to the standardisation and conversion processes, it was necessary to carry out some pre-processing steps. Some of these interventions concerned all the resources and involved, in particular: i) the conversion of superscripts into accented letters and distinction of high and low accents; ii) the removal of proper nouns, such as named entities (e.g. *Petrarca*) and trade names (e.g. *Xerox*); iii) the exclusion of abbreviations (e.g. *Dott*. instead of *doctor*); iv) the exclusion of multiword expressions, which included nouns, adjectives, adverbs and prepositions (e.g. the expression *a ferro e fuoco*); v) the removal of unadapted loanwords (e.g. word processor).

Loanwords, multi words and proper nouns require different treatment and will therefore be the subject of future work.

4.1. Standardisation

The following paragraphs describe the interventions undertaken to make the models of the considered resources homogeneous in terms of morphology and semantic relations.

4.1.1 Morphology

As can be seen from Section 3, the models and reference vocabularies of LexicO, M-GLF and treebanks differ from each other, often profoundly. This divergence between linguistic information representation systems is also motivated by the different approach used to represent linguistic data.

In fact, M-GLF and LexicO can be included in the category of lexicographic resources, whereas the standard used for treebanks, based on the UD paradigm, pertains to the annotation of linguistic corpora.

In order to standardise the vocabularies, we decided to use LexInfo, an inventory of types, values and properties designed to describe linguistic data categories (LexInfo n.d.). LexInfo comprises morphological properties, such as gender, number, mood, grammatical categories (POS), and semantic relations, such as synonymy, hypernymy, and so on. This choice is justified primarily by the alignment of this vocabulary with the OntoLex-Lemon model (W3C 2016) used to represent CompL-it, as LexInfo serves as the reference linguistic ontology for resources created with this model. Additionally, LexInfo complies with other standards related to the OntoLex-Lemon model, including OLiA (Ontologies of Linguistic Annotation) (Chiarcos and Sukhareva 2015; OLiA n.d.), a repository of linguistic categories specific to annotated corpora. Ultimately, the selection of the LexInfo vocabulary was largely driven by the need to produce a lexical resource that is as interoperable as possible with other re-

sources based on the OntoLex-Lemon model, in accordance with the Linked Data paradigm.

This difference in the representation of linguistic information in the models of the three sources occurs mainly at the level of the association between POS and morphological traits. In fact, LexicO and M-GLF are based on a model and specific vocabularies of labels, which are characterised by a very fine-grained categorisation of POS. In the case of the treebanks, the UPOS has a coarser grain: for example, the combination of UPOS and trait "possessive" with the value "yes" in the treebanks has been mapped to a specific LexInfo POS. For example, the entry *mio* (mine) appears in treebanks with the following annotation: PRON for the UPOS, with the feature "Poss=Yes". In CompL-it, this entry has been described with POS "possessivePronoun", according to the LexInfo vocabulary. More in general, the vocabularies of LexicO, M-GLF and the treebanks were mapped into LexInfo, according to the following scenarios: i) direct mapping between POS, if available (as was often the case for LexicO and M-GLF); ii) conversion of POS and trait combinations present in the treebanks into a LexInfo POS; iii) conversion into OLiA[2] or proposal of an *ad hoc* label, if the trait was not present in LexInfo. The conversion tables have been made available on (Github n.d.b).

4.1.2 Semantics

In CompL-it, 137 types of semantic relations derived from LexicO have been included. These relations are categorised into eight classes, listed below.

- Four classes are related to the four *qualia* roles taken from the Generative Lexicon theory[3], namely:
  - Formal: the role that describes the entity conveyed by the sense in relation to other entities. An example of a relation associated with the formal role is hyponymy, e.g., *gatto-mammifero* (cat-mammal);
  - Agentive: the role that provides information about the origin of an entity. An example of a relation associated with the agentive role is *caused by*, e.g., *infezione-batterio* (infection-bacterium);
  - Telic: the role that specifies a function of an entity. An example of a relation associated with the telic role is *Object of activity*, linking an object to a certain event, such as *libro-leggere* (book-to read);

[2] OLiA has been used in the conversion of two traits in M-GLF for "Diminutive" and "Augmentative".

[3] For an overview of the theory and the relationship between qualia roles and relations, see (Sciolette, Giovannetti, and Marchi 2023). Following the terminology in the PSC documentation, entities refer to the concept expressed by the sense, conveyed by a specific entry. These entities can be connected to each other through semantic relations. Semantic relations are also classified according to qualia roles.

  - Constitutive: the role that describes the composition of an entity; an example of a relation associated with the constitutive role is meronymy, as *senatore-senato* (senator-senate).
- A derivational class, reserved for relations concerning senses that undergo a change in grammatical category, e.g., from adjective to noun, as *triste-tristezza* (sad-sadness).
- A class related to polysemy relations, as listed in Malmgren (1988). An example of a regular polysemy class is *Substance-Colour*, as seen in the sense of *turchese* (turquoise), which can refer to both the gemstone and the colour.
- Two classes not documented in the original PSC model, namely synonymy, e.g. *ciclone-uragano* (cyclone-hurricane), and metaphor, e.g. *leone* (lion) to relate the sense of a brave man to the sense of the animal.

These few examples convey the image of a system of relations aimed at defining meaning according to very fine-grained categories, as exemplified in the case of meronymy, which distinguishes senses related to parts of a set, components of a group, and, as a subclass, followers of a certain movement, as for example *Marxista* (Marxist).

At present, it has proved particularly complex to find exact correspondences between the semantic relations described in the reference ontologies (LexInfo and OLiA) for the OntoLex-Lemon model. In some cases, it was necessary to define relations from scratch.

The need to update lexical resources in Linked Data formats was also felt in the past and led to the creation of some resources conforming to previous versions of the OntoLex-Lemon model (Del Gratta et al. 2015; Villegas and Bel 2015). However, it was not possible to reuse these resources either because they did not include updates to the OntoLex-Lemon model, or because they did not include a mapping with other reference models, such as LexInfo.

To ensure maximum interoperability, where possible, relations formalised from scratch were linked to the corresponding reconstructed resources in previous versions of the OntoLex-Lemon model, with the seeAlso relation (W3C 2005).

For the construction of a vocabulary of CompL-it relations, the following were also considered: i) equivalences, where possible, with LexInfo, such as in the case of synonymy; ii) additional properties, not previously defined by other resources, but reconstructed through documentation and analysis of the resource.

The vocabulary definition phase also had a direct effect on the enrichment interventions of the resource. For example, the mapping with LexInfo made it possible to define an additional relation, hypernymy, as the inverse of hyponymy (which translates the *isA* relation present in the LexicO model); since hy-

pernymy is the inverse of hyponymy, even if the relation is not described in the source resource, it was still possible to infer a number of instances. This happened for all relations of which we could formalise additional properties from the study of the documentation (as in the case of *causes*, inverse of *causedBy*).

4.2. Conversion to Linked Data

Once the data from the three lexical sources had been standardised, they were converted into the Linked Data format. After introducing the reference model adopted, an example of converted data is provided.

