

Atti del XV Convegno Annuale **AIUCD**

Digitale *e Public Engagement*

Pratiche e prospettive nelle Digital
Humanities

Cagliari 3-4-5 giugno 2026

a cura di

Cristina **Marras** | Andrea **Pergola** | Giampaolo **Salice**

ISBN 9791298618817



Copyright ©2026 AIUCD

Associazione per l'Informatica Umanistica e la Cultura Digitale



Il presente volume e tutti i contributi sono rilasciati sotto licenza Creative Commons Attribution ShareAlike 4.0 International license (CC-BY-SA 4.0). Ogni altro diritto rimane in capo ai singoli autori.

This volume and all contributions are released under the Creative Commons Attribution Share-Alike 4.0 International license (CC-BY-SA 4.0). All other rights retained by the legal owners.

A cura di: Cristina Marras; Andrea Pergola; Giampaolo Salice (2026). *Digitale e Public Engagement: pratiche e prospettive nelle Digital Humanities*, Atti del XV Convegno Annuale AIUCD, Cagliari 3-5 giugno 2026, Università degli Studi di Cagliari.

Ultimo accesso agli URL in data 13 maggio 2026.

Last URL access May, 13 2026.

Il presente volume è stato modificato in data 31 luglio 2026 per errata corrige. La revisione riguarda l'aggiunta dell'indicizzazione del nome e del ruolo di Paolo Bonora quale co-autore del contributo «Un grafo per la ricerca storico-filologica in archivio: il caso del fondo Giuseppe Albini».

This volume was updated on 31 July 2026 to incorporate an erratum. The revision consists of the addition of Paolo Bonora to the index and author attribution as co-author of the contribution «Un grafo per la ricerca storico-filologica in archivio: il caso del fondo Giuseppe Albini».

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 del Convegno AIUCD 2026 A disponibile online:

<https://www.aiucd2026.unica.it/companion/>

The AIUCD 2026 Conference Program is available online:

<https://www.aiucd2026.unica.it/companion/>

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 2026.

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 2026.

Gli Atti del Convegno AIUCD 2026 sono pubblicati come raccolta dei contributi forniti direttamente dagli autori e dalle autrici. I file sono stati raccolti e assemblati senza interventi redazionali significativi da parte dei curatori.

The proceedings of the AIUCD 2026 Conference are published as a collection of contributions provided directly by the authors. The files have been collected and compiled without significant editorial intervention by the editors.

Il logo di AIUCD 2026 A opera di Raffaele Argiolas

The AIUCD 2026 logo was designed by Raffaele Argiolas

La copertina A stata realizzata da Giampaolo Salice

The cover was created by Giampaolo Salice

GENERAL CHAIRS

Giampaolo Salice | Università degli Studi di Cagliari | DH UNICA

Cristina Marras | CNR-ILIESI | AIUCD

COMITATO ORGANIZZATORE / ORGANIZING COMMITTEE

Il Comitato Organizzatore è composto da Giampaolo Salice (Università degli Studi di Cagliari), Cristina Marras (CNR-ILIESI), Raffaele Argiolas, Alessandro Capra, Giommara Carboni, Andrea Pergola, Eleonora Todde (DH UNICA Università degli Studi di Cagliari), Christian D'Agata (Università di Catania), Francesca Frontini (CNR-ILC), Simone Rebora (Università di Verona), Marco Rospoher (Università di Verona), Laura Stochino (ISSASCO), Matteo Tatti e Alice Nozza (Associazione Itzokor).

Iscrizioni / Conference Registration

Eleonora **Todde** – DH UniCa – Università degli Studi di Cagliari

Atti / Conference Proceedings

Andrea **Pergola** – DH UniCa – Università degli Studi di Cagliari

Comunicazione / Communication

Raffaele **Argiolas** – DH UniCa – Università degli Studi di Cagliari

Promozione / Promotion

Giommara **Carboni** – DH UniCa – Università degli Studi di Cagliari

Tecnologo dell'Evento / Event Technologist

Alessandro **Capra** – DH UniCa

Organizzazione locale

Dipartimento di Lettere, Lingue e Beni Culturali, Università degli Studi di Cagliari, Ufficio Amministrativo:
Serena Serra, Giulia Cadoni, Claudia Serri

COMITATO SCIENTIFICO / SCIENTIFIC COMMITTEE

Il Comitato Scientifico del convegno AIUCD 2026 è composto dai Chairs di Programma e dai Chairs delle cinque aree tematiche.

Program Chairs

Giampaolo Salice | Università degli Studi di Cagliari | DH UNICA

Cristina Marras | CNR-ILIESI | AIUCD

DH e co-costruzione del sapere con le comunità: sfide, metodi, strumenti / *DH and co-construction of knowledge with communities: challenges, methods, tools*

Arianna Ciula | King's Digital Lab, London

Greta Franzini | EURAC, Bolzano

Archivi ed edizioni: descrizioni aumentate, accessibilità e sistemi informativi / *Archives and editions: augmented descriptions, accessibility, and information systems*

Federico Valacchi | Università di Macerata

Paolo Monella | Università Kore Enna | AIUCD

Testualità digitali: prospettive, sviluppi e sperimentazioni / *Digital textualities: perspectives, developments, and experimentations*

Simone Rebora | Università di Verona | AIUCD

Alessandro Adamou | Biblioteca Hertziana

Rappresentazione di Dati e Conoscenza / *Data and Knowledge Representation*

Francesca Tomasi | Alma Mater Studiorum – Università di Bologna

Angelo Mario Del Grosso | CNR-ILC

Memorie, storia e patrimoni culturali digitali / *Tangible and Intangible Digital Heritage*

Michela Tardella | CNR-ILIESI

Enrica Salvatori | Università di Pisa

COMITATO DI PROGRAMMA / PROGRAM COMMITTEE

Stefano Allegrezza (Università di Macerata), Laura Antonietti (Université de Versailles Saint Quentin en Yvelines - Université Paris-Saclay), Alessio Antonini (Open University), Liborio P. Barbarino (Università di Catania), Nicola Barbuti (Università di Bari "Aldo Moro"), Stefano Bazzaco (Università di Verona), Andrea Bellandi (CNR-ILC), Giulia Benotto (CNR-ILC), Monica Berti (Leipzig University), Mario A. Bochicchio (Università di Bari "Aldo Moro"), Andrea Bolioli (Ricercatore indipendente), Marco Bombieri (Università di Verona), Flavia Bruni (Università di Chieti-Pescara), Marina Buzzoni (Università Ca' Foscari di Venezia), Alberto Campagnolo (KU Leuven), Vittore Casarosa (CNR-ILC), Raffaele Cioffi (Università di Napoli Federico II), Vincenzo Colaprice (Università di Torino), Francesca Congiu (Università degli Studi di Cagliari), Giuseppe Consolo (Università degli studi di Napoli, Federico II), Christian D'Agata (Università di Catania), Giulia D'Agostino (TU Darmstadt), Elisa D'Argenio (HUN-REN Hungarian Research Centre for Linguistics), Davide Dainese (Alma Mater Studiorum Università di Bologna), Stefano Dall'Aglio (Università Ca' Foscari Venezia), Marilena Daquino (Alma Mater Studiorum Università di Bologna), Mauro De Bari (Università di Bari "Aldo Moro"), Angelo M. Del Grosso (CNR-ILC), Matteo Di Cristofaro (UniMore), Francesca Di Donato (CNR-ILC), Giorgia Di Marcantonio (Università di Macerata), Giorgio Maria Di Nunzio (Università di Padova), Roberto Evangelista (CNR-ISPF), Stefano Ferilli (Università di Bari "Aldo Moro"), Franz Fischer (Università Ca' Foscari di Venezia), Greta H. Franzini (EURAC), Francesca Frontini (CNR-ILC), Daniele Fusi (Università Ca' Foscari di Venezia), Mariangela Giglio (Alma Mater Studiorum Università di Bologna), Tiago Luis Gil (University of Brasilia), Michela Giordano (Università degli Studi di Cagliari), Luca Giovannini (University of Potsdam), Milena Giuffrida (Università di Catania), Giovanna Granata (Università degli Studi di Cagliari), Edmondo Grassi (Università Telematica San Raffaele Roma), Miryam Grasso (Università di Catania), Piergiovanna Grossi (Università di Verona), Marina M. Guglielmi (Università degli Studi di Cagliari), Alessandro Iannella (Università degli Studi di Milano), Sabrina Iorio (Università Napoli Federico II), Michele Lacriola (Università di Siena), Alessandro Laruffa (ISSTOR), Anna Mambelli (UniMore), Francesco Mambrini (Università Cattolica del Sacro Cuore), Francesco Mameli (Università degli Studi di Cagliari), Lorenzo Mancini (CNR-ILIESI), Anna Maria Marras (Università di Torino), Cristina Marras (CNR-ILIESI), Pietro Mazzarisi (Università di Trieste), Barbara McGillivray (King's College London), Federico V. Meschini (Università della Tuscia), Alessio Miaschi (CNR-ILC), Giulia Miglietta (Università del Salento), Paolo Monella (Università Kore di Enna), Rossana Morriello (Università degli Studi di Firenze), Giulia Murgia (Università degli Studi di Cagliari), Enrico Natale (infoclio.ch), Sebastiana Nocco (CNR-ISEM), Serge Noiret (AIPH), Giuseppe Palazzolo (Università di Catania), Mafalda Papini (CNR-ILC), Enrico Pasini (Università di Torino), Giulia Pedonese (CNR-ILC), Paola Peratello (Università Ca' Foscari Venezia), Andrea Pergola (Università degli Studi di Cagliari), Ginevra Peruginelli (CNR-IGSG) University of Groningen, Federico Pianzola (University of Groningen), Fabio C. Pinna (Università degli Studi di Cagliari), Igor Pizzirusso (AIPH), Tiziana Pontillo (Università degli Studi di Cagliari), Valeria Quochi (CNR-ILC), Lisa Reggiani (CNR-ILIESI), Giulia Renda (Alma Mater Studiorum Università di Bologna), Pietro Restaneo (CNR-ILIESI), Dario Rodighiero (University of Groningen), Roberto Rosselli Del Turco (Università di Torino), Federica Rovelli (Università di Pavia), Giampaolo Salice (Università degli Studi di Cagliari), Enrica Salvatori (Università di Pisa), Emilio M. Sanfilippo (CNR-ISTC), Eva Sassolini (CNR-ILC), Manfredi Scanagatta (UniMore), Andrea Schimmenti (Alma Mater Studiorum Università di Bologna), Flavia Sciolette (CNR-ILC), Alessia Scognamiglio (CNR-ISPF), Luigi Serra (CNR-ISEM), Pietro Sichera (CNR-ILIESI), Daria Spampinato (CNR-ISTC), Giulia Speranza (Università di Napoli "L'Orientale"), Rachele Sprugnoli (Università di Parma), Francesco V. Stella (Università di Siena), Timothy Tambassi (Università Ca' Foscari di Venezia), Mirko Tavosanis (Università di Pisa), Francesca Tomasi (Alma Mater Studiorum Università di Bologna), Simona Turbanti (Università di Milano), Nicoletta Usai (Università degli Studi di Cagliari), Marco Venuti (Università di Catania), Gennaro Vessio (Università di Bari "Aldo Moro"), Gabriele Vezzani (Università di Verona), Simone Zenzaro (CNR-ILC).

