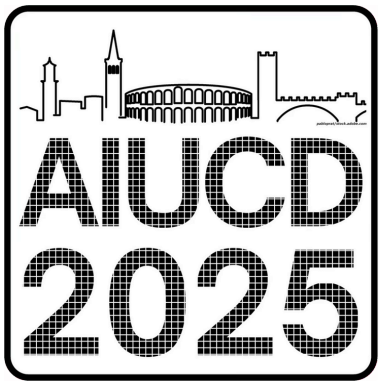

Atti del XIV Convegno Annuale

Diversità, Equità e Inclusione: Sfide e Opportunità per l'Informatica Umanistica nell'Era dell'Intelligenza Artificiale

Verona :: 11-13 giugno 2025

A cura di:
Simone Rebora • Marco Rospocher • Stefano Bazzaco

UNIVERSITÀ di VERONA
Dipartimento di LINGUE E LETTERATURE STRANIERE

daih

ASSOCIAZIONE per l'INFORMATICA UMANISTICA e la CULTURA DIGITALE

ISBN 978-88-942535-9-7

A cura di: Simone Rebora; Marco Rospocher; Stefano Bazzaco (2025). Diversity, Equity, and Inclusion: Challenges and Opportunities for Digital Humanities in the Age of Artificial Intelligence, Proceedings del XIV Convegno Annuale AIUCD, Verona 11-13 giugno 2025, Università di Verona.

Ultimo accesso agli URL in data 8 maggio 2025.

Si prega di notificare all'editore ogni omissione o errore si riscontri: segreteria [at] aiucd.org
Please notify the publisher of any omissions or errors found: segreteria [at] aiucd.org

Il programma della conferenza AIUCD 2025 è disponibile online
https://aiucd2025.dlls.univr.it/detailed-schedule/
The AIUCD 2025 Conference Program is available online
https://aiucd2025.dlls.univr.it/en-gb/detailed-schedule/

I contributi pubblicati nel presente volume hanno ottenuto il parere favorevole da parte di valutatori esperti della materia, attraverso un processo di revisione anonima mediante double-blind peer review, effettuata dai membri del Comitato di Programma sotto la supervisione del Comitato Scientifico di AIUCD 2025.
All the papers published in this volume have received favourable reviews by experts in the field of DH, through an anonymous double-blind peer review, carried out by the members of the Programme Committee under the supervision of the Scientific Committee of AIUCD 2025.

Gli atti del convegno AIUCD 2025 sono pubblicati come raccolta di contributi in formato PDF forniti direttamente dagli autori e dalle autrici. I file sono stati raccolti e assemblati senza interventi redazionali da parte dei curatori.
The proceedings of the AIUCD 2025 conference are published as a collection of PDF contributions provided directly by the authors. The files have been collected and compiled without editorial intervention by the editors..

Il logo di AIUCD 2025 include l'immagine "Verona Dark Line Simple Minimalist Skyline With White Background" di @pabloprat/stock.adobe.com, ottenuta tramite la licenza Adobe Stock dell'Università di Verona.
The AIUCD 2025 logo includes the image "Verona Dark Line Simple Minimalist Skyline With White Background" by @pabloprat/stock.adobe.com, used under the Adobe Stock license of the University of Verona.

Il background della copertina è stato creato con tecniche di AI generativa con lo strumento "Magic Media" disponibile su Canva, usando un prompt con il tema del convegno.
The background of the cover was created using generative AI techniques with the "Magic Media" tool available on Canva, using a prompt based on the conference theme.

Comitato Organizzatore / *Organizing Committee*

General Chairs
Simone Rebora (Università degli Studi di Verona)
Marco Rospocher (Università degli Studi di Verona)

Local Chair
Anna Cappellotto (Università degli Studi di Verona)

Registration Chair
Giorgia Pomarolli (Università degli Studi di Verona)

Proceedings Chair
Stefano Bazzaco (Università degli Studi di Verona)

Sponsorship Chair
Matteo Lissandrini (Università degli Studi di Verona)

Publicity Chair
Sabrina Piccinin (Università degli Studi di Verona)

Comitato Scientifico / *Scientific Committee*

Program Chairs
Simone Rebora (Università degli Studi di Verona)
Marco Rospocher (Università degli Studi di Verona)

Digital Humanities e inclusione / *Inclusive DH*
Stefano Bazzaco (Università degli Studi di Verona)
Massimo Salgaro (Università degli Studi di Verona)

Archivi ed Edizioni Digitali / *Archives and Digital Editions*
Elisa Cugliana (Cologne Center for eHumanities)
Christian D'Agata (Università di Catania)

Metodi Computazionali / *Computational Methods*
Rachele Sprugnoli (Università degli Studi di Parma)
Sara Tonelli (Fondazione Bruno Kessler)

Rappresentazione di Dati e Conoscenza / *Data and Knowledge Representation*
Francesco Mambrini (Università Cattolica del Sacro Cuore)
Elena Spadini (Universität Bern)

Preservazione della Memoria e del Patrimonio Digitale / *Preservation of Memory and Digital Cultural Heritage*
Monica Berti (Universität Leipzig)
Daria Spampinato (Istituto di Scienze e Tecnologie della Cognizione-CNR)