4.2.1 The OntoLex-Lemon model

In the context of the representation and publication of lexical data as knowledge graphs and/or as Linguistic Linked Open Data, the OntoLex-Lemon model has become a *de facto* standard. This model was created with the aim of supporting the linguistic foundation of a given ontology by adding information on how ontological entities are lexicalised in different languages. However, OntoLex-Lemon can also be used as a lexicographic model to represent linguistic entities without any concept they denote being defined. OntoLex-Lemon is inspired by many other models, in particular the Lexical Markup Framework (LMF) (Francopoulo et al. 2006), LexInfo (Cimiano et al. 2011) — aligned with DatCatInfo (DatCatInfo n.d.) — and LIR (Linguistic Information Repository) (Montiel-Ponsoda et al. 2008).

Figure 1 represents the core of the model, called *ontolex*. The rectangles represent the classes of the model, the arrows with full heads represent the properties of the objects, and the arrows with empty heads represent the subclass relationships.

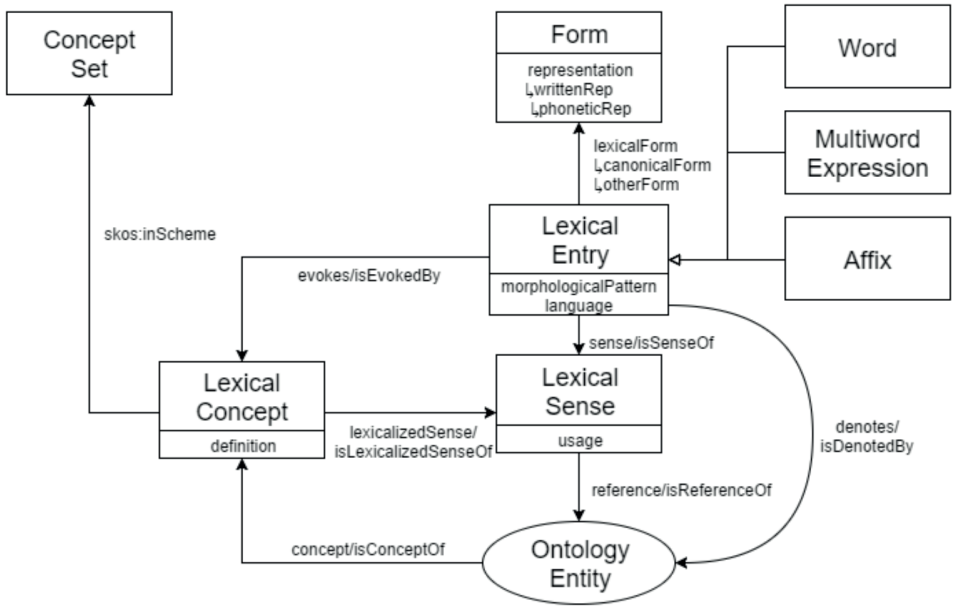


Figure 1: The core model of OntoLex-Lemon (ontolex). Picture taken from the W3C OntoLex Final Community Report at (W3C 2016).

ontolex is based on the definition of three fundamental classes: i) LexicalEntry, that «represents a unit of analysis of the lexicon that consists of a set of forms that are grammatically related and a set of base meanings that are associated with all of these forms»; ii) Form, that «represents one grammatical realisation of a lexical entry»; iii) LexicalSense, that «represents the lexical meaning of a lexical entry when interpreted as referring to the corresponding ontology element, if it is given». With reference to Figure 1, it is necessary to emphasise that *ontolex* also allows us to express the fact that a given lexical entry evokes a certain mental concept or refers to an entity with a formal interpretation defined in an ontology. Therefore, OntoLex-Lemon introduces a fourth element, the LexicalConcept class, which represents a mental abstraction, concept or unit of thought that can be lexicalised by a given collection of meanings.

The rest of the architecture of OntoLex-Lemon is divided into 4 modules, each representing a different linguistic aspect, namely: i) decomposition (decomp), i.e. the process of describing which elements constitute a multiword or compound; ii) the lexical and semantic relations between lexical entries and lexical senses respectively (vartrans); iii) the syntactic behaviour of lexical entries (synsem); iv) the description of the metadata of the lexical resource (lime). However, in this paper, our conversion work will mainly use *ontolex*, dealing with neither composition nor syntactic aspects in particular.

It is important to emphasise that the model abstracts from specific linguistic theories or category systems used to describe the properties of lexical entries and their syntactic behaviour. The re-use of existing category systems or linguistic ontologies is therefore strongly encouraged. In our case, as recommended by the community that developed the model and as described in section 4.1.1, the LexInfo model was used, which offers a rich vocabulary of linguistic categories and relationships for morphology, syntax and semantics.

4.2.2 Representing data in RDF OntoLex-Lemon

The conversion of the standardised data coming from the three sources into OntoLex-Lemon was performed by an algorithm in two steps: i) conversion of the linguistic information according to the formalisation described in the core *ontolex* module of the model; ii) serialisation of the data into Turtle[4]. The obtained lexicon was then loaded into Ontotext GraphDB (Ontotext n.d.), a semantic repository compliant with RDF and SPARQL (W3C 2013). Below is an example of an RDF OntoLex-Lemon representation of a CompL-it lexical entry in Turtle format.

```
:coniglio_entry a ontolex:Word;
      lexinfo:partOfSpeech lexinfo:noun;
      ontolex:canonicalForm coniglio_lemma;
      ontolex:otherForm coniglio_form_1;
      ontolex:sense coniglio_sense_1, coniglio_sense_2, co-
      niglio_sense_3.
:coniglio_lemma a ontolex:Form;
      lexinfo:gender lexinfo:masculine;
      lexinfo:number lexinfo:singular;
      ontolex:writtenRep "coniglio"@it, "rabbit"@en.
:coniglio_form_1 a ontolex:Form;
      lexinfo:gender lexinfo:masculine;
      lexinfo:number lexinfo:plural;
      ontolex:writtenRep "conigli"@it, "rabbits"@en.
:coniglio_sense_1 a ontolex:LexicalSense;
      skos:definition "mammifero della famiglia dei Lepori-
      di, con pelame di vario colore, lunghe orecchie, occhi
```

[4] Turtle is a serialisation format for RDF data types (W3C 2014).

```
      grandi e sporgenti e grossi incisivi"@it, "Mammal of
      the Leporidae family, with variously colored fur, long
      ears, large, protruding eyes and large incisors"@en;
      lexinfo:hyponym mammifero_sense;
      simple:polysemyAnimalFood coniglio_sense_3.
:coniglio_sense_2 a ontolex:LexicalSense;
      skos:definition "persona timida e molto paurosa"@it,
      "shy and very
      fearful person"@en;
      lexinfo:hyponym persona_sense;
      simple:metaphor coniglio_sense_1.
:coniglio_sense_3 a ontolex:LexicalSense;
      skos:definition "carne dell'omonimo animale"@it, "meat
      of the animal"@en.
```

In this example, the lexical entry *coniglio* (rabbit) is associated with two forms, one of which is defined as the canonical form (the lemma) and the other suitable for representing the plural form *conigli* (rabbits), both of which are equipped with the appropriate morphological traits. The lexical entry is also associated, via the *ontolex:sense* relation, with three lexical senses, each of which has a natural language definition. Furthermore, the first two senses are also endowed with semantic relations that link them to other lexical senses. For example, *rabbit_sense_2* is defined as a hyponym of *mammal_sense*.

4.3. CompL-it in numbers

In this section, the CompL-it lexicon is described from a quantitative perspective, both by enumerating the entities and relations it comprises and by comparing it with lexicographic resources available for the Italian language of a similar nature.