Enti organizzatori / Organisers

Il XV convegno annuale dell'Associazione per l'Informatica Umanistica e la Cultura Digitale (AIUCD) è organizzata dal DH UNICA - Centro Interdipartimentale per l'Umanistica Digitale e dal Dipartimento di Lettere, Lingue e Beni Culturali dell'Università di Cagliari in collaborazione con AIUCD.

Sommario

PREMESSA

I-IV

Cristina Marras, Andrea Pergola, Giampaolo Salice

[1/A] INTELLIGENZA ARTIFICIALE GENERATIVA E RIFLESSIONE TEORICA

Una Agenda sull'AI generativa per Digital Humanities Fabio Ciotti	6
L'errore dell'IA come euristica per una deliberazione assistita Valerio De Luce; Tiberio Uricchio	14
Verso una memoria artificiale? Il progetto Johannes tra questioni tecnologiche, archeologiche e morali Stefano Bertoldi	21
AI confini della realtà: il pensiero umano nell'epoca della sua riproducibilità generativa Daniele Silvi	31

[1/B] CORPORA DIGITALI: REQUISITI E SOLUZIONI GESTIONALI

<i>Verba volant, emails manent</i> : una proposta di metodologia per l'acquisizione e la preservazione degli archivi di posta elettronica di personaggi illustri Alberto Abis, Stefano Allegrezza, Fabrizio Stupino	38
Sfide alla sostenibilità dei progetti di digitalizzazione: i casi studio degli archivi manicomiali calabresi Emanuela Nicole Donato, Maria Natalia Federico, Grazia Serratore	45
A technical solution for a digital cultural heritage corpus: Notes on collaboration, requirements and sustainability Arianna Ciula, Miguel Vieira, Geoffroy Noel, Neil Jakeman, Zihao Lu, Tiffany Ong, Simona Stoyanova, Jonathan Prag, Alessia Coccato, Ryan Heuser	53

[1/C] CO-PRODUZIONE DI ARCHIVI E COLLEZIONI DI COMUNITÀ

L'Archivio Digitale di Vigne Nuove. Costruire un archivio di comunità, tra teoria e pratica Manfredi Scanagatta	62
Participatory Digital Archives as a Method for Co-constructing Cultural Memory: The Bunjevkje Women's Platform Sandra Iršević PhD	68
Il ritorno dei partigiani sardi. Una collezione digitale co-prodotta Walter Falgio, Laura Stochino, Matteo Quarantiello	73
Dalla materialità delle fonti al dato condiviso: metodi ibridi per un archivio digitale partecipativo Tiziana Pasciuto	78

[2/A] LARGE LANGUAGE MODELS: APPLICAZIONI UMANISTICHE

Scholarly Opinion Mining using LLMs and Knowledge Graphs: the case of the Van den vos Reynaerde	87
Andrea Schimmenti, Joris van Zundert, Fabio Vitali, Marieke van Erp	
Un LLM per le Scienze Umane e Sociali? ReSearch_SSH come esperimento europeo	95
Adam Faci, Alessio Miaschi, Anne Combe, Pascal Cuxac ⁴ , Francesca Frontini, Nicolas Larrousse, Stéphane Pouyllau	
L'AI a servizio della Musicologia: costruzione di thesauri musicologici con il supporto dei Large Language Models	102
Manuela Grillo, Paolo Bonora	
Per una formazione del filologo computazionale del futuro: Domain Specific Language e Large Language Model	107
Christian D'Agata, Angelo Mario Del Grosso, Federico Boschetti	

[2/B] OPEN SCIENCE, SOSTENIBILITÀ E PRATICHE APERTE

Are Digital Humanities really committed to open? An exploratory study on the availability of methodological workflows and open peer review practices	116
Silvio Peroni	
Monitorare la transizione: l'Osservatorio sulle pratiche di Open Science nelle SSH come strumento di Public Engagement	122
Pamela Barletta, Marta Caradonna, Nicola Giampietro, Alice Orrù, Federico Silvestri	
Nel 25° anniversario di Wikipedia: l'ecosistema Wikimedia e le Digital Humanities tra pratiche collaborative e sostenibilità della conoscenza	130
Piergiiovanna Grossi	
Una pausa e tre domande. La crisi della memoria nell'era dell'accesso	137
Enrica Salvatori, Federico Valacchi	

[2/C] MODELLI DIGITALI PER GESTIRE E CONDIVIDERE FONTI MULTIDISCIPLINARI

Digital Repatriation in Practice: Sharing Ethnographic Photography with Communities of Origin at the Náprstek Museum	142
Daniela Šnapková	
Una storia da ascoltare. Descrivere, mappare, connettere: un modello per la valorizzazione degli archivi musicali	147
Martina Gremignai, Luca Andrea Ludovico	
LETSDigit: Archivio digitale della letteratura a Trieste	154
Alina Jill Simeone, Cristina Fenu, Marina Buzzoni, Roberto Rosselli Del Turco	
La mostra <i>Coerenza In Coerenza</i> (1984) oltre la fotogrammetria: ricostruzione digitale tramite reti neurali	161
Filippo Yahia Masri, Maria Carmeliti	

[3/A] AI E NLP PER TESTI E RECUPERO DELL'INFORMAZIONE

Creating a new gender-fair Italian morphological classifier based on UmBERTo fine-tuning	168
Irene Caiazza, Giovanna Maria Dimitri	
Assessing LLM Capabilities in Cultural Heritage: Recent Trends and Gaps in Benchmarks	175

Remo Grillo, Gianmarco Spinaci, Lukas Klic, Giovanni Colavizza

L'intelligenza artificiale per il potenziamento dell'information retrieval nei discovery tool delle biblioteche 183

Ilaria Belvedere, Simona Turbanti

Sinergie ludonarrative: l'evoluzione del racconto videoludico tra divulgazione, tecnica ed intelligenza artificiale 191

Giulia Angelini, Tommaso Mazzoli

[3/B] DH PER STORIA, DIRITTO E ARTE

Archivi ibridi di persona e preservazione forense dei materiali nativi digitali 198

Mariangela Giglio, Adele Gorini

Digital Humanities e studi neogreci: il progetto della Biblioteca Digitale dell'Osservatorio "Mario Vitti" 204

Georgios Katsantonis

artresearch.net: A Research Infrastructure for Exploring Interpretive Plurality in Art-Historical Photo Archives 211

Marilena Daquino, Francesca Mambelli, Alessandro Adamou, Pietro Maria Liuzzo, Stefanie Schneider, Spyros Koulouris, Artem Kozlov, Rafael Brundo Uriarte

Legislative Changes Concerning the Reproduction of Images of Italian Cultural Heritage Assets between 2023 and 2025 217

Marco Ciurcina, Piergiiovanna Grossi

[3/C] ANALISI COMPUTAZIONALE DEL TESTO LETTERARIO

Flaiano postcoloniale? Un approccio computazionale alla *vexata quaestio* di *Tempo di uccidere* 226

Silvia Lilli, Daniel Raffini

Verso una rappresentazione del macrotesto: filologia d'autore e analisi quantitativa nei racconti di Fausta Cialente 234

Emmanuela Carbè

Analisi fluidodinamica dei campi semantici in testi letterari: il Numero di Reynolds come modello di turbolenza testuale 241

Pietro Sichera, Carla Perrone, Antonio Sichera

Dentro l'Officina del Testo: Modellare Strutture Letterarie Complesse 248

Enrica Bruno

[4/A] STUDI LETTERARI DIGITALI E DISTANT READING

Tracing Intertextuality at Scale. A Network Study of Poetic Influence 255

Gabriele Vezzani

The Landscape of Italianistica: Italian Literary Studies and Digital Humanities in Semantic Space 261