Comitato di programma / *Program committee*

Stefano Allegrezza (Università di Macerata); Laura Antonietti (Université de Versailles Saint Quentin en Yvelines (Université Paris-Saclay); Luigi Bambaci (École pratique des hautes études, PSL); Liborio P. Barbarino (Università di Catania); Nicola Barbuti (Università degli Studi di Bari Aldo Moro (Dipartimento di Ricerca e Innovazione Umanistica); Sofia Baroncini (Leibniz Institute of European History); Andrea Bellandi (Institute for Computational Linguistics (CNR); Mario A. Bochicchio (University of Bari, Dep.t of Computer Science); Andrea Bolioli (Independent researcher); Marco Bombieri (University of Verona); Paolo Bonora (Università di Bologna); Flavia Bruni (Università di Chieti-Pescara); Marina Buzzoni (Ca' Foscari University of Venice); Alberto Campagnolo (KU Leuven); Anna Cappellotto (Università di Verona); Emmanuela Carbé (Università di Siena); Vittore Casarosa (ISTI-CNR); Raffaele Cioffi (Università di Napoli Federico II); Fabio Ciotti (Università di Roma Tor Vergata); Vincenzo Colaprice (University of Turin); Giuseppe Consolo (Università degli studi di Napoli, Federico II); Elisa Conti (Università di Catania); Salvatore Cristofaro (CNR ISTC); Giulia D'Agostino (TU Darmstadt); Elisa D'Argenio (HUN-REN Hungarian Research Centre for Linguistics); Enrico Daga (The Open University); Stefano Dall'Aglio (Università Ca' Foscari Venezia); Marilena Daquino (University of Bologna); Mauro De Bari (University of Bari Aldo Moro); Angelo M. Del Grosso (CNR-ILC); Matteo Di Franco (Università di Napoli Federico II); Giorgio Maria Di Nunzio (University of Padua); Stefano Ferilli (University of Bari); Lorenzo Ferroni (Università degli Studi di Verona); Franz Fischer (Ca' Foscari Università Ca' Foscari); Greta H. Franzini (Eurac Research); Francesca Frontini (CNR-ILC); Daniele Fusi (VeDPH, Stuttgart University); Mariangela Giglio (University of Bologna); Tiago Luis Gil (University of Brasilia); Luca Giovannini (University of Potsdam); Milena Giuffrida (Università di Catania); Edmondo Grassi (Università Telematica San Raffaele Roma); Miryam Grasso (Università di Catania); Piergiovanna Grossi (Università di Verona); Fahad Khan (CNR-ILC); Michele Lacriola (Università di Siena); Maurizio Lana (Univ. del Piemonte Orientale); Federica Lazzerini (Università degli Studi di Torino); Eleonora Litta (Università Cattolica del Sacro Cuore, Milano); Dominique Longrée (ULiège); Diego Mantoan (University of Palermo); Anna Maria Marras (University of Turin); Cristina Marras (CNR); Pietro Mazzarisi (University of Trieste); Barbara McGillivray (King's College London); Federico V. Meschini (Tuscia University); Alessio Miaschi (Istituto di Linguistica Computazionale "A. Zampolli" (CNR-ILC), Pisa); Andrea Micheletti (University of Padua); Giulia Miglietta (Università del Salento); Paolo Monella (Università Kore di Enna); Johanna Monti (Università degli Studi di Napoli "L'Orientale"); Rossana Morriello (Università degli Studi di Firenze); Gloria Mugelli (ILC CNR); Serge Noiret (AIPH (Associazione Italiana di Public History) (European University Institute); Giuseppe Palazzolo (Università di Catania); Mafalda Papini (CNR-ILC); Enrico Pasini (UniTO/CNR-ILIESI); Giulia Pedonese (CNR (Istituto di Linguistica Computazionale "Antonio Zampolli"); Paola Peratello (Università Ca' Foscari Venezia); Federico Pianzola (University of Groningen); Chantal Pivetta (Lund University (Sweden)); Igor Pizzirusso (AIPH); Giulia Re (ILC-CNR); Giulia Renda (University of Bologna); Dario Rodighiero (University of Groningen); Roberto Rosselli Del Turco (Università di Torino); Enrica Salvatori (Università di Pisa); Emilio M. Sanfilippo (CNR); Eva Sassolini (CNR-ILC); Andrea Schimmenti (University of Bologna); Flavia Sciolette (CNR-ILC); Pietro Sichera (ILIESI-CNR); Daniele Silvi (Università di Roma 'Tor Vergata'); Giulia Speranza (University of Naples L Orientale); Francesco V. Stella (UNISI); Timothy Tambassi (Ca' Foscari University of Venice); Mirko Tavosanis (Università di Pisa); Francesca Tomasi (Università di Bologna); Simona Turbanti (University of Milano); Marco Venuti (Università di Catania); Gennaro Vessio (University of Bari Aldo Moro); Gabriele Vezzani (Università di Verona / RWTH Aachen University); Fabio Vitali (University of Bologna).

Enti organizzatori / *Organisers*

AIUCD; Università di Verona: Dipartimento di Lingue e Letterature Straniere; Digital Arena for Inclusive Humanities (DAIH).

Sommario

Prefazione

Digital Humanities and Inclusion

Archives and Digital Editions

Computational Methods

Data and Knowledge Representation

Preservation of Memory and Digital Cultural Heritage

Prefazione

AIUCD 2025, il XIV Convegno annuale dell'Associazione per l'Informatica Umanistica e la Cultura Digitale, è dedicato al tema "Diversità, Equità e Inclusione: Sfide e Opportunità per l'Informatica Umanistica nell'Era dell'Intelligenza Artificiale". Il convegno si configura come un'occasione privilegiata per riflettere su tematiche di cruciale rilevanza in un'epoca in cui l'integrazione tra tecnologia e discipline umanistiche appare sempre più necessaria. In un mondo in cui l'intelligenza artificiale sta ridefinendo i confini del sapere, diventa fondamentale interrogarsi su come guidare questa trasformazione secondo principi di inclusività ed equità. Il tema delle Digital Humanities in relazione all'inclusione richiama l'attenzione sull'importanza di una rappresentazione diversificata nel campo delle scienze umanistiche e sottolinea il potenziale degli strumenti digitali e computazionali nella democratizzazione dell'accesso al sapere e alla cultura.

Il Convegno AIUCD 2025 è organizzato dalla Digital Arena for Inclusive Humanities (DAIH), centro di ricerca interdisciplinare del Dipartimento di Lingue e Letterature Straniere dell'Università degli Studi di Verona. Istituito nel 2023 nell'ambito del progetto di eccellenza 2023–2027 "Inclusive Humanities. Prospettive di sviluppo nella ricerca e nella didattica delle lingue e letterature straniere", il centro rappresenta la naturale evoluzione del precedente progetto di eccellenza del Dipartimento, incentrato su "Le Digital Humanities applicate alle lingue e letterature straniere" (2018-2022). DAIH si configura come catalizzatore di collaborazioni interdisciplinari, con obiettivi di ricerca all'intersezione tra informatica – in particolare intelligenza artificiale – e discipline umanistiche, con un focus specifico sugli studi linguistici e letterari. Il centro promuove l'avanzamento della conoscenza scientifica e l'innovazione nell'ambito delle tecnologie digitali e dei metodi computazionali applicati agli studi umanistici, contribuendo alla costruzione di una società più equa, inclusiva e diversificata.