From a morphological standpoint, the resource is composed of 101,795 lexical entries (comprising a total of 791,541 word forms), classified with 36 POS categories and described with morphological traits. Figure 2 depicts a Venn diagram representing the different dimensions, in terms of lexical entries, of the three source resources and their intersections. As observed, the most significant contribution of words comes from M-GLF. However, both LexicO and the treebanks contribute significantly with a total of 47,069 forms and 9,028 additional lexical entries.

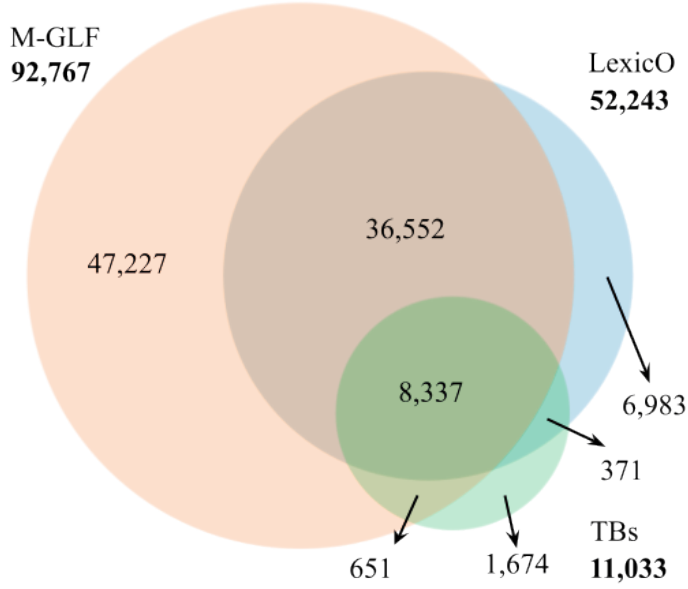


Figure 2: Representation, by the number of lexical entries, of the three source resources and their various intersections.

Going more into the specific linguistic content, Table 2 shows the distribution of word forms by POS: as expected, most are verbal forms (69% of the total). The *other* class encompasses the 32 POS not explicitly specified, such as, for example, number, conjunction, determiner, and preposition.

POS	LexicO	M-GLF	TBs	CompL-it
verb	345,307	526,635	10,741	545,342
noun	76,396	115,562	11,723	136,744
adj.	45,712	98,603	7,083	103,869
adv.	742	2,818	845	3223
other[5]	935	670	926	2,364
total	469,092	744,288	31,318	791,542

Table 2: Distribution of word forms by POS

[5] This category includes the following POS: adposition, article, auxiliary, cardinalNumeral, conjunction, coordinatingConjunction, definiteArticle, demonstrativeDeterminer, demonstrativePronoun, determiner, exclamativeDeterminer, exclamativePronoun, fusedPreposition, indefiniteArticle, indefiniteDeterminer, indefinitePronoun, interjection, interrogativeAdverb, interrogativeDeterminer, interrogativePronoun, numeral, numeralDeterminer, numeralPronoun, particle, personalPronoun, possessiveAdjective, possessiveDeterminer, possessivePronoun, pronoun, relativeDeterminer, relativePronoun, subordinatingConjunction.

As for the data related to the semantic layer, CompL-it describes 55,713 word senses connected to each other through 137 types of semantic relations, totalling 86,577 instances. Table 3 shows a distribution of the 10 most numerous types of semantic relation instances.

Semantic relation	**# instances**	**an example**
hyponym	43,069	*medicina*, *scienza* (medicine, science)
approximateSynonym	5,666	*sciocco*, *stupido* (foolish, stupid)
usedFor	3,291	*matita*, *scrivere* (pencil, to write)
partMeronym	3,159	*giorno*, *settimana* (day, week)
partHolonym	3,159	*cinghiale*, *grugno* (boar, snout)
createdBy	2,857	*quadro*, *dipingere* (painting, to paint)
ObjectOfTheActivity	1,366	*bistecca*, *mangiare* (steak, to eat)
memberMeronym	1,318	*segretario*, *partito* (secretary, party)
ResultingState	1,063	*bruciare*, *bruciato* (to burn, burnt)
memberHolonym	979	*stormo*, *uccello* (flock, bird)
other	20,255	-
total	86,577	

Table 3: Distribution of semantic relations instances.

To provide an overview of the dimensions and richness of linguistic information conveyed by CompL-it, we finally present, in Table 4, a comparison with other lexical resources available for Italian[6].

[6] Data updated at the time of writing.

	entries	forms	senses/synsets	semantic relations instances	semantic relations types
LexicO	71,021	469,708	56,870 senses	89,340	137
IWN	48,416	-	49,350 synsets	138,385	83
MWN	41,491	-	32,673 synsets	45,593	14
OmegaWiki	30,258[7]	-	23,417 senses	66,005[8]	41
SIMPLELex-IT	7,022	26,560	-	-	-
Morph-it!	34,968	504,906	-	-	-
CompL-it	101,795	791,541	56,870 senses	86,577[9]	137

Table 4: Concise comparison of some of the main freely available resources containing lexical data for the Italian language.

5. CompL-it: access and use

The resource, in addition to being available for download, can be queried through a dedicated web interface (KLAB n.d.). This interface, shown in Figure 3, allows the user to select a series of precompiled SPARQL queries (visible on the left), modify one of them using the right panel, or formulate a new query from scratch.

As an example, the figure includes a precompiled query that allows for displaying all meanings of the verb *fare* (to do). If selected, the interface queries the resource and returns 7 senses of that verb, displaying their definitions and some examples. Using the corresponding SPARQL query shown in the right panel, it is possible to modify the label *fare* (highlighted in the figure) to insert another Italian verb, click the execute query button at the top right, and view the meanings of that verb in CompL-it.

In addition to its presentation as a linguistic resource in itself, as it is freely distributed, numerically rich and conforms to Linked Open Data standards, CompL-it can also be described in relation to the uses that can be made of it for information management and retrieval tasks.

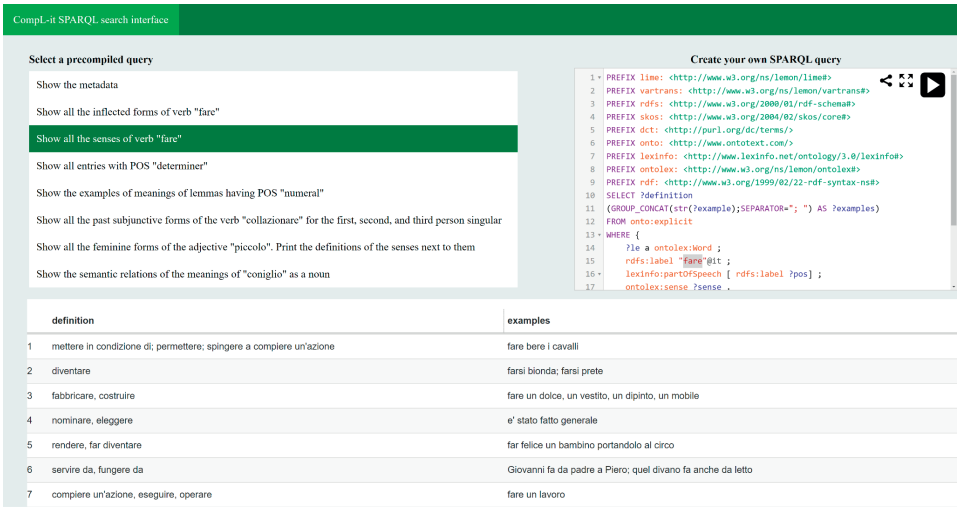
For instance, a computational lexicon can be used in full-text search approaches in which queries can fully exploit the morphological information contained therein and the complex and articulated system of semantic rela-

[7] This number includes both dictionary entries and encyclopaedic data (such as named entities).

