# Corpus de textos del “Diccionario del Español Medieval electrónico” (DEMel)

Caroline Müller[1], Robert Stephan[2], Rafael Arnold[3], Ulrike Henny-Krahmer[4] y Karsten Labahn[5]

Universidad de Rostock 2026

https://doi.org/10.18453/rosdok_id00005463



**Resumen**

El conjunto de datos contiene 200 textos completos de fuentes españolas medievales de los siglos X a principios del siglo XV codificados en XML/TEI. Los textos se basan en digitalizaciones mediante OCR así como en ediciones ya disponibles en versión digital del *Hispanic Seminary of Medieval Studies* (HSMS), del *Archivo Digital de Manuscritos y Textos Españoles* (ADMYTE) y de la *Biblioteca Virtual Miguel de Cervantes*, y abarcan una amplia variedad de géneros medievales. Están disponibles en los formatos XML/TEI, HTML y TXT. El corpus de textos fue preparado en el marco del proyecto *Diccionario del Español Medieval electrónico* (DEMel) con el fin de vincular las documentaciones de la base de datos con sus textos fuente. Aproximadamente el 60 % de las documentaciones primarias del DEMel han sido anotadas en los textos de este conjunto de datos, lo que facilita su uso para la investigación del español medieval.

**Zusammenfassung**

Das Datenset enthält 200 XML/TEI-kodierte Volltexte mittelalterlicher spanischer Quellen des 10. bis frühen 15. Jahrhunderts. Sie basieren auf OCR-Digitalisierungen sowie bereits in digitaler Form vorliegenden Editionen des *Hispanic Seminary of Medieval Studies* (HSMS), des *Archivo Digital de Manuscritos y Textos Españoles* (ADMYTE) und der *Biblioteca Virtual Miguel de Cervantes* und decken ein breites Spektrum mittelalterlicher Gattungen ab. Die Texte stehen in den Formaten XML/TEI, HTML und TXT zur Verfügung. Sie wurden im Projekt *Diccionario del Español Medieval electrónico* (DEMel) für die Verknüpfung der Belege aus der Datenbank mit ihren Quellentexten aufbereitet. Insgesamt sind ca. 60 % der im DEMel enthaltenen Primärbelege in den Texten dieses Datensets annotiert, was ihre Nutzung für die Erforschung des mittelalterlichen Spanisch erleichtert.

---

[1] https://orcid.org/0000-0002-8591-7859
[2] https://orcid.org/0000-0001-7605-7415
[3] https://orcid.org/0009-0003-1581-9086
[4] https://orcid.org/0000-0003-2852-065X
[5] https://orcid.org/0000-0002-8482-807X

**Abstract**

The dataset contains 200 full texts of medieval Spanish sources from the 10th to the early 15th century encoded in XML/TEI. The texts are based on OCR digitizations as well as editions already available in digital form from the *Hispanic Seminary of Medieval Studies* (HSMS), the *Archivo Digital de Manuscritos y Textos Españoles* (ADMYTE), and the *Biblioteca Virtual Miguel de Cervantes*, and cover a wide range of medieval genres. They are available in XML/TEI, HTML, and TXT formats. The corpus was prepared within the project *Diccionario del Español Medieval electrónico* (DEMel) to link the attestations of the database with their source texts. Approximately 60% of the primary attestations included in DEMel are annotated in the texts of this dataset, facilitating their use for research on medieval Spanish.

# 1. El proyecto DEMel y su corpus de textos medievales

Este conjunto de datos forma parte del proyecto *Diccionario del Español Medieval electrónico* (DEMel)[6]. El proyecto surgió con el objetivo de digitalizar y hacer públicamente accesible el extenso archivo de datos recopilado en la Universidad de Heidelberg en el marco del *Diccionario del español medieval* (DEM), bajo la dirección de Bodo Müller. Iniciado en 1971, el DEM se convirtió en una obra lexicográfica de referencia para el estudio del español medieval, aunque quedó inconclusa cuando el proyecto fue suspendido a finales de 2007 por razones financieras. El archivo original comprende unas 865 000 fichas en papel, de las cuales alrededor de 700 000 documentan formas léxicas medievales con su contexto de uso, fuente y datación. Tras su digitalización y registro sistemático en el marco del proyecto, financiado por la *Deutsche Forschungsgemeinschaft* (DFG) en tres fases entre 2016 y 2026, el DEMel ofrece acceso libre a más de 32 000 lemas con aproximadamente 700 000 documentaciones extraídas de más de 600 obras literarias y no literarias del siglo X hasta principios del XV.

