ZoneRW: verso un'integrazione con Kraken ed eScriptorium per il riconoscimento e la gestione avanzata delle regioni di interesse

Pietro Sichera[1], Angelo Mario Del Grosso[2], Laura Mazzagufo[3], Daria Spampinato[4]
[1]CNR ILIESI, Italia pietro.sichera@cnr.it
[2] CNR ILC, Italia - angelomario.delgrosso@cnr.it
[3]CNR ISTC, Italia - laura.mazzagufo@istc.cnr.it
[4]CNR ISTC, Italia - daria.spampinato@cnr.it

ABSTRACT (ITALIANO)
L'identificazione delle regioni di interesse (*region of interest*, ROI) nei documenti facsimilari è essenziale per un corretto processo di digitalizzazione e per lo studio sia di testi manoscritti sia di testi a stampa. Questo contributo presenta un'estensione di funzionalità del software ZoneRW (nato nel contesto dell'edizione scientifica digitale *Bellini Digital Correspondence*), per integrare il tool Kraken, dedicato al rilevamento automatico delle zone e al riconoscimento automatico del testo, e l'ambiente digitale eScriptorium per la gestione avanzata del processo di digitalizzazione a partire da repertori di immagini di documenti testuali. Il workflow proposto utilizza il formato XML-PAGE per garantire interoperabilità e scalabilità, permettendo contestualmente di definire, modificare ed esportare le regioni di interesse verso eScriptorium. Inoltre, le nuove funzionalità di ZoneRW consentono di collegare i documenti di descrizione di immagini mediante protocollo IIIF (manifest IIIF). Il contributo evidenzia e discute le differenze tra i formati XML-PAGE e XML-ALTO nonché le prospettive per futuri sviluppi dello strumento ZoneRW nel contesto delle Digital Humanities.
Parole chiave: ZoneRW; regioni di interesse; Kraken; eScriptorium; digitalizzazione.

ABSTRACT (ENGLISH)
ZoneRW: Towards Integration with Kraken and eScriptorium for Advanced Recognition and Management of Regions of Interest
The detection of regions of interest (ROI) in facsimile documents is essential for a proper digitization process and for the study of both manuscript and printed texts. This contribution presents a feature extension of the ZoneRW software (created in the context of the *Bellini Digital Correspondence* digital scholarly edition), to integrate the Kraken tool, dedicated to automatic zone detection and automatic text recognition, and the eScriptorium digital environment for advanced management of the digitization process from image repositories of textual documents. The proposed workflow uses the XML-PAGE format to ensure interoperability and scalability, while simultaneously allowing regions of interest to be defined, edited, and exported to eScriptorium. In addition, the new ZoneRW capabilities allow image description documents to be linked via IIIF protocol (manifest IIIF). The paper highlights and discusses the differences between XML-PAGE and XML-ALTO formats as well as prospects for future developments of the ZoneRW tool in the Digital Humanities context.
Keywords: ZoneRW; regions of interest; Kraken; eScriptorium; digitization.

1. INTRODUZIONE
Il riconoscimento delle regioni di interesse consente di delimitare aree rilevanti presenti all'interno di documenti. Il riconoscimento e la classificazione in automatico di tali regioni, insieme al relativo contenuto, facilita la trascrizione e l'interpretazione delle fonti digitalizzate.
ZoneRW,[1] (Sichera P., 2023) strumento sviluppato originariamente nel contesto delle attività volte alla realizzazione dell'edizione scientifica digitale *Bellini Digital Correspondence*,[2] è una applicazione stand-alone progettata per la gestione CRUD (Create, Read, Update, Delete) delle zone codificate nei file XML-TEI dell'edizione, mediante un'interfaccia user-friendly e intuitiva. La caratteristica principale di ZoneRW prevedeva inizialmente il trasferimento delle coordinate delle zone di interesse in un file XML-TEI; successivamente l'applicativo si è dimostrato uno strumento estremamente flessibile, che ha consentito di estendere il suo utilizzo verso nuove funzionalità. Grazie a questa flessibilità, ZoneRW si è evoluto in un