Maria Levchenko

L'Edicola della Storia di Trieste: NLP e *distant reading* Per il confronto tra periodici storici italiani e sloveni nella Trieste del 1902 267

Cristina Fenu, Giovanni Pinna

Literary Tropes as Prescriptive Devices: A Case Study from BookTok 274

[4/B] MUSEI E ACCESSO DIGITALE

- Rigenerare un patrimonio desueto con il phygital: la Casa Museo Bartolo Longo 282
Mauro De Bari
- Musei senza confini: il digitale come infrastruttura di memoria partecipata nei musei delle aree interne. Il GeoMuseo Monte Arci Stefano Incani di Masullas e il Museo Paleontologico PARC di Genoni 290
Paola Palmas
- Pubblici remoti: la Realtà Virtuale come infrastruttura culturale di accesso 296
Elisa Smeraldo

[4/C] FORMAZIONE, DIDATTICA E LABORATORI DH

- Gli Hunger Games del LabCD. Un diario di sopravvivenza 304
Enrica Salvatori, Vittore Casarosa, Francesco Ricciardi
- It's Buildin' Time: A Project-based Approach for Teaching Data Management to Cultural Heritage Graduate Students 309
Sebastian Barzaghi, Alba d'Elia, Esther Montserrat Giordano, Giulia Guidarelli, Gianluca Petrosillo, Elisabetta Sabattini, Fiammetta Sabba, Lucia Sardo
- Integrazione di tecnologie per la traduzione e CLIL nella creazione di risorse didattiche per il sardo 319
Gianfranco Fronteddu, Igor Deiana

[5/A] LINKED OPEN DATA, MODELLI SEMANTICI, INFRASTRUTTURE

- Reti sociali e linked data: persone e co-occorrenze documentarie nel Portale delle fonti per la storia della Repubblica italiana 329
Herbert Natta, Gianluca Rossi, Roberta Maggi
- Web Portals on Linked Open Data: Testing Sampo-UI in Projects on Early Modern Bibliographic Data and Chinese Calligraphy and Arts 336
Arianna Moretti, Katarina Lučić, Iiro Lassi Ilmari Tiihonen, Jonas Paul Fischer
- Risorse Terminologiche e Linked Open Data: il *Glossario delle Infrastrutture di Ricerca* (GIR) come Caso di Studio 345
Lucia Francalanci, Alessia Scognamiglio, Giulia Pedonese, Michele Mallia, Irene Falini, Fahad Khan
- Orchestrazione di servizi e scenari avanzati: fruizione e necessità delle comunità di ricerca nel marketplace di H2IOSC 352
Pietro Sichera, Cristina Marras, Enrico Pasini, Vittoria Fabiani, Paolo Ongaro, Michele Scapicchi, Chiara Di Pietro

[5/B] PATRIMONIO CULTURALE: SISTEMI INFORMATIVI INTEGRATI

- Architettura e *Digital Humanities*. Conoscenza e valorizzazione del Patrimonio attraverso la costruzione di sistemi informativi integrati 361
Donatella Rita Fiorino, Caterina Giannattasio, Elisa Pilia, Valentina Pintus, Andrea Pirinu, Marcello Schirru
- Tra testo e architettura digitali per lo studio delle trasformazioni storico-artistiche dei monasteri benedettini 368

Gianmario Guidarelli, Paolo Borin, Sonia Cavicchioli, Chantal Pivetta, Renato Canearo,	
Dalla pergamena al virtuale. Manoscritti miniati nello spazio tridimensionale	375
Valeria Minisini, Bruno Fanini, Giorgio Gosti	
Modeling and visualizing palimpsests with IIIF Presentation API 3.0	382
Panagiotis Leontaridis, Giacomo Marchioro, Glen Robson	

[5/C] CROWDSOURCING E PUBLIC ENGAGEMENT

<i>Tag ANIMALx</i> : Crowdsourcing the Non-Human in Cultural Heritage Collections	389
Daniela Teixeira Gomes, Carla Vieira, Joana Vieira Paulino	
From perception to memory: Public Engagement and Neurohumanistic Approaches to the Collective Reconstruction of Archaeological Heritage	396
Grazia Solenne, Vincenza Ferrara, Elisa Corrà	
Comunicare la ricerca archeologica nell'ecosistema digitale: pratiche, metodi e sperimentazioni su Instagram	404
Marco Demuru	
Public engagement nell'era dei synthetic media: per una co-interpretazione critica tra Digital Humanities e Media Education	411
Gabriele Prosperi, Mario A. Bochicchio	

[6/A] MODELLAZIONE TESTUALE E LINGUISTICA COMPUTAZIONALE

Verso un dizionario narrativo computazionale degli usi linguistici in aree ad alta vulnerabilità sociale	420
Michela Bandini, Silvia Piccini, Andrea Bellandi, Emiliano Giovannetti	
Verso la creazione di una Treebank per il sardo	426
Nicoletta Puddu, Manuela Sanguinetti, Luigi Talamo	
A Learning-by-Doing Approach to Spoken Data Collection: The Case of the LPSP Corpus	432
Claudia Roberta Combei, Gaia Eleonora Di Raimondo, Sofia Maestri, Benedetta Romanazzi, Elena Scotti	
Digital Humanities e Letteratura Elettronica in Italia: marginalità e convergenze	440
Giuseppe Arena	

[6/B] CO-CREAZIONE, RICERCA PARTECIPATIVA, COMUNITÀ

Metodi e strumenti di sperimentazione tra co-creazione, Digital Humanities, e ricerca demo-etno-antropologica: la digitalizzazione come pratica relazionale ed etica.	448
Eleonora De Longis, Matteo Cova, Giovanni Bertelli	
Integrare strumenti digitali e conoscenze della comunità per una progettazione urbana resiliente al clima: l'iniziativa Dundrum by Design	455
Chiara Cocco, Mattia Leone, Antonio Savino, Miruna Popa, Sara Tedesco	
HCI e AI per il Public Engagement: casi di studio per la valorizzazione del patrimonio culturale	462
Samuel Aldo Iacolina, Manuela Angioni, Valentina Marotto, Francesca Mura, Piergiorgio Palla	

The Automation of Research in Digital Humanities: Artificial Researchers, Digital Archives, and the Democratization of Scholarly Access	468
Roberto Di Quirico	

[6/C] AI E AUTOMAZIONE PER ARCHIVI E PATRIMONIO

Dal problema storico alle IA: Il Database on the Slave Trade between the Mediterranean and the Atlantic (15 th -16 th Centuries)	477
Salvatore Spina, Carlo Taviani, Jörg Hörnschemeyer	
Peirce Interprets Peirce: Digitization, Automation, and Interpretation in Charles Peirce's Manuscripts	485
Alessandro Adamou, Sebastian Feil, Davide Picca, Carlo Teo Pedretti, Dario Rodighiero,	
Lorenzo Zangari	
A New Semantic Model for Mythologiae: from Data to Interpretation	494
Bianca La Manna	

[7/A] ONTOLOGIE, FAIR DATA E MODELLAZIONE DELLA CONOSCENZA

The ATLAS Guidelines for producing FAIR research products in the Digital Humanities	502
Chiara Martignano, Giorgia Rubin, Sebastiano Giacomini, Alessia Bardi, Marina Buzzoni, Marilena Daquino, Riccardo Del Gratta, Angelo Mario Del Grosso, Franz Fischer, Roberto Rosselli Del Turco, Francesca Tomasi	
RATIO: strumenti per la creazione e la gestione di risorse e collezioni FAIR nell'ambito delle scienze umane	509
Federico Silverstri, Lorenzo Mancini, Laura Baggiani, Marco Bagiacchi	
Making Uncertainty Explicit: The Preservation of Interpretive Complexity across Digital Humanities Project Workflows	516
Tommaso Battisti, Valentina Pasqual, Giulia Renda, Marilena Daquino	
Iconclass as a Semantic Framework for the Analysis of the Sacred Space	524
Polina Voronova, Alessandro Adamou	

[7/B] GIS E SPAZIALIZZAZIONI DIGITALI

Gli <i>Historical GIS</i> e la storia forestale: potenzialità e applicazioni	531
Vincenzo Colaprice	
Le nazioni genovesi nell'Atlante digitale di Storia Marittima del Regno di Sardegna (ASMSA). Lavori in corso	539
Giampaolo Salice, Giommaria Carboni	
Integrating Romanian Toponymy in a GIS Background: dictionary forest names vs. OpenStreetMap forest data	547
Roxana Patras, Ana Odochiciuc, Mihai-Bogdan Atanasiu, Constantin Răchită	
Digital Humanities per narrare i territori: un WebGIS sui porti minori della Sardegna dal medioevo alla contemporaneità. Una proposta di implementazione collaborativa	554
Luigi Serra, Sebastiana Nocco	

[7/C] MODELLI, METODI E PROSPETTIVE DELL'EDIZIONE DIGITALE

Roman and Justinianic Legal Terminology in Thirteenth-Century Western European Diplomatic Sources	562
---	-----

