Revista de Humanidades Digitales

La digitalización del *Diccionario Técnico de la Música* de Felipe Pedrell

The digitization of Felipe Pedrell's Diccionario Técnico de la Música

Eugenia GALLEGO
eugenia.gallego@universidadviu.com
Universidad Internacional de Valencia
https://orcid.org/0000-0002-9159-5817

María ORDIÑANA GIL
mordinana@universidadviu.com
Universidad Internacional de Valencia
https://orcid.org/0000-0001-9606-3572

David PÉREZ-SEVILLA PÉREZ-MEDRANO
david.perez-sevilla@alum.unirioja.es
Universidad de La Rioja
https://orcid.org/0009-0007-0758-022X

Teresa CASCUDO GARCÍA-VILLARACO
teresa.cascudo@unirioja.es
Universidad de La Rioja
https://orcid.org/0000-0002-8926-068X

Rosario LÓPEZ GÓMEZ
rosario.lopez@unirioja.es
Universidad de La Rioja
https://orcid.org/0000-0002-9069-395X

RESUMEN

Los diccionarios del siglo XIX representan una fuente inestimable para comprender la evolución del lenguaje y la cultura de la época. Sin embargo, la digitalización de estos textos puede resultar una tarea ardua debido a la complejidad de la tipografía y la presencia de errores tipográficos. El método aplicado en estudios anteriores ha estado basado en la lectura de cada fuente y en el análisis del contexto de publicación. Se ha obviado la utilización de las nuevas posibilidades abiertas por las Humanidades Digitales, así como la aplicación de los resultados obtenidos para responder a las necesidades de este campo de investigación. En este artículo se describe el proceso seguido para el estudio y la digitalización del *Diccionario técnico de la música* de Felipe Pedrell (1894) a través de la aplicación de herramientas informáticas y de gestión de la información digitalizada. Por un lado, se han explorado plataformas de inteligencia artificial, en concreto las herramientas de *machine learning* Transkribus y Nanonets, con el objetivo de entrenar modelos propios de reconocimiento de escritura y transcripción.

ABSTRACT

Nineteenth-century dictionaries are an invaluable resource for understanding the linguistic and cultural evolution of that period. However, digitizing these texts can be a formidable challenge due to the complexity of their typography and the prevalence of typographical errors. The methodology employed in prior studies has relied on scrutinizing each source and analyzing its publication context. The potential offered by Digital Humanities has been neglected, as well as leveraging the findings to address the demands of this research field.

This paper delineates the approach undertaken for the study and digitization of Felipe Pedrell's *Technical Dictionary of Music* (1894), utilizing computational tools and digital information management strategies. It explores artificial intelligence platforms, particularly machine learning tools such as Transkribus and Nanonets, aimed at training bespoke models for handwriting recognition and transcription.

RHD 9 (2024)

ISSN
2531-1786

Revista de Humanidades Digitales http://revistas.uned.es/index.php/RHD/index
Recibido 29/04/2024 – Aceptado 18/06/2024

Por otro lado, se han seguido metodologías clásicas de digitalización y reconocimiento de caracteres y se ha realizado un proceso de análisis, filtrado y tratamiento posterior a través de algoritmos que mejoran el proceso de transcripción. Finalmente, se ha desarrollado un servicio web de consulta del *Diccionario* que recoge los resultados de este trabajo y lo convierte en un recurso de calidad en el ámbito de la investigación musicológica y en la preservación de documentos históricos.

PALABRAS CLAVE

Digitalización, *machine learning*, diccionario, musicología, Humanidades Digitales.

Additionally, it adheres to conventional digitization and character recognition methodologies, followed by an analysis, filtering, and enhancement process using algorithms to refine the transcription procedure. Ultimately, a web-based consultation service for the *Dictionary* has been developed, compiling the outcomes of this endeavor to provide a high-quality resource for musicological research and the preservation of historical documents.

KEYWORDS

Digitization, Machine Learning, Dictionary, Musicology, Digital Humanities.

1. DESCRIPCIÓN DEL PROYECTO

La figura de Felipe Pedrell (1841-1922) es sobradamente conocida, tanto en su vertiente de compositor como en la de musicólogo. Considerado el creador de la musicología española moderna (Carreras, 2001) fundó la *Ilustración musical hispano-americana* (1888-1896), así como el *Diccionario técnico de la música* (1894) y el *Diccionario biográfico y bibliográfico de músicos y escritores de música españoles, portugueses e hispano americanos antiguos y modernos* (1897). El primero, fuente principal de este proyecto, se incluye dentro de la "tradición de obras que se denominan *diccionarios de música*" que se inicia en la segunda mitad del siglo XIX (Quilis Merín, 2019). La labor lexicográfica pedrelliana no ha sido objeto de atención en el ámbito de la musicología. De hecho, tampoco la lexicología y terminologías musicales han suscitado demasiado interés a juzgar por el reducido número de publicaciones dedicadas al estudio de dicha tradición. En 1970, José Subirá, quien desarrolló una relevante labor como lexicógrafo especializado en el dominio musical, abordó la cuestión del vocabulario de la música en español de forma panorámica (Subirá Puig, 1970). Más recientemente, desde la musicología, Justiniano López se ha centrado en las voces musicales del *Diccionario de autoridades* (1726-1739) (Justiniano López, 2019). En el campo de la lingüística, existen estudios monográficos sobre el *Diccionario de la música, técnico, histórico y bibliográfico* de María Luisa Lacal (1899) (Quilis Merín, 2019), sobre los lemas con la marca "música" incluidos en el DRAE (López-Vallejo y López Aguirre, 2021) o sobre los diccionarios técnicos de la música del siglo XIX (Rubio Amondarain, 2023).

Los diccionarios del siglo XIX representan una fuente muy rica para comprender la evolución del lenguaje y la cultura de la época, por lo que vale la pena, no solo abordar su estudio, sino facilitar su consulta haciendo uso de las herramientas que nos ofrece actualmente la informática. Este es el marco en el que se mueve el proyecto nacional "Léxico en español y Ontología de la Música" (LexiMus), que trata de cubrir una doble laguna: no solo la de los estudios sobre diccionarios, sino, sobre todo, la de la escasa aplicación de herramientas informáticas y de gestión de la información digitalizada en la investigación musicológica. El Subproyecto 2 de LexiMus, del que este

artículo es un resultado preliminar, persigue digitalizar de forma estructurada los diccionarios técnicos de música publicados en español en el siglo XIX; elaborar una librería de términos musicales (*termbank*); elaborar un modelo conceptual digital a partir de este corpus y realizar una aplicación web en formato abierto que lo haga accesible a todas las personas interesadas.

El método aplicado en los estudios citados ha estado basado en la lectura de cada fuente y en el análisis del contexto de publicación, así como en la utilización de reproducciones en papel o en la transcripción, por medios humanos, del texto que contienen. Esta es la situación corriente en el ámbito de la terminología y lexicología de temática musical y de carácter histórico. En este momento, todavía no se ha generalizado en estos campos la utilización de las nuevas posibilidades abiertas por las Humanidades Digitales, así como la aplicación de los resultados obtenidos para responder a las necesidades de este campo de investigación. En este momento, tenemos disponibles, en diversos repositorios institucionales, tanto diccionarios técnicos de la música como enciclopedias y diccionarios generales que contienen términos musicales. Las fuentes están digitalizadas en repositorios abiertos, tales como la Biblioteca Digital Hispánica o la Biblioteca Virtual de la Filología Española. Sin embargo, los PDF allí disponibles son muy mejorables y, desde luego, insuficientes para poder aplicar sobre ellos herramientas que, de manera automática o semiautomática, mejoren los procesos de consulta y de extracción de información. En consecuencia, el primer escollo con el que se enfrentaron los miembros del equipo del Subproyecto 2 de LexiMus fue el de conseguir transformar estas fuentes en documentos estructurados y manipulables por los ordenadores. La digitalización estructurada de estos textos –de forma a transformarlos en bases de datos o documentos en texto plano libres de errores– puede resultar ardua por la complejidad y variedad de la maquetación y de la tipografía, así como por los errores tipográficos que también pueden contener en ocasiones. Otro problema adicional es el que se deriva de la conveniencia de normalizar la ortografía atendiendo a las reglas actualmente vigentes.

En este artículo se describe el proceso seguido en el desarrollo de LexiMus para intentar sortear estas dificultades. En particular, describe el trabajo realizado para alcanzar los objetivos del mencionado Subproyecto 2, residido en la Universidad de La Rioja que, a su vez, fueron anticipados, por miembros del mismo equipo, en un proyecto propio de la Universidad Internacional de Valencia (VIU). En este punto cabe señalar que, si bien la búsqueda y aplicación de herramientas y tecnologías se ha adecuado a los datos a extraer y al formato de fuente específico del *Diccionario*, se han tomado como referencia trabajos modélicos sobre estructuración automática y ediciones digitales, tales como García Serrano y Castellanos González (2017) y Mateos Frühbeck (2021).

Por otra parte, en lo que concierne al uso de ChatGPT[1], en el momento mismo en que se puso a disposición de la comunidad investigadora, comenzaron a probarse las posibilidades que abría, especialmente en el ámbito de la lingüística y, por supuesto, también de la lexicografía (de Schryver y Joffe, 2023).

[1] Accesible desde: https://www.openai.com/chatgpt.

2. OBJETIVOS

Los objetivos planteados en esta fase preliminar del proyecto y cuyo resultado exponemos en este artículo son:

- Digitalizar de forma estructurada los diccionarios técnicos de la música publicados en español en el siglo XIX, empezando por el Diccionario técnico de la música de Felipe Pedrell (1894).
- Generar un servicio de consulta electrónico vía web.