La call for papers del convegno AIUCD 2025 ha sollecitato contributi su cinque assi tematici, uno dei quali esplicitamente centrato sul tema dell'inclusione: Digital Humanities e inclusione, Archivi ed edizioni digitali, Metodi computazionali, Rappresentazione dei dati e della conoscenza, e Preservazione della memoria e del patrimonio digitale. Sono pervenute 126 proposte, di cui 26 presentate con il contributo di studiosi e studiose affiliati a istituzioni straniere, a testimonianza della crescente dimensione internazionale e interdisciplinare del convegno annuale.

Le proposte sono state valutate attraverso un processo di revisione a doppio cieco, gestito da un comitato di programma composto da 10 track chair e 94 revisori, oltre ai general chair. Ogni contributo è stato assegnato ad almeno tre revisori indipendenti, ottimizzando le preferenze espresse nella fase di bidding e il carico di lavoro complessivo. In totale, sono state raccolte 366 valutazioni, che hanno permesso ai track chair di ciascun asse tematico di formulare raccomandazioni informate ai general chair in merito all'accettazione e alla forma di presentazione dei contributi. A seguito del ritiro di sei proposte, le decisioni finali hanno portato alla seguente selezione:

- 65 contributi accettati per presentazione orale,
- 37 contributi accettati come poster,
- 18 contributi rifiutati.

Le proposte accettate sono così distribuite nelle varie track:

- Digital Humanities e inclusione: 21
- Archivi ed edizioni digitali: 32
- Metodi computazionali: 14
- Rappresentazione dei dati e della conoscenza: 21
- Preservazione della memoria e del patrimonio digitale: 14

Come nelle edizioni precedenti, il convegno è stato preceduto dal ciclo di incontri online "Aspettando AIUCD 2025", organizzato in collaborazione con l'iniziativa Digital Spritz del Dipartimento di Lingue e Letterature Straniere dell'Università di Verona, che ha ospitato gli interventi di Rachele Sprugnoli (Università Cattolica del Sacro Cuore), Susanna Allés-Torrent (University of Miami), Giulia Pedonese e Francesca Frontini (CNR-ILC).

Il programma del convegno è ulteriormente arricchito dalla partecipazione di due keynote speaker d'eccezione:

- Evelyn Gius (fortext lab, Institute of Linguistics and Literary Studies, Technical University of Darmstadt), con un intervento dal titolo "Measuring What Matters — or, The Temperature of Literary Texts",
- Viviana Patti (Dipartimento di Informatica, Università di Torino), con un intervento dal titolo "Absit iniuria verbis".

Concludiamo questa prefazione con un sentito ringraziamento a tutte le persone e gli enti che hanno reso possibile la realizzazione di AIUCD 2025. Un ringraziamento speciale va al Direttivo AIUCD, e in particolare alla presidente Marina Buzzoni, al segretario Paolo Monella e alla tesoriera Francesca Frontini, per il costante supporto istituzionale e la preziosa collaborazione in tutte le fasi preparatorie del convegno. Desideriamo esprimere la nostra gratitudine al Comitato di Programma, composto da 10 track chair e 94 revisori, per la disponibilità e l'accuratezza nella valutazione delle proposte, così come a tutte le persone che hanno risposto alla call for papers e partecipato al convegno, contribuendo a rendere il programma ricco e stimolante. Un grazie sentito va al Comitato Organizzatore, formato dai membri del consiglio direttivo DAIH, per l'instancabile impegno nella gestione di ogni fase dell'evento: dal coinvolgimento degli sponsor (Matteo Lissandrini), alla comunicazione e promozione dell'evento (Sabrina Piccinin), dalla gestione delle iscrizioni (Giorgia Pomarolli) al coordinamento delle attività logistiche locali (Anna Cappellotto). Ringraziamo inoltre il personale della segreteria amministrativa del Dipartimento di Lingue e Letterature Straniere e le studentesse dell'Università di Verona che, attraverso i loro tirocini, hanno offerto un prezioso supporto all'organizzazione. Infine, rinnoviamo la nostra gratitudine a tutti gli enti e le aziende che, in varie forme, hanno sostenuto AIUCD 2025 e ne hanno reso possibile la realizzazione.

Verona, maggio 2025

Simone Rebora, Marco Rospocher e Stefano Bazzaco

Entità in relazione: policies, soluzioni tecnologiche e modelli lessicali per un (eco)sistema informativo integrato

Herbert Natta[1], Michela Tardella[2], Eleonora Lattanzi[3], Gianluca Rossi[4], Roberta Maggi[5]
[1] Istituto di Matematica applicata e tecnologie informatiche IMATI-CNR, Italy – herbert.natta@ge.imati.cnr.it
[2] Istituto per il Lessico intellettuale italiano ed europeo ILIESI-CNR, Italy – michela.tardella@cnr.it
[3] Istituto per il Lessico intellettuale italiano ed europeo ILIESI-CNR, Italy – eleonora.lattanzi@iliesi.cnr.it
[4] Istituto di Matematica applicata e tecnologie informatiche IMATI-CNR, Italy – gianluca.rossi@ge.imati.cnr.it
[5] Istituto di Matematica applicata e tecnologie informatiche IMATI-CNR, Italy – maggi@area.ge.cnr.it

