Modellazione, interoperabilità e riuso in DiScEPT

Tiziana Mancinelli[1], Hansmichael Hohenegger[2], Federico Boschetti[3], Angelo Mario Del Grosso[4], Eleonora De Longis[5], Gloria Mugelli[6]

[1] Istituto Italiano di Studi Germanici, Italia - mancinelli@studigermanici.it
[2] Istituto Italiano di Studi Germanici, Italia - hohenegger@studigermanici.it
[3] Istituto di Linguistica Computazionale CNR, Italia – federico.boschetti@unive.it
[4] Istituto di Linguistica Computazionale CNR, Italia - angelomario.delgrosso@cnr.it
[5] Istituto Italiano di Studi Germanici, Italia – delongis@studigermanici.it
[6] Istituto di Linguistica Computazionale CNR, Italia - gloria.mugelli@gmail.com

ABSTRACT (ITALIANO)

Il progetto DiScEPT (Digital Scholarly Editions and Parallel Translations Platform) ha l'obiettivo di creare un ambiente per la produzione e la pubblicazione di edizioni digitali. Il multilinguismo è il punctum saliens di questo progetto per una scelta scientifica, ma anche perché si propone di favorire l'accessibilità, l'inclusione e il dialogo interculturale almeno nell'ambito della comunità inerente alla filologia digitale nonché della traduttologia. Le traduzioni non sono solo il superamento delle barriere linguistiche, ma strumenti di arricchimento critico, opportunità per lo studio semantico, storico e culturale dei testi. Integrando strumenti avanzati, come il riconoscimento automatico dei caratteri (HTR e OCR) e l'allineamento linguistico basato su BERT-Align, DiScEPT propone flussi di lavoro dinamici per la creazione di edizioni digitali con traduzioni allineate. La metodologia si fonda su modelli multipli che considerano la materialità dei documenti, la critica testuale e soprattutto l'analisi dei corpora paralleli. Superando l'idea di una piattaforma monoblocco, il progetto adotta un ambiente distribuito, basato sull'integrazione di piattaforme esistenti tramite API.

Parole chiave: edizione scientifica digitale; TEI-XML; Allineamento di corpora; Linked Open Data.

ABSTRACT (ENGLISH)

Modeling, interoperability, and reuse in DiScEPT. The DiScEPT (Digital Scholarly Editions and Parallel Translations Platform) project aims to create an environment for the production and publication of digital editions. Multilingualism is the key focus of this project, not only as a scientific choice but also to promote accessibility, inclusion, and intercultural dialogue, at least within the community of publishers and translators. Translations are not merely a way to overcome language barriers but are critical tools for enriching textual analysis, offering opportunities for semantic, historical, and cultural studies. By integrating advanced tools such as Handwritten Text Recognition (HTR, OCR) and linguistic alignment using BERT-Align, DiScEPT offers dynamic workflows for creating digital editions with aligned translations. The methodology is based on multiple models that consider document materiality, textual criticism, and, particularly, the analysis of parallel corpora. Moving beyond the idea of a monolithic platform, the project adopts a distributed environment, relying on the integration of existing platforms through APIs.

Keywords: digital scholarly editing; linked open data; corpora alignment, TEI-XML;

1. INTRODUZIONE

Il progetto DiScEPT parte dall'esigenza di creare edizioni scientifiche digitali multilingue in grado di mettere in relazione testi originali e loro traduzioni già edite attraverso l'integrazione di strumenti e pratiche digitali già consolidate, limitando le aggiunte di componenti soltanto ai casi in cui ci siano specifiche esigenze di studio. L'obiettivo principale è costruire un ecosistema flessibile e aperto che consenta di analizzare e rappresentare le connessioni linguistiche e culturali, migliorando così la qualità editoriale e promuovendo una maggiore consapevolezza delle dinamiche traduttive.

Il focus sul multilinguismo è un elemento distintivo di DiScEPT, che non solo affronta la necessità di superare le barriere linguistiche, ma enfatizza anche il valore critico delle traduzioni e di trasmissione della conoscenza. Le traduzioni non vengono considerate come semplici strumenti per comunicare un contenuto, ma come veri e propri strumenti di arricchimento culturale, di diffusione del sapere e analisi comparativa. Questo approccio inoltre favorisce una comprensione più profonda delle dinamiche linguistiche e delle scelte traduttive, offrendo nuove opportunità per la ricerca e l'editoria scientifica.

