LIX Congresso Internazionale SLI, Bologna, 10-12 settembre 2026

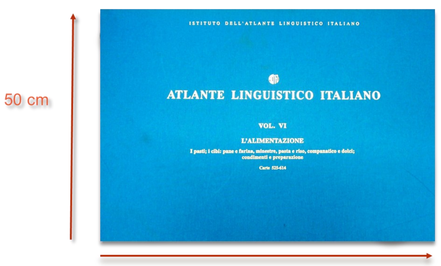
# Per una valorizzazione del dato geolinguistico alla luce della digitalizzazione dell'Atlante Linguistico Italiano

*Stefano Fiori*[1], *Matteo Rivoira*[1], *Alberto Ghia*[1], *Lorenzo Ferrarotti*[2]
[1]Università di Torino, [2]Università degli studi di Bergamo

## Cos'è l'ALI

L'Atlante Linguistico Italiano (ALI) è la più ambiziosa impresa geolinguistica condotta in territorio Italiano, fondata nel **1924** in collaborazione tra l'**Università di Torino** e la **Società Filologica Friulana**. Le campagne di inchiesta hanno avuto luogo tra il **1925** e il **1965**, con una breve interruzione durante la guerra, e hanno coinvolto più di 1000 località. Negli anni, l'ALI è stato diretto da Matteo Bartoli, Benvenuto Terracini, Giuliano Bonfante, Corrado Grassi, Arturo Genre, Lorenzo Massobrio e, attualmente, da Matteo Rivoira.

- https://www.atlantelinguistico.it




50 cm

70 cm

## La pubblicazione dei volumi

Il primo volume viene pubblicato nel **1995**, dopo una lunga fase di sistemazione dell'archivio cartaceo. Grazie alla collaborazione con l'Istituto Poligrafico e Zecca di Stato (**IPZS**), fin **dall'inizio la redazione è parzialmente digitalizzata** e si serve di una *suite* di programmi appositamente sviluppati che seguono tutte le fasi della produzione delle carte linguistiche, dalla digitazione dei dati all'impaginazione **[1]**. Dal 2010, i volumi degli atlanti sono prodotti interamente dalla redazione nella sede di Torino.

## I numeri dell'ALI

- **727** inchieste svolte tra il 1925 e il 1942 da U. Pellis, **282** tra il 1953 e il 1965 da 6 raccoglitori: T. Franceschi, C. Grassi, M. Melillo, G. Piccitto, G. Tropea, R. Giacomelli.
- **947** punti cartografati (ma molte più località effettive!), che comprendono numerose **località alloglotte.**
- **~1.700** informatori.
- **8309** domande (voci di questionario).
- **~ 2-3 milioni** di risposte stimate **[2]**
- Ricco archivio iconografico: **~2500** illustrazioni e disegni, quasi **9000 fotografie**.
- **10** volumi pubblicati su **20 previsti** (1995-2023).

Link al **fondo fotografico Ugo Pellis**, digitalizzato dalla Società Filologica Friulana, sul sito Teche Friulane.

## Il progetto «DigitALI»

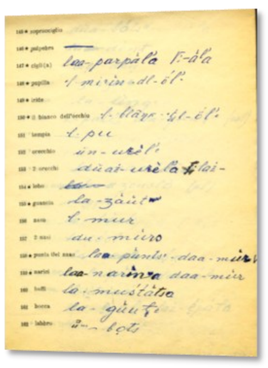
- Iniziato nel 2023 grazie al PE PNRR CHANGES, in seno al WP5 dello Spoke 3 «Digital libraries, archives and philology» (capofila Università di Bergamo), terminato ufficialmente nel 2025. Ha come obiettivo principale la **digitalizzazione dei volumi pubblicati dell'ALI**, tramite la creazione di un database strutturato, l'integrazione degli altri materiali d'archivio (**verbali d'inchiesta, illustrazioni e fotografie**) e lo sviluppo di un'**interfaccia cartografica** per la loro consultazione online.
- L'ALI si allinea così al crescente interesse per la pubblicazione online di risorse geolinguistiche **[3]**, sposando la prospettiva di poter utilizzare l'**atlante linguistico come un corpus [4]**.
- **Opportunità:** la diffusione di tecnologie *open-source* per la mappatura e la conservazione dei dati incoraggia la transizione digitale.
- **Ostacoli: l'assenza di standard condivisi** per la rappresentazione dei dati geolinguistici rischia di generare una proliferazione di risorse isolate che non dialogano fra di loro, impedendo la condivisione dei dati e il rispetto dei **principi FAIR [5]**.