ABSTRACT (ITALIANO)
L'attività di ricerca che qui presentiamo è stata svolta nell'ambito del progetto *Portale delle fonti per la storia della Repubblica italiana*, un'iniziativa connotata da importanti implicazioni teoriche e metodologiche, ma anche da notevoli risvolti civili. Attraverso la collaborazione tra tre istituti del Consiglio Nazionale delle Ricerche - CNR, diversi enti pubblici e dodici istituti di cultura privati afferenti all'Associazione delle istituzioni di cultura italiane - AICI, si è cercato di elaborare un sistema concettuale e tecnologico capace di integrare risorse fra loro eterogenee, sia nei metodi di descrizione che nei formati e nelle tecnologie di trasmissione e condivisione dei dati. Il presente contributo si concentrerà sulle soluzioni individuate per gestire tale varietà, in particolare nel processo di acquisizione dei dataset provenienti dalle istituzioni di cultura private, come la progettazione e lo sviluppo di una *pipeline* ETL flessibile, capace di ricondurre la varietà dei dati in *input* al modello logico del database del *Portale*, nonché l'elaborazione di un vocabolario controllato dei livelli di descrizione archivistica.
Parole chiave: patrimonio culturale; descrizione archivistica; sistemi informativi; vocabolari controllati

ABSTRACT (ENGLISH)
Entities in relationship: policies, technological solutions and lexical models for an integrated information (eco)system - The research activity presented here was carried out within the project *Portale delle fonti per la storia della Repubblica italiana*, an initiative characterized by important theoretical and methodological implications, but also by notable civil ones. Through the collaboration between three institutes of the italian National Research Council, several public bodies and twelve private cultural institutes belonging to the Associazione delle istituzioni di cultura italiane - AICI, an attempt was made to develop a conceptual and technological system capable of integrating heterogeneous resources, both in the description methods and in the formats and technologies for data transmission and sharing. This contribution will focus on the solutions identified to manage this variety, in particular in the process of acquiring datasets from private cultural institutions, such as the design and development of a flexible ETL pipeline, capable of bringing the variety of input data back to the logical model of the database, as well as the development of a controlled vocabulary of the archival description levels.
Keywords: cultural heritage; archival description; information systems; controlled vocabularies

1. INTRODUZIONE
Il progetto *Portale delle fonti per la storia della Repubblica italiana*, ha rappresentato una sfida tanto culturale quanto tecnologica, con importanti implicazioni civili, teoriche e metodologiche, e che vede la collaborazione tra tre istituti del CNR, enti pubblici, e dodici istituti di cultura privati afferenti all'Associazione delle istituzioni di cultura italiane - AICI (cfr. Tardella *et al*, 2024).
Il progetto nasceva dalla ferma convinzione di creare uno strumento rivolto non solo ad un'utenza esperta, ma anche e soprattutto al pubblico cosiddetto "generalista", composto da non addetti ai lavori, dagli studenti delle scuole, da chi si approccia per la prima volta al mondo degli archivi.
Conseguenza naturale di questa vocazione, che potremmo dire civile, e di questa apertura al dialogo con tipologie di utenti dai profili eterogenei, è stata la creazione di uno strumento semplice e di immediata comprensione, che garantisse l'apertura e la condivisione dei dati, delle descrizioni e delle tecnologie, ma che tenesse in considerazione le esperienze pregresse e coeve in campo archivistico.

Uno strumento semplice, ma tecnologicamente evoluto che permettesse di fruire del patrimonio culturale, con particolare riguardo a quello archivistico, conservato dagli enti pubblici[1] e dagli istituti privati e messo a disposizione di questo specifico progetto, attraverso il quale si mira a rendere più agevole la consultazione dei documenti degli organi costituzionali e degli apparati amministrativi dello Stato, insieme a quelli prodotti e conservati dalle associazioni private. L'impegno principale è stato rivolto all'elaborazione di un sistema concettuale e tecnologico capace di integrare risorse caratterizzate da una notevole eterogeneità, sia nei metodi di descrizione che nei formati e nelle tecnologie di trasmissione e condivisione dei dati. In questo scenario, il *Portale* ambisce a diventare un ecosistema nel quale giungere ad una integrazione dei dati e dei sistemi, superando quel "particolarismo informativo" che caratterizza l'eterogeneità e la frammentarietà delle informazioni e che si riscontra anche nella proliferazione dei siti per la ricerca archivistica (Cfr. Cacioli, 1996). Per perseguire questo obiettivo, i gruppi di lavoro, a carattere fortemente interdisciplinare (si compongono infatti di archivisti, bibliotecari, informatici, filosofi ed esperti di linguaggi), hanno sviluppato strategie volte all'armonizzazione di patrimoni archivistici, dei modelli concettuali e dei formati di rappresentazione dei dati. Il presente contributo si concentra, in particolare, sulle soluzioni teoriche, metodologiche e tecnologiche adottate per l'integrazione dei dati in un unico modello logico e per la elaborazione di un vocabolario controllato dei livelli di descrizione archivistica.

2. LE ISTITUZIONI DI CULTURA, I FONDI, I SISTEMI INFORMATIVI

La partecipazione delle istituzioni di cultura afferenti all'AICI si è svolta secondo i criteri definiti dall'accordo quadro stipulato tra l'Associazione stessa e il CNR nel luglio del 2022. L'accordo, finalizzato a favorire la cooperazione tra le parti per la realizzazione di iniziative progettuali - in aree tematiche di comune interesse e, soprattutto, nella collaborazione al progetto *Portale delle fonti per la storia della Repubblica italiana* - prevedeva l'indizione di un bando per la presentazione di manifestazioni di interesse. La selezione delle proposte pervenute si è basata su una serie di criteri, secondo i quali i fondi acquisibili dovevano rappresentare i diversi orientamenti politici, sociali e culturali dell'Italia repubblicana, nonché la varietà tipologica delle fonti (archivistiche e bibliografiche). Per rispondere agli obiettivi di progetto, sono stati ritenuti di notevole interesse gli archivi contenenti materiali relativi all'attività istituzionale del soggetto produttore e, in particolare, quelli non ancora digitalizzati. Sono stati inoltre presi in considerazione sia la consistenza dei fondi in rapporto ai tempi del progetto e alle risorse finanziarie disponibili sia, ai fini della sperimentazione tecnologica, i software utilizzati e i formati di interscambio.
Le istituzioni di cultura selezionate e i fondi proposti per la partecipazione al progetto sono indicati nella tabella 1.