Estos objetivos se encuadran en los generales del Suproyecto 2 que forma parte de LexiMus:

- Elaborar una librería de términos musicales a partir de dicha digitalización.
- Elaborar un modelo conceptual digital que parta del glosario resultante.
- Difundir los resultados de las investigaciones en congresos y publicaciones académicas.

3. METODOLOGÍA

En esta primera fase del proyecto se ha trabajado en una doble vía. Por un lado, la utilización de programas de inteligencia artificial (IA) para la transcripción de documentos históricos. El uso de estos programas permite el procesamiento de imágenes digitalizadas, el reconocimiento del diseño y la estructura de los documentos y la extracción de la información según patrones aprendidos. Por lo tanto, a través de herramientas de *machine learning* para el procesado del lenguaje como Transkribus y Nanonets, se han entrenado modelos propios de IA que reconocen el tipo de escritura del *Diccionario técnico de la música* y generan una transcripción lo más fiel a la fuente original. Se ha realizado una exploración de este *software* y los resultados se exponen en los apartados cuatro y cinco de este artículo. En paralelo al uso de técnicas de aprendizaje automático para el desarrollo de modelos de reconocimiento de caracteres, se ha seguido también un proceso tradicional de digitalización de diccionarios, en el que, partiendo de un procesado de imágenes, se realiza un análisis de OCR y se diseñan algoritmos para la extracción del texto en base a patrones identificados. Se realiza seguidamente un tratamiento de la información y un posterior procesado. Los resultados en esta línea de trabajo han permitido, además, la creación de un sistema web de consulta de los términos del *Diccionario* digitalizado y completar esta búsqueda con características más avanzadas que sí permiten un contenido digitalizado como la búsqueda sugerida, la clasificación de términos, la búsqueda relacional, etc.

4. TRANSKRIBUS. RECONOCIMIENTO DE CARACTERES A TRAVÉS DE UN MODELO DE INTELIGENCIA ARTIFICIAL

La extracción de datos de las fuentes y la creación de una base de datos siempre es uno de los pasos más costosos en cuanto tiempo requerido, más aún si la fuente no está editada y se trata de manuscritos. Existe una herramienta que tiene a la vez un aplicativo web y una versión cliente para instalar en el ordenador que está alcanzando una gran popularidad y que fue premiada en los DH Awards de 2022. Se trata de Transkribus, una plataforma de IA para la transcripción de documentos históricos[2]. Transkribus (Kahle et al., 2017) permite reconocer automáticamente el texto, el diseño y la estructura de los documentos con ayuda de la IA, más concretamente a través del motor PyLaia con tecnología HTR (handwritten text recognition ó reconocimiento de texto manuscrito). HTR es una tecnología similar a OCR, pero no se centra en letras individuales, sino que escanea y procesa la imagen de líneas enteras e intenta descodificar estos datos. La principal diferencia para el usuario es que la etapa de análisis del trazado está integrada en el motor OCR, mientras que en el HTR es un paso independiente. Por ello, con esta herramienta se pueden entrenar modelos de IA para que se ajusten a documentos específicos (Colutto et al., 2019; Perdiki, 2022). Para crear un modelo propio que reconozca un determinado estilo de escritura, la documentación de Transkribus recomienda entrenar entre 5000 y 15000 palabras (aproximadamente 25-75 páginas), dependiendo del formato del texto. En Ströbel, Clematide y Volk (2020) se estudia la cantidad de datos y épocas (número de veces que el algoritmo de aprendizaje necesita para completar el conjunto de datos) que son necesarios para un entrenamiento satisfactorio de la red neuronal. Otros estudios indican el porqué y cuándo se requieren más épocas y datos y el cuidado que hay que tener para que no se produzca un sobreajuste de la red que empeore los resultados (Rabus, 2019). Cabe tener en cuenta que la precisión de HTR no es completa, pero es posible obtener tasas de error de palabras y caracteres muy bajas. Todo modelo cuya tasa de error esté por debajo del 10 % se considera satisfactorio pues significa que el 90 % de los caracteres de una transcripción generada automáticamente son correctos. Si se trabaja con un texto impreso se suele necesitar una menor cantidad de palabras para el entrenamiento. Asimismo, existen modelos públicos que pueden emplearse para la transcripción o utilizarlos de base. Por otro lado, además del reconocimiento textual, Transkribus también permite enriquecer el material con metadatos fácilmente exportables en múltiples formatos incluidos PDF y XML.

4.1. Etapas de creación de un modelo de IA con Transkribus

Para esta investigación, el proceso de concebir un modelo HTR de transcripción del texto consistió en tres etapas. En primer lugar, se evaluaron los límites de la plataforma Transkribus y se han explorado las posibilidades de lograr los objetivos propuestos. Para ello se ha entrenado al

[2] Accesible desde: https://www.transkribus.org/.

programa con las primeras treinta páginas del *Diccionario*, transcritas a mano, creando varios modelos de reconocimiento de prueba, donde se cambiaron parámetros y orden de las estructuras del entrenamiento, para evidenciar las posibilidades en formato imprenta, especialmente con grafía del siglo XIX. Una vez que el programa fue evaluado y se instauró un método válido para un mayor reconocimiento y mejor entrenamiento, se subió el *Diccionario* completo, digitalizado previamente utilizando ORC. Seguidamente, a partir de un entrenamiento previo mixto, utilizando tanto la base transcrita a mano como una base preexistente denominada Transkribus Print M1 (que contaba con un entrenamiento previo de 5068310 palabras de imprenta y un índice de CER de entrenamiento del 2.00 %), se generaron las transcripciones necesarias para el entrenamiento del programa. Los errores detectados se corrigieron manualmente, lo que permitió la creación de un modelo de reconocimiento a partir de dichos recursos. El índice de CER (Character Error Rate) conseguido en el entrenamiento es de 1.00 % (Figura 1), con un entrenamiento de trescientas intentos, dado que aprende a base de iteraciones (Epochs) (Figura 2).

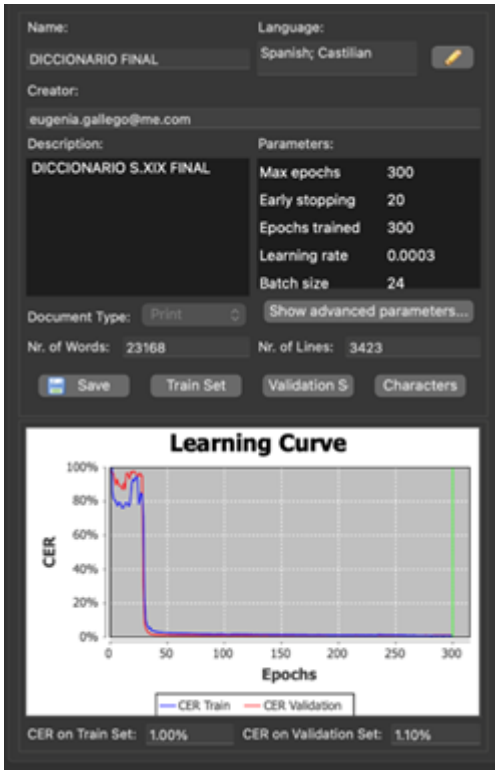


Figura1. Figura SEQ Figura * ARABIC 2. CER del modelo.
Fuente: elaboración propia.

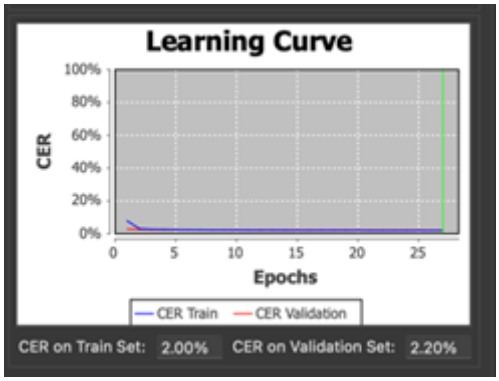


Figura 2. Figura SEQ Figura * ARABIC 1. Curva de aprendizaje del modelo.
Fuente: elaboración propia.

En cuanto a la evaluación de Transkribus, la experiencia previa con otros textos de imprenta desempeñó un papel crucial. Transkribus, en realidad, es una plataforma que promete resultados positivos con documentos impresos, principalmente debido a su uniformidad en comparación con otras obras. Sin embargo, es difícil lograr transcripciones fiables en un rango de 100 %. En este caso, hemos alcanzado un nivel de precisión cercano al 99 % en el texto correcto, pero no una escritura libre de errores. Aunque muchos de estos errores podrían no ser significativos, la transcripción no sería fiable si no se lleva a cabo la revisión humana de los resultados.

Otra observación propia es que Transkribus funciona mucho mejor cuando se entrena con un solo texto como fuente primaria de información.

Las imágenes digitalizadas de las fuentes provienen del portal Biblioteca Digital Hispánica y tienen una resolución superior a 300 ppp. Para que el programa funcione correctamente, se cargan, en primer lugar, las imágenes digitalizadas de las páginas en la plataforma, a continuación se dividen las imágenes en secciones y líneas de texto (*layout analysis*) y después se procede a la transcripción de algunas páginas de la fuente (*ground truth*) que se consideran válidas para servir de fundamento. Gracias al entrenamiento del programa y la creación de un modelo de HTR, es posible realizar la transcripción automática de las páginas restantes.