**Altri componenti del DB:**

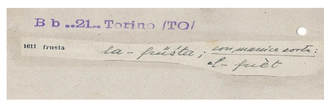
- **Verbali d'inchiesta**, digitalizzati a partire dai volumi pubblicati **[8]**, trascrivendo una selezione delle informazioni, integrabili in seguito. Contengono i profili socio-biografici degli informatori e notizie sulle località visitate.
- **Coordinate dei punti**, geolocalizzate manualmente con GoogleEarth.
- **Testo del questionario**, adattato da un file Excel già presente nell'archivio della redazione e ampliato con informazioni desunte dalla versione pubblicata **[9]**.
- **Fotografie**, tabella estrapolata dalla digitalizzazione dei verbali che consente di mettere in relazione ciascuna fotografia con il luogo in cui è stata scattata.

## Redazione dei volumi cartacei

taccuino d'inchiesta



scheda manoscritta



- Le **schede** sono le risposte originali ritagliate dai taccuini, sui quali rimane la copia-carbone; tutte le schede di una medesima voce sono raggruppate in **mazzette**.
- Tutte le risposte relative a una voce (salvo eccezioni) sono digitate in un file di testo, contrassegnati dal numero del volume e della voce e dalla dicitura «ELAB»; eventuali note e riferimenti fotografici sono digitati in file separati con la dicitura «NOTE» e «RIFO».

schermata del programma ALIdigi

File di testo in **codifica ALIcode**

```
0048ba1001  %4as%4.triva%4|s%4. , l%4/ as%4.triva
0049ca011fue%4\%6_t
0050bb210la fru%4"%3|s%4@ta ,  %4el fue%4\t! %4U
0052bb220 %4al fua%4\t
0053bc170 %4Pl fue%4\t
```

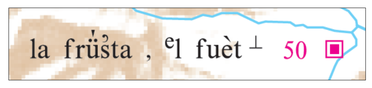
```
bb210   50   la frûšta , ˡel fuèt ⊥
```

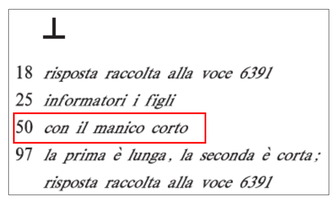
```
0018ba050#ccrisposta raccolta alla
0025bb110#ccinformatori i figli#e
0050bb210#cccon il manico corto#e
0097cb190#ccla prima e%4\ lunga!, l
0151bc220#ccdei carrettieri#e
```

```
bb210   50   con il manico corto
```

- ALIcode è un *beta-code* **[6]** basato su **combinazioni di caratteri ASCII** che conserva senza perdita di informazioni la notazione fonetica dell'ALI; molto efficace, ma **non segue le norme Unicode**.

## Resa cartografica

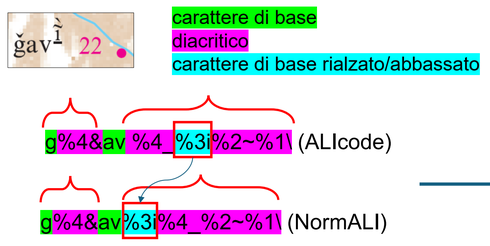



## Retro-standardizzazione

**1. Normalizzazione di ALIcode**

Uno script Python, tramite espressioni regolari, produce stringhe **NormALI**, una versione «normalizzata» di ALIcode che segue le norme Unicode (p.es., i diacritici, sia sottoscritti che sovrascritti, devono sempre sguire i caratteri di base)



**2. Traslitterazione**

Eseguita sulla base di una tabella di corrispondenze (**«profilo ortografico»**) tra codifiche ALIcode, Unicode e HTML, con il pacchetto R **qlcData**, creato appositamente per la conversione di ortografie e sistemi di trascrizione **[7]**.

NormALI viene traslitterato in:

- **Unicode**, per l'**interoperabilità** e la conversione in **IPA** (in lavorazione…)
- **HTML**, per la riproduzione delle stringhe originali (può rappresentare il corsivo e i caratteri sollevati e abbassati)

**3. Ristrutturazione**

Uno script R svolge alcune operazioni, tra cui le più importanti sono:

- estrazione delle **informazioni complementari** (marche d'uso, simboli, sigle identificative dei rilievi e degli informatori), assegnate a campi specifici.
- separazione delle **risposte multiple come record individuali** (con conservazione delle stringhe originali).
- aggiunta di ulteriori campi (voce, volume, id delle risposte…)
- unione di tutti i file relativi a un singolo volume (ALI_05, ALI_06, ALI_07…)
- procedimento analogo per note e riferimenti fotografici

| punto integer | codice character varying (5) | orig_unicode character varying (2000) | risposta character varying | tag character varying (1000) |
|---|---|---|---|---|
| 50 | bb210 | la frûšta , èl fuèt ⊥ | la frûšta | [null] |
| 50 | bb210 | la frûšta , èl fuèt ⊥ | èl fuèt | [null] |

| punto integer | codice character varying (5) | note_unicode character varying (1000) | nota character varying (1) | voce character varying (12) |
|---|---|---|---|---|
| 50 | bb210 | con il manico corto | N | 1611 |