ISTITUTO STURZO	Fondo Francesco Bartolotta
	Archivio Giulio Andreotti (2 serie)
FONDAZIONE GRAMSCI	Archivio Mosca
FONDAZIONE UGO SPIRITO E RENZO DE FELICE	Fondo Nino Tripodi
FONDAZIONE LUIGI MICHELETTI	Fondo Repubblica Sociale italiana
FONDAZIONE PASTORE	Fondo Giulio Pastore
	Fondo Lamberto Giannitelli
FONDAZIONE LELIO E LISLI BASSO	Fondo Ada Alessandrini
	Fondo Lelio Basso (2 serie)
ISTITUTO NAZIONALE FERRUCCIO PARRI	Fondo Ferruccio Parri
	Carte Ferruccio Parri
FONDAZIONE GRAMSCI TORINO	Fondo Partito comunista italiano - Federazione di Torino
	Fondo Unione donne italiane

[1] L'Archivio Storico della Presidenza della Repubblica e l'Archivio Storico della Camera hanno partecipato attraverso dataset già disponibili ed esposti come linked open data. L'Archivio Storico del Senato ha invece condiviso dati in formato Csv, relativi in particolare all'anagrafica dei Senatori. Per informazioni di dettaglio si veda Tardella *et al.*, 2024.

FONDAZIONE GRAMSCI EMILIA ROMAGNA	Fondo Triumvirato insurrezionale
	Fondo Giuseppe Dozza
ARCHIVIO AUDIOVISIVO DEL MOVIMENTO OPERAIO E DEMOCRATICO	Collezioni film e audiovisivi
FONDAZIONE GIUSEPPE E SALVATORE TATARELLA	Fondo Maselli Campagna
	Fondo Fotografico
FONDAZIONE UGO LA MALFA	Fondo Fotografico
	Fondo La voce della Donna
	Fondo Ugo La Malfa
	Fondo La Malfa - Appendice

Si tratta di nuclei archivistici profondamente difformi gli uni dagli altri, sia per quanto riguarda la storia della sedimentazione delle carte, sia in relazione al loro trattamento archivistico. Realtà variegate cui corrispondeva un eterogeneo livello di descrizione e di maturazione tecnologica. Per sintetizzare, dati diversi con diversi gradi di analiticità descrittiva, prodotti con software diversi.
Le relazioni tra il CNR e le singole Istituzioni sono state normate attraverso convenzioni bilaterali, nelle quali sono stati formalizzati gli aspetti progettuali e fiscali, con particolare riguardo alle attività e agli obiettivi da raggiungere, ai tempi e alle condizioni di svolgimento e ai termini del supporto finanziario relativo alle attività condotte per il raggiungimento dei comuni obiettivi scientifico-culturali. Erano parte integrante delle stesse convenzioni anche alcuni allegati tecnici, contenenti: le informazioni di dettaglio relative al piano di progetto e ai relativi fondi proposti dalle Istituzioni; le linee guida per la gestione e la rendicontazione delle attività, e, soprattutto, le linee guida per la descrizione archivistica, la digitalizzazione e le policies di condivisione dei dati e degli oggetti digitali, fondamentali queste ultime per la definizione di criteri di descrizione e di modalità di digitalizzazione dei documenti condivisi.

3. PROCEDURE DI ACQUISIZIONE DEI DATI

L'eterogeneità delle fonti, degli strumenti di descrizione e dei formati di interscambio ha richiesto la progettazione e sviluppo di un sistema di componenti software (*pipeline*), per l'estrazione, trasformazione e caricamento (ETL) dati, flessibile, capace di ricondurre la varietà dei dati in *input* al modello logico del database del *Portale*, minimizzando la perdita di contenuto informativo.
L'analisi dei dati ricevuti ha messo in evidenza, nell'ampio margine di variabilità, pattern ricorrenti. In particolare, i formati e le strutture sono risultati dipendenti dagli strumenti di descrizione utilizzati e riconducibili a tre tipologie: documenti Word (.doc/.docx), cartelle di lavoro Excel (.xls/.xlsx), XML strutturati in base a diversi schemi, principalmente EAD e diverse configurazioni di xDams, la piattaforma per il trattamento, la gestione e la fruizione di archivi storici utilizzata da alcuni degli Istituti di cultura.
In aggiunta alle naturali differenze negli schemi dei dataset, che hanno richiesto la predisposizione di strumenti di *mapping*, sono state riscontrate generali difformità nelle modalità di rappresentazione dell'informazione temporale, sia date puntuali sia intervalli, e delle voci di indice, che hanno richiesto specifiche procedure di trasformazione, finalizzate, nel primo caso, a normalizzare la rappresentazione degli estremi cronologici e, nel secondo, ad agevolare l'identificazione degli agenti relazionati.
Sulla base di queste analisi preliminari, sono stati sviluppati algoritmi *ad hoc* per i diversi *dataset* acquisiti, basati su un modello architetturale comune. Si tratta di micro-applicazioni, sviluppate in linguaggio python e composte di: script principale (*importer*) eseguibile, contenente l'intera procedura ETL (dalla lettura del/dei dataset sorgenti alla trasformazione alla generazione di output, prima tabellari, per permettere la verifica dei risultati, procedendo infine all'inserimento dei dati nel database del *Portale*), file con le funzioni richiamate dall'*importer*, modificate in relazione al formato di input e alle procedure di trasformazione necessarie, file di setup, con la configurazione delle costanti (parametri di connessione, percorso dei dataset in input, eventuali parametri di mapping), file di mapping, con dizionari contenenti la corrispondenza tra la struttura di input e il modello della base dati del *Portale*, costruiti sulla base delle indicazioni definite dagli esperti di dominio.