Para poder vincular las documentaciones primarias con sus textos fuente, en la tercera fase del proyecto (2023–2026) se compiló un corpus del español medieval que abarca una amplia parte de las fuentes primarias empleadas en el DEM. Los textos que se publican en este conjunto de datos representan una amplia variedad de géneros medievales y están disponibles en los formatos XML/TEI, HTML y TXT.

Bajo la dirección de Rafael Arnold, Ulrike Henny-Krahmer y Karsten Labahn, esta tarea fue llevada a cabo por Caroline Müller y Robert Stephan. Caroline Müller fue responsable del procesamiento y control de calidad de los textos en formato XML/TEI, mientras que Robert Stephan implementó su transformación a los formatos HTML y TXT. Participaron como ayudantes estudiantiles en el procesamiento de los textos y la anotación de las documentaciones: Trinidad Almendra Arias Villalobos, Hanna Bott, Hagar El Souhagy Louarmassi, Elisabeth Gebauer, Anna Magdalena Goldmann-Klein, Dominika Jabłońska, Marie Kisser, Freyja Peters, Erik Renz, Laura Sackritz, Friederike-Sophie Schubert y Julia Siebert.

Las siguientes secciones describen cómo se construyó el corpus y qué elementos TEI se emplearon en la codificación, antes de presentar la estructura de archivos y carpetas de esta publicación.

[6] El portal web del proyecto puede consultarse en https://demel.uni-rostock.de.

# 2. Creación del corpus de textos

## 2.1. Compilación del corpus

Del corpus DEM, compuesto por más de 600 obras, se seleccionaron 200 para el presente corpus de textos. La selección se basó en diversos criterios: por un lado, se dio prioridad a las fuentes primarias que ya estaban disponibles en formato de texto completo digital a través de proveedores externos y que podían reutilizarse. Por otro lado, se generaron textos a partir de archivos de imagen mediante reconocimiento óptico de caracteres (OCR). A la hora de seleccionar las fuentes adecuadas, se dio prioridad a la relación coste-beneficio: debían existir archivos de imagen de la edición principal utilizada en el DEM, las fuentes debían presentar una alta densidad de documentaciones y el diseño de la edición debía tener una estructura sencilla. Para el reconocimiento de texto se utilizó inicialmente OCR4all[7]. Para ello, se generaron datos de entrenamiento a partir de modelos disponibles y se corrigieron manualmente, con el fin de entrenar posteriormente un modelo específico para el español medieval, el cual se ajustó ligeramente para cada edición. Esto permitió obtener textos de alta calidad. Únicamente los caracteres especiales, como los corchetes y las comillas angulares, requirieron un procesamiento manual posterior. Sin embargo, OCR4all presentaba deficiencias en el reconocimiento de la maquetación, especialmente en lo que respecta al orden de lectura, y no permitía entrenar modelos propios de maquetación. Dado que estas limitaciones afectaban sistemáticamente a la calidad del corpus, se tomó la decisión de cambiar a Transkribus[8] para el procesamiento de los textos restantes. Los modelos públicos de Transkribus proporcionaron resultados satisfactorios, por lo que no fue necesario ningún entrenamiento adicional. Solo hubo que volver a corregir manualmente los paréntesis.

El corpus final comprende 200 obras:

- 109 textos generados mediante OCR
- 44 textos de la *Digital Library of Old Spanish Texts* del *Hispanic Seminary of Medieval Studies* (HSMS)[9]
- 40 textos del *Archivo Digital de Manuscritos y Textos Españoles* (ADMYTE)
- 7 textos de la *Biblioteca Virtual Miguel de Cervantes*[10] en formato HTML

## 2.2. Procesamiento de los textos

Dado que los textos completos procedían de fuentes heterogéneas y se presentaban en distintos formatos, fue necesario procesar y convertirlos a un formato uniforme antes de proceder a la anotación de las documentaciones. El objetivo consistía en extraer, mediante script o manualmente, el texto continuo medieval de cada fuente, convertirlo al formato XML conforme a las directrices de la *Text Encoding Initiative* (TEI)[11] y, en la medida de lo posible, marcar las principales estructuras textuales.