Tamás Kovács, Anguelos Nicolaou, Johannes Laroche, Georg Vogeler	
In Search of Lost Time: an Innovative Model for Representing Autograph Texts	567
Giulia Baldelli, Daniele Fusi, Matteo Zupancic, Franz Fischer, Claus Zittel	
<i>Naples Dante Project</i> . Un ambiente integrato per l'analisi bibliografica, filologica e iconografica	573
Gennaro Ferrante, Andrea La Veglia, Daniele Fusi, Stefano Angelo Rizzo, Angelo Eugenio Mecca, Daniele Duranti, Anna Sviridova	
La curatela biblioteconomica delle raccolte digitali MAB: linee gestionali, requisiti, workflow	580
Nicola Barbuti	

[8/A] GRAFI, LINKED DATA E MAPPING

DOG: an open-source ecosystem for mapping, narrating, and exploring humanities data	588
Alessandro Laruffa, Alessandro Capra	
Beyond Spatial Data: Exploring Gender and Property in the Catasto Generale Toscano through the Florentia Illustrata Semantic Platform	596
Erica Andreose, Remo Grillo, Gianmarco Spinaci	
Per la storia del mercato dell'arte tra otto e novecento: raccolta, modellazione e valorizzazione dei dati sugli antiquari italiani	602
Valentina Rossetti, Marilena Daquino, Francesca Mambelli, Ludovica Pannitto, Francesca Tomasi	
Un grafo per la ricerca storico-filologica in archivio: il caso del fondo Giuseppe Albini	609
Lucia Giagnolini, Mohamed Iheb Ouerghi, Valentina Pasqual, Cecilia Tamagnini, Paolo Bonora, Francesca Tomasi	

[8/B] GEOGRAFIE LETTERARIE E MULTIMEDIALI

'Spazializzare' la narrativa ecocritica italiana: la piattaforma dell'Atlante digitale della letteratura ecologica italiana	617
Valentina Pasqual	
MeMo: mappe digitali, memoria letteraria e public engagement nel Mezzogiorno	624
Laura Giurdanella; Giuseppe Palazzolo	
Una proposta di approccio computazionale alle geografie del cinema italiano	631
Alberto Savi, Luca Giovannini	
Il campo semantico del fuoco nella Commedia: un caso di studio sulla pertinenza lessicale e sulla visualizzazione	638
Ruoci Song	

[8/C] ANNOTAZIONI E MARKUP

Ritagli di teatro. Annotazione semantica della rassegna stampa del Teatro Sant'Erasmus	644
Paolo Bonora	
Annotazione semantica delle formule poetiche del Medioevo germanico	651
Roberto Rosselli Del Turco, Letizia Vezzosi	
Ça y est: Annotating Arabic Texts for Teaching	658

Dal markup all'interpretazione: interrogare il linguaggio della critica verista in un corpus XML/TEI	665
Denise Bruno, Salvatore Cristofaro, Antonio Di Silvestro, Laura Mazzagufò, Daria Spampinato, Giuseppe Zappalà	

[9] SESSIONE POSTER

Archivi letterari nativi digitali: modelli descrittivi ed edizioni sperimentali	673
Elena Barchielli, Simon Willemin, Elena Spadini.	
Lexicon Philosophicum: una rivista digitale per la condivisione della conoscenza umanistica	679
Pamela Barletta, Maria Cristina Dalfino, Pietro Restaneo	
Intelligenza artificiale, reference digitale e accessibilità: l'esperienza di Alphabetic	684
Flavia Bruni – Elisabetta Castro	
Grafoteca e la costruzione digitale della memoria manoscritta: curatela, metadattazione e reti di risorse	691
Amalia Carrano	
Valorizzazione digitale del patrimonio culturale: Public-Private Partnership e partecipazione sociale	697
Annalisa Ciliberti	
I Vangeli digitali e accessibili, per un'editoria religiosa universale e senza barriere	703
Giulia D'Arcangelo	
Implementing a controlled vocabulary for medieval administrative acts: theoretical and methodological considerations	709
Chiara De Bastiani	
L'Archivio del Centro Conservazione e Restauro "La Venaria Reale": un caso studio per la diffusione dei dati tecnico-scientifici	717
Stefania De Blasi, Lorena Palmieri, Chiara Pipino	
Strumenti digitali per i percorsi di visita della Palazzina di Caccia di Stupinigi: Residenza Sabauda e Museo dell'Ammobiliamento	722
Stefania De Blasi	
Strategie di protezione, governance e sostenibilità degli archivi orali	727
Rosaria De Luca, Elvira Mercatanti, Monica Monachini	
FAIR Memories: A Workflow for the Preservation and Collaborative Reuse of Analog Oral History Collections	734
Nike del Quercio, Costanza Paolillo, Laurent Fintoni	
Ontologycore: A Formal Knowledge Graph of Internet Aesthetics	740
Anouk Flinkert, Ekaterina Krasnova, Shiho Nakamura	
Il ruolo del Test Plan per la validazione: il caso d'uso del Marketplace H2IOSC	747
Vittoria Fabiani	
Partecipazione, mediazione e conoscenza: Àndalas de Cultura quale ecosistema digitale del patrimonio culturale della Sardegna	754
Carolina Floris	
From Violin Lessons to Linked Open Data. Reconstructing Tartini's <i>Scuola delle Nazioni</i> through Digital Prosopography and Digital Editions	762
Selina Galka, Marcella Tambuscio, Rolf Wissmann, Cristina Scuderi, Georg Vogeler	

Addestramento di un modello di HTR in Transkribus per testi plurilingui di età moderna: il caso del <i>Ripulimento della lingua sarda</i> (Madau)	768
Michela Incollu, Giulia Murgia	
Modelli digitali interattivi per gli archivi teatrali storici: mappare le relazioni semantiche al Teatro Metastasio di Prato (1827–1860)	775
Matilde Innocenti	
Tra fotografia e segno: la trascrizione informatizzata dei registri storici Alinari	779
Pamela Krzemien	
Biblioteche filosofiche private e memoria culturale: http://picus.unica.it/	786
Andrea Lamberti, Giovanna Granata	
Beyond Archives: A Three-Stage Workflow for Relational Digital Libraries of Written and Oral Memory Sources	791
Giulia Lembo	
A Workflow-Centred Approach to Managing Semantic Artifacts as Linked Open Data	800
Michele Mallia, Fahad Khan, Valeria Quochi	
SEBASTIAN: uno strumento per l'organizzazione, l'analisi spaziale e la visualizzazione dei dati storici	807
Christian Marcantonio, Giuseppe Consolo	
Raccontando e valorizzando il patrimonio culturale. Dalle prime esperienze ai progetti futuri del CNR-ISEM	814
Maria Grazia Mele R., Samuel Aldo Iacolina, Luigi Serra, Giovanni Serrelli	
La Digital Library e l'Ecosistema digitale per la cultura come opportunità di <i>public engagement</i>	821
Federico Meschini, Biancamaria Hermanin, Antonella Negri, Giuliano Romalli	
Strategie di engagement nella creazione e sostenibilità di un parco letterario: i risultati del progetto CHANGES	827
Sara Obbiso, Nicola Mariniello, Davide Bagnaresi, Riccardo Stracuzzi, Giuliana Benvenuti, Alessandro Iannucci	
Lo studio integrato del patrimonio materiale e immateriale per la ricostruzione delle pratiche storiche di gestione delle risorse ambientali nell'Alpujarra (Sierra Nevada): il progetto Sinergy	835
Alessandro Panetta	
<i>Reveduti, Correcti, Aprobati et Confirmati</i> : The Digital Edition of the Statutes of Ascoli	841
Michela Parma, Fabio Mariani	
Trascrivere automaticamente i registri catastali sardi del primo Novecento con Transkribus: risultati di un caso di studio	849
Andrea Pergola, Cecilia Tasca	
Ricostruzione virtuale di un patrimonio disperso: le collezioni di disegni a Genova nel XIX secolo	856
Giulia Pilosu	
Dai workflow descrittivi ai workflow applicativi tramite AEON e l'IA	860
Francesco Pinna, Emiliano Degl'Innocenti	
<i>Gamification</i> per la valorizzazione dei beni musicali. Un'ipotesi di raccolta di dati catalografici in <i>crowdsourcing</i>	867

Marcello Ranieri	
Tra riconoscimento disciplinare e prospettiva pubblica: un modello logico-applicativo per le conoscenze e le competenze nelle DH	874
Lisa Reggiani	
Mappe, scale e interpretazione. Per una cartografia digitale semantica dello spazio narrativo	883
Lorenzo Sabatino	
Verso un grande corpus diacronico dell'italiano: il contributo degli esempi del GDLI	893
Eva Sassolini, Sebastiana Cucurullo, Marco Biffi e Simonetta Montemagni	
Verso una classificazione funzionale dei prompt basata sulle intenzioni cognitive esplicitate dagli studenti nell'interazione con l'IA	901
Daniele Scala, Salvatore Varriale	
Il videogioco come strumento di mediazione interculturale	909
Alessia Stocco	
Edizione digitale commentata di <i>Le libere donne di Magliano</i> di Mario Tobino: primi passi, metodi e prospettive	917
Fabio Zarroli	
Il restauro testuale dei papiri nel progetto GreekSchools: CoPhiEditor e l'integrazione di modelli linguistici	923
Simone Zenzaro, Federico Boschetti, Angelo Mario Del Grosso, Graziano Ranocchia	
MAGIC-MEMO: pratiche condivise di digitalizzazione per la memoria scritta di Montecassino	931
Elvira Zambardi, Ylenia Nardone, Giuseppina Civitillo	



Trascrivere automaticamente i registri catastali sardi del primo Novecento con Transkribus: risultati di un caso di studio

Andrea Pergola¹, Cecilia Tasca²

¹ Università degli Studi di Cagliari, Italia - andrea.pergola@unica.it

² Università degli Studi di Cagliari, Italia - tasca.cecilia@tiscali.it

ABSTRACT (ITALIANO)

L'articolo presenta i risultati del processo di digitalizzazione e trascrizione automatica di una serie di registri catastali sardi dei primi del Novecento attraverso l'utilizzo dell'applicativo Transkribus. Tale processo ha permesso la creazione di tre modelli personalizzati (tabella, layout, trascrizione) sviluppati attraverso l'etichettatura e la trascrizione manuale di 21 registri, che hanno portato a un significativo miglioramento dei modelli base proposti dalla piattaforma, consentendo di ottenere dataset strutturati interrogabili con un margine di errore contenuto.