In particolare, è stata utilizzata la libreria *pandas* per l'elaborazione dei formati tabellari, la libreria *xml* per il processamento degli XML e la libreria *docx* per i documenti Word.
La normalizzazione delle date è stata sviluppata mediante progressiva generalizzazione e adattamento di specifiche funzioni, definite nel file dedicato e basate sulla libreria *datetime*, in base alle caratteristiche dei dataset pervenuti. Le funzioni sono strutturate in una principale pubblica che, ricevendo in input la stringa contenente la data e l'indicazione relativa al fatto che si tratti di data iniziale o finale, restituisce in output la data correttamente formattata, richiamando due funzioni private: una che procede alla sostituzione delle stringhe relative ai mesi con la corrispondente notazione numerica, l'altra che integra giorno e/o mese dove assenti e riformattando la stringa in base al pattern richiesto, riconoscendo l'eventuale indicazione di intervalli temporali (e distinguendo quindi la data iniziale da quella finale).
Il trattamento delle voci di indice ha invece richiesto un'elaborazione più complessa, dovuta, da un lato, all'eterogeneità delle fonti e, dall'altro, alla loro rilevanza per l'attivazione di relazioni semantiche trasversali all'interno del sistema Portale. L'approccio metodologico adottato ha previsto l'entificazione dei nomi di enti/persone/famiglie/congressi (agenti) e luoghi mediante la creazione di una scheda dedicata per ogni voce. In questo modo è stato possibile sia riferire le risorse acquisite a un'entità univoca sia dare modo agli operatori di arricchire la semplice denominazione con altri dati relativi all'entità rappresentata.
Per trattare la variabilità delle voci di indice si è però resa necessaria un'elaborazione in più fasi, inclusiva sia di procedure automatiche sia di una validazione *expert based* dei risultati parziali.
In particolare, i casi più ricorrenti sono stati: i) voci di indice distinguibili nella struttura del dataset tramite specifici elementi (es. <controlaccess>), tipologia distinguibile (es. <persname>, <corpname>, <geogname>) e denominazione formattata in modo uniforme (es. 'Cognome, Nome'); ii) voci di indice distinguibili, ma non è distinguibile la tipologia e/o la denominazione non è formattata in modo uniforme.
Per la prima condizione si è generata la scheda corrispondente, riportando gli elementi della denominazione nei relativi campi e mantenendo l'identificativo (dove presente) del sistema di origine; per la seconda, si sono elaborate le voci di indice tramite librerie specifiche (quali *names_dataset*) per ricostruire, tramite ranking[2], la probabilità che un record corrisponda a nome o cognome di persona o a denominazione di ente. Prima di importare i dati, il risultato è verificato manualmente per minimizzare l'errore. Come si evince, se generalmente le voci di indice sono identificabili nei dataset importati[3], il riconoscimento della tipologia o della formattazione della denominazione, funzionale a strutturare il dato coerentemente con il sistema di destinazione (distinguendo per esempio nome, cognome, qualifica per le persone, denominazione e ulteriori denominazioni per gli enti, denominazione, luogo e data di svolgimento di congressi).
Il numero complessivo delle voci di indice importate è di 32.014 agenti, dei quali 23.614 persone, 8.355 enti, 38 congressi, 7 famiglie, e 9.938 luoghi, dei quali 6.524 riconducibili a comuni italiani, 45 stati, 19 regioni italiane e 3.354 non classificati (es. microtoponimi, città o regioni estere, ecc.).
Per gli agenti, nonostante la strategia di deduplicazione adottata in fase di importazione, che ha previsto la verifica della corrispondenza con agenti già precedentemente importati (in questi casi non è stata creata una nuova scheda, ma attivato un nuovo collegamento a schede esistenti), sono risultate inserite nel sistema 1.828 denominazioni univoche associate a più di un identificativo.
In questi casi, si è deciso di mantenere la duplicità dei record, uniformandoli però attraverso la relazione allo stesso identificativo in sistemi di rappresentazione esterni (VIAF, Geonames, Wikidata). Questa procedura di arricchimento, che ha riguardato l'intero corpus di voci importate, è stata sviluppata attraverso un flusso di lavoro semiautomatico, che ha previsto, in una prima fase, la definizione ed esecuzione di uno script di *matching*, basato sull'interrogazione delle API utilizzando come parametri gli attributi noti delle entità (toponimo, denominazione, nome e cognome, qualifica, ecc.). Le associazioni ottenute, sono state poi oggetto, in seconda fase, di una revisione *expert based*, volta a risolvere ambiguità, duplicazioni e falsi positivi.
La necessità, da un lato, di garantire la coerenza dei dati importati rispetto alle fonti e, dall'altro, di metterli a sistema in uno strumento di descrizione integrato, nel quale i dati potessero convivere

[2] Nei casi di formattazione non uniforme, la denominazione è stata tokenizzata, ogni token è stato classificato come first name (FN), last name (LN), mixed (M) in base al ranking generato da names_dataset, componendo così, per ogni stringa, un pattern, ulteriormente elaborato considerando il posizionamento del token nella stringa, la lunghezza del token e la presenza di elementi tipici dell'onomastica (es. 'Di', 'De' tipici di alcuni cognomi italiani).

[3] Per i documenti Word (.doc/.docx) le voci di indice sono state isolate o attraverso l'analisi preliminare del documento e l'individuazione di specifiche caratteristiche di formattazione o attraverso l'elaborazione dei marcatori di indice (index markers).

generando nuovo contenuto informativo, ha richiesto inoltre la definizione di vocabolari condivisi rispetto ai quali procedere alla normalizzazione dei dati importati.