[8] Semantic relations are defined between "concepts" and not between senses of a specific language.

[9] There are fewer instances of semantic relations in CompL-it than in LexicO because proper nouns were not extracted from the latter resource (as specified at the beginning of section 4) and the associated semantic relations were excluded with them.

tions between the words of the lexicon (especially synonymy and hyponymy). Better results are naturally obtained if searches are carried out on linguistically pre-analysed texts (at least with POS tagging) to reduce ambiguity in the results. In the following section we provide an example of a full-text search supported by a computational lexicon. For more details, see (Giovannetti et al. 2022).



	definition	examples
1	mettere in condizione di; permettere; spingere a compiere un'azione	fare bere i cavalli
2	diventare	farsi bionda; farsi prete
3	fabbricare, costruire	fare un dolce, un vestito, un dipinto, un mobile
4	nominare, eleggere	e' stato fatto generale
5	rendere, far diventare	far felice un bambino portandolo al circo
6	servire da, fungere da	Giovanni fa da padre a Piero; quel divano fa anche da letto
7	compiere un'azione, eseguire, operare	fare un lavoro

Figure 3: The query interface with the example of search for the senses of the Italian verb fare (to do).

Finally, we would also like to emphasise that representing data according to the LOD paradigm can bring some advantages, namely: (i) the federation mechanism with other datasets potentially allows the integration and improvement of queries results for more enriched searches, e.g., linking with etymological datasets (see Section 6); (ii) the addition of a semantic layer to the data through ontologies, allows the implicit knowledge in the dataset to be inferred and exploited in queries to the text, e.g., exploiting the transitivity of synonymy or hypernymy.

5.1. An example: lexicon-based search of the Babylonian Talmud

In this example, we show a query of the Italian translation of the Babylonian Talmud. This translation is being carried out by expert translators in the context of the Babylonian Talmud Translation Project (PTTB n.d.) using *Traduco*, a computer-aided translation tool developed at the CNR-ILC (Giovannetti et al. 2016). The search function of the tool allows the text to be accessed with complex queries, which can include information conveyed by the Italian lexicon related to morphology and semantics.

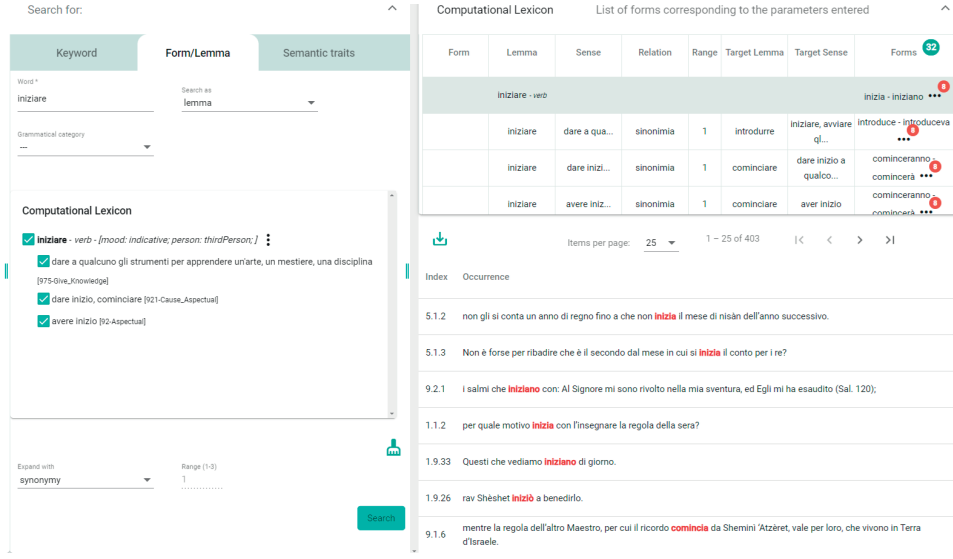


Figure 4: An example of query and relative results with the verb iniziare.

Considering the lemma *iniziare* (to start) it is possible to insert both morphological and semantic traits into the query (Fig. 4). By adding, for example, a restriction on the indicative mood and the third person, it is possible to extract all the contexts of the Talmud with the forms of the verb *iniziare* characterised by these traits. On the semantic side, we can also expand the query to include all lemmas having at least one sense as synonym to one of the (three) senses available in the lexicon for *iniziare*: the lexicon returns *introdurre* (to introduce) and *cominciare* (to begin). With the aforementioned morphological and semantic restrictions, the system is able to return Talmudic contexts containing, for example, the form *inizia* (singular, present indicative), but also *comincia*, *cominciano* (plural, present indicative), and *iniziò* (singular, past indicative).

6. Conclusions and perspectives

In this article, we presented CompL-it, a new computational lexicon for contemporary Italian. In the first part, we described the state of the art for this type of resource and outlined a landscape that, although rich, presents some challenges, such as the heterogeneity of data formats and linguistic models. Next, we described the three resources from which the lexicon was constructed: a computational lexicon (LexicO), a lemmatised form list obtained from a morphological analyser (M-GLF) and a set of treebanks. The three sources were based on different formats and models, which made it necessary

to standardise the data, also in accordance with the standards used by the Linguistic Linked Open Data community. Standardisation was followed by a conversion phase, which led to the final version of the lexicon in the form of Linguistic Linked Open Data according to the OntoLex-Lemon model.

We have also described the resource on a quantitative basis, also comparing CompL-it with some of the lexicographic resources available for Italian. The lexicon has been released as an open resource, is freely downloadable and can be consulted through a SPARQL interface. Finally, it was shown, through an example of a search on the Italian text of the Babylonian Talmud, how such a resource can be usefully exploited to provide linguistic-semantic access to textual corpora.

By its very nature, the editing of a lexicon can never be called a finished work. In the immediate future, CompL-it will first be further enriched in the semantic layer from the data that have not yet been extracted from LexicO (including templates and semantic traits). Subsequently, a merging methodology similar to the one adopted for morphology will also be applied to the semantic layer, in particular by considering available semantic resources such as ItalWordNet. A version of CompL-it will also be released, albeit limited to the morphological layer, conforming to the Universal Dependencies model. The resource will also be further extended to include cliticised forms, multiword forms and forms generated with suffixes, as in the case of *papà-papino* (dad, daddy). Finally, other linguistic layers will be considered, such as syntax (including data related to syntax-semantics interface) and phonetics, starting by leveraging on such data already available in LexicO.

Acknowledgement

Scientific publication produced thanks to the agreement between the National Research Council – Institute of Computational Linguistics and the PTTB S.c.a r.l. – Babylonian Talmud Translation Project.