Transkribus es una herramienta que se apoya en un servicio cliente-servidor, siendo los recursos en la nube de acceso libre. En contraste, la obtención de las transcripciones generadas mediante el reconocimiento automático está sujeta a ciertas restricciones.

Cabe mencionar que nuestra fuente estaba conformada por un formato de imprenta de doble columna, con algunas imágenes entre ellas. Este hecho complicó la correcta interpretación de cada línea de texto por lo que tuvimos que recurrir a la función interna de Transkribus llamada *page to page layout analysis* (P2PaLA), que nos permitió segmentar las dos columnas de texto y posteriormente, asignar (y, en algunos casos, reasignar) un orden de lectura de las líneas correspondiente a nuestro *Diccionario*. La base para la creación de un modelo propio, como hemos mencionado, está conformada por diez páginas transcritas a mano, junto con el modelo preexistente Transkribus M1 print. El modelo resultante se ha denominado “Diccionario s.XIX final” y está a disposición de los otros usuarios de Transkribus de manera pública.

Sus principales ventajas son:

- *Aprendizaje automático*. Transkribus se beneficia del aprendizaje automático para mejorar la precisión de la transcripción, lo que lo hace adecuado para textos antiguos y complejos.
- *Código abierto*. La herramienta es de código abierto y constantemente mejorada por la comunidad académica, lo que permite su adaptación y personalización.
- *Corrección automática*. La función de corrección automática ahorra tiempo en la corrección de errores tipográficos, acelerando el proceso de transcripción.
- *Interfaz de usuario intuitiva*. La interfaz de usuario amigable facilita la revisión y edición de transcripciones.

Sus principales desventajas son:

- *Dependencia de una mínima calidad de la imagen*. Haga clic aquí para escribir texto. Las imágenes de muy baja resolución o dañadas pueden resultar en transcripciones menos precisas.
- *Idioma específico*. La efectividad de Transkribus varía según el idioma y la tipografía del texto original. Puede requerir un entrenamiento más específico para idiomas menos comunes dado que no puedes utilizar bases compartidas.
- *Curva de aprendizaje*. Para aprovechar al máximo las capacidades de Transkribus, los usuarios pueden requerir tiempo para familiarizarse con la aplicación y su proceso de transcripción.
- *Coste económico*. Esta tecnología no es gratuita y, por lo tanto, es crucial entender su coste económico y evaluar su rentabilidad. Inicialmente, se dan créditos de prueba aunque después, para un uso amplio, es necesario su adquisición.
- *Estructuración deficiente*. Se da una estructuración deficiente para el caso del diccionario, dada la necesidad de extraer términos.

5. NANONETS. RECONOCIMIENTO DE CARACTERES A TRAVÉS DE UN MODELO DE INTELIGENCIA ARTIFICIAL

Dadas las dificultades en la extracción única de los términos del *Diccionario* desde la herramienta Transkribus, se ha empleado un software de OCR avanzado y aprendizaje profundo llamado Nanonets, una avanzada plataforma de procesamiento de lenguaje natural (NLP). Nanonets es una herramienta que emplea inteligencia artificial y que ofrece soluciones de aprendizaje automá-

tico y procesamiento de imágenes[3]. Está diseñada para la gestión de documentos de empresas, tales como facturas, DNI, etc., tanto para la digitalización de los documentos como para la extracción de información y ha sido utilizada también para la transcripción de manuscritos antiguos (Harish, 2024). Al igual que Transkribus, Nanonets requiere de un entrenamiento previo, aunque, en este caso, solo de la estructura o campos a extraer. Así despliega modelos de inteligencia artificial personalizados y puede extraer datos rápidamente e intuitivos de nuevos documentos.

El proceso de extracción de términos en el *Diccionario* de Pedrell con Nanonets implicó varios pasos clave:

- *Preparación de datos*. Se eligieron páginas que contenían términos del *Diccionario* en formatos variados, aunque destacando la negrita del término.
- *Entrenamiento del modelo*. Mediante el entrenamiento previo de Nanonets, se le enseñó a reconocer la estructura específica de los campos relevantes, en este caso, seleccionando las palabras de términos en situación de negrita.
- *Extracción de términos*. Una vez entrenado, Nanonets fue capaz de extraer los términos del *Diccionario* de manera rápida y precisa y exportarlo a un fichero Excel.
- *Validación y refinamiento*. Se verificó la precisión de la extracción y se realizaron ajustes para mejorar la calidad de los resultados.
- *Implementación continua*. Nanonets puede ser implementado de manera continua para la extracción de nuevos términos del *Diccionario* a medida que se actualizan los documentos.

Sus principales ventajas son:

- *Procesamiento de documentos más rápido y eficiente*. Nanonets puede procesar documentos no estructurados más rápido y con mayor precisión que el procesamiento manual.
- *Mejor calidad de datos*. Nanonets puede extraer información relevante de los documentos con mayor precisión que el procesamiento manual, lo que resulta en una mejor calidad de los datos recopilados.
- *Ahorro de tiempo*. Posee una enorme velocidad de procesamiento.
- *Escalabilidad*. Puede procesar grandes volúmenes de documentos.

Sus principales desventajas son:

- *Coste económico elevado*. Una de las principales desventajas de Nanonets es su elevado coste económico. La implementación de esta plataforma de inteligencia artificial suele

[3] Accesible desde: https://nanonets.com/.

requerir una inversión significativa en términos de licencias. Esto puede ser una barrera financiera para organizaciones académicas o individuos que no cuenten con un presupuesto adecuado.

- *Dificultad para adaptarse a los cambios de estructura del documento*. Nanonets es altamente efectivo cuando se utiliza en documentos con una estructura predefinida. Sin embargo, tiene dificultades para adaptarse a cambios en la estructura de los documentos. Si los términos del diccionario se presentan en diferentes formatos o si los documentos varían en su organización, puede requerir una reconfiguración y reentrenamiento del modelo, lo que consume tiempo y recursos adicionales.
- *Errores en la transcripción*. A pesar de su precisión general, Nanonets no está exento de errores en la transcripción. Los errores pueden surgir debido a la calidad de los documentos de entrada, la legibilidad del texto, la caligrafía o la presencia de palabras o caracteres poco comunes. Estos errores pueden afectar la calidad de los datos extraídos y requerir una revisión manual adicional.
- *Desorden en la extracción*. Los datos extraídos a menudo no se presentan de manera ordenada. Los términos pueden aparecer en una lista sin una estructura específica, lo que puede dificultar la posterior organización y análisis de los datos. Esto hace que se requieran esfuerzos adicionales para clasificar y etiquetar los términos, lo que aumenta la carga de trabajo.
- *Dependencia de datos de entrenamiento*. La efectividad de Nanonets depende en gran medida de los datos de entrenamiento proporcionados.

6. RECONOCIMIENTO DE CARACTERES POR EL MÉTODO TRADICIONAL. CREACIÓN DE UN SERVICIO WEB

En esta segunda línea de trabajo se ha llevado a cabo la digitalización del *Diccionario técnico de la música* siguiendo métodos tradicionales. A través de una serie de librerías en lenguaje Python (versión 3.8)[4], se ha procesado la imagen digital de cada página del *Diccionario* y se han utilizado técnicas de OCR (Optical Character Recognition) (Mori, 1992) para la identificación de caracteres (Spherber, 2018). Seguidamente, se ha realizado la extracción de la información a través de la utilización de algoritmos iterativos para identificar patrones sobre el texto (Li et al., 2014; Kavčič Čolić y Hari, 2024). Finalmente, se ha realizado un tratamiento y procesado de la información extraída. El proceso de digitalización en su totalidad ha implicado un exhaustivo trabajo de extracción de información, con el fin de asegurar la máxima fidelidad y precisión en la transferencia del texto original al transcrito.

[4] Accesible desde: https://www.python.org.

En esta línea de trabajo, se ha dado un paso más allá de la digitalización al diseñar una web de consulta del *Diccionario,* lo que permite un acceso abierto de gran interés. Cuánto más sencillo, cómodo y rápido sea este sistema de consulta, más usuarios captará. Por tanto, la digitalización del *Diccionario* ha permitido, por un lado, conservar su contenido de manera duradera y, por otro, facilitar también su acceso final a través de un navegador, lo que brinda la posibilidad de explorar eficientemente los términos y definiciones contenidos, en este caso, en un diccionario del siglo XIX. La dificultad de esta línea de trabajo se encontraba, por un lado, en el análisis y extracción de la información proveniente del formato original, ya que, aunque escaneadas sus páginas y, por tanto, se tenía acceso a los caracteres del libro mediante cualquier sistema de OCR, los textos obtenidos eran de baja calidad y no seguían reglas y patrones bien definidos que permitieran una automatización rápida. Se ha requerido la identificación de esos patrones para ir extrayendo y separando el término de la definición, las palabras asociadas a la definición, el idioma origen del término y otros aspectos cuyo interés requiriera un procesado posterior. Todo este proceso conforma la primera fase de análisis y extracción que se detalla a continuación. La segunda fase corresponde a la creación de la web que permita la búsqueda avanzada de términos y su clasificación. Se detallan a continuación los aspectos más importantes que se han considerado para cada una de las fases mencionadas.