Il progetto rappresenta anche uno spazio per ripensare i flussi di lavoro e i modelli editoriali, considerando i diversi formati e materiali che possono essere coinvolti, dalle immagini facsimilari ai testi strutturati come XML-TEI. DiScEPT si distingue per il suo approccio centrato sul riuso e sull'interoperabilità, evitando la creazione di silos informatici e incoraggiando la condivisione e la sostenibilità dei dati.

Un elemento chiave è l'utilizzo di tecnologie avanzate come il riconoscimento automatico dei caratteri (HTR) tramite strumenti come eScriptorium e l'allineamento linguistico automatizzato con BERT-Align. Questi strumenti permettono di confrontare testi e traduzioni, evidenziando scelte semantiche, storiche e culturali sia tramite analisi da parte dell'utente/editor, sia attraverso strumenti automatici di allineamento linguistico come BERT-Align, che supportano l'individuazione delle corrispondenze testuali.
DiScEPT non è solo un progetto tecnologico, ma anche un'iniziativa interdisciplinare che valorizza il multilinguismo per ampliare l'accesso a fonti e dati. Inoltre, contribuisce alla conservazione del patrimonio culturale globale, traducendo e interpretando testi in diverse lingue, e promuove l'uso di tecnologie linguistiche avanzate per affrontare sfide complesse.

2. MODELLI E WORKFLOW PER LE EDIZIONI DIGITALI: UN LABORATORIO DI EDIZIONI MULTILINGUE.

DiScEPT è progettato per consolidare metodologie e pratiche derivate da diversi progetti nel campo dell'ecdotica digitale e dei corpora multilingue allineati. L'obiettivo principale è proporre e integrare strumenti e modelli già affermati nella filologia digitale in un ecosistema che copra tutte le fasi del processo editoriale, dalla produzione alla diffusione.
Il progetto sviluppa diversi tipi di workflow, adattandosi al materiale di partenza: il lavoro può iniziare da immagini facsimilari, testi digitali in XML-TEI o trascrizioni critiche. Nel caso in cui il flusso di lavoro parta da immagini facsimilari prive di trascrizione testuale, vengono impiegate tecnologie avanzate di riconoscimento automatico dei caratteri (HTR), come eScriptorium, garantendo trascrizioni digitali precise e coerenti.
Infine, DiScEPT si distingue per tre modelli principali di descrizione delle edizioni:

- Materialità del documento: Analisi e valorizzazione del documento, della mise en page e della trascrizione del testo.
- Tipologie di edizioni e critica testuale: modello di edizione di ogni singolo documento, tenendo presente la tradizione del testo e le metodologie provenienti dalla filologia e dall'ecdotica.
- Modello linguistico per corpora paralleli: Creazione di corpora per studi comparativi, favorendo analisi approfondite delle dinamiche linguistiche e culturali.

Un focus particolare è dedicato alla materialità del documento, attraverso la possibilità di includere, rappresentare e analizzare immagini digitali facsimilari. La trascrizione non ha il solo fine di digitalizzare il testo, ma viene arricchita con annotazioni e dettagli sulla struttura fisica del testo come sezioni del testo, layout della pagina e anomalie grafiche. Le immagini, elaborate tramite tecnologie HTR, diventano quindi parte integrante del workflow, garantendo che il legame tra contenuto e forma sia preservato e valorizzato in tutte le fasi del processo editoriale. La peculiarità e la sfida di DiScEPT consistono nel collegare i diversi modelli di rappresentazione testuale (materialità, critica testuale, corpora paralleli) e, allo stesso tempo, nel renderli formalmente espliciti e interoperabili, garantendo la tracciabilità tra struttura editoriale, la relazione fra testi originali e traduzioni, contenuto e metadati.