References

BabelNet. n.d. “BabelNet | Il Più Grande Dizionario Enciclopedico e Rete Semantica Multilingue.” Accessed December 3, 2024. https://babelnet.org/.

Bamman, David, and Gregory Crane. 2010. “Computational Linguistics and Classical Lexicography.” In *Changing the Center of Gravity*, edited by Melissa Terras and Gregory Crane, 297-322. Gorgias Press. https://doi.org/10.31826/9781463219222-015.

Bartolini, Roberto. 2016. “IWN-LOD.” http://hdl.handle.net/20.500.11752/ILC-66.

Basili, Roberto, Silvia Brambilla, Danilo Croce, and Fabio Tamburini. 2017. "Developing a Large Scale FrameNet for Italian: The IFrameNet Experience." In *Proceedings of the Fourth Italian Conference on Computational Linguistics CLiC-It 2017*, edited by Roberto Basili, Malvina Nissim and Giorgio Satta, 59-64. Torino: Accademia University Press. https://doi.org/10.4000/books.aaccademia.2364.

Battista, Marco, and Vito Pirrelli. 1999. "Una Piattaforma di Morfologia Computazionale per l'analisi e la Generazione delle Parole Italiane." ILC-CNR Technical Report.

Bel, Nuria, Federica Busa, Nicoletta Calzolari, Elisabetta Gola, et al. 2000. "SIMPLE: A General Framework for the Development of Multilingual Lexicons." In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, edited by M. Gavrilidou, G. Carayannis, S. Markantonatou, S. Piperidis and G. Stainhauer. Athens, Greece: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2000/pdf/61.pdf.

Brown, Susan Windisch, Julia Bonn, Ghazaleh Kazeminejad, Annie Zaenen, James Pustejovsky, and Martha Palmer. 2022. "Semantic Representations for NLP Using VerbNet and the Generative Lexicon." *Frontiers in Artificial Intelligence* 5 (April):821697. https://doi.org/10.3389/frai.2022.821697.

Chen, Hsinchun, Tak Yim, David Fye, and Bruce Schatz. 1995. "Automatic Thesaurus Generation for an Electronic Community System." *Journal of the American Society for Information Science* 46 (3): 175-93.

Chiarcos, Christian, and Maria Sukhareva. 2015. "OLiA – Ontologies of Linguistic Annotation." Edited by Sebastian Hellmann, Steven Moran, Martin Brümmer, and John P. McCrae. *Semantic Web* 6 (4): 379-86. https://doi.org/10.3233/SW-140167.

Chiarcos, Christian, Elena-Simona Apostol, Besim Kabashi, and Ciprian-Octavian Truică. 2022. "Modelling Frequency, Attestation, and Corpus-Based Information with OntoLex-FrAC." In *Proceedings of the 29th International Conference on Computational Linguistics*, edited by Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, et al., 4018-27. Gyeongju, Republic of Korea: International Committee on Computational Linguistics. https://aclanthology.org/2022.coling-1.353.

Chiari, Isabella. 2012. "Il Dato Empirico in Lessicografia: Dizionari Tradizionali e Collaborativi a Confronto." *Bollettino Di Italianistica* II (January): 94-125.

Cimiano, Philipp, Christian Chiarcos, John P. McCrae, and Jorge Gracia. 2020. *Linguistic Linked Data: Representation, Generation and Applications*. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-30225-2.

Cimiano, Philipp, Paul Buitelaar, John McCrae, and Michael Sintek. 2011. "LexInfo: A Declarative Model for the Lexicon-Ontology Interface." *Journal of Web Semantics* 9 (1): 29-51. https://doi.org/10.1016/j.websem.2010.11.001.

CLARIN. n.d. "ParlaMint: Comparable and Interoperable Parliamentary Corpora | CLARIN ERIC." Accessed December 3, 2024. https://www.clarin.eu/parlamint.

CLARIN-IT. n.d.a. "CompL-It." Accessed December 3, 2024. https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/handle/20.500.11752/ILC-1007.

CLARIN-IT. n.d.b. "LexicO." Accessed December 3, 2024. https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/handle/20.500.11752/ILC-977.

CLARIN-IT. n.d.c. "MAGIC - Generated Lemmatized Forms." Accessed December 3, 2024. https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/handle/20.500.11752/ILC-1002.

CLARIN VLO. n.d. "Virtual Language Observatory." Accessed September 30, 2024. https://www.clarin.eu/content/virtual-language-observatory-vlo.

Dankova, Klara, Maria Teresa Zanola, and Silvia Calvi. 2022. "Pan-Latin Textile Fibres Vocabulary." http://hdl.handle.net/20.500.11752/OPEN-975.

DatCatInfo. n.d. "Welcome to DatCatInfo." Accessed December 3, 2024. https://datcatinfo.net/.

De Mauro, Tullio. 1980. *Guida all'uso delle parole: parlare e scrivere semplice e preciso per capire e farsi capire*. Libri di base. Roma: Editori Riuniti.

De Mauro, Tullio, a cura di. 2016. *Il Nuovo Vocabolario Di Base Della Lingua Italiana*. December 23, 2016. https://www.dropbox.com/scl/fi/zg2y99x-qik4k11nj19fgi/nuovovocabolariodibase.pdf?rlkey=s0uf8ggv11kf44i-p6a2ldz16n&e=1&dl=0.

Del Gratta, Riccardo, Francesca Frontini, Anas Fahad Khan, and Monica Monachini. 2015. "Converting the PAROLE SIMPLE CLIPS Lexicon into RDF with Lemon." *Semantic Web* 6 (4): 387-92. https://doi.org/10.3233/SW-140168.

ELEXIS. n.d. "ELEXIS European Lexicographic Infrastructure." Accessed September 30, 2024. https://elex.is/.

Francopoulo, Gil, Monte George, Nicoletta Calzolari, et al. 2006. "Lexical Markup Framework (LMF)." In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, edited by Nicoletta Calzolari, Khalid Choukri, Aldo Gangemi, Bente Maegaard, Joseph Mariani, Jan Odijk, and Daniel Tapias. Genoa, Italy: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2006/pdf/577_pdf.pdf.

Frontini, Francesca, Riccardo Del Gratta, and Monica Monachini. 2016. "Geodomain WordNet ITA ENG V 1.0." http://hdl.handle.net/20.500.11752/ILC-68.

Giovannetti, Emiliano, Davide Albanesi, Andrea Bellandi, and Giulia Benotto. 2016. "Traduco: A Collaborative Web-Based CAT Environment for the Interpretation and Translation of Texts." *Digital Scholarship in the Humanities* 32 (suppl_1): i47-62. https://doi.org/10.1093/llc/fqw054.

Giovannetti, Emiliano, Davide Albanesi, Andrea Bellandi, Simone Marchi, Mafalda Papini, and Flavia Sciolette. 2022. "The Role of a Computational Lexicon for Query Expansion in Full-Text Search." In *Proceedings of the Eighth Italian Conference on Computational Linguistics CliC-It 2021*, edited by Elisabetta Fersini, Marco Passarotti, and Viviana Patti, 162-68. Accademia University Press. https://doi.org/10.4000/books.aaccademia.10638.

Github. n.d.a. "The Ontolex Module for Frequency, Attestation and Corpus Information." Accessed December 3, 2024. https://github.com/acoli-repo/frac-addenda/blob/master/index.md.