6.1. Fase de análisis y extracción de la información

La fase de análisis comprende: 1) la organización de las definiciones del libro; 2) la identificación de patrones que permitan obtener los diferentes términos junto a sus aclaraciones, definiciones y referencias; y 3) la creación de un programa que automatice este proceso. En cuanto a la extracción de las definiciones, se han implementado rutinas que identifican esos patrones en el texto para obtener partes bien diferenciadas. Un ejemplo claro ha sido identificar la gran variedad de alternativas que se muestran en el *Diccionario* para representar un término y su definición. Finalmente, la información extraída se ha almacenado en una base de datos, para luego poder relacionarla y utilizarla para alimentar la página web de consulta. Para el diseño de la web, se ha optado por una base de datos de tipo relacional. Una vez se dispone de una base de datos bien estructurada, se ha seguido con el diseño de la página web y la implementación de la lógica necesaria para crear un sistema de búsqueda eficiente.

6.2. Fase de diseño y creación de la interfaz web

Una vez concretado el diseño de la interfaz a través de un prototipo, se optó por utilizar un método de desarrollo incremental, en el que se implementaran las funcionalidades más básicas, añadiendo progresivamente más funcionalidades hasta llegar al resultado final: la conexión de la información generada en la fase 1, que estaba almacenada en bases de datos, con el sistema de búsqueda que proporcionó la interfaz web.

6.3. Tecnologías utilizadas

Las tecnologías empleadas en este proyecto se enumeran a continuación:

- Python, como lenguaje de programación. Se trata de un lenguaje potente y conocido por muchos programadores y de fácil utilización para la identificación de patrones en texto. Es un lenguaje muy versátil y rápido para encontrar soluciones a problemas y existe bastante documentación en Internet.
- Visual Studio Code, como entorno de desarrollo. Es una interfaz de desarrollo que permite muchas variantes y extensiones[5].
- XAMPP, como gestor de base de datos. Permite desplegar en local servidores de Apache y MySQL. phpMyAdmin[6] (incluido en XAMPP) es utilizado para realizar consultas SQL, crear tablas con los atributos necesarios, claves primarias, y definición de tipos[7].
- SQL, para almacenar la información en la base de datos[8].
- Figma, es una herramienta de diseño y prototipado que se utiliza para crear el diseño de interfaces[9].
- HTML, utilizada para crear y estructurar el contenido de la página web.
- CSS, para crear los estilos en las páginas.
- JavaScript, que agrega funcionalidad a la Web[10].
- Flask (Ronacher, 2024), para el desarrollo web con Python y la conexión con la base de datos.
- API de ChatGPT, para hacer la revisión ortotipográfica final.

6.4. Programa de extracción de definiciones

En este apartado se detalla el estudio previo del *Diccionario* de cara a su digitalización. Se explica la estructura de la información asociada al término y su definición, así como los pasos llevados a cabo para la extracción de la información de las páginas y los problemas encontrados.

6.4.1. Estructura del Diccionario

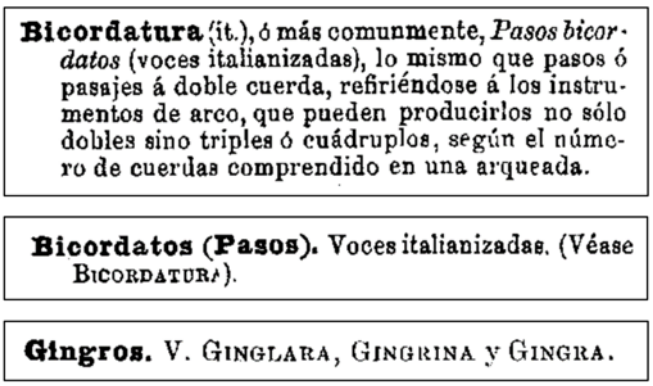
- El *Diccionario* está organizado en orden alfabético.

[5] Accesible desde: https://code.visualstudio.com/.
[6] Accesible desde: https://www.phpmyadmin.net/.
[7] Accesible desde: https://www.apachefriends.org/index.html.
[8] Accesible desde: https://www.microsoft.com/es-es/sql-server/sql-server-downloads.
[9] Accesible desde: https://www.figma.com/es-es/.
[10] Accesible desde: https://developer.mozilla.org/en-US/docs/Web/JavaScript.

- Separadores de secciones: cuando la primera inicial de una palabra cambia de letra, se incluye una página con una imagen estética centrada en la parte superior de la letra siguiente. Esto sirve como un separador visual para indicar el cambio de sección.
- Páginas de términos: contiene un encabezado y un pie de página que se tienen que eliminar. El resto de la página está dividido en dos columnas que contienen los diferentes términos del *Diccionario*.

Esta estructura proporciona una guía clara para la organización del contenido del libro y para la extracción de datos relevantes a almacenar en una base de datos que sirvan para la creación de la página web.

Aunque algunos términos del *Diccionario* siguen una estructura clásica, compuesta por el término seguido de su definición, en la mayoría de los casos la definición es mucho más compleja, con puntos y cambios de línea. Otras veces, la información atómica no se limita a un solo término y su definición, sino que está compuesta de varios términos, pudiendo ser esta simple o compuesta. En otros casos se alude al idioma del término y otras aclaraciones entre paréntesis, como por ejemplo el contexto en el que ese término es utilizado, o posibles variantes relacionadas con la raíz del término o términos. Y en otros se hace alusión a términos sinónimos en los cuáles no existe un patrón único (Ejemplo: véase, V. o el propio término en cursiva), como puede apreciarse en la figura 3.



Bicordatura (it.), ó más comunmente, *Pasos bicordatos* (voces italianizadas), lo mismo que pasos ó pasajes á doble cuerda, refiriéndose á los instrumentos de arco, que pueden producirlos no sólo dobles sino triples ó cuádruplos, según el número de cuerdas comprendido en una arqueada.

Bicordatos (Pasos). Voces italianizadas. (Véase BICORDATURA).

Gingros. V. GINGLARA, GINGRINA y GINGRA.

Figura 3. Tipos de términos. Fuente: elaboración propia.

Por tanto, ha sido necesario un análisis detallado del formato de las entradas del Diccionario y desarrollar algoritmos específicos para extraer los datos de manera efectiva. Las indicaciones en la definición para ver otros términos, como se ha podido comprobar, complican la extracción automatizada, ya que requieren una búsqueda adicional para asegurar dicha relación.

6.4.2. Extracción de la información

Para el proceso de extracción de la información se identificaron tres subetapas: la obtención del texto del *Diccionario*, la división del texto por términos y el almacenamiento en la base de datos.

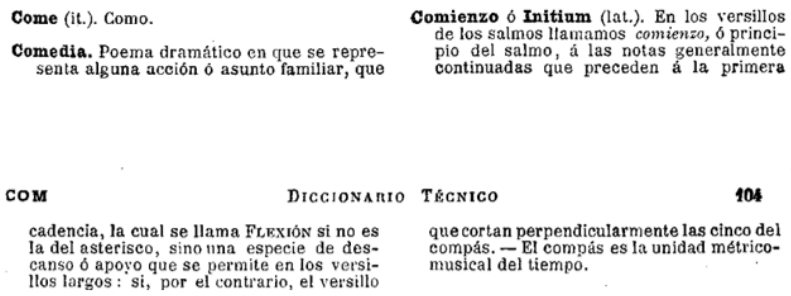
En la extracción del texto se probó la utilización de imágenes en formato .JPG y el uso de la librería Tesseract (Kay, 2007), un motor de reconocimiento óptico de caracteres que permite leer una amplia variedad de formatos de imagen y convertirlos a texto. Sin embargo, el resultado obtenido fue muy lejano del esperado. Al tratarse de un diccionario antiguo, el reconocimiento de los caracteres es muy pobre y nos encontramos con grandes dificultades. Tampoco se disponía en esta librería de un OCR robusto que permitiera ajustar sus configuraciones para mejorar la precisión de reconocimiento, por lo que iba a ser necesaria una revisión manual profunda y corrección de los errores del OCR.

La siguiente opción que tuvimos en cuenta fue usar un lector de PDF. Para ello utilizamos la librería en Python PyPDF2. Se creó un script iterativo que recorriera todas las páginas y fuera capturando el texto y guardándolo en ficheros texto (Figura 4). Aunque el resultado mejoraba el anterior proceso nos dimos cuenta de que perdíamos otras características importantes del *Diccionario*, como las negritas y las cursivas. Estas características, en particular la negrita, nos facilitaba el trabajo de distinguir entre el término principal y su definición. Es por ello que, finalmente, optamos por la API de PDF.co utilizando la opción de extraer el texto en formato XML y cuya etiqueta *fontStyle="Bold"* nos ayudaría en el proceso de identificación del término.

```python
def uploadFile(fileName):
    """Uploads file to the cloud"""

    # 1. RETRIEVE PRESIGNED URL TO UPLOAD FILE.

    # Prepare URL for 'Get Presigned URL' API request
    url = "{}/file/upload/get-presigned-url?contenttype=application/octet-stream&name={}".format(
        BASE_URL, os.path.basename(fileName))

    # Execute request and get response as JSON
    response = requests.get(url, headers={ "x-api-key": API_KEY })
    if (response.status_code == 200):
        json = response.json()

        if json["error"] == False:
            # URL to use for file upload
            uploadUrl = json["presignedUrl"]
            # URL for future reference
            uploadedFileUrl = json["url"]

            # 2. UPLOAD FILE TO CLOUD.
            with open(fileName, 'rb') as file:
                requests.put(uploadUrl, data=file, headers={ "x-api-key": API_KEY, "content-type": "application/octet-stream" })

            return uploadedFileUrl
        else:
            # Show service reported error
            print(json["message"])
    else:
        print(f"Request error: {response.status_code} {response.reason}")

    return None


if __name__ == '__main__':
    main()
```

Figura 4. Convertir texto a XML. Fuente: elaboración propia.