Un altro elemento distintivo di DiScEPT è il sistema di allineamento linguistico, che combina metodi manuali e automatici per analizzare le relazioni tra testi originali e traduzioni. Basato su tecnologie come BERT-Align, l'allineamento automatico individua corrispondenze tra segmenti testuali, evidenziando scelte semantiche, stilistiche e culturali. Quando necessario, nei casi più complessi, è previsto un intervento manuale per affinare i risultati.
L'allineamento considera diverse configurazioni relazionali tra i segmenti di testo. Nel caso di uno a uno, ogni segmento del testo originale corrisponde direttamente a uno nella traduzione, comune nei testi tecnici o strutturati. La configurazione molti a uno si verifica quando più segmenti originali confluiscono in uno tradotto, tipico nelle traduzioni sintetiche o nei testi semplificati. Al contrario, la relazione molti a molti rappresenta un livello di complessità maggiore, in cui più segmenti originali e tradotti sono interconnessi, riflettendo riformulazioni stilistiche, aggiunte o omissioni.

Questo consente di tracciare le dinamiche storiche e culturali che influenzano il processo traduttivo, offrendo una comprensione più profonda delle trasformazioni testuali. La combinazione di strumenti automatici e interventi manuali migliora non solo la qualità dell'allineamento, ma anche l'analisi comparativa e gli studi filologici avanzati.
Grazie a questo approccio modulare e integrato, DiScEPT promuove un ambiente di ricerca multidisciplinare, valorizzando la collaborazione internazionale e l'accesso inclusivo alle risorse digitali.

3. CHE COS'È DISCEPT E LA SUA EVOLUZIONE

DiScEPT è un ecosistema per la creazione di edizioni digitali basato sull'integrazione di progetti open source e tecnologie aperte, in grado di gestire dati e metadati complessi. Questo processo ha permesso di sviluppare un'architettura modulare composta da diversi componenti:

- Editor testuale basato su DSL (Domain-specific language): Facilita la codifica tramite convenzioni familiari ai filologi, convertibili in XML-TEI attraverso l'uso di un parser ANTLR e XSLT. Può ricevere i dati della trascrizione tramite API e permetterne così una codifica ulteriore (per esempio per la costruzione di un apparato critico);
- Dialogo fra software HTR e sistema di editing basato su DSL;
- Allineamento testuale: Utilizza metodologie stand-off per creare corrispondenze tra testi originali e traduzioni a diversi livelli di granularità, promuovendo lettura sinottica e analisi linguistica. Allineamento di corpora multilingue automatico basato su tecnologie come Bert-Align.

Pubblicazione e visualizzazione: Basata su TEI-Publisher e tecnologie come ODD e WebComponent, offre funzionalità end-to-end dalla creazione alla pubblicazione delle edizioni digitali, garantendo riuso e interoperabilità con comunità come e-editiones.

4. INTEROPERABILITÀ TRA PIATTAFORME

Posizionare Facendo tesoro del concetto di Virtual Research Environment (VRE) distribuito, DiScEPT abbandona l'idea ambiziosa di nuova piattaforma (o, per meglio dire, "just another platform") integrata per le edizioni scientifiche digitali e abbraccia invece il paradigma di integrazione tra piattaforme esistenti e comunicanti tramite API.
Il modello di VRE distribuito è stato teorizzato nei primi lustri degli anni Duemila all'interno di progetti pioneristici come Bamboo (https://academic.oup.com/dsh/article/29/3/326/2938127), che purtroppo "failed to move to the implementation phase" (https://digital.humanities.ox.ac.uk/project/project-bamboo), in parte per motivi contingenti e in parte perché troppo avanzato rispetto ai limiti tecnologici dell'epoca, sia hardware, per le connessioni di rete, sia software, per gli inefficienti protocolli di comunicazione.
L'evoluzione delle tecnologie di comunicazione macchina a macchina senza l'intermediazione di interfacce utente, ci permette oggi di poter gestire in modo molto più efficiente flussi di dati tra piattaforme altamente specializzate e fra loro indipendenti.