[1] La versione di ZoneRW è in fase di test. La versione stabile attualmente pubblicata è disponibile al link https://github.com/pierpaolosichera/ZoneRW (cons. 26/01/2025)

[2] *Bellini Digital Correspondence* - https://bellinicorrespondence.cnr.it/ (cons. 26/01/2025)

componente in grado di gestire l'interoperabilità tra strumenti diversi. In particolare, la prima estensione è rivolta all'integrazione con Kraken,[3] un sistema open source per il rilevamento automatico delle zone di interesse e il riconoscimento ottico del testo, e, in secondo luogo, con eScriptorium,[4] una piattaforma per la gestione avanzata dei dati provenienti da campagne di digitalizzazione e il supporto di workflow collaborativi. Queste nuove caratteristiche hanno ampliato notevolmente le potenzialità del software.
L'estensione proposta consente di collegare i suddetti strumenti grazie al formato XML-PAGE,[5] uno standard per la descrizione del layout, della struttura e dei contenuti testuali presenti in documenti digitalizzati. Questo approccio garantisce interoperabilità e scalabilità nonché la possibilità di definire, modificare ed esportare zone di interesse verso l'ambiente eScriptorium. Inoltre, ZoneRW permette di associare ad ogni immagine il relativo documento manifest IIIF (International Image Interoperability Framework),[6] migliorando l'accessibilità e la condivisione delle risorse digitali.
Il beneficio di questa integrazione è duplice: da un lato, semplificare il workflow di digitalizzazione riducendo l'intervento manuale; dall'altro, migliorare l'accuratezza del riconoscimento automatico del testo grazie all'uso combinato di ZoneRW e Kraken. Il presente contributo discute inoltre le differenze tra i due formati di riferimento per la rappresentazione di risorse testuali in ambienti OCR/HTR, vale a dire XML-PAGE e XML-ALTO,[7] evidenziando i vantaggi offerti dal primo formato rispetto al secondo nell'ambito delle Digital Humanities.

2. STATO DELL'ARTE

Diverse soluzioni software dedicate al rilevamento automatico delle zone di interesse in documenti scritti si sono affermate nel tempo (Pinche A., 2023), con approcci che spaziano dall'elaborazione manuale alla completa automazione (Clérice T., 2022). Tra le più note si annovera TEIZoner (Dumont B., 2022),[8] uno strumento online che permette di segmentare i documenti in zone e generare file XML-TEI (Burnard L., 2014). Tuttavia, TEIZoner non consente di memorizzare i risultati della segmentazione in modo permanente per eventuali ulteriori modifiche: questo limite riduce l'efficienza dei workflow che richiedono interazioni ripetute, attività collaborative o l'integrazione con altri strumenti. In questo contesto, ZoneRW si distingue offrendo funzionalità CRUD per la gestione avanzata delle zone, permettendo agli utenti di apportare modifiche dinamiche e persistenti alle informazioni. Un altro strumento online di annotazione è Recogito,[9] che consente di creare un account dove memorizzare i propri progetti. Lo spazio di memorizzazione però è molto limitato (200MB), e rende necessario, nel caso di progetti di grandi dimensioni, l'impiego di un server IIIF per il caricamento delle immagini da annotare. Un ulteriore tool da segnalare è IMA[10] (Image annotation tool), un sistema di annotazione per singole immagini. Il tool esporta in formato XML-TEI, ma non gestisce le <surface> multiple.
Le piattaforme digitali per la trascrizione, l'annotazione e l'analisi di testi, come Transkribus[11] ed eScriptorium, sono focalizzati sull'attività di trascrizione automatica di testi (ATR). Pur eccellendo nell'analisi testuale, queste piattaforme richiedono, almeno inizialmente, un intervento manuale per la configurazione delle regioni di interesse, specialmente in contesti complessi. Strumenti come Kraken e Tesseract[12] hanno mostrato notevoli progressi nel rilevamento automatico delle zone e nel riconoscimento testuale (Kay A., 2007), ma non offrono nativamente funzionalità grafiche moderne[13] per la modifica delle zone di interesse, così come non supportano nativamente workflow completamente integrati con formati di rappresentazione standard come XML-PAGE o XML-ALTO.