Superando otros obstáculos relativos al uso de esta librería, se consiguió disponer finalmente de los textos del *Diccionario* en formato XML desde el cual comenzar la identificación de patrones que permitieran la extracción de la información. Lo más básico a realizar fue la separación del término principal, en negrita, del resto del texto asociado al mismo, y su separación del siguiente término en la página. La definición contendría, además, términos relacionados, idioma, etc. Se automatizó el proceso en un script en Python que recorrió todas las páginas y separó ambas partes. Seguidamente comenzó la fase de filtrado, con la eliminación, por ejemplo, de todos los guiones, puntos, comas o espacios en blanco que puedan modificar cualquier palabra o término o que la cortaran al cambiar de línea (estamos tratando con libros con escritura en formato clásico de separación de sílabas por guiones). Por otro lado, había definiciones que se encontraban entre dos páginas (final de una página y comienzo de la siguiente), como puede observarse en el ejemplo de la figura 5. Dado el caso, había que conseguir que ese contenido se asociara al término de la página anterior, eliminando igualmente el encabezado o pie de página de las páginas digitalizadas.



Come (it.). Como.

Comedia. Poema dramático en que se representa alguna acción ó asunto familiar, que

Comienzo ó **Initium** (lat.). En los versillos de los salmos llamamos *comienzo*, ó principio del salmo, á las notas generalmente continuadas que preceden á la primera

COM DICCIONARIO TÉCNICO 104

cadencia, la cual se llama FLEXIÓN si no es la del asterisco, sino una especie de descanso ó apoyo que se permite en los versillos largos: si, por el contrario, el versillo

que cortan perpendicularmente las cinco del compás. — El compás es la unidad métrico-musical del tiempo.

Figura 5. Ejemplo de definición separa en distintas hojas.
Fuente: Elaboración propia.

A lo largo del proceso, se implementaron una serie de criterios para separar los diferentes campos y disponer de la información en trozos o partes que facilitaran su almacenamiento por separado en la base de datos. Todo ello sirvió para nutrir una página web y poder realizar búsquedas por término, por clasificación, términos relacionados, etc., de manera rápida y ágil. Esos criterios determinaron la creación de una serie de variables que se utilizarían para separar la información y poder así almacenarla adecuadamente en una base de datos:

- termino_final. Esta variable contiene el texto referente al término de la definición. Es la variable que se añade como resultado a la base de datos.
- termino_next. Esta variable accede al supuesto siguiente término del texto, a partir del que nos encontramos. Su propósito es comprobar si el siguiente supuesto término es realmente un término o parte de la definición anterior. En relación con su resultado, establecerá el valor de la variable *nextBold* como true o como false.

- termino_anterior. Almacena el término anterior registrado respecto al que nos encontramos, por si el texto que estamos tratando es parte del término anterior y no se trata de un nuevo término.
- Paréntesis. Esta variable contiene el texto que se encuentra entre paréntesis de la definición. Es la variable que se añade como resultado a la base de datos.
- Idioma. Esta variable contiene el texto que hace referencia al idioma de la definición. Es la variable que se añade como resultado a la base de datos.
- Definición. En esta variable, se van almacenando todas las partes de la definición del término, tras eliminar la información almacenada en el resto de variables.
- definicion_final. En esta variable modificamos el contenido de la variable "definición" para dejarlo de forma adecuada. Es la variable que se añade como resultado a la base de datos.
- upper_words. Esta variable contiene el texto que se encuentra en mayúsculas de la definición, y que puede ser un término relacionado. Es la variable que se añade como resultado a la base de datos.
- nextBold. Esta variable booleana nos permite avisar si el siguiente texto a procesar es de un nuevo término o se trata de una continuación.
- is_bold. Esta variable booleana nos permite avisar si la palabra se encuentra en negrita o no.

El procesamiento llevado a cabo en el texto permitió identificar patrones, extraer la información adecuada a almacenar en cada variable y su posterior volcado en una base de datos relacional. Además, en una fase posterior, se diseñó un script de Python que permitió corregir, y actualizar según los casos, la ortotipografía y separar palabras que habían quedado juntas a pesar del trabajo previo de filtrado. No obstante, cabe señalar que dicho trabajo de revisión automática se completó con una revisión humana, porque algunas de las actualizaciones en el campo de los términos no eran correctas. La principal fortaleza del sistema de extracción tradicional que acabamos de describir es que permite la elaboración, siguiendo un proceso relativamente simple, de listados de términos o *termbanks* que, a su vez, conforman glosarios que, en un futuro y para alcanzar los objetivos generales de LexiMus, será posible cruzar con otras herramientas lexicográficas digitales disponibles, tales como WordNet, siguiendo pautas desarrolladas en el ámbito de la terminología (Cabré 2000ab, 2001ab, Cabré 2003; Bobillo, Gómez-Romero y Araúz, 2012; Faber et al., 2005; Fellbaum y Hicks, 2019).

6.5. Análisis y diseño de la base de datos

Para el diseño de la base de datos, tras una exhaustiva exploración de las distintas formas de presentación de los términos y sus definiciones, se determinaron cuatro entidades o tablas distintas:

- Definición. Contiene los diferentes datos de cada una de las definiciones.
- Idioma. Contiene todos los diferentes idiomas encontrados en el libro, que están asociados a las definiciones.
- Clasificación. Contiene los diferentes tipos de clasificaciones en los que podemos agrupar los términos.
- Usuario. Recoge los usuarios que pueden acceder a las modificaciones de los términos desde la página web.

Para cada una de estas tablas se diseñaron los diferentes atributos que podían ser relevantes. Por ejemplo, para la tabla definición, se atribuyeron siete campos importantes, que son identificador, término, idioma, aclaración, definición, términos relacionados y clasificación, siendo idioma, términos relacionados y clasificación campos para almacenar listas de identificadores que se encuentren ubicados en las otras tablas, de ahí que se consiga una relación de la información que facilite su rápida consulta por términos, por clasificación, por idioma, etc.

Diseñada la base de datos, se creó su estructura con el programa phpMyAdmin y se comenzó el almacenamiento de los datos. Para ello, se llevaron a cabo acciones de inserción de los datos obtenidos del "Programa de extracción de definiciones" realizando inserciones de términos nuevos y actualizaciones de los ya introducidos para añadir o modificar campos previamente almacenados. Otras iteraciones sobre la información obtenida del programa, permitió identificar el contenido completo a insertar en la tabla idiomas, en la tabla de clasificación, etc.

6.6. Sistema Web: Diccionario Técnico de la Música

En esta fase se llevó a cabo un análisis exhaustivo para determinar los requisitos y objetivos que queríamos que cumpliera el sistema de consultas web. Se muestran a continuación los requisitos más importantes:

- Que el sistema de búsqueda permita insertar palabras clave y proporcione sugerencias automáticas mientras se escribe para completar la consulta.
- Que los resultados aparezcan de forma ordenada y clara.
- Que en los resultados se disponga de un acceso rápido a términos relacionados con el término de búsqueda.

- Que se pueda modificar/completar la información relativa al término y la información asociada al mismo, así como la eliminación o inserción de información nueva.
- Que exista un usuario administrador de la web que permita la gestión de usuarios para dotar a los usuarios de privilegios de consulta y/o edición de contenido.

Además, se determinó como debía ser la interfaz gráfica de la web, su estructura y presentación gráfica, así como la estructura y apariencia de los diferentes formularios dedicados a la visualización, inserción y/o modificación de la información. Se analizó igualmente la mejor manera de presentar la información creando un sistema de navegación entre páginas, búsqueda de términos por letra, la propia página de inicio de sesión, etc.

Para el desarrollo de la interfaz web, la estructura del contenido de la página web y el propio contenido a mostrar, se utilizó el lenguaje HTML. Se usaron hojas de estilo (CSS) y las librerías Bootstrap y Fontawesome, así como JQuery para mejorar significativamente la apariencia, la funcionalidad y la experiencia de usuario en el sitio web, optimizando el desarrollo y manteniendo la consistencia visual y la eficiencia en el código. JavaScript permitió implementar la lógica de programación y algunas partes de la presentación (*front-end*). El diseño de plantillas con CSS se ha utilizado para controlar el diseño, la presentación y el estilo visual de las páginas web, en términos de color, tamaño, posición y otros atributos visuales. De la librería Bootstrap se han seleccionado plantillas y componentes predefinidos (botones, formularios, navegación, etc.) basados en HTML, CSS y Javascript para la presentación *responsive* de las distintas páginas, es decir, que permita un ajuste automático al tamaño de la pantalla del dispositivo utilizado (ordenador de escritorio, tablet, móvil, etc.).

El desarrollo final de la web de consulta puede observarse en los siguientes ejemplos funcionales. La página de inicio (Figura 6) muestra un esbozo de toda la funcionalidad del sistema. En la imagen siguiente se puede apreciar la ficha técnica del *Diccionario,* seguido de una breve descripción de la plantilla de búsqueda que, al pulsarla, lleva a dicha funcionalidad. En la misma página principal se puede acceder a una búsqueda por clasificaciones, así como la búsqueda por términos de diccionario (búsqueda por letras).