È il caso, in DiScEPT, della connessione fra eScriptorium, piattaforma sviluppata da INRIA (https://gitlab.inria.fr/scripta/escriptorium) per l'Handwritten Text Recognition (HTR) e TEIPublisher (https://teipublisher.com), un toolbox di instant publishing di documenti XML-TEI per la piattaforma eXist-DB (https://exist-db.org).
Benché in DiScEPT sia prevista la comunicazione bidirezionale, attualmente è stata sviluppata la parte di comunicazione da eScriptorium a eXist-DB, per trasferire tramite API i dati testuali relativi alle trascrizioni (semi)automatiche dall'ambiente di editing focalizzato sulla correzione degli errori di HTR all'ambiente di editing sviluppato appositamente per DiScEPT su eXist-DB. Tale ambiente è basato su Domain-Specific Languages (DSL) con esportazione in XML-TEI per rendere i contenuti finali fruibili tramite TEIPublisher (la demo è in corso di sviluppo; quando sarà pronta il codice sorgente si troverà all'indirizzo: https://github.com/CoPhi/cophi-discept).

Il componente riguardante l'allineamento dei testi in DiScEPT (https://istituto-italiano-di-studi-germanici.github.io/DiScEPT/) si distingue per due approcci complementari: manuale e automatico, ciascuno mirato a garantire la massima precisione nell'identificazione delle relazioni tra testo originale e traduzioni.
Nel caso dell'allineamento manuale, il flusso di lavoro inizia con una fase di tokenizzazione, ancora in fase di perfezionamento, che suddivide il testo in unità linguistiche fondamentali. L'editor offre un'interfaccia utente intuitiva, che consente per chi la utilizza di selezionare manualmente segmenti di testo e abbinarli alle traduzioni corrispondenti. Un aspetto importante di questo processo è la possibilità di espandere il modello di riferimento in TEI, in modo da includere anche RDF, per migliorare la capacità di citazione e l'interoperabilità tra risorse digitali.
Per quanto riguarda l'allineamento automatico, DiScEPT si avvale di modelli linguistici avanzati, come Bert-Align, applicato specificamente alle lingue italiana e tedesca. Questo sistema è in fase di sviluppo e mira a sfruttare l'intelligenza artificiale per riconoscere e allineare automaticamente i segmenti di testo tra versione originale e traduzione.

Il modello Bert-Align, basato su reti neurali, consente un allineamento molto affidabile, con potenziali applicazioni per una vasta gamma di lingue. Alla luce delle riflessioni proposte da Hämmerl, Libovický e Fraser (2024), l'integrazione di modelli come BERT-Align si colloca all'interno di un quadro teorico in cui l'allineamento cross-linguistico viene inteso non solo come corrispondenza formale, ma come costruzione di un ponte semantico tra sistemi linguistici diversi. In questa prospettiva si inseriscono anche gli sviluppi proposti da Reimers e Gurevych (2020), che mostrano come modelli monolingui per il confronto semantico di frasi possano essere adattati con successo a contesti multilingue, attraverso strategie che permettono di trasferire le capacità apprese da una lingua all'altra. A queste soluzioni si affiancano strumenti recenti che rendono disponibili, in modo modulare, diversi algoritmi per il confronto tra segmenti testuali, permettendo di combinare in modo flessibile la somiglianza semantica con il calcolo della distanza formale. La capacità di bilanciare informazioni condivise e specificità linguistiche, emersa come centrale in tale studio, rappresenta una direzione promettente per lo sviluppo futuro di DiScEPT, ad esempio attraverso il fine-tuning controllato per domini o generi testuali differenti.
La capacità di bilanciare informazioni condivise e specificità linguistiche, emersa come centrale in tale studio, rappresenta una direzione promettente per lo sviluppo futuro di DiScEPT, ad esempio attraverso il fine-tuning controllato per domini o generi testuali differenti. L'allineamento, sia automatico che manuale, genera un unico file TEI in cui i testi sono collegati tramite un markup stand-off, strutturato all'interno del tag <tei:corpus>. I due approcci sono complementari e flessibili, consentendo agli utenti di scegliere il metodo più adatto al materiale da analizzare. La loro combinazione unisce la precisione manuale all'efficienza automatica, migliorando l'accuratezza e la coerenza nell'analisi dei testi tradotti. Questo sistema promuove una comprensione più approfondita delle scelte traduttive e delle dinamiche linguistiche.