Parole chiave: Transkribus; HTR; Sardegna; Archivistica.

ABSTRACT (ENGLISH)

Automatic Transcription of Early Twentieth-Century Sardinian Cadastral Registers Using Transkribus: A Methodological Case Study

The article presents the results of the digitization and automatic transcription process applied to a series of early twentieth-century Sardinian cadastral registers using the Transkribus platform. This process enabled the creation of three customized models (table, layout, and transcription), developed through the manual labeling and transcription of 21 registers. These models significantly improved the baseline configurations provided by the platform, allowing the generation of structured and searchable datasets with a limited margin of error.

Keywords: Transkribus; HTR; Sardinia; Archives.

1. INTRODUZIONE

Negli ultimi anni, i sistemi di Handwritten Text Recognition (HTR) sono stati utilizzati con sempre maggior frequenza in progetti di analisi, descrizione e valorizzazione dei patrimoni documentari. In questo contesto, Transkribus, sviluppata nell'ambito dei progetti Transcriptorium (Sánchez et al., 2013) e READ, oggi READ-COOP (Terras et al., 2025), si è affermata come piattaforma di riferimento per la trascrizione automatica di documentazione storica manoscritta (Muehlberger et al., 2019; Nockels et al., 2022)¹, grazie alla sua capacità di trasformare immagini digitalizzate in risorse interrogabili (Romein et al., 2025). Seppur ancora frammentati, diversi casi di studio (Carboni & Molino Lova, 2025; Iorio, 2019; Loss et al., 2025; Schwarz-Ricci, 2022) dimostrano un uso sempre più sistematico di Transkribus in contesti archivistici che può facilitare enormemente l'interrogabilità di fondi documentari consistenti attraverso la loro trasformazione in basi dati ricercabili. Questa possibilità non ha solo impatti diretti sulla ricerca storica e, più in generale, sulle pratiche di accesso alle informazioni in ambito digitale, ma può agevolare concretamente azioni di descrizione archivistica di fondi di grande entità. A questo proposito, il presente contributo fornisce informazioni sull'esperienza portata avanti da un gruppo di ricerca dell'Università degli Studi di Cagliari² su un *corpus* di registri catastali sardi dei primi del Novecento, con lo scopo di permettere una valutazione dei risultati raggiunti, ossia la creazione di tre modelli personalizzati (tabella, layout, trascrizione) che ha permesso di elaborare una base di dati con un margine di errore contenuto.

¹ Sebbene Transkribus si sia attestata nel tempo come piattaforma di riferimento, va ricordata anche la piattaforma web Escriptorium, piattaforma open-source per il riconoscimento di testo manoscritto (HTR) nei documenti storici, integrata con il motore Kraken (Kiessling et al., 2019; Sicchera et al., 2025).

² Il gruppo di ricerca, sotto il coordinamento di Giampaolo Salice e Cecilia Tasca, è composto da Bianca Fadda, Mariangela Rapetti, Eleonora Todde, Andrea Pergola e Beatrice Schivo. Hanno collaborato alle attività di ricerca anche Stella Barbarossa e, nelle ultime fasi, Giommara Carboni e Davide Pisanu.

2. CORPUS E OBIETTIVI

L'utilizzo dell'applicativo è stato il fulcro di un'indagine archivistica volta allo studio di documenti utili all'individuazione di territori soggetti ad uso civico in Sardegna³, realizzata in seno al Centro di Umanistica Digitale dell'Università degli Studi di Cagliari (DH.UniCA)⁴. L'attività di ricerca è stata condotta sui registri catastali del fondo Ufficio Tecnico Erariale (UTE) dell'Archivio di Stato di Cagliari⁵: unità manoscritte, prodotte da diversi Comuni della Sardegna e dunque caratterizzate da una forte eterogeneità in termini di formato, qualità di scrittura e stato di conservazione (presenza di pieghe, bruciature, disallineamenti, differenze di contrasto). Il contenuto informativo di questi registri è organizzato in strutture tabellari più o meno regolari, con colonne comprensive di dati numerici e testuali relativi ad appezzamenti di terreni e ai loro possessori, alla loro localizzazione geografica, alla qualità delle colture, alla stima economica e, in ultimo, a eventuali osservazioni. L'attività è stata realizzata attraverso diverse fasi, condotte in collaborazione con la Società Linkalab S.r.l. di Cagliari, che ha curato gli aspetti prettamente informatici della sperimentazione e supportato con un ausilio formativo di base⁶. Questa modalità di lavoro ha permesso, non solo, di prendere dimestichezza con il software e i suoi processi, ma ha anche dato modo al gruppo di ricerca di verificare in maniera graduale il livello di risposta dei modelli forniti dal software, specialmente su questo specifico tipo di documentazione che, pur prodotta dalla medesima tipologia di soggetto produttore, presenta spesso difformità grafiche e organizzative. La sperimentazione ha portato al raggiungimento di un triplice obiettivo: migliorare il margine di errore dei modelli base di Transkribus attraverso la creazione di modelli personalizzati, portando a ridurre quanto più possibile la correzione manuale dei risultati ottenuti; riuscire a produrre un file strutturato unico ed esportabile – necessario per procedere a interrogazioni puntuali della documentazione analizzata–; creare i presupposti per procedere con la trascrizione automatica di ulteriore documentazione di tipologia simile a quella di partenza con tempi maggiormente rapidi rispetto alla trascrizione manuale.

3. METODOLOGIA E WORKFLOW

L'attività di trascrizione automatica è stata impostata in un flusso di lavoro in più fasi, pensato per rispondere alle caratteristiche specifiche del corpus: registri catastali manoscritti, fortemente eterogenei per qualità materiale, organizzazione grafica e fisionomia scrittoria. La scelta di utilizzare Transkribus come strumento di lavoro principale risponde non solo alla sua consolidata affidabilità scientifica, ma anche alla sua recente adozione come standard nazionale per i 102 Archivi di Stato italiani⁷. Questa piattaforma offre un ecosistema integrato che permette di gestire in un unico workflow la complessa segmentazione delle strutture tabulari sarde e la successiva trascrizione dei campi testuali, facilitando la creazione di database pronti per la ricerca avanzata: analisi del layout, riconoscimento di strutture tabellari, trascrizione automatica del testo manoscritto, correzione manuale ed esportazione dei risultati

³ Il progetto "Indagine archivistica nei fondi medievali e moderni, acquisizione, digitalizzazione, trascrizione documenti reperiti attraverso uno studio paleografico e diplomatistico, creazione mappa geolocalizzata, individuazione territori soggetti ad uso civico in Sardegna", è stato finanziato dall'Assessorato della Pubblica Istruzione, Beni Culturali, Spettacolo, Sport e Informazione della Regione Autonoma della Sardegna, L.R. 21 febbraio 2023, n. 1 (Legge di stabilità 2023), art. 1, comma 3, Tabella D); responsabile scientifico Prof. Giampaolo Salice, Dipartimento di Lettere, Lingue e Beni culturali. Si tratta della prosecuzione del progetto "Dai Feudi agli Usi civici: la mappa archivistica per la storia della proprietà in Sardegna", anch'esso finanziato dalla Regione Autonoma della Sardegna L.R. 7/07 annualità 2017) a valere sul Fondo di sviluppo e coesione 2014-2020 (Patto per lo sviluppo della Regione Sardegna - Area tematica 3 Linea d'azione 3.1 "Interventi di sostegno alla ricerca"), responsabile scientifico Prof.ssa Cecilia Tasca, Dipartimento di Lettere, Lingue e Beni culturali, conclusosi nel dicembre 2022. Hanno collaborato alle attività di ricerca Andrea Pergola e Myriam Mereu.

⁴ (DH.UniCa, s.d.).

⁵ Costituito da 6.616 unità (cc. 3917, regg. 2458, b. 1) relative ai secoli XIX e XX, il fondo contiene la documentazione prodotta in seguito alla misurazione dei terreni realizzata per la formazione del primo catasto terreni, e comprende, oltre al materiale cartografico, i registri delle matricole dei possessori, i registri delle partite e i sommari (Archivio di Stato di Cagliari, s.d.).

⁶ Si ringrazia inoltre Linkalab per la predisposizione di un report dettagliato sugli aspetti tecnici delle operazioni condotte, parzialmente utilizzato per la stesura di questo contributo (Homepage, 2024).

⁷ A questo proposito, l'ICAR nel luglio 2025 ha dato il via al progetto Transkribus, che ha l'obiettivo di «fornire ai 102 Archivi di Stato italiani e ad ogni ricercatore la possibilità di accedere in modo più semplice e rapido alla lettura di documenti trascritti automaticamente in formato digitale». (Archivi | Via al progetto «Transkribus», 2025).

in formati riutilizzabili. Questa integrazione è risultata particolarmente adatta per la documentazione oggetto del presente studio, che non presenta un problema esclusivamente grafico, ma anche strutturale e tabellare.