Cuando se lleva a cabo la búsqueda por término en el buscador (Figura 7), el sistema ofrece sugerencias del término que se está escribiendo que le ayuden al usuario a completar su búsqueda. Tras la inserción o selección del término a buscar y al pulsar *Intro* o el icono de *lupa*, se obtiene el término buscado, su definición y características añadidas y, si lo hubiere, se muestra también la información de otros términos relacionados, facilitando un enlace a su definición (Figura 8). La búsqueda por clasificación (Figura 9) permite el acceso a todos los términos categorizados. Estos se muestran en un sistema de navegación de páginas para agilizar la consulta. En la figura 9 puede observarse una clasificación de términos por instrumento. Finalmente, la búsqueda por letras

(Figura 10) permite, como en cualquier diccionario, la búsqueda por letras del alfabeto. Si la información ocupa varias páginas se organiza todo en un sistema de navegación por páginas que facilita la carga de los datos haciéndola más rápida.

Toda la información que se muestra en la web es fácilmente editable, aspecto clave para la revisión y corrección. La edición es únicamente viable para aquellos usuarios con perfil de administrador o de editor de contenidos. La modificación de los datos se puede hacer en el propio término, su definición, su idioma, así como agregarlo a diferentes clasificaciones, como puede apreciarse en la figura 11. A diferencia de una simple web de consulta, la creación del sistema de gestión de roles de usuarios para la edición de la información garantiza que se puedan llevar a cabo procesos de revisión manual y de corrección de errores hoy día todavía necesarios, a pesar del avance en las técnicas de digitalización en textos antiguos.

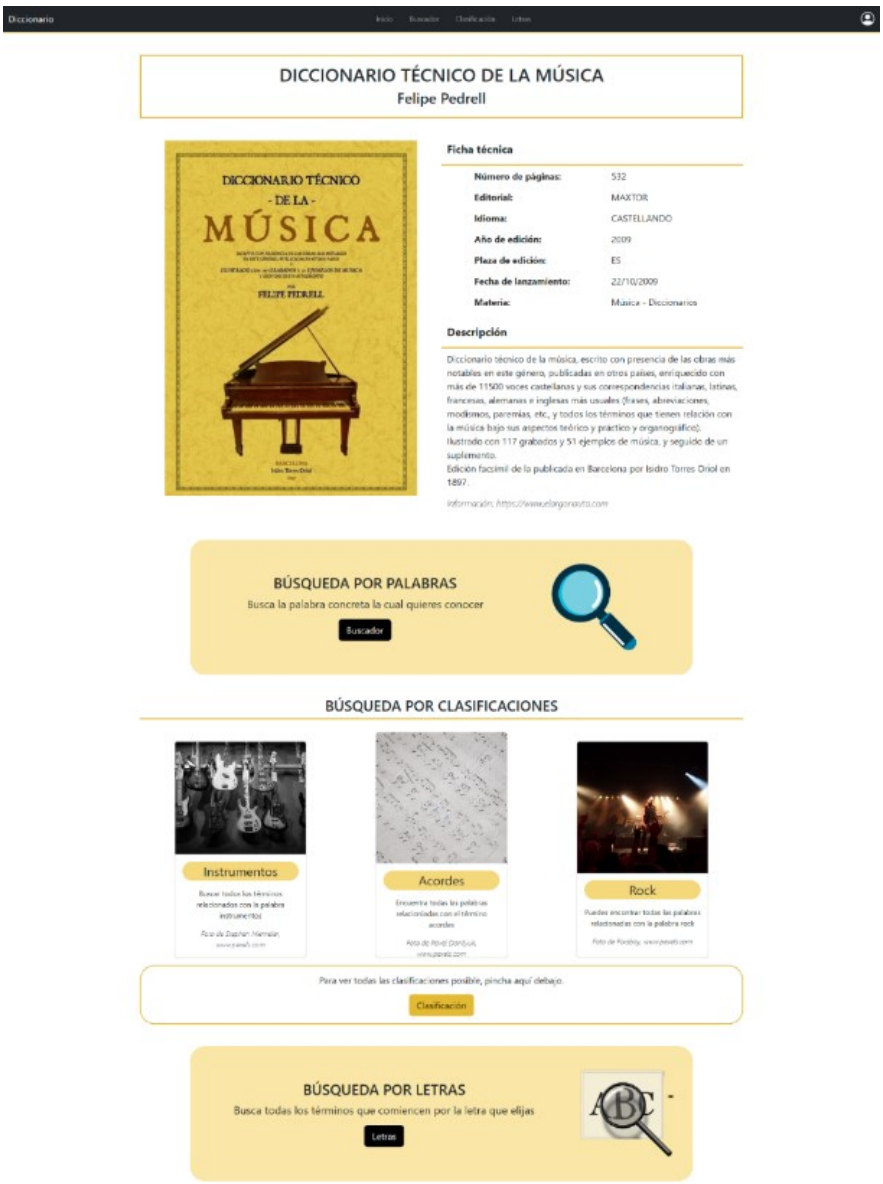


Figura 6. Resultado de la interfaz – página de inicio.
Fuente: elaboración propia.

Figura 7. Resultado de la interfaz – página de buscador – autocompletado.
Fuente: elaboración propia.

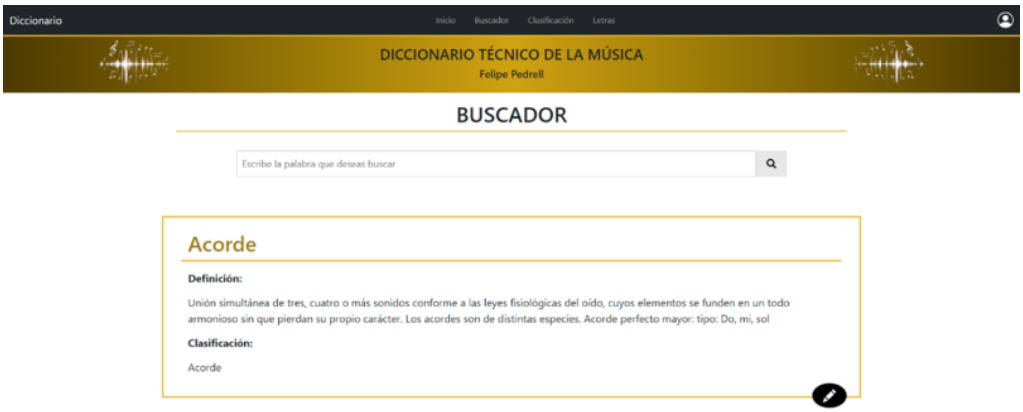


Figura 8. Resultado de la interfaz – página de buscador.
Fuente: elaboración propia.

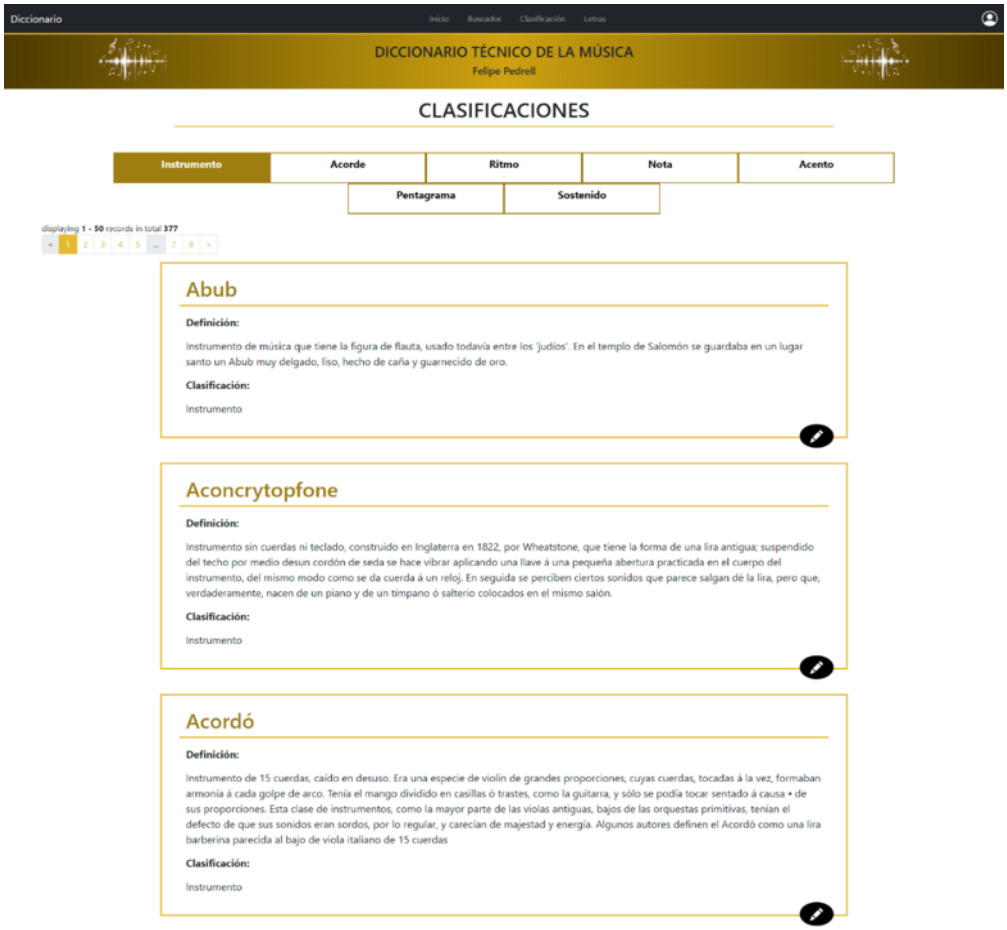


Figura 9. Resultado de la interfaz – página de buscador por clasificación.
Fuente: elaboración propia.