---

[3] https://kraken.re/main/index.html (cons. 26/01/2025)

[4] https://escriptorium.readthedocs.io/en/latest/ (cons. 26/01/2025)

[5] https://github.com/PRImA-Research-Lab/PAGE-XML (cons. 26/01/2025)

[6] https://iiif.io/guides/using_iiif_resources/ (cons. 26/01/2025)

[7] https://loc.gov/standards/alto/ (cons. 26/01/2025)

[8] http://teicat.huma-num.fr/zoner.php (cons. 26/01/2025)

[9] https://recogito.pelagios.org/ (cons. 10/04/2025)

[10] https://ima.coders.tools/ (cons. 26/01/2025)

[11] https://www.transkribus.org/ (cons. 26/01/2025)

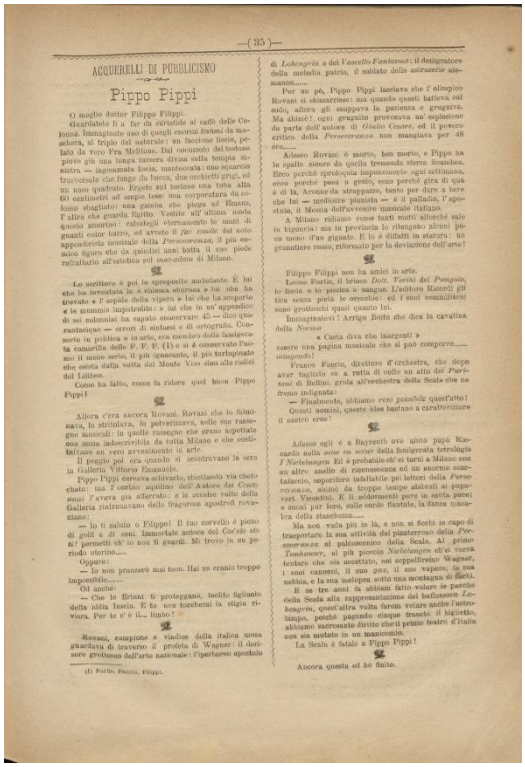
[12] https://github.com/tesseract-ocr/tesseract (cons. 26/01/2025)

[13] Ad esempio, Tesseract-OCR QT4 è una semplice GUI per Tesseract ma non viene aggiornata dal 2011 https://github.com/zdenop/tesseract-ocr-qt4gui (cons. 26/01/2025)

L'estensione di ZoneRW supera tali limiti grazie alla sua flessibilità e alla capacità di integrarsi con sistemi come Kraken ed eScriptorium. La possibilità di generare file XML-PAGE o XML-ALTO, entrambi standard consolidati nel panorama della digitalizzazione, consente a ZoneRW di garantire una piena interoperabilità nell'ambito di progetti complessi di ambito DH e perseguire al meglio i principi FAIR per la scienza aperta (Dumouchel, S & al., 2020).