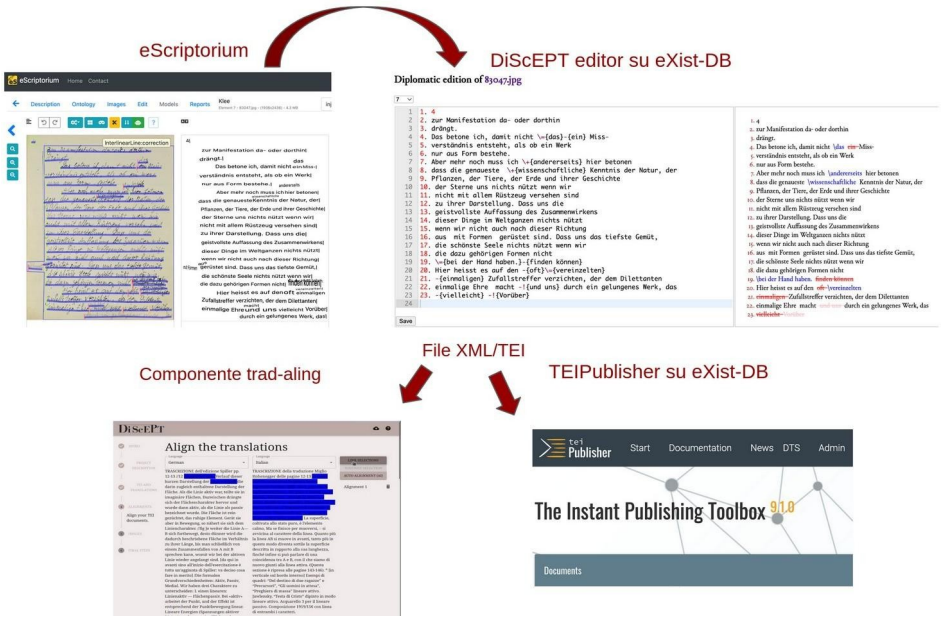

Figura 1. Flusso di lavoro

Il progetto DiScEPT dunque rappresenta un passo avanti nella creazione di edizioni digitali multilingue, l'attenzione sul multilinguismo non è solo un valore scientifico, ma anche un obiettivo per favorire l'accesso globale ai testi e promuovere l'inclusione interculturale e di diffusione del patrimonio culturale.

RINGRAZIAMENTI

L'attuale prototipo di DiScEPT si avvale di servizi offerti da H2IOSC Project - Humanities and cultural Heritage Italian Open Science Cloud funded by the European Union – NextGenerationEU – NRRP M4C2 - Project code IR0000029 - CUP B63C22000730005.
ILC4CLARIN, nodo B di CLARIN-IT, ospiterà i dati e il software del progetto per la long-term preservation. Il CLARIN Knowledge Centre DiPText-KC mette inoltre a disposizione del progetto DiScEPT il proprio Helpdesk (https://diptext-kc.clarin-it.it/helpdesk).

BIBLIOGRAFIA

Boschetti, F. (2008). Alignment of Variant Readings for Linkage of Multiple Annotations. *Proceedings of the ECAL 2007 Electronic Corpora of Ancient Languages*, Prague 16–17 November 2007. Zemánek, P. (Ed.). 11-24. http://usj.ff.cuni.cz/system/files/Boschetti-Ch-2007.pdf.

Boschetti, Federico; Taddei Andrea, Bambaci, Luigi; Del Grosso Angelo Mario; Mugelli, Gloria; Khan, Fahad; Bellandi, Andrea. "Collaborative and Multidisciplinary Annotations of Ancient Texts: the Euporia System". In The Ancient World Goes Digital Case Studies on Archaeology, Texts, Online Publishing, Digital Archiving, and Preservation, Edited by: Bigot Juloux, Vanessa; di Ludovico, Alessandro; Matskevich, Sveta. Leiden - Boston: Brill, 2023

Boschetti, Federico; Rigobianco, Luca; Quochi, Valeria. "Domain-Specific Languages for Epigraphy: The Case of ItAnt". In CLARIN Annual Conference Proceedings. (2023): pp. 80-84

Buzzetti, D., & McGann, J. (2006). *Electronic Textual Editing: Critical Editing in a Digital Horizon*. *Electronic Textual Editing*. Burnard L. et al. (Eds.), Modern Language Association of America. http://www.tei-c.org/About/Archive_new/ETE/Preview/mcgann.xml.