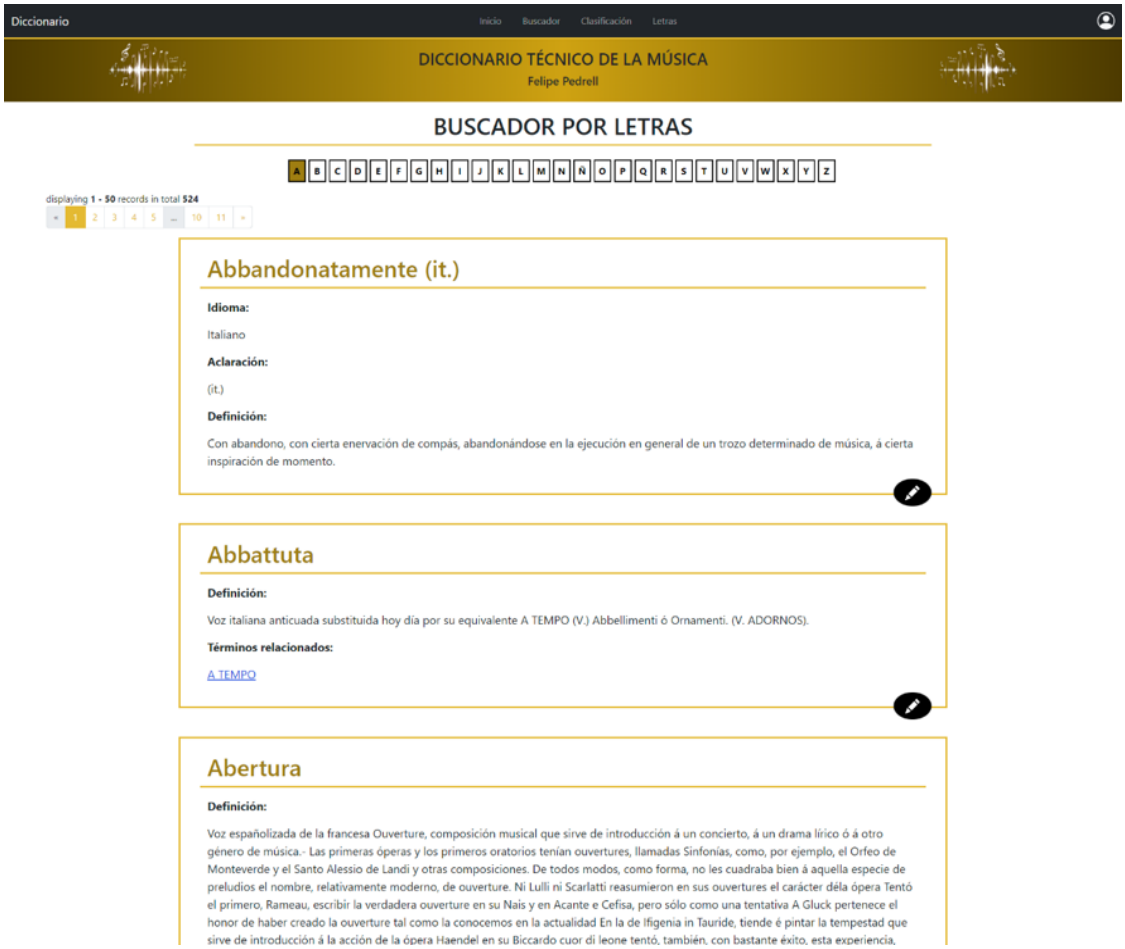


Figura 10. Resultado de la interfaz – página de buscador por letras. Fuente: elaboración propia.

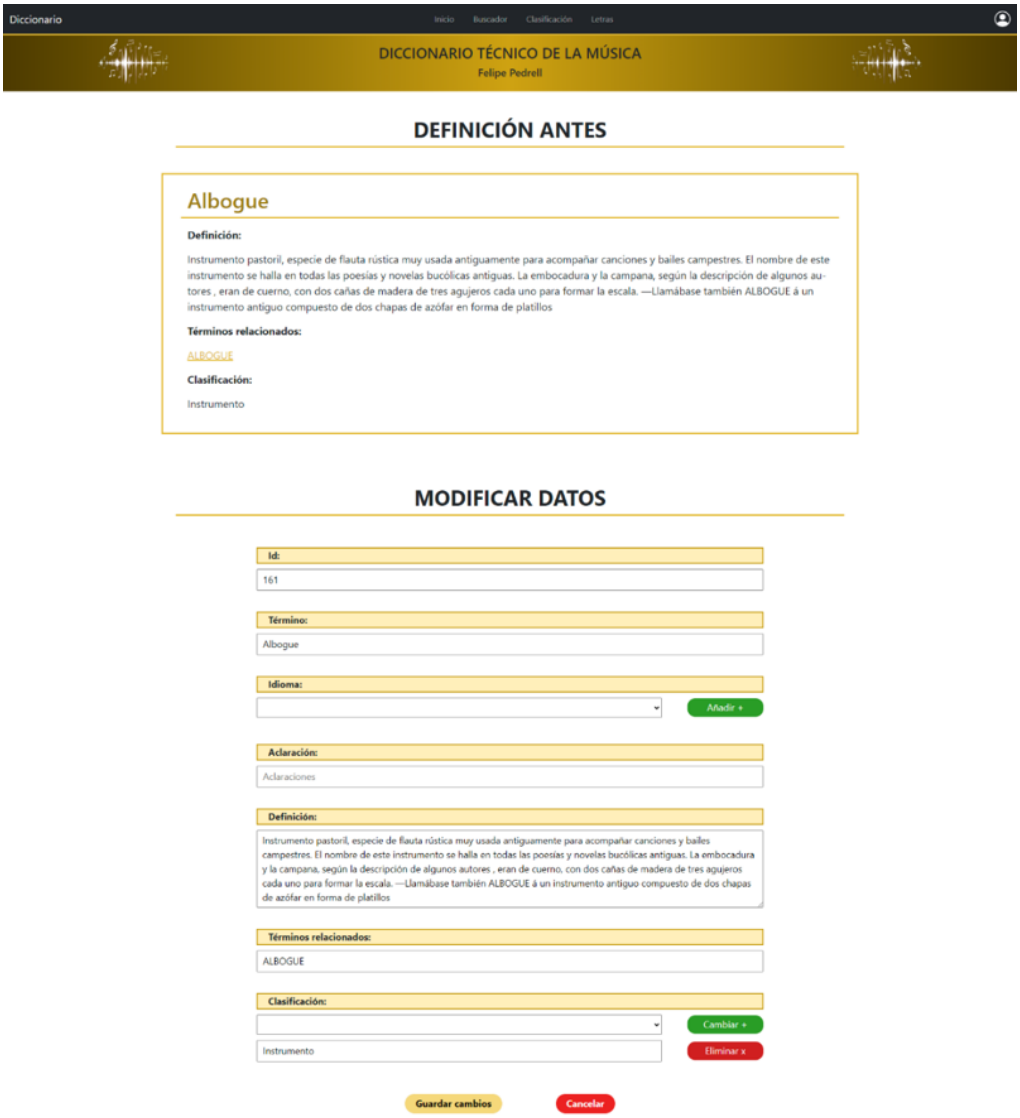


Figura 11. Resultado de la interfaz – página de modificar términos.
Fuente: elaboración propia.

7. CONCLUSIONES

La inversión de tiempo en la búsqueda y prueba de herramientas y tecnologías es un paso crucial en el proceso de transcripción de documentos históricos. La elección de soluciones adecuadas según la fuente de los documentos es esencial. En este contexto, se ha utilizado Transkribus como una herramienta de procesamiento automático que puede acelerar significativamente el proceso de transcripción en comparación con la transcripción manual. Dicha herramienta no solo ahorra tiempo, sino también recursos valiosos. La precisión del reconocimiento de escritura de Transkribus ha sido una ventaja importante. Se ha podido comprobar que, para reducir significativamente la probabilidad de errores, se requiere de una correcta configuración, adecuada a los textos específicos, que sirvan como base para la fase de entrenamiento del modelo. No obstante, los resultados preliminares obtenidos reducen enormemente los costes asociados a las correcciones posteriores.

En paralelo a la exploración de técnicas de inteligencia artificial se ha llevado a cabo la digitalización del *Diccionario* y un trabajo minucioso de extracción de la información siguiendo metodologías clásicas de análisis y filtrado de textos. A través de la identificación de patrones, se ha extraído y procesado la información, intentando asegurar la máxima fidelidad y precisión con el texto original, a pesar del problema de la falta de calidad de los textos, impresos en un diccionario con formato de escritura del siglo XIX.

Tenemos que destacar que a pesar de los resultados obtenidos en ambas líneas de trabajo sigue siendo esencial una revisión manual, especialmente en casos en los que la precisión es fundamental, ya que incluso un pequeño porcentaje de error (como el uno por ciento mencionado con Transkribus) puede tener un impacto significativo en la calidad de la transcripción.

La búsqueda de programas que permitan la exportación a formatos específicos, como XML y CSV, es otro aspecto crítico en el proceso de transcripción. Estos formatos son ampliamente utilizados en la gestión y análisis de datos y, como hemos podido comprobar, facilitan la posterior manipulación y el acceso a la información. La elección de herramientas que faciliten la exportación a estos formatos es una consideración importante para garantizar la accesibilidad y utilidad de los datos transcritos.

Con todo ello, podemos afirmar que la optimización del proceso de transcripción de documentos históricos y manuscritos implica la búsqueda y selección cuidadosa de herramientas y tecnologías que ahorren tiempo y recursos, así como que garanticen la precisión en el reconocimiento de escritura.