Il lavoro è stato preceduto da un'operazione di acquisizione digitale dei registri in formato TIFF⁸. Poiché Transkribus non consente il caricamento diretto di questo formato, è stato necessario realizzare, sempre con il supporto tecnico della Società Linkalab, uno script Python per la conversione massiva dei file in PDF e PNG.

Una volta predisposte le immagini, i documenti sono stati organizzati in collezioni distinte sulla base del Comune di provenienza. Tale scelta ha risposto a un'esigenza sia tecnica sia metodologica: da un lato, ha consentito una gestione ordinata del corpus, dall'altro, ha permesso di strutturare con maggior precisione la variabilità interna dei documenti, evitando che il training si concentrasse su un unico tipo di mano, di impaginazione o di stato di conservazione. Includere registri diversi per provenienza, struttura e condizioni materiali, ha permesso di aumentare la capacità di generalizzazione dei modelli pur a discapito di una loro maggiore precisione.

Il flusso di lavoro è stato articolato realizzando gradualmente tre modelli distinti, ciascuno dei quali rispondeva a una funzione specifica. Il primo, il modello tabellare, era destinato al riconoscimento della struttura delle tabelle presenti nei registri. Nei documenti analizzati, infatti, le informazioni sono organizzate in griglie regolari, composte da righe e colonne contenenti dati testuali e numerici. La segmentazione tabellare è stata trattata come un passaggio autonomo, preliminare rispetto alla trascrizione vera e propria, perché in questo tipo di documentazione un errore nella rilevazione della struttura può rendere instabile l'associazione tra cella, contenuto e posizione e compromettere irrimediabilmente l'interrogabilità del dato.

Il secondo modello, dedicato al layout, è stato impiegato per l'individuazione delle linee testuali e numeriche e per la definizione delle *baseline* su cui poi si fonda il riconoscimento dei caratteri. Questa fase non va confusa con la trascrizione, poiché riguarda l'individuazione della struttura grafica del testo e non la lettura del contenuto. Il modello layout ha quindi avuto il compito di stabilire con maggiore precisione l'assetto delle linee, riducendo gli errori causati da righe spezzate, campi disallineati o elementi grafici poco leggibili. In presenza di registri con cifre, apici, simboli di ripetizione e altre convenzioni amministrative ricorrenti, questa fase si è rivelata decisiva per limitare i falsi positivi e migliorare la qualità dell'input destinato al modello di trascrizione (Schwarz-Ricci, 2022).

Il terzo modello, infine, è stato dedicato alla trascrizione automatica del testo manoscritto. L'avvio del training è stato effettuato a partire da modelli base disponibili nella piattaforma, individuati in base alla loro rassomiglianza, tipologica e cronologica, con le grafie del corpus. Una scelta di partenza, dunque, non arbitraria, ma guidata dalla volontà di ridurre la distanza tra base model e documentazione di riferimento, limitando il rischio di trasferire al nuovo training configurazioni troppo lontane dalle caratteristiche reali dei registri. Questo modello è stato l'ultimo applicato alla lavorazione, perché riconosce i caratteri a partire dalle linee identificate dal modello layout e li colloca all'interno delle sezioni individuate dal modello tabella. La trascrizione automatica si basa su sistemi di Handwritten Text Recognition (HTR), che non coincidono con l'OCR. A differenza dei sistemi OCR, che scompongono il testo in singoli glifi indipendenti, la tecnologia HTR interpreta l'andamento orizzontale della scrittura come un flusso unitario (Bazzaco, 2024). Questo approccio permette al software di non limitarsi al riconoscimento delle lettere isolate, ma di analizzare l'intera riga di testo, sfruttando modelli probabilistici per decodificare le grafie talvolta compresse o irregolari tipiche dei compilatori dei registri catastali. L'addestramento dei tre modelli è avvenuto in modo progressivo, con un incremento graduale del numero di immagini annotate e trascritte manualmente. Questa strategia è stata adottata per due ragioni principali. In primo luogo, perché consentiva al gruppo di ricerca un maggior controllo sulla qualità delle annotazioni e di correggere gli errori nelle fasi iniziali, prima che si propagassero nelle versioni successive dei modelli. In secondo luogo, perché permetteva di misurare con maggiore chiarezza il contributo di ciascun ciclo di addestramento al miglioramento delle prestazioni. Le immagini impiegate per l'addestramento sono state selezionate in modo da rappresentare la maggiore varietà possibile di

⁸ Il lavoro di reperimento, schedatura, acquisizione digitale e metadattazione dell'imponente documentazione reperita negli Archivi di Stato di Cagliari e di Sassari, nell'Archivio centrale dello Stato di Roma e nell'Archivio dell'Assessorato dell'Agricoltura di Cagliari, che ammonta allo stato attuale a oltre 20.000 file digitalizzati, è stato realizzato da Stella Barbarossa (Borsa di ricerca n. 09/2023, Dipartimento di Lettere, Lingue e Beni culturali) a partire dal 1° dicembre 2023.

condizioni: registri di Comuni diversi, mani differenti, tabelle regolari e irregolari, pagine in buono stato e pagine danneggiate da strappi, pieghe, bruciature, macchie o disallineamenti dovuti alla digitalizzazione.

Data l'estrema variabilità delle mani e dello stato di conservazione dei registri dell'UTE, la creazione di un ground truth⁹ accurato è stata la fase più delicata. Per i modelli tabellare e layout, le annotazioni sono state inizialmente prodotte con il supporto dei modelli base della piattaforma e poi corrette manualmente; per il modello di trascrizione, a causa dell'unicità della scrittura manoscritta, nonostante la loro rassomiglianza con le grafie del corpus, i modelli di trascrizione di base sviluppati dalla community di Transkribus, hanno spesso dato risultati non accettabili. Per questo motivo, dunque, nelle prime fasi di utilizzo dell'applicativo è spesso necessario procedere con una profonda revisione manuale dei risultati ottenuti con i modelli di base o direttamente con una trascrizione manuale del materiale. Questa scelta rispondeva alla necessità di disporre di dati di addestramento il più possibile accurati, poiché la qualità del *ground truth* incide in modo determinante sull'efficacia del modello finale. I dati così prodotti hanno costituito la base per successive iterazioni di training e validazione.

Il processo si è svolto in due fasi principali. Nella prima, il training è stato condotto su un campione iniziale di documenti, con il duplice obiettivo di verificare la risposta dei modelli base e misurare la distanza tra questi e la documentazione di riferimento. Nella seconda fase, i modelli sono stati riaddestrati su un corpus più ampio, comprendente un numero consistente di immagini provenienti da più registri e più Comuni. Questa procedura ha consentito una progressiva riduzione del *Character Error Rate* (CER)¹⁰ per il modello di trascrizione e un miglioramento della qualità di segmentazione per i modelli tabella e layout. Nonostante ciò, il modello tabellare ha mantenuto un margine di errore più elevato rispetto agli altri due, soprattutto nelle pagine con condizioni conservative peggiori e nelle tabelle più irregolari, a conferma del fatto che la struttura documentaria rappresenta, in questo caso, la componente più complessa da automatizzare.

Completato l'addestramento, i documenti sono stati esportati da Transkribus in formato Excel e sottoposti a una fase di post-processing. Questa fase non è stata considerata un semplice intervento correttivo finale, ma una parte integrante del workflow, indispensabile per trasformare i risultati dell'HTR in un dataset realmente interrogabile. I file esportati sono stati aggregati in un unico foglio di calcolo e normalizzati, soprattutto per quanto riguarda toponimi, abbreviazioni e alcune forme ricorrenti di valore numerico. Il lavoro di normalizzazione ha permesso di ridurre ulteriormente le ambiguità residue e di rendere il dataset adatto a successive interrogazioni e analisi quantitative.

4. RISULTATI RAGGIUNTI

4.1 Modello tabella

Il modello tabella personalizzato per i registri catastali sardi (Table Model v2) è stato creato utilizzando come base di partenza il modello "Amsterdam Tables" fornito da Transkribus.

Durante la fase uno, il team di Linkalab ha addestrato un primo modello, tracciando le tabelle su 50 immagini estratte da registri catastali di 10 Comuni del Marchesato di Quirra¹¹. È stata successivamente addestrata una seconda versione del modello, con l'applicazione dei risultati a 100 immagini di ulteriori registri. Nella fase due, il modello è stato applicato su 21 unità archivistiche di registri catastali (una media fra 40 e 170 carte per un totale di 1.120 file immagine); i risultati sono stati corretti manualmente dal team di UniCa¹².

⁹ Nel contesto del machine learning e, in particolare, con questo termine si indicano i testi utilizzati dal software per confrontare e affinare le proprie trascrizioni dei documenti digitalizzati. Il ground truth può essere realizzato completamente a mano dall'utente oppure generato in modo semi-automatico, partendo da un modello già esistente (ad esempio un public model sviluppato da altri gruppi di ricerca) e correggendone gli errori. Questo materiale testuale costituisce la base su cui il sistema apprende, per cui le trascrizioni devono essere quanto più accurate possibile, altrimenti il modello apprenderà erroneamente. (*Preparazione dei dati*, s.d.)

¹⁰ (*Character Error Rate e curva di apprendimento*, s.d.).