Cushman, Ellen. "Supporting Manuscript Translation in Library and Archival Collections: Toward Decolonial Translation Methods". In Libraries and Archives in the Digital Age, edited by Susan L. Mizruchi. London: Palgrave Macmillan, 2020.

Daquino, Marilena; Giovannetti, Francesca; Tomasi, Francesca. "Linked Data Per Le Edizioni Scientifiche Digitali. Il Workflow Di Pubblicazione dell'edizione Semantica Del Quaderno Di Appunti Di Paolo Bufalini". In *Umanistica Digitale* 3 (7). https://doi.org/10.6092/issn.2532-8816/9091

Del Grosso, A. M., Capizzi, E., Cristofaro, S., Seminara, G., & Spampinato, D. (2019). *Promoting Bellini's Legacy and the Italian Opera by Scholarly Digital Editing His Own Correspondence*. Poster presented at the TEI Conference and Member's Meeting. What is Text, really? TEI and beyond. Graz, Austria. https://zenodo.org/badge/DOI/10.5281/zenodo.3461673.svg.

Del Grosso, A. M., & Spampinato, D. (Eds.). 2023. *Bellini Digital Correspondence*. CNR Edizioni, 2023. ISBN: 978-88-8080-562-5.

Dombrowski, Q. (2014). What Ever Happened to Project Bamboo? *Literary and Linguistic Computing*, 29(3), 326–339. https://doi.org/10.1093/llc/fqu026

Hämmerl, Johanna; Libovický, Jindřich; Fraser, Alexander. "Understanding Cross-Lingual Alignment—A Survey." *Findings of the Association for Computational Linguistics*: ACL 2024. https://aclanthology.org/2024.findings-acl.649

Martignano, Chiara. 2021. "A Conceptual Model to Encourage the Development and Reuse of Apps for Digital Editions". *Umanistica Digitale* 5 (10):71-88. https://doi.org/10.6092/issn.2532-8816/12620.

Pierazzo, E. (2015). Digital Scholarly Editing: Theories, Models and Methods. Farnham, Surrey:Ashgate. Pozzo, Riccardo; Gatta Timon; Hohenegger, Hansmichael; Kuhn, Jonas; Pichler, Axel; Turchi, Marco and

Genabith, Josef van. "Aligning Immanuel Kant's Work and its Translations". In *CLARIN: The*

Infrastructure for Language Resources. Edited by Fišer, D. and Witt, A. Berlin, Boston: De Gruyter, 2022, pp. 727-746. https://doi.org/10.1515/9783110767377-029

Reimers, Nils, and Iryna Gurevych. 2020. "Making Monolingual Sentence Embeddings Multilingual Using Knowledge Distillation." In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics*. https://arxiv.org/abs/2004.09813.

Viglianti, Raffaele. Why TEI Stand-off Markup Authoring Needs Simplification. Journal of the Text Encoding Initiative vol. 10 (2019). https://journals.openedition.org/jtei/1838. doi:https://doi.org/10.4000/jtei.1838.

Spadini, Elena and Turska, Magdalena. XML-TEI Stand-off Markup: One Step Beyond. Digital Philology: *A Journal of Medieval Cultures* vol. 8 (2019) pp. 225-239. doi:https://doi.org/10.1353/dph.2019.0025.

Suzgun, Mirac, Stuart M. Shieber, and Dan Jurafsky. 2023. "String2string: A Modern Python Library for String-to-String Algorithms." https://arxiv.org/abs/2304.14395.

Wu, Shuo, Xiaodan Yin, and Yue Zhang. (2023). "Improving Sentence-Level Alignment with Contrastive Learning in Multilingual Contexts." In *Proceedings of the 2023 Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics*.