3. IMPLEMENTAZIONE DEL WORKFLOW

L'implementazione del workflow proposto si basa sull'integrazione di ZoneRW con Kraken ed eScriptorium. ZoneRW è stato sviluppato utilizzando la piattaforma RAD per database 4th Dimension (4D),[14] che offre tra gli altri un supporto nativo per XML tramite la libreria Xerces della Apache Foundation. La gestione delle strutture XML avviene tramite lo standard DOM (Document Object Model), che consente di accedere e manipolare in maniera nativa ed efficiente ogni elemento della struttura XML. Questo approccio permette non solo di navigare l'albero XML, ma anche di aggiornare, riorganizzare o aggiungere elementi in tempo reale senza necessità di operazioni intermedie complesse. La gestione dinamica dei nodi <surface>, <graphic> e <zone> avviene con estrema precisione, ottimizzando i tempi di elaborazione e garantendo la piena coerenza dei dati. Le regioni di interesse vengono rappresentate utilizzando il formato SVG (Scalable Vector Graphics), che garantisce un'accurata visualizzazione e modifica delle coordinate direttamente sull'immagine. L'interfaccia grafica di ZoneRW è progettata per essere user-friendly e reattiva, permettendo all'utente di lavorare agevolmente su immagini ad alta risoluzione. ZoneRW accetta due modalità di input: immagini accoppiate a file XML-TEI, oppure cartelle contenenti esclusivamente immagini. Nel primo caso, ZoneRW utilizza il tag <surface> per rappresentare la pagina e i relativi tag <graphic> e <zone> per identificare visivamente le regioni di interesse. Ogni zona può essere creata o modificata manualmente tramite un'interfaccia grafica intuitiva, che consente di aggiungere o correggere coordinate e identificativi delle zone senza necessità di intervento esterno. Nel secondo caso, quando vengono caricate cartelle di immagini senza XML associato, l'utente ha la possibilità di lanciare da ZoneRW un'istanza di Kraken per il rilevamento automatico delle zone di testo e il riconoscimento OCR. Kraken, un potente strumento open source basato su Python, utilizza modelli di deep learning per segmentare il layout delle immagini e riconoscere il testo, anche su documenti complessi.



ACQUERELLI DI PUBBLICISMO

Pippo Pippi

Fig.1 Riproduzione fotografica di una pagina della rivista «La Farfalla» (anno 1876, n.2, fascicolo 5, p. 35)

[14] https://www.4d.com (cons. 26/01/2025)

I file generati includono le coordinate delle zone e i dati di riconoscimento testuale, pronti per essere importati in ZoneRW per ulteriori modifiche o validazioni.