4. I VOCABOLARI CONTROLLATI

La definizione delle entità relative al dominio archivistico utilizzate nel *Portale* ha posto la necessità di utilizzare un vocabolario controllato. Come è noto, un vocabolario controllato è costituito da una serie di lemmi organizzati e "definisce concetti (temi, stili artistici, autori) che vengono utilizzati come valori nei metadati. [...] Un vocabolario rappresenta quindi una lista chiusa di valori controllati consentiti per un elemento" (Guerrini & Possemato, 2015 : 99).
Per la realizzazione del *Portale delle fonti* è stato richiesto ai partner di conferire descrizioni molto analitiche e quindi in molti casi vengono presentati i singoli documenti che compongono i complessi documentari, come ad esempio fotografie, manifesti, cartoline etc.
Si è però ravvisato come la terminologia applicata dai diversi istituti culturali non fosse sempre semanticamente univoca, nonostante l'ambito disciplinare specifico[4], ma presentasse delle sfumature che dovevano essere uniformate per poter procedere all'attività di importazione dei dati e alla loro successiva esposizione come linked open data. Termini legati alla struttura multilivellare degli archivi si presentavano infatti caratterizzati da usi dipendenti da diverse interpretazioni e, pertanto, applicati secondo sensi e in modalità differenti, come, ad esempio, "Complesso archivistico", che rimanda ad insieme di documenti di varia natura, oppure di concetti quale "Collezione" o "Subfondo".
Questa distorsione semantica si ripercuote anche nelle diverse definizioni date a questi stessi concetti nell'ambito dei diversi standard, che risentono dell'ambiente culturale e, inevitabilmente, del contesto storico di produzione.
Sono stati quindi individuati e definiti[5] i seguenti 14 lemmi: Classe/Categoria, Collezione/Raccolta, Fondo, Unità archivistica, Sottofascicolo/Inserto, Unità documentaria, Altro livello, Complesso archivistico, Subfondo, Serie, Sottogruppo, Sottoserie, Articolazione interna al fascicolo, Parte di un documento.
Tale lavoro si colloca pienamente nella prospettiva di arricchire il modulo archivistico dell'ontologia ArCo[6] e permettere così l'esposizione nel *Portale* dei dati relativi alle risorse che sono state tipizzate tramite il vocabolario[7].
Lo studio ha preliminarmente preso in considerazione le tipologie in uso a livello nazionale validate dall'Istituto centrale per gli archivi (ICAR) e le ha confrontate rispetto a quanto prescritto da Encoded Archival Description[8], standard che garantisce l'interoperabilità fra diversi sistemi archivistici.
Successivamente è stata effettuata una ricerca nella letteratura scientifica per poter attribuire delle definizioni univoche alle tipologie da accettare nel vocabolario controllato. Dopo aver consolidato il vocabolario è stato necessario renderlo operativo anche nel sistema GeCa (Maggi, 2023; Natta, 2024) verificando quali tipologie fossero già presenti e aggiornandolo in modo da poter trattare in modo corretto i dati descrittivi importati dai diversi dataset.
Il *Portale delle fonti*, come abbiamo visto, viene popolato da dati derivati da sorgenti differenti: solo in alcuni casi sono già disponibili come linked open data mentre quelli provenienti dagli Istituti culturali che partecipano al progetto devono subire un processo di lodificazione. In questo frangente si sono riscontrate alcune criticità per quello che riguarda i dati a valle della importazione nell'infrastruttura GeCa. Il sistema utilizza diversi standard internazionali per la lodificazione a seconda del dominio specifico. La prima problematica affrontata riguarda il fatto che i termini del vocabolario implementato devono avere una corrispondenza con l'ontologia relativa alle risorse archivistiche, Records in contexts (RiC-O), versione 1.0[9].
Questa ontologia è basata su un modello concettuale che prevede una entità "Record Resource"[10] che si

[4] Si noti già la polisemia del termine "archivio", che può essere riferito al complesso di documenti, all'istituto di conservazione o a quella parte di un edificio destinata al deposito della documentazione e che richiede dunque una disambiguazione. Per una estesa riflessione sul termine archivio si vedano Ciandrini *et al*, 2023; Carucci, 2010; ISAD(G), 2000:10-11.
[5] Le fonti di riferimento per le definizioni in italiano sono il Glossario della Direzione Generale Archivi, il volume di Paola Carucci e Mariella Guercio, *Manuale di archivistica* (Roma, 2021) e le Norme Internazionali per la Descrizione Archivistica. Per la versione in inglese, sono stati invece utilizzati il Dictionary of Archives terminology, l'Encoded Archival Description Tag Library e la versione inglese delle le Norme Internazionali per la Descrizione Archivistica.
[6] Vedi: http://wit.istc.cnr.it/arco (cons: 07/01/2025).
[7] In tal senso il passo successivo sarà quello di rispettare le raccomandazioni previste per la pubblicazione dei dati in *linked open data*. Si veda: W3C, 2014.
[8] Vedi: https://www.loc.gov/ead/EAD3taglib/EAD3-TL-eng.html#attr-level (cons: 07/01/2025).
[9] Vedi: https://www.ica.org/resource/records-in-contexts-ontology/ (cons: 07/01/2025).
[10] Rispetto allo standard di descrizione precedente, ISAD(G), questa entità "is conceptually comparable to unit of description" (*Records In Contexts – Conceptual Model Version 1.0:* 20).

distingue in Record Set, Record, Record Part. In particolare, per i Record set esiste una ulteriore categorizzazione che nell'ontologia di Records in contexts definisce quattro tipi[11] (ICA, Expert group on archival description, 2023):

- Fonds, ovvero "the whole of the records, regardless of form or medium, organically created and/or accumulated and used by a particular person, family, or corporate body in the course of that creator's activities and functions";
- Series, ovvero "documents arranged in accordance with a filing system or maintained as a unit because they result from the same accumulation or filing process, or the same activity; have a particular form; or because of some other relationship arising out of their creation, receipt, or use. A series is also known as a records series";
- Collection, ovvero "an artificial assemblage of documents accumulated on the basis of some common characteristic without regard to the provenance of those documents. Not to be confused with an archival fonds";
- File, ovvero "an organized unit of documents grouped together either for current use by the creator or in the process of archival arrangement, because they relate to the same subject, activity, or transaction. A file is usually the basic unit within a record series".

Il 14 lemmi del vocabolario sopra elencati sono stati dunque allineati e implementati nell'infrastruttura e ricondotti alle differenti entità.
L'altra problematica in fase di risoluzione riguarda invece il rapporto tra Records in contexts e ArCo, l'ontologia del *Portale delle fonti*, e il modo in cui quest'ultima deriva queste sei tipizzazioni senza perdere l'informazione presente nel dato originario, che invece ne distingue dodici.