Github. n.d.b. "CompL-It Mapping Tables." Accessed December 3, 2024. https://github.com/klab-ilc-cnr/Tables-for-mapping-of-Italian-lexicon-CompL-it.

Global WordNet Association. n.d. "Main Page." Accessed December 3, 2024. http://globalwordnet.org/.

Grella, Matteo. 2018a. "Italian Content Words V3." http://hdl.handle.net/11372/LRT-2894.

Grella, Matteo. 2018b. "Italian Function Words V3." http://hdl.handle.net/11372/LRT-2893.

Hmeidi, Ismail, Mahmoud Al-Ayyoub, Nizar A. Mahyoub, and Mohammed A. Shehab. 2016. "A Lexicon Based Approach for Classifying Arabic Multi-Labeled Text." *International Journal of Web Information Systems* 12 (4): 504-32. https://doi.org/10.1108/IJWIS-01-2016-0002.

Hodge, Gail. 2000. *Systems of Knowledge Organization for Digital Libraries: Beyond Traditional Authority Files*. Washington, DC: Digital Library Federation, Council on Library and Information Resources.

ILC4CLARIN CNR. 2016. "PAROLE-SIMPLE-CLIPS." http://hdl.handle.net/20.500.11752/ILC-88.

Khan, Fahad, Ana Salgado, Isuri Anuradha, et al. 2024. "CHAMUÇA: Towards a Linked Data Language Resource of Portuguese Borrowings in Asian Languages." In *Proceedings of the 9th Workshop on Linked Data in Linguistics @ LREC-COLING 2024*, edited by Christian Chiarcos, Katerina Gkirtzou, Maxim Ionov, Fahad Khan, John P. McCrae, Elena Montiel Ponsoda, and Patricia Martín Chozas, 44-48. Torino, Italia: ELRA and ICCL. https://aclanthology.org/2024.ldl-1.6.

KLAB. n.d. "CompL-It SPARQL Search Interface." Accessed December 3, 2024. https://klab.ilc.cnr.it/CompL-it-SPARQL-interface/.

Kyjánek, Lukáš, Zdeněk Žabokrtský, Jonáš Vidra, and Magda Ševčíková. 2021. "Universal Derivations v1.1." http://hdl.handle.net/11234/1-3247.

LexInfo. n.d. "About the Ontology." Accessed September 30, 2024. https://lexinfo.net/.

LLOD. n.d. "Linguistic Linked Open Data." Accessed December 3, 2024. https://linguistic-lod.org/.

Mallia, Michele, Michela Bandini, Andrea Bellandi, et al. 2024. "DigItAnt: A Platform for Creating, Linking and Exploiting LOD Lexica with Heterogeneous Resources." In *Proceedings of the 9th Workshop on Linked Data in Linguistics @ LREC-COLING 2024*, edited by Christian Chiarcos, Katerina Gkirtzou, Maxim Ionov, Fahad Khan, John P. McCrae, Elena Montiel Ponsoda, and Patricia Martín Chozas, 55-65. Torino, Italia: ELRA and ICCL. https://aclanthology.org/2024.ldl-1.8.

Malmgren, Sven-Göran. 1988. "On Regular Polysemy in Swedish." In *Studies in Computer-Aided Lexicology*, 179-200. Data Linguistica 18. Stockholm: Almqvist & Wiksell.

Mambrini, Francesco, and Marco Carlo Passarotti. 2023. "The LiLa Lemma Bank: A Knowledge Base of Latin Canonical Forms." *Journal of Open Humanities Data* 9 (November):1-5. https://doi.org/10.5334/johd.145.

Mazzei, Alessandro. 2016. "Building a Computational Lexicon by Using SQL." In *Proceedings of the Third Italian Conference on Computational Linguistics CLiC-It 2016*, 200-04. Napoli: Accademia University Press. https://doi.org/10.4000/books.aaccademia.1808.

Meijssen, Gerard. 2014. "OmegaWiki." http://hdl.handle.net/11372/LRT-853.

Miller, George A. 1995. "WordNet: A Lexical Database for English." *Communications of the ACM* 38 (11): 39-41. https://doi.org/10.1145/219717.219748.

Montiel-Ponsoda, Elena, Wim Peters, Mauricio Espinoza, Asunción Gómez-Pérez, and Margherita Sini. 2008. "Multilingual and Localization Support for Ontologies." Technical Report 2.4.2. http://neon-project.org/deliverables/WP2/NeOn_2008_D242.pdf.

Morph-it! 2018. "Resources:Morph-It." Last Modified May 03. https://docs.sslmit.unibo.it/doku.php?id=resources:morph-it.

MultiWordNet. n.d. "NLP Research Group - MultiWordNet." Accessed December 3, 2024. https://nlplab.fbk.eu/tools-and-resources/lexical-resources-and-corpora/multiwordnet.

Navigli, Roberto, and Simone Paolo Ponzetto. 2012. "BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network." *Artificial Intelligence* 193 (December): 217-50. https://doi.org/10.1016/j.artint.2012.07.001.

OLiA. n.d. "Ontologies of Linguistic Annotation (OLiA) | Olia." Accessed December 3, 2024. https://acoli-repo.github.io/olia/.

Ontotext. n.d. "Ontotext GraphDB." Accessed December 3, 2024. https://www.ontotext.com/products/graphdb/.

Passarotti, Marco Carlo, and Francesco Mambrini. 2021. "Linking Latin: Interoperable Lexical Resources in the LiLa Project." In *Building New Resources for Historical Linguistics*, edited by Erica Biagetti, Chiara Zanchi and Silvia Luraghi, 103-24. https://doi.org/10.5281/zenodo.5994271.

Pianta, Emanuele, Luisa Bentivogli, and Christian Girardi. 2002. "MultiWordNet: Developing an Aligned Multilingual Database." In *Proceedings of the First International Conference on Global WordNet*.

Pirrelli, Vito, and Marco Battista. 2000. "The Paradigmatic Dimension of Stem Allomorphy in Italian Verb Inflection." *Italian Journal of Linguistics* 12 (2): 307-80.

Prakash, T. Nikil, and Amalanathan Aloysius. 2021. "Textual Sentiment Analysis Using Lexicon Based Approaches." *Annals of the Romanian Society for Cell Biology* 25 (4): 9878–85.

Princeton University. n.d. "WordNet." Accessed December 3, 2024. https://wordnet.princeton.edu/.

PTTB. n.d. "Progetto Traduzione Talmud Babilonese." Accessed December 3, 2024. https://www.talmud.it/it/.

Pustejovsky, James. 1995. *The Generative Lexicon*. The MIT Press. https://doi.org/10.7551/mitpress/3225.001.0001.

Realiter. n.d. "Home | Realiter." Accessed December 3, 2024. https://www.realiter.net/.

Roventini, Adriana, Antonietta Alonge, Francesca Bertagna, et al. 2003. "'ItalWordNet': Building a Large Semantic Database for the Automatic Treatment of Italian." *Linguistica computazionale: XVIII/XIX, 1998/1999*, 745-91. https://doi.org/10.1400/18178.

Roventini, Adriana, Rita Marinelli, and Francesca Bertagna. 2016. "ItalWordNet v.2." http://hdl.handle.net/20.500.11752/ILC-62.