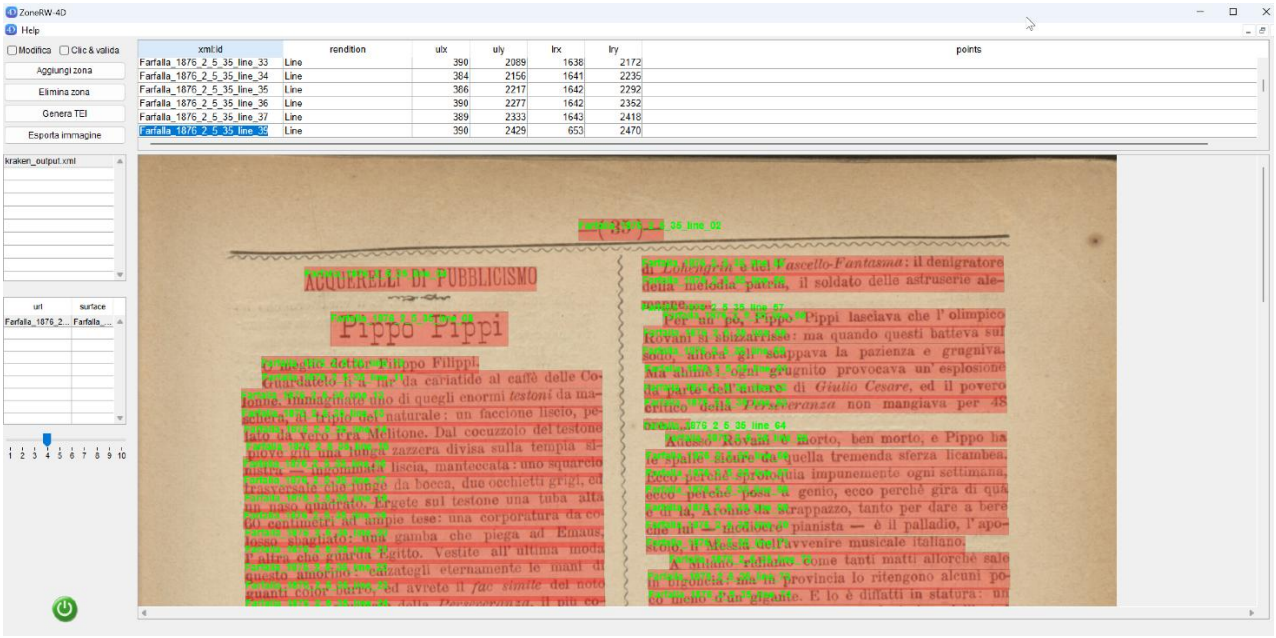


Fig.2 Le regioni di interesse della stessa pagina de «La Farfalla» importate in ZoneRW

ZoneRW consente di modificare, arricchire e correggere le zone rilevate e importate da Kraken. L'utente può inoltre associare ogni immagine a un manifest IIIF, un file in formato JSON standardizzato per descrivere e organizzare risorse digitali ospitate su server compatibili con il protocollo IIIF. Questo protocollo è progettato per facilitare la gestione e la condivisione di immagini ad alta risoluzione, fornendo metadati dettagliati e percorsi di accesso strutturati alle risorse. L'adozione dei manifest IIIF elimina la necessità di caricare manualmente le immagini in eScriptorium, consentendo invece un accesso diretto e referenziato alle immagini già ospitate su server IIIF. Tale approccio non solo migliora l'accessibilità e la tracciabilità delle risorse, ma ottimizza anche il workflow riducendo i tempi e i costi associati alla duplicazione dei dati.
Una volta completata l'elaborazione, ZoneRW esporta le informazioni sulle zone in formato XML-PAGE che, grazie alla sua capacità di rappresentare layout complessi, è il principale standard supportato da eScriptorium, permettendo un'integrazione diretta con quest'ultimo. eScriptorium consente di esportare i dati anche nel formato XML-ALTO, che è però maggiormente orientato alla descrizione del testo.
Ogni file XML-PAGE (vd. Fig.3) descrive una pagina digitale, tipicamente un'immagine derivata dalla digitalizzazione di un documento fisico, attraverso una struttura gerarchica ricca di informazioni.

```xml
<?xml version="1.0" encoding="UTF-8"?>
<PcGts xmlns="http://schema.primaresearch.org/PAGE/gts/pagecontent/2019-07-15">
  <Metadata>
    <Creator>ZoneRWProcessor</Creator>
    <Created>2025-01-01T12:00:00</Created>
    <LastChange>2025-01-01T12:30:00</LastChange>
  </Metadata>
  <Page imageWidth="2480" imageHeight="3508" imageFilename="ilmomento.png">
    <TextRegion id="region_1">
      <Coords points="100,100 500,100 500,200 100,200"/>
      <TextEquiv>
        <Unicode>Primo esempio di testo.</Unicode>
      </TextEquiv>
    </TextRegion>
    <TextRegion id="region_2">
      <Coords points="600,300 1000,300 1000,400 600,400"/>
      <TextEquiv>
        <Unicode>Secondo esempio di testo.</Unicode>
      </TextEquiv>
    </TextRegion>
  </Page>
</PcGts>
```