CONCLUSIONI

L'acquisizione e integrazione di dati da fonti eterogenee in un unico sistema informativo ha richiesto l'adozione di strategie operative, soluzioni tecnologiche e scelte teorico-metodologiche volte a massimizzare l'attivazione di relazioni latenti e, in generale, l'interoperabilità del sistema, minimizzando la perdita di contenuto informativo e l'ambiguità semantica.
La strada aperta da questo lavoro ha il chiaro e naturale obiettivo, orientato dal paradigma dei *linked open data*, di formalizzare le scelte semantiche effettuate nella definizione del vocabolario, integrandole stabilmente nell'ontologia del *Portale delle fonti*, basata su ArcO, la rete di ontologie per la descrizione del patrimonio culturale a livello nazionale.
L'auspicabile prosecuzione del progetto, oltre a prevedere un ampliamento rispetto all'arco cronologico preso in esame (1943-1953) e quindi ai contenuti, dischiude anche ulteriori prospettive.
Il *Portale* si è rivolto principalmente al patrimonio archivistico proponendo fonti custodite da Istituti culturali che si presentavano, seppur in modo vario e a un diverso livello di conformità agli standard, come dataset riconducibili a modelli di descrizione archivistica. Risulterebbe interessante un'integrazione di queste fonti, costituite per la maggior parte da documenti originali, con le risorse bibliografiche, tra cui gli studi e le ricerche condotte sin dalla nascita della Repubblica. In particolare si possono prendere in considerazione le pubblicazioni del Quirinale, della Camera e del Senato, come ad esempio la "Bibliografia del Parlamento"[12], che comprende anche i singoli contributi analitici presenti nelle pubblicazioni, suddivisi per argomento. Alla modellazione dei dati secondo ontologie specifiche del dominio bibliografico, così come fatto rispetto al dominio archivistico, si aggiungerebbe la possibilità di relazionare le risorse con il catalogo del Servizio bibliotecario nazionale per consentire all'utente il loro reperimento, nel caso non sia accessibile una loro versione in formato digitale.
Un'ulteriore evoluzione del Portale potrebbe riguardare i prodotti di natura divulgativa, audio e video, realizzati specificamente per questo progetto. Si tratta di una serie di videointerviste e podcast che contestualizzano le risorse documentali e propongono alcune tematiche fondamentali per tracciare la storia della Repubblica italiana. Questi prodotti potrebbero quindi essere oggetto di una metadatazione dettagliata e dello sviluppo di funzionalità che permettano la realizzazione di sottotitoli in più lingue, per gli audiovisivi, e la trascrizione degli interventi.Queste prospettive dunque rientrano nell'orizzonte di un progetto che non si limita a mettere a disposizione una raccolta di fonti per la storia politica e istituzionale nazionale relativa alla seconda metà del Novecento, ma mira a fornire strumenti di approfondimento orientati ad ampliare la fruizione da parte di un pubblico diversificato, favorendo e migliorando l'accessibilità alle risorse e al loro contenuto.

[11] Questi tipi riprendono quanto definito da ISAD(G), cui si rimanda.
[12] Vedi: https://storia.camera.it/bpr/ (cons: 09/04/2025).

BIBLIOGRAFIA

Cacioli, M. (Eds.). 1996. Gli archivi dei partiti politici: atti dei seminari di Roma, 30 giugno 1994, e di Perugia, 25-26 ottobre 1994. Roma: Ministero per i beni culturali e ambientali, Ufficio centrale per i beni archivistici.

Carucci, P. (2010). Le fonti archivistiche: ordinamento e conservazione, Roma: Carocci.

Ciandrini, P., Lattanzi, E., Maggi, R., Tardella, M. (2023). Archivi e contaminazioni disciplinari: dai linguaggi ai modelli, dai metodi alle tecniche. Migrazioni e contaminazioni tra le scienze. Metodi e linguaggi interdisciplinari. Laureti, S., Marras, C., & Peddis, D. (Eds.). Collana PLURIMI – IV, 2023. CNR Edizioni. ISBN 978 88 8080. DOI 10.36173/PLURIMI-2023-4

Guerrini, M., & Possemato, T. (2015). Linked data per biblioteche, archivi e musei. Editrice Bibliografica. ISBN 978-88-7075-830-6

ICA, Expert group on archival description (November 2023). Records in Contexts conceptual model. Version 1.0. International Council on Archives
https://www.ica.org/app/uploads/2023/12/RiC-CM-1.0.pdf

ISAD(G), 2000. General international standard archival description adopted by the Committee on Descriptive Standards, Stockholm, 19-22 September, Ottawa.
https://icar.cultura.gov.it/fileadmin/risorse/docu_standard/RAS_2003_1.pdf

Maggi, R., Pasciuto, T., Mazzoleni, M., Artese, M.T., Gagliardi, I., & Albertoni R. (2023). GECA 3.0 - A new tool for cataloguing and enjoying cultural heritage. La memoria digitale. Proceedings del XII Convegno Annuale AIUCD, Siena, 5-7/06/2023, 978-88-942535-7-3.

Natta, H., Rossi, G., Maggi, R. (2024). Luoghi comuni: metodi e strategie di sviluppo software in ambito GLAM, dalle voci di autorità all'esplorazione cartografica. Me.Te. Digitali. Mediterraneo in rete tra testi e contesti. Proceedings del XIII Convegno Annuale AIUCD, Catania 28-30 maggio 2024, Università di Catania. Di Silvestro, & A., Spampinato D. (Eds.). ISBN 978-88-942535-8-0. DOI 10.6092/unibo/amsacta/7927

Pavone, C., & D'Angiolini, P. (1981-1994). Guida generale degli archivi di Stato italiani. Roma: Ministero per i beni culturali e ambientali, Ufficio centrale per i beni archivistici, voll. 1-4.

Tardella, M., Maggi, R., Lodi, G., Albertoni, R., Natta, H., Rossi, G., Pasciuto, T., Ciandrini, P., Sinopoli, L., Artese, M.T., Gagliardi, I., Lattanzi, E., Ventroni, S., Tizzoni, E., Russo, A., & Porena, M. (2024). Un futuro per la memoria. Strumenti, modelli e sinergie per l'integrazione dei dati nel Portale delle fonti per la storia della Repubblica italiana. Me.Te. Digitali. Mediterraneo in rete tra testi e contesti. Proceedings del XIII Convegno Annuale AIUCD, Catania 28-30 maggio 2024, Università di Catania. Di Silvestro, & A., Spampinato D. (Eds.). ISBN 978-88-942535-8-0. DOI 10.6092/unibo/amsacta/7927

W3C (2014). Best Practices for Publishing Linked Data. W3C Working Group Note. Hyland, B., Atemezing, G., Villazón-Terrazas, B. (Eds.). World Wide Web Consortium
https://www.w3.org/TR/ld-bp/