Ruimy, Nilda, Monica Monachini, Raffaella Distante, et al. 2002. "CLIPS, a Multi-Level Italian Computational Lexicon: A Glimpse to Data." In *Proceedings of the Third International Conference on Language Resources and Evaluation, LREC 2002, May 29-31, 2002, Las Palmas, Canary Islands, Spain*. European Language Resources Association. http://www.lrec-conf.org/proceedings/lrec2002/sumarios/197.htm.

Sabatini, Francesco. 2006. "La Storia dell'Italiano nella Prospettiva della Corpus Linguistics." In *Proceedings of the 12th EURALEX International Congress*, edited by Cristina Onesti, Elisa Corino and Carla Marello, 31-37. Torino: Edizioni dell'Orso.

Sanguinetti, Manuela, and Cristina Bosco. 2015. "PartTUT: The Turin University Parallel Treebank." In *Harmonization and Development of Resources and Tools for Italian Natural Language Processing within the PARLI Project*, edited by Roberto Basili, Cristina Bosco, Rodolfo Delmonte, Alessandro Moschitti, and Maria Simi, 589: 51-69. Studies in Computational Intelligence. Springer. https://doi.org/10.1007/978-3-319-14206-7_3.

Sciolette, Flavia, Emiliano Giovannetti, and Simone Marchi. 2023. "LexicO: An Italian Computational Lexicon Derived from Parole-Simple-Clips." *Umanistica Digitale* 7 (15): 169-93. https://doi.org/10.6092/issn.2532-8816/15176.

Sciolette, Flavia. 2024. "Modeling Linking between Text and Lexicon with OntoLex-Lemon: A Case Study of Computational Terminology for the Babylonian Talmud." In *Proceedings of the 9th Workshop on Linked Data in Linguistics @ LREC-COLING 2024*, edited by Christian Chiarcos, Katerina Gkirtzou, Maxim Ionov, Fahad Khan, John P. McCrae, Elena Montiel Ponsoda, and Patricia Martín Chozas, 103-7. Torino, Italia: ELRA and ICCL. https://aclanthology.org/2024.ldl-1.13.

Sérasset, Gilles. 2015. "DBnary: Wiktionary as a Lemon-Based Multilingual Lexical Resource in RDF." *Semantic Web* 6 (4): 355-61. https://doi.org/10.3233/SW-140147.

Shiri, Ali. 2015. "Semantic Access and Exploration in Cultural Heritage Digital Libraries." In *Cultural Heritage Information: Access and Management*, edited by Ian Ruthven and Gobinda G. Chowdhury, 177-96. Facet Publishing.

Simi, Maria, Cristina Bosco, and Simonetta Montemagni. 2014. "Less Is More? Towards a Reduced Inventory of Categories for Training a Parser for the Italian Stanford Dependencies." In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, edited by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, 83-90. Reykjavik, Iceland: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2014/pdf/818_Paper.pdf.

SimpleLEX-IT. n.d. "Alexmazzei/SimpleLEX-IT: SimpleLEX-IT Is the Computational Lexicon Employed into SimpleNLG-IT." Accessed December 3, 2024. https://github.com/alexmazzei/SimpleLEX-IT.

Smith, Jocelyn C. 1997. "The Use of Lexicons in Information Retrieval in Legal Databases." In *Proceedings of the Sixth International Conference on Artificial Intelligence and Law - ICAIL '97*, 29-38. Melbourne, Australia: ACM Press. https://doi.org/10.1145/261618.261625.

Thompson, Paul, John McNaught, Simonetta Montemagni, et al. 2011. "The BioLexicon: A Large-Scale Terminological Resource for Biomedical Text Mining." *BMC Bioinformatics* 12 (1): 397. https://doi.org/10.1186/1471-2105-12-397.

Universal Dependencies. n.d.a. "Introduction." Accessed December 3, 2024. https://universaldependencies.org/it/overview/introduction.html.

Universal Dependencies. n.d.b. "Universal Dependencies." Accessed December 3, 2024. https://universaldependencies.org/.

Universal Dependencies. n.d.c. "UD Guidelines." Accessed December 3, 2024. https://universaldependencies.org/guidelines.html.

Universal Dependencies. n.d.d. "UD_Italian-ISDT." Accessed December 3, 2024. https://github.com/UniversalDependencies/UD_Italian-ISDT/tree/master.

Universal Dependencies. n.d.e. "UD_Italian-VIT." Accessed December 3, 2024. https://github.com/UniversalDependencies/UD_Italian-VIT/.

Universal Dependencies. n.d.f. "CoNLL-U Format." Accessed December 3, 2024. https://universaldependencies.org/format.html.

Universal Dependencies. n.d.g. "UD_Italian-ParTUT." Accessed December 3, 2024. https://github.com/UniversalDependencies/UD_Italian-ParTUT/.

Universal Dependencies. n.d.h. "UD_Italian-ParlaMint." Accessed December 3, 2024. https://github.com/UniversalDependencies/UD_Italian-ParlaMint/.

Vetere, Guido, Alessandro Oltramari, Isabella Chiari, Elisabetta Jezek, Laure Vieu, and Fabio Massimo Zanzotto. 2011. "Senso Comune, an Open Knowledge Base of Italian Language." *Traitement Automatique Des Langues* 52 (3): 217-43.

Villegas, Marta, and Núria Bel. 2015. "PAROLE/SIMPLE 'Lemon' Ontology and Lexicons." Edited by Sebastian Hellmann, Steven Moran, Martin Brümmer, and John McCrae. *Semantic Web* 6 (4): 363-69. https://doi.org/10.3233/SW-140148.

W3C. 2005. "UsingSeeAlso - W3C Wiki." Last Modified January 13. https://www.w3.org/wiki/UsingSeeAlso.

W3C. 2013. "SPARQL 1.1 Overview." https://www.w3.org/TR/sparql11-overview/.

W3C. 2014. "RDF 1.1 Turtle." https://www.w3.org/TR/turtle/.

W3C. 2016. "Lexicon Model for Ontologies: Community Report, 10 May 2016." https://www.w3.org/2016/05/ontolex/.

WikiMedia. 2022. "OmegaWiki - Meta." Last Modified December 03. https://meta.wikimedia.org/wiki/OmegaWiki.

Wikipedia. 2024a. "Verbi Incoativi - Wikipedia." Last Modified January 26. https://it.wikipedia.org/wiki/Verbi_incoativi.

Wikipedia. 2024b. "Verbi Irregolari Italiani - Wikipedia." Last Modified November 6. https://it.wikipedia.org/wiki/Verbi_irregolari_italiani.

Žabokrtský, Zdeněk, Nyati Bafna, Jan Bodnár, et al. 2022. "Universal Segmentations 1.0 (UniSegments 1.0)." http://hdl.handle.net/11234/1-4629.

Zanchetta, Eros, and Marco Baroni. 2005. "Morph-It! A Free Corpus-Based Morphological Resource for the Italian Language." In *Proceedings of Corpus Linguistics Conference Series 2005 (ISSN 1747-9398)*, 1: 1-12. University of Birmingham.

Zanola, Maria Teresa, Klara Dankova, Claudio Grimaldi, and Anna Serpente. 2023. "Pan-Latin Lexicon of Collars and Sleeves in Fashion and Costume." http://hdl.handle.net/20.500.11752/OPEN-987.