Fig.3 File in formato XML-PAGE

Il file contiene metadata che documentano il processo di creazione, tra cui informazioni sulla provenienza dell'immagine, i dettagli tecnici del processo di digitalizzazione (ad esempio, la risoluzione o il software utilizzato) e altri dati descrittivi.
La struttura del tag <Page> rappresenta l'elemento principale, che descrive l'immagine della pagina stessa con informazioni come dimensioni (gli attributi @imageWidth e @imageHeight) e un riferimento all'immagine sorgente (attributo @imageFilename). Questo elemento contiene una successione ordinata di <TextRegion>, che rappresentano le diverse zone di testo presenti nella pagina. Ogni <TextRegion> è ulteriormente arricchito da elementi figli come <Coords>, che definisce le coordinate geometriche delle zone di testo mediante poligoni o rettangoli, e <TextEquiv>, che include il contenuto testuale associato ad ogni zona. Allo stesso livello di <TextRegion> il tagset PAGE supporta una rappresentazione di altre tipologie di regioni, come <ImageRegion> per le immagini contenute nella pagina, <SeparatorRegion> per le linee di separazione, e <TableRegion> per le tabelle, ognuno sempre con le relative coordinate.
XML-ALTO (vd. Fig.4), invece, è maggiormente focalizzato sulla descrizione testuale e sullo stile dei caratteri, risultando ridondante rispetto alle necessità di esportazione di ZoneRW.

```xml
<?xml version="1.0" encoding="UTF-8"?>
<alto xmlns="http://www.loc.gov/standards/alto/ns-v4#">
  <Description>
    <MeasurementUnit>pixel</MeasurementUnit>
    <sourceImageInformation>
      <fileName>ilmomento.png</fileName>
    </sourceImageInformation>
    <OCRProcessing ID="ocr_1">
      <ocrProcessingStep>
        <processingSoftware>
          <softwareName>ZoneRWProcessor</softwareName>
          <softwareVersion>1.0</softwareVersion>
        </processingSoftware>
      </ocrProcessingStep>
    </OCRProcessing>
  </Description>
  <Layout>
    <Page WIDTH="2480" HEIGHT="3508" PHYSICAL_IMG_NR="1" ID="page_1">
      <PrintSpace>
        <TextBlock ID="block_1" HPOS="100" VPOS="100" WIDTH="400" HEIGHT="100">
          <TextLine>
            <String CONTENT="Primo esempio di testo."/>
          </TextLine>
        </TextBlock>
        <TextBlock ID="block_2" HPOS="600" VPOS="300" WIDTH="400" HEIGHT="100">
          <TextLine>
            <String CONTENT="Secondo esempio di testo."/>
          </TextLine>
        </TextBlock>
      </PrintSpace>
    </Page>
  </Layout>
</alto>
```

Fig.4 Lo stesso file mostrato in Fig.3, in formato XML-ALTO

4. CONCLUSIONI E SVILUPPI FUTURI

ZoneRW si è dimostrato uno strumento versatile e adattabile per la gestione delle zone di interesse nei documenti digitalizzati. La sua capacità di integrare funzionalità di rilevamento automatico con quelle di modifica manuale dei dati lo rende uno strumento molto efficace all'interno di progetti complessi di ambito DH. L'integrazione con strumenti quali Kraken ed eScriptorium, attraverso il formato XML-PAGE, garantisce un flusso di lavoro fluido e interoperabile, migliorando l'efficienza e riducendo il carico operativo per gli utenti, seguendo principi di scienza aperta e FAIR. Il supporto per i manifest IIIF rappresenta un ulteriore avanzamento nella gestione e accessibilità delle risorse digitali.
Come per qualsiasi prodotto software, esistono sono ampie opportunità per migliorare ulteriormente ZoneRW, ampliandone le potenzialità.
Un primo sviluppo previsto è il supporto del formato XML-ALTO, sia in fase di importazione delle ROI che in esportazione delle stesse.
Uno sviluppo futuro importante consiste nella realizzazione di una versione web-based di ZoneRW, che permetterebbe agli utenti di accedere alle sue funzionalità direttamente da un browser senza necessità di