## Database relazionale

- La retro-standardizzazione produce file in **formato tabellare (TSV)** in codifica **UTF-8**, incamerati in un database **PostgreSQL** con estensione **PostGIS** per la gestione delle geometrie puntuali e l'integrazione con software quali QGIS. L'ambiente di sviluppo della webmap si serve invece di un database **SQLite** (singolo file, leggero, nessun bisogno di connessioni server).
- L'architettura relazionale si adatta bene alla natura fortemente interconnessa dell'archivio ALI, ma non impedisce affatto diverse serializzazioni dei dati (p.es. JSON).

| punto integer | codice character varying (1000) | str character varying (1000) | norm character varying (1000) | tag character varying (1000) |
|---|---|---|---|---|
| 668 | fh040 | [null] | prədùčča | [null] |
| 668 | fh040 | *ai | na ũrićća | [null] |

| nome_luogo | punto | codice | num_inf | nome_cognome | eta | sesso | anno_nascita |
|---|---|---|---|---|---|---|---|
| Sora | 668 | fh040 | 1 | Giovanni Evangelista | 53 | M | 1901 |
| Sora | 668 | fh040 | 2 | Vincenzo Mattacchione | 72 | M | 1882 |

- La ristrutturazione dell'archivio ha permesso di istituire **stretto collegamento tra risposte dialettali e dati degli informatori**, consentendo una consultazione integrata; si ottiene così una **lettura profonda e «verticale» di dati apparentemente piatti e bidimensionali**, senza dover passare continuamente dalla carta linguistica ai verbali.

## Cosa è stato fatto

- Attualmente, il database comprende **8 volumi digitalizzati (2, 3, 5, 6, 7, 8, 9 e 10)**; i materiali del volume 1 stanno venendo ridigitati *ex novo* e lo stesso avverrà con quelli del volume 4, poiché non è stato possibile recuperare i file originali della redazione. Il processo di retro-standardizzazione, comunque, è da subito replicabile sui dati dei volumi attualmente in lavorazione (11 e 12). Oltre ai dati linguistici, che costituiscono il nucleo centrale,
- Nel corso dei lavori sono state utilizzate **esclusivamente tecnologie *open-source***, al fine di massimizzare la trasparenza e la riproducibilità delle soluzioni adottate.
- Alcuni articoli hanno già esplorato le possibilità di ricerca offerte dall'archivio digitale: in particolare, sono state condotte ricerche sulle forme dialettali contrassegnate da **marche d'uso [10]**, per esempio <*v*> per «forma usata dai vecchi» oppure <*borgh*> e <*rust*> per le forme usate, rispettivamente, in città e in campagna. Per diversi punti d'inchiesta, la rapida aggregazione di dati provenienti da diverse carte ha prodotto un quadro d'insieme che getta sprazzi sulle dinamiche di diffusioni di correnti linguistiche nell'Italia della prima metà del secolo scorso; una consultazione ordinaria delle carte dell'atlante avrebbe richiesto molto tempo per giungere a questo risultato. La facilità di estrazione delle risposte dal database ha molto agevolato anche l'analisi lessicale di una carta unica dell'ALI, «testa e croce» **[11]**, per la quale lo scaricamento diretto dei dati in formato tabellare ha grandemente agevolato la tipizzazione delle risposte.

## Cosa resta da fare

- Il progetto DigitALI ha prodotto dati *open* a 3 stelle, anzi, 2 stelle e mezza, poiché utilizzano un formato aperto e non proprietario (CSV), segue gli standard W3C, ma non sono ancora di fatto in rete! Inoltre, anche se fossero in rete, non sono ancora dotati di identificatori uniformi (URI/URL) o di DOI **[12].**
- La geolinguistica digitale sconta un certo ritardo rispetto ad altre aree come la lessicografia, che si è dotata da tempo di standard per la rappresentazione sul Web come *Linked Open Data* (p.es. OntoLex). Come detto**, manca ancora uno standard riconosciuto per la rappresentazione dei dati geolinguistici**. Una proposta molto matura è stata fatta già una decina d'anni fa **[13]**, ma non è stata largamente adottata e andrebbe rivalutata.
- Per DigitALI, i prossimi passi consisteranno in:
  1. **Collegamento con cataloghi di riferimento**, sull'esempio di VerbaAlpina, p. es. **GeoNames** per i nomi dei punti d'inchiesta; sull'aspetto semantico, un'opzione potrebbe essere **Concepticon**, oppure **WordNet**, che consentirebbe un accesso ragionato ai materiali dell'atlante **[14]**.
  2. **Tipizzazione delle forme dialettali**, fondamentale per la creazione di un indice generale per l'ALI e utile al collegamento semantico con altri cataloghi (cfr. punto precedente)
  3. **Trascrizione in IPA.** La notazione fonetica dell'ALI è un sistema aperto e votato alla registrazione di minime differenze di pronuncia (nei volumi digitalizzati ci sono **158 varianti della vocale <*a*>**!), dunque impossibile da far corrispondere esattamente con l'IPA; tuttavia, una trascrizione di servizio, pur comportando una perdita d'informazione, consentirebbe analisi dialettometriche e, di nuovo, il collegamento a cataloghi come CLTS (Cross-Linguistic Transcription Systems(