Rubriche

AIDAinformazioni
ISSN 1121–0095
ISBN 979-12-5965-456-4
DOI 10.57574/596545647
pag. 151-154 (luglio-dicembre 2024)

Non solo libri

Claudio Gnoli*

Il 6 e 7 febbraio 2023 si è tenuto all'Università la Sapienza il convegno "Look beyond: indicizzazione per soggetto delle risorse non librarie"[1], seguito da una sessantina di partecipanti e organizzato dal gruppo CILW (Catalogazione, indicizzazione, linked open data e web semantico) dell'Associazione Italiana Biblioteche insieme a ISKO Italia, che aveva già toccato temi affini in un evento del 2012 (ISKO Italia n.d.).

Ci insegna la teoria dei prototipi di Eleanor Rosch, citata non a caso da due dei relatori, che noi tendiamo ad associare molti termini del linguaggio all'idea di un loro tipo classico, sicché *uccello* ci fa pensare più facilmente a un passero che a uno struzzo: ebbene, allo stesso modo *documento* ci fa pensare solitamente a un testo scritto e non a un arbusto vivo esposto in un orto botanico, sebbene anche quest'ultimo sia un membro della classe dei documenti, al pari della famosa antilope nello zoo di Suzanne Briet (dell'idea di documento avevamo già discusso nel fascicolo 3-4 del 2009). Così come certe insegne di fornai propongono "non solo pane", i cataloghi possono indicizzare "non solo libri".

Al di là del fatto che il loro formato sia tradizionale o digitale, i documenti possono avere una forma testuale ma anche visiva, sonora, tridimensionale e così via. In questi casi un'indicizzazione *derivata* con termini già presenti nel documento non è possibile, proprio perché esso non è fatto di parole. Come identificare allora il loro soggetto? La relatrice invitata Athena Salaba, seguendo i recenti modelli concettuali dell'IFLA, propone di applicare a questi documenti l'iconologia del noto storico dell'arte Erwin Panofsky. Un'immagine, secondo Panofsky, contiene tre livelli di significato: l'*isness* ovvero la natura del documento, l'*ofness* ovvero le cose in esso rappresentate e l'*aboutness* (talvolta tradotto con *circalità*) ovvero ciò che con queste cose si vuole esprimere, soli-

* Biblioteca della scienza e della tecnica, Università degli Studi di Pavia, Italia. claudio.gnoli@unipv.it.

1 Programma e materiali disponibili su (Associazione Italiana Biblioteche n.d.).

tamente considerato corrispondente al soggetto dei documenti testuali (Hjørland 2017). In una statua di Martin Luther King, *l'isness* è il pezzo di granito scolpito, *l'ofness* è Martin Luther King e *l'aboutness* sono i diritti civili.

Osserviamo che queste informazioni vengono in genere tradotte a loro volta in indici testuali, per esempio nella stringa `diritti civili` oppure `antilopi`, perché il testo alfanumerico è la modalità fondamentale con cui funziona la documentazione, anche nella nostra epoca digitale basata sul codice ASCII dei calcolatori; altrimenti possiamo ricorrere a tecniche diverse di *multimedia information retrieval* (MMIR), quale la ricerca di particolari forme e colori, che sono attualmente oggetto della ricerca informatica: ne ha scritto ampiamente negli ultimi anni Roberto Raieli.

I documenti che hanno la forma di oggetti tridimensionali, come una scultura, uno strumento musicale o un fossile, sono esposti e fruiti in istituzioni tradizionalmente chiamate *musei*. Tuttavia Salaba ha considerato la tendenza recente a varare "libraries of things" che propongono, oltre a libri, oggetti svariati compresi strumenti musicali che gli utenti possono provare: qui il termine *biblioteca* è forse suggerito, più che dal tipo di documento, dalle modalità di fruizione comprendenti una "consultazione" o un prestito, che nei musei sono invece inusuali.

In effetti può accadere che la funzione di documento venga assunta dagli oggetti più diversi, come ho avuto modo di constatare interessandomi di strumenti delle tradizioni popolari (Gnoli 2010): oltre che da esemplari degli strumenti stessi, infatti, succede di ottenere informazioni rilevanti anche da una statuina di presepio o da una marionetta della famiglia torinese Lupi, in quanto esse imbracciano una cornamusa in cui si vedono i bordoni infilati in punti diversi della sacca, che ne tradiscono un'origine in Italia settentrionale (dove questi strumenti sono poi in gran parte scomparsi) e non centro-meridionale (dove le zampogne tuttora diffuse hanno tutti i bordoni fuoriuscenti da un unico foro). Altre notizie su questi strumenti ci giungono dai suonatori ritratti in affreschi di chiese lombarde, piemontesi e liguri, e in un caso recente perfino dalla scena dipinta sul sipario ottocentesco del teatro Carbonetti di Broni, in Oltrepò Pavese, dove si vede uno strumento conico che fa pensare agli oboi popolari del vicino Appennino delle Quattro Province.

In tutti questi casi, però, la scoperta delle informazioni è stata casuale, perché nessuno si era dato la pena di indicizzare ogni particolare del presepio o del sipario. Entra qui in gioco la questione della rilevanza del *tema di base* (affine alla *aboutness*) del documento e quella dei suoi diversi *temi particolari*, che secondo Alberto Cheti si possono registrare o meno anche tenendo conto del loro interesse per gli utenti. La rilevanza, tuttavia, è difficile da predire in anticipo: come si può immaginare che un affresco sacro verrà un giorno studiato da un etnomusicologo?

Se pure quei particolari fossero stati indicizzati, lo sarebbero stati probabilmente in un archivio diverso dai cataloghi delle biblioteche, rendendo per ora utopistica una ricerca sul concetto *oboi* veramente cross-mediale (usando un termine caro all'architetto dell'informazione Luca Rosati) che recuperi informazioni tanto da libri quanto da affreschi e sipari…

Non mancano le iniziative per l'indicizzazione del nostro immenso patrimonio culturale, comprese quelle presentate a Roma; esse tuttavia sono ancora disperse e non sempre interoperabili fra loro. A collegarle potrebbe essere il linguaggio trasversale dei linked open data (LOD) nel quale codificare sia le registrazioni bibliografiche che gli indici di altri tipi di documenti. Ma solo a patto di affiancare all'interoperabilità tecnica anche un'interoperabilità concettuale, capace di identificare gli oboi trattati in un video documentario e quelli dipinti su un sipario con codici comuni: e a questo scopo non potremo fare a meno dei buoni vecchi sistemi di organizzazione della conoscenza, quali tesauri e classificazioni (o se preferite, ontologie e tassonomie).

Riferimenti bibliografici

Associazione Italiana Biblioteche. n.d. "Look beyond." Consultato il 13 novembre 2024, https://www.aib.it/eventi/look-beyond-ita/.

Gnoli, Claudio. 2010. "Classification Transcends Library Business." *Knowledge Organization* 37 (3): 223-29.

Hjørland, Birger. 2017. "Subject (of documents)." *Knowledge Organization* 44 (1): 55-64. Also available in *ISKO Encyclopedia of Knowledge Organization*, edited by Birger Hjørland and Claudio Gnoli, isko.org/cyclo/subject#3.

ISKO Italia. n.d. "Organizzare la conoscenza in musei, teatri e archivi multimediali." Consultato il 13 novembre 2024, http://www.iskoi.org/doc/nbm.