¹¹ I Comuni di Arbus, Assemini, Genuri, Guspini, Iglesias, Las Plassas, Lunamatrona, Muravera, Pabillonis e Pauli Arborei.

¹² Nel rispetto dell'ordine dei 20 Comuni selezionati, la correzione è stata realizzata sui primi 35.965 record (su un totale di 90.387) del file excel in cui il team di Linkalab ha riversato i contenuti dei files immagine. Sempre a cura del team di Unica sono stati realizzati alcuni apparati esemplificativi: anzitutto l'elenco normalizzato delle abbreviazioni riscontrate nei diversi registri e l'elenco dei numerosissimi toponimi, per il quale ci si è avvalsi delle informazioni

Al termine dell'addestramento, il modello tabella presentava ancora un errore relativamente alto (CER del 40–50%), causato principalmente dalla forte varietà di condizioni sia materiali che digitali (contrasto irregolare dell'immagine, pieghe, bruciature e strappi di pagina, disallineamenti tra righe e colonne relativi alle operazioni di digitalizzazione). Nonostante ciò, il confronto con il modello base "Amsterdam Tables" ha evidenziato un miglioramento netto: il modello di base spesso non rilevava righe, aggiungeva colonne non presenti ai margini della pagina e mostrava disallineamenti marcati rispetto all'immagine, diversamente dal modello personalizzato in cui, ad eccezione di casi critici, la struttura tabellare risulta coerente e utilizzabile come base per i passaggi successivi.

Nel complesso, il modello tabella, pur non raggiungendo ancora livelli di accuratezza pienamente soddisfacenti, ha risposto positivamente - soprattutto nei casi in cui la qualità delle immagini era migliore - nell'identificare la struttura delle tabelle presenti nei registri, segmentando correttamente righe e colonne anche in presenza di layout irregolari, delimitazioni grafiche non sempre visibili e contenuti multi-riga all'interno della stessa cella, riducendo sensibilmente il lavoro di correzione richiesto per ottenere tabelle leggibili e interrogabili.

4.2 Modello layout

Il modello layout personalizzato (Baseline model v1) è stato implementato utilizzando quello base "Universal Lines", progettato come modello generico per il rilevamento di linee su una vasta gamma di tipologie documentarie, ma non specificamente ottimizzato per tabelle miste testo/numero. Nonostante questi presupposti, al termine dell'addestramento il modello ha risposto positivamente alla documentazione sarda, raggiungendo un CER intorno al 9,5% nel riconoscimento dei campi testuali e numerici. Si tratta di un valore che può essere considerato complessivamente buono, considerato che nei registri catastali sono spesso presenti diffusamente cifre, apici e simboli di difficile distinzione. In alcune colonne, specialmente quelle in cui erano presenti esclusivamente dati numerici, il modello di base "Universal Lines" produceva dei falsi positivi individuando linee inesistenti.

Il modello personalizzato ha invece eliminato i falsi positivi e migliorato significativamente la copertura delle colonne numeriche, con errori residui, limitati al mancato riconoscimento di doppie virgolette in celle consecutive (utilizzate per indicare la ripetizione su colonna di una stessa cifra) e di alcuni valori in colonne di stima. Questo risultato ha ridotto considerevolmente il numero di correzioni manuali necessarie per ottenere linee coerenti su cui applicare il modello di trascrizione.

4.3 Modello di trascrizione

Il modello di trascrizione personalizzato (Transcription model v5) è stato sviluppato con un training sui modelli base "Italian - 20th Century Minutes of Mediobanca's Board of Directors & Exec. Committee"¹³ e "Transkribus Italian Handwriting M1"¹⁴ che la piattaforma presenta con un CER teorico di partenza rispettivamente del 2,38% e del 6,7%. Alla loro prima applicazione, i modelli hanno presentato un CER molto elevato (50% e 60%) con errori rilevanti su quasi tutte le tipologie di campo, con trascrizioni poco affidabili dei nomi di persona, dei toponimi e dei valori numerici, tanto da risultare inadatti per un uso archivistico diretto su questo specifico tipo di registri. Per questo motivo, l'addestramento è avvenuto talvolta partendo da una revisione manuale dei risultati ottenuti con il modello di base, talaltra trascrivendo in maniera totalmente manuale i registri, in modo da fornire al sistema indicazioni quanto più corrette per il raggiungimento del *ground truth* necessario e adattarlo alle specificità grafiche e linguistiche dei registri catastali sardi. Anche in questo caso, il lavoro è stato graduale nella fase uno, con un addestramento del modello conseguito trascrivendo manualmente un numero sempre maggiore di immagini (10 di tre Comuni; 25 di 3 Comuni; 50 di 5 Comuni; 80 di 10 Comuni) e applicando i risultati a 100 immagini relative a registri di 20 Comuni differenti¹⁵.

normalizzate presenti nel Dizionario dei toponimi della Sardegna disponibile on line attraverso il Sardegna Geoportale e il Servizio ricerca toponimi (Portale Toponimi 1.0.5, s.d.)

¹³ Modello creato da Silvia Carboni per trascrivere i verbali manoscritti del Consiglio di Amministrazione e del Comitato Esecutivo di Mediobanca. Il Training Set comprende 13 diverse mani e 90.729 parole (circa 50 pagine per ciascuna mano). È stato utilizzato come modello base uno preesistente e non pubblico, addestrato su 25 pagine di ground truth per ciascuna delle 6 mani più frequenti nei verbali.

¹⁴ Modello generico, sviluppato su diverse tipologie di scritture manoscritte in lingua italiana del XVI-XIX secolo. Si tratta di un modello generico addestrato su un dataset diversificato.

Successivamente a questi cicli di training, il modello personalizzato, basato sul modello base Transkribus "Italian Handwriting M1" ha raggiunto un CER tra l'8 e il 10%. Sebbene possa sembrare un valore ancora relativamente alto, va considerato che si tratta di un margine di errore di riconoscimento che si riferisce a documenti stilati da diverse persone, quindi con grafie differenti. In generale, pur essendo ancora perfettibile, il modello personalizzato individua e trascrive correttamente le parole nelle intestazioni e nei titoli di colonna e gran parte dei contenuti testuali e numerici, con errori concentrati specialmente su numeri, apici e alcune grafie particolarmente irregolari. Dal punto di vista operativo, ciò permette di utilizzare le trascrizioni come base per ricerche per parola chiave e analisi quantitative, limitando l'intervento umano alla revisione di casi problematici e alla normalizzazione dei dati.

5. CONCLUSIONI

L'esperimento ha mostrato che anche con un numero limitato di unità è possibile ottenere modelli personalizzati in grado di migliorare sensibilmente le prestazioni rispetto ai modelli base di Transkribus, in particolare per il riconoscimento delle linee e per la trascrizione. Analizzando i risultati raggiunti, non emerge tanto la possibilità di automatizzare integralmente il trattamento dei registri catastali, quanto piuttosto la possibilità di costruire un flusso di lavoro semi-automatico in cui l'HTR riduce in modo consistente il carico di lavoro umano, pur non eliminando la necessità di controllo, correzione e normalizzazione.

In questo senso, il valore principale dei modelli personalizzati non consiste nella sostituzione di chi trascrive la documentazione, ma nella possibilità di rendere possibile la trasformazione dei registri in un dataset interrogabile e riusabile. Relativamente ai base model, inoltre, è utile evidenziare che seppur la loro scelta può essere stata metodologicamente decretata in base alla prossimità tipologica, cronologica e grafica rispetto ai documenti da trattare non sempre può garantire buone prestazioni sul nuovo corpus. In conclusione, la sperimentazione condotta sui registri catastali sardi del primo Novecento ha mostrato che l'impiego di Transkribus può produrre risultati utili anche in presenza di documentazione archivistica complessa.

I risultati ottenuti consentono di distinguere con maggiore chiarezza il comportamento dei tre moduli impiegati. Il modello layout e il modello di trascrizione, una volta personalizzati, hanno raggiunto livelli di accuratezza sufficienti per sostenere attività di interrogazione, ricerca per parola chiave e analisi quantitativa su larga scala, pur mantenendo un margine di errore non trascurabile nei campi numerici e nei segni abbreviativi. Il modello tabella, invece, si è confermato il segmento più problematico dell'intero flusso di lavoro, poiché la varietà dei layout, la presenza di celle irregolari e le condizioni materiali dei registri hanno inciso in modo significativo sulla stabilità della segmentazione.

Dal punto di vista archivistico, il guadagno principale risiede nella diminuzione del margine di errore che porta all'ottenimento di trascrizioni sufficientemente accurate che permettono di svolgere interrogazioni mirate e analisi su larga scala.

Tale esperienza, dunque, suggerisce come sia ancora necessario combinare i modelli HTR con strategie avanzate di normalizzazione. Al netto di queste limitazioni, l'uso di Transkribus e dei modelli personalizzati permette di trasformare semplici digitalizzazioni in un *dataset* strutturato interrogabile e integrabile con altre basi dati e si rivela, dunque, una soluzione praticabile ed efficace per l'interrogazione dei dati derivanti da registri catastali storici ad alta complessità. Questa stessa metodologia, a partire dalla definizione del *ground truth* all'addestramento dei modelli, fino alla normalizzazione dei dati esportati, può essere riapplicata, adattandola, a tipologie di registri amministrativi di tipologia simile o ad altri contesti archivistici.