installazioni locali. Questo approccio aumenterebbe notevolmente la diffusione dello strumento, rendendolo accessibile a una comunità di utenti più ampia e facilitando la collaborazione in progetti distribuiti.
Un'altra possibile direzione di sviluppo riguarda l'esposizione di interfacce di programmazione (API) compatibili con il formato richiesto da sistemi di integrazione e orchestrazione di servizi, quale ad esempio il MarketPlace in corso di sviluppo all'interno di H2IOSC,[15] il progetto che federa i nodi italiani delle infrastrutture europee CLARIN, DARIAH, E-RIHS e OPERAS. Il MarketPlace di H2IOSC avrà la capacità di orchestrare API (Sichera & al. 2024) provenienti da sorgenti e provider (endpoint) differenti, facilitando la condivisione e l'interoperabilità delle risorse digitali tra diversi sistemi e progetti. Questa implementazione consentirebbe a ZoneRW di integrarsi agevolmente con altre piattaforme e applicazioni disponibili sul MarketPlace.
In conclusione, ZoneRW si configura non solo come uno strumento per la gestione delle regioni di interesse, ma come un utile componente in un ecosistema di strumenti interoperabili, contribuendo alla valorizzazione e alla fruizione delle risorse digitalizzate in contesti accademici, archivistici e museali. Il suo sviluppo continuo risponde alle esigenze in evoluzione delle Digital Humanities, promuovendo la sostenibilità e l'innovazione nei progetti di trascrizione, codifica e valorizzazione di risorse testuali.

BIBLIOGRAFIA

Burnard, L. (2014). What is the Text Encoding Initiative? (1-). OpenEdition Press. https://doi.org/10.4000/books.oep.426

Clérice, T. (2022). You Actually Look Twice At it (YALTAi): Using an object detection approach instead of region segmentation within the Kraken engine. ArXiv, abs/2207.11230. https://api.semanticscholar.org/CorpusID:251018572

Cristofaro, S., Del Grosso, A.M., Mazzagufo, L., Sichera, P., & Spampinato D. (2023). "Bellini Digital Correspondence: A Model for Making Collaborative Digital Scholarly Editions." In *2023 7th IEEE Congress on Information Science and Technology (CiSt)*, 615–20. IEEE. https://doi.org/10.1109/CiSt56084.2023.10409920

Del Grosso, A. M., Spampinato, D. (Eds.). (2023). Bellini Digital Correspondence. CNR Edizioni, <https://bellinicorrespondence.cnr.it>. ISBN: 978-88-8080-562-5.

Dumont, B. (2022). "Review of 'TEI Critical Apparatus Toolbox: Web-based tools for ongoing XML-TEI editions'." RIDE 15. doi: 10.18716/ride.a.15.4.

Dumouchel, S., Blotière, E., Breitfuss, G., Chen, Y., Donato, F. D., Eskevich, M., Forbes, P., Georgiadis, H., Gingold, A., Gorgaini, E., Moranville, Y., Pohle, S., Paoli, S. de, Petitfils, C., & Toth-Czifra, E. (2020). GOTRIPLE: A User-Centric Process to Develop a Discovery Platform. Information, 11(12), 563. https://doi.org/10.3390/info11120563

Kay, A. (2007). Tesseract: An open-source optical character recognition engine. Linux J., 2007(159), 2.

Pinche, A. (2023). Generic HTR Models for Medieval Manuscripts. The CREMMALab Project. J. Data Min. Digit. Humanit., 2023. https://api.semanticscholar.org/CorpusID:264424158

Sichera P. (2023). pierpaolosichera/ZoneRW: v0.5-beta (7th IEEE CiSt'23) (v0.5-beta). Zenodo. https://doi.org/10.5281/zenodo.5599509

Sichera, P., Marras, C., & Pasini, E. (2024). Orchestrazione API per workflow applicativi nell'ambito delle Digital humanities. Zenodo. https://doi.org/10.5281/zenodo.14187534

[15] https://www.h2iosc.cnr.it/ (cons. 26/01/2025)