## Bibliografia

**[1]** Matteo Rivoira. 2016. L'archivio dell'ALI: informatizzazione dei dati e prospettive di sviluppo. In Francesco Avolio & Antonino Cigno (edd.), *Archivi etnolinguistici multimediali. Dalla formazione alla gestione e al dialogo col territorio*, *Pescara, 6 ottobre 2012, Atti del convegno.* Pescara, Museo delle genti d'Abruzzo, 49-61. **[2]** Federica Cugno, Matteo Rivoira & Giovanni Ronco. 2019. L'Atlante Linguistico Italiano. *Romance Philology* 74, 191-216. **[3]** Thibault, André et al. (eds.): *Nouveaux regards sur la variation dialectale – New Ways of Analyzing Dialectal Variation*. Strasbourg, Éditions de Linguistique et de Philologie. **[4]** Carolina Julià Luna. 2021. Los atlas como corpus lingüísticos: más allá de la variación geolectal. In Adolfo Arejta (ed.), *Diversidad: vocablos y voces de las variantes lingüísticas. II Congreso Internacional de Dialectología de Euskaltzaindia*, 13-15. **[5]** Sebastiana Cucurullo, Simonetta Montemagni, Rubino Saccoccio, Eva Sassolini. 2025. The Challenge of Obsolescence of Digital Archives in Cultural Heritage. A Case Study. *2025 IEEE 8th Congress on Information Science and Technology (CiSt)*, Marrakech, 506-512. **[6]** Beatrice Colcuc & Florian Zacherl. 2022. Tools for data processing and visualization in the project VerbaAlpina. In Nicole Palliwoda (ed.), *Data Processing and Visualization in Variational Linguistics/Dialectology*. Szeged/Hamburg, University of Szeged: 55–70. **[7]** Steven Morano & Michael Cysouw. 2018. *The Unicode Cookbook for Linguists: Managing Writing Systems Using Orthography Profiles*. Berlin: Language Science Press. **[8]** Ugo Pellis et alii. 1995. *Atlante Linguistico Italiano. Verbali delle inchieste*, a c. di L. Massobrio *et al.*, Roma: IPZS, 2 voll. **[9]** Matteo Bartoli *et al.* 1971. *Questionario dell'Atlante Linguistico Italiano. I,a – Testo*, a c. di A. Genre *et al.* Torino, Istituto dell'Atlante Linguistico Italiano. **[10]** Massimo Cerruti, L. Ferrarotti, S. Fiori, M. Rivoira & Barbara Turchetta. 2025. Il progetto DigitALI: problemi e prospettive della digitalizzazione dell'Atlante Linguistico Italiano (ALI). *Linguistik Online*, *137*(5), 37-67. **[11]** Lorenzo Ferrarotti, Stefano Fiori, Alberto Ghia, Matteo Rivoira, Barbara Turchetta & Massimo Cerruti (in stampa). *Testa o croce* tra tipizzazione lessicale, geolinguistica e storia linguistica. Un caso di studio nel contesto della digitalizzazione dell'ALI. **[12]** Thomas Krefeld & Stephan Lücke. 2023. FAIRNESS – Medien im methodologischen Zentrum der Geolinguistik. Version 3 (22.06.2023, 16:20). In Thomas Krefeld, Stephan Lücke & Christina Mutter (edd.), *Berichte aus der digitalen Geolinguistik (II): Vernetzung und Nachhaltigkeit*. (Korpus im Text 9), Version 30. **[13]** Emanuele di Buccio, Giorgio Maria Di Nunzio & Gianmaria Silvello. 2014. A linked open data approach for geolinguistics applications. *International Journal of Metadata, Semantics and Ontologies* 9(1), 29-41. **[14]** Paola Marongiu & Simonetta Montemagni. 2026. Towards Semantic Access and Interoperability in Digital Dialectal Atlases. A Case Study. *Proceedings of Workshop on Dialects in NLP – A Resource Perspective (DialRes)*, 274-286.