Además, llevar a cabo procesos que relacionen la información extraída, como por ejemplo la clasificación de términos, las relaciones entre los términos de búsqueda, la búsqueda de familia de términos, etc., han complicado también la extracción automatizada, requiriendo de numerosas iteraciones para procesar el texto extraído. Como conclusión al trabajo realizado, se prevé necesario un análisis más detallado del formato de las entradas del *Diccionario* y desarrollar algoritmos específicos para extraer los datos de manera más efectiva. Asimismo, explorar el uso de técnicas de aprendizaje automático más específicas que permitan identificar y extraer información de estructuras no estándar, combinado con métodos tradicionales de reconocimiento óptico de caracteres y tratamiento de textos permitirían superar las dificultades encontradas y lograr una extracción exitosa de los datos.

Finalmente, queremos destacar que la creación de un sistema web de consulta brinda la posibilidad de explorar de forma rápida y eficiente los términos y definiciones contenidos en el *Diccionario*. Aunque se han implementado funcionalidades adicionales a la de la simple consulta de términos, como la búsqueda avanzada, la categorización de términos y la posibilidad de búsqueda por letras, consideramos que un conocimiento más profundo de la información que se maneja dotaría a la web de mayor funcionalidad y construiríamos, por tanto, un servicio de consulta de alta calidad para el ámbito de la investigación y la preservación de documentos históricos.

8. BENEFICIOS

Los resultados de este trabajo darán continuidad a los alcanzados en trabajos anteriores y permitirán desarrollar metodologías y herramientas acordes al actual entorno digital. De forma adicional, la elaboración del glosario y la construcción de un modelo conceptual digital, siguiente objetivo del proyecto en el que se enmarca este trabajo, nos proporcionará una herramienta fundamental para la historia de la música. El uso de técnicas como la minería de datos, la estilometría o el marcado semántico, permitirán el análisis de textos verbales históricos y colecciones textuales digitalizadas relacionados con este ámbito de la cultura humana y, su resultado, almacenado en una base de datos, hará posible poner este tipo de fuentes a disposición no sólo de la comunidad científica, sino también de los intérpretes, superando así las barreras estrictamente académicas.

RECONOCIMIENTOS

Este trabajo ha sido financiado por la Agencia Española de Investigación y los Fondos Europeos de Desarrollo Regional en el proyecto LEXIMUSP2 PID2022-139589NB-C32 (MCIN/AEI/10.13039/501100011033/FEDER,UE) y por la Universidad Internacional de Valencia (VIU) en el proyecto interno PII2023_63 *Digitalización, estructuración y elaboración de un modelo conceptual del Diccionario Técnico de la Música de Felipe Pedrell*, de la convocatoria 2023 de proyectos de investigación.

REFERENCIAS BIBLIOGRÁFICAS

Bobillo, F., Gómez-Romero, J., y Araúz, P. L. (2012). Fuzzy Ontologies for Specialized Knowledge Representation in WordNet. International Conference on Information Processing and Management of Uncertainty.

Cabré, M. T. (2000a). Elements for a theory of terminology: Towards an alternative paradigm. *Terminology*, *6*(1), 35-57.

Cabré, M. T. (2000b). Sur la représentation mentale des concepts: bases pour une tentative de modélisation. En H. Béjoint y P. Thoiron (Eds.), *Le sens en terminologie* (pp. 20-39). Presses Universitaires de Lyon.

Cabré, M. T (2001a). Consecuencias metodológicas de la propuesta teórica (I). *En La terminología científico-técnica: reconocimiento, análisis y extracción de información formal y semántica* (pp. 27-36). Institut Universitari de Lingüística Aplicada (Universitat Pompeu Fabra).

Cabré, M. T. (2001b). Sumario de principios que configuran la nueva propuesta teórica. En M. T. Cabré y J. Feliu (Eds.), *La terminología científico-técnica: reconocimiento, análisis y extracción de información formal y semántica* (pp. 17-26). Universitari de Lingüística Aplicada (Universitat Pompeu Fabra).

Cabré, M. T. (2003). Theories of terminology: their description, prescription and explanation. En M. T. Cabré, *Terminology: international journal of theoretical and applied issues in specialized communication*, 9(2), 163-200.

Carreras, J. J. (2001). Hijos de Pedrell: La historiografía musical española y sus orígenes nacionalistas (1780-1980). *Il Saggiatore Musicale*, *8*(1), 121-169.

Colutto, S., Kahle, P., Guenter, H., y Muehlberger, G. Transkribus: A Platform for Automated Text Recognition and Searching of Historical Documents. En Proceedings of the 2019 15th International Conference on EScience (EScience), San Diego, CA, USA, 24-27 September 2019 (pp. 463-466).

de Schryver, G.-M., y Joffe, D. (2023). The end of lexicography, welcome to the machine: on how ChatGPT can already take over all of the dictionary maker's tasks. http://hdl.handle.net/1854/LU-01GWSGV0M8HYZVNXV8SSZ0VXB

Faber, P, Márquez Linares, C., y Vega Expósito, M. (2005). Framing Terminology: A Process-Oriented Approach. *Meta: journal des traducteurs / Meta: Translators' Journal*, *50*(4).

Fellbaum, C. D., y Hicks, A. (2019). When WordNet Met Ontology. Ontology Makes Sense.

García Serrano, A. y Castellanos González, Ángel. (2017). Representación y organización de documentos digitales: detalles y práctica sobre la ontología DIMH. *Revista de Humanidades Digitales*, 1, 314-344. https://doi.org/10.5944/rhd.vol.1.2017.17155

Harish, R., Raghavendra Rao, G. N. (2024). Transcription of Ancient Indian Manuscripts Through Artificial Intelligence—Current Status of Technology and the Way Forward. En H. Sharma, A. Chakravorty, S. Hussain, R. Kumari (Eds.), *Artificial Intelligence: Theory and Applications. AITA 2023. Lecture Notes in Networks and Systems* (vol. 844). Springer. https://doi.org/10.1007/978-981-99-8479-4_2

Justiniano López, J. C. (2019). La música en la Real Academia Española, el diccionario y la institución. De Autoridades al Diccionario de la lengua española de 2014. En J. Sanmartín Sáez y M. Quilis Merín (Coords.), *Retos y avances en lexicografía: los diccionarios del español en el eje de la variación lingüística* (pp. 187-202).

Kahle, P., Sebastian Colutto, G. H., y Muhlberger, G. (2017). 'Transkribus – A Service Platform for Transcription, Recognition and Retrieval of Historical Documents'. En *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)* (pp. 19-24). IEEE. https://doi.org/10.1109/ICDAR.2017.30

Kavčič Čolić, A., y Hari, A. (2024). Improving accessibility of digitization outputs: EODOPEN project research findings. Digital Library Perspectives.

Kay, A. (2007). Tesseract: an open-source optical character recognition engine. Linux Journal, *159*(2).

Li, Y., Jia, W., Shen, C., y van den Hengel, A. (2014). *Characterness: An indicator of text in the wild. IEEE Transactions on Image Processing*, *23*(4), 1666–1677. https://doi.org/10.1109/TIP.2014.2302896

López Vallejo, M. A., y López Aguirre, R. (2021). La terminología musical en la 23.ª edición del Diccionario de la Lengua Española: hallazgos y desencuentros. *Tonos Digital*, 41, 1-34. http://hdl.handle.net/10201/11104

Mateos Frühbeck, N. (2021). Las ediciones digitales de PROLOPE: entre la edición digital y la edición digitalizada. *Revista de Humanidades Digitales*, 6, 236-251. https://doi.org/10.5944/rhd.vol.6.2021.3065

Mori, S., Suen, C. Y., y Yamamoto, K. (1992). Historical Review of OCR Research and Development.

Pedrell, F. (1900). Diccionario técnico de la música (2ª ed.). Isidro Torres Oriol.

Perdiki, E. (2022). Review of 'Transkribus: Reviewing HTR training on (Greek) manuscripts'. RIDE, 15. doi: 10.18716/ride.a.15.6.

Quilis Merín, M. (2019). Luisa Lacal, la primera lexicógrafa española, y su Diccionario de la música, técnico, histórico, bio-biográfico (Madrid, 1899). *Revista Argentina de Historiografía Lingüística*, *11*(1), *47-75*.

Rabus, A. (2019). Recognizing Handwritten Text in Slavic Manuscripts: A Neural-Network Approach Using Transkribus. *Scripta & E-Scripta*,19, 9-32.

Ronacher, A. (2024). Flask: The Python Micro Framework for Building Web Applications*. *GitHub*. https://github.com/pallets/flas

Rubio Amondarain, M. (2023). La disciplina musical en la lexicografía española del siglo XIX. [Trabajo de Fin de Máster. Universidad Nacional de Educación a Distancia].

Spherber G. (2018). A gentle introduction to OCR. *Towards Data Science*. https://towardsdatascience.com/a-gentle-introduction-to-ocr-ee1469a201a

Ströbel, P. B., Clematide, S., y Volk, M. (2020). How Much Data Do You Need? About the Creation of a Ground Truth for Black Letter and the Effectiveness of Neural OCR. En *Proceedings of the 12th Language Resources and Evaluation Conference* (pp. 3551–3359). European Language Resources Association. https://www.aclweb.org/anthology/2020.lrec-1.43

Subirá Puig, J. (1970). Un panorama histórico de la lexicografía musical. *Anuario Musical: Revista de Musicología del CSIC*, *25*, 125-142.