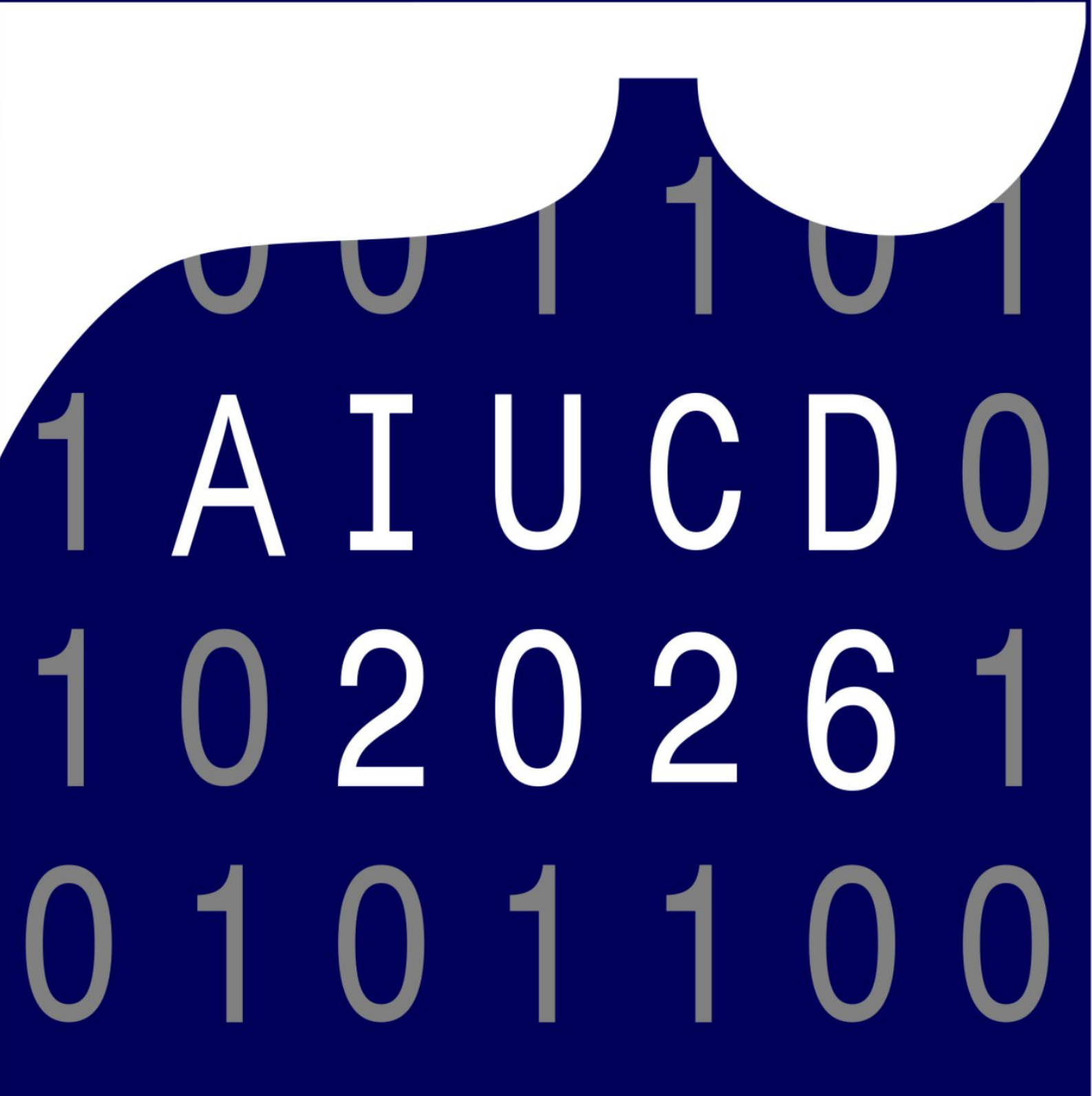
BIBLIOGRAFIA

Archivio di Stato di Cagliari. (s.d.). SIAS. Recuperato 7 febbraio 2026, da <https://sias-archivi.cultura.gov.it/cgi-bin/pagina.pl?TipoPag=strumcorr&Chiave=%2033254&RicProgetto=as%2dcagliari>

Bazzaco, S. (2024). La trascrizione automatica di documenti a stampa antichi. Appunti per un modello di riconoscimento della tipografia in corsivo. *DigItalia*, 19(1), 63–86.
<https://doi.org/10.36181/digitalia-00094>

¹⁵ I Comuni di Assemini, Barrali, Decimoputzu, Elmas, Goni, Gonnese, Guamaggiore, Lunamatrona, Monastir, Pirri, Portoscuso, Pula, S. Pietro Pula, Salto di Quirra, Sarroch, Segaiu, Setzu, Seurgus, Sicci e Villasimius.

- Carboni, S., & Molino Lova, T. (2025). L'utilizzo dell'Handwritten Text Recognition nei sistemi di consultazione degli archivi: Il caso dell'Archivio Storico Mediobanca "Vincenzo Maranghi" e Transkribus. *I quaderni de Il mondo degli archivi*, 1. <https://www.ilmondodegliarchivi.org/i-quaderni-de-il-mondo-degli-archivi-1-2025-lutilizzo-dellhandwritten-text-recognition-nei-sistemi-di-consultazione-degli-archivi-il-caso-dellarchivio-storico-me/>
- DH.UniCa. (s.d.). DH UNICA. Recuperato 7 febbraio 2026, da <https://dh.unica.it/>
- Homepage. (2024, settembre 27). Linkalab. <https://linkalab.it/>
- Iorio, S. (2019). L'utilizzo della piattaforma Transkribus nell'Archivio Storico del Banco di Napoli: Il "Progetto Pandetta". *Quaderni dell'Archivio Storico - nuova serie online*, 1(1), 195–208.
- Kiessling, B., Tissot, R., Stokes, P., & Stökl Ben Ezra, D. (2019). eScriptorium: An Open Source Platform for Historical Document Analysis. *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, 2, 19–19. <https://doi.org/10.1109/ICDARW.2019.10032>
- Loss, E., Guernaccini, F., & Carassai, M. (2025). From Manuscript to Metadata: Experiments on Handwritten Text Recognition, Tagging and Importation for the Memoriali series (1265-1452). *JLIS.It*, 16(2), 59–85. <https://doi.org/10.36253/jlis.it-641>
- Muehlberger, G., Seaward, L., Terras, M., Ares Oliveira, S., Bosch, V., Bryan, M., Colutto, S., Déjean, H., Diem, M., Fiel, S., Gatos, B., Greinöcker, A., Grüning, T., Hackl, G., Haukkovaara, V., Heyer, G., Hirvonen, L., Hodel, T., Jokinen, M., ... Zagoris, K. (2019). Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study. *Journal of Documentation*, 75(5), 954–976. <https://doi.org/10.1108/JD-07-2018-0114>
- Nockels, J., Gooding, P., Ames, S., & Terras, M. (2022). Understanding the application of handwritten text recognition technology in heritage contexts: A systematic review of Transkribus in published research. *Archival Science*, 22(3), 367–392. <https://doi.org/10.1007/s10502-022-09397-0>
- Romein, C. A., Rabus, A., Leifert, G., & Ströbel, P. B. (2025). Assessing advanced handwritten text recognition engines for digitizing historical documents. *International Journal of Digital Humanities*, 7(1), 115–134. <https://doi.org/10.1007/s42803-025-00100-0>
- Sánchez, J. A., Mühlberger, G., Gatos, B., Schofield, P., Depuydt, K., Davis, R. M., Vidal, E., & de Does, J. (2013). tranScriptorium: A european project on handwritten text recognition. *Proceedings of the 2013 ACM symposium on Document engineering, DocEng '13*, 227–228. <https://doi.org/10.1145/2494266.2494294>
- Schwarz-Ricci, V. I. (2022). Handwritten Text Recognition per registri notarili (secc. XV-XVI): Una sperimentazione. *Umanistica Digitale*, (13), 171–181. <https://doi.org/10.6092/issn.2532-8816/14926>
- Sichera, P., Grosso, A. M. D., Mazzagufo, L., & Spampinato, D. (2025). ZoneRW: verso un'integrazione con Kraken ed eScriptorium per il riconoscimento e la gestione avanzata delle regioni di interesse. In S. Rebora, M. Rospocher, & S. Bazzaco (A c. Di), *Diversità, Equità e Inclusione: Sfide e Opportunità per l'Informatica Umanistica nell'Era dell'Intelligenza Artificiale, Proceedings del XIV Convegno Annuale AIUCD2025*. AIUCD.
- Terras, M., Anzinger, B., Gooding, P., Mühlberger, G., Prien, M., Nockels, J., Romein, C. A., Stauder, A., & Stauder, F. (2025). The artificial intelligence cooperative: READ-COOP, Transkribus, and the benefits of shared community infrastructure for automated text recognition. *Open Research Europe*, 5, 16. <https://doi.org/10.12688/openreseurope.18747.2>
- Via al progetto «Transkribus»: Trascrizione automatica documenti tramite intelligenza artificiale. (2025, luglio 1). Archivi. https://archivi.cultura.gov.it/archivio-notizie/notizia?tx_news_pi1%5Baction%5D=detail&tx_news_pi1%5Bcontroller%5D=News&tx_news_pi1%5Bnews%5D=2064&cHash=a028de822f75028c8b6fcb2c167b600c



UNICA

UNIVERSITÀ
DEGLI STUDI
DI CAGLIARI



ASSOCIAZIONE per
l'INFORMATICA UMANISTICA
e la CULTURA DIGITALE