---

[7] https://www.ocr4all.org
[8] https://www.transkribus.org
[9] Las ediciones del HSMS fueron publicadas en Zenodo (https://doi.org/10.5281/zenodo.18931376) en abril de 2026 bajo una licencia CC BY NC SA 4.0, si bien en el proyecto se emplearon aún las versiones disponibles en https://www.hispanicseminary.org, ya que la publicación en Zenodo se produjo cuando el proyecto estaba llegando a su fin.
[10] https://www.cervantesvirtual.com
[11] https://tei-c.org

### 2.2.1. De HSMS y ADMYTE a XML/TEI

Los textos completos del HSMS y de ADMYTE se basan en las convenciones de transcripción del HSMS, descritas en el *Manual of Manuscript Transcription for the Dictionary of the Old Spanish Language*[12] de David Mackenzie y Ray Harris-Northall y disponibles en versión abreviada en Zenodo[13]. La conversión de los textos completos a XML/TEI se realizó en gran medida de forma automática mediante script. Las intervenciones manuales solo fueron necesarias en casos excepcionales, por ejemplo cuando el número de etiquetas de apertura no coincidía con el de cierres "}".

No se pretendió establecer un mapeo completo de todas las etiquetas HSMS a sus equivalentes XML, sino obtener un texto continuo depurado sin elementos adicionales como títulos de columna, reclamos y pies de ilustración, y con las marcas paleográficas (letras voladas, corchetes para añadidos, etc.) resueltas. En concreto, se realizaron los siguientes cambios en el paso de HSMS y ADMYTE a XML/TEI:

**Etiquetas eliminadas con su contenido:**

Los elementos que no pertenecían al texto continuo fueron suprimidos por completo: `{ILL.}`, `{IN.}`, `{HD.}`, `{MIN.}`, `{DIAG.}`, `{BLANK.}`, `{CW.}`, `{SG.}`, `{RMK.}`, `{SYMB.}`

**Etiquetas de lengua eliminadas (contenido conservado):**

Los pasajes en otras lenguas se conservaron pero se eliminaron sus etiquetas: `{ARB.}`, `{BAS.}`, `{CAT.}`, `{ENG.}`, `{FRN.}`, `{GAL.}`, `{GER.}`, `{GRK.}`, `{HEB.}`, `{ITL.}`, `{LAT.}`, `{PRT.}`, `{PRV.}`

**Etiquetas conservadas (con equivalente TEI):**

| Etiqueta HSMS | Elemento TEI |
|---|---|
| `{RUB.}` | `<hi rend="rubric">` |
| `{GL.}` | `<gloss>` |
| `{AD.}` | `<add>` |
| `[fol.]` | `<pb/>` |

**Caracteres especiales eliminados:**

`[ ]`, `< >`, `<< >>`, `^`, `#2`, `\`, `*`, `$`, `|`, `[+]`, `%`

**Combinaciones de caracteres ASCII resueltas (ejemplos):**

| Original | Resultado |
|---|---|
| `c'` | ç |
| `n~` | ñ |
| s' | σ |

Además, se eliminaron los paréntesis redondos junto con su contenido, ya que marcan supresiones editoriales. Los paréntesis dobles se convirtieron en simples, pues en el original representan paréntesis en el texto. Los fragmentos textuales, separados por `+`, `%2` o `%3` de acuerdo con la estructura del manuscrito, se reunificaron.

---

[12] https://www.hispanicseminary.org/docs/HSMS-manual.pdf
[13] https://doi.org/10.5281/zenodo.18931376

El resultado son textos continuos de las fuentes medievales depurados de caracteres especiales. Dado que las convenciones de transcripción del HSMS se centran principalmente en una reproducción paleográficamente precisa y no contemplan el marcado de unidades estructurales como capítulos o párrafos, los textos procesados tampoco contienen ninguna otra estructuración, salvo un `<div>` y un `<p>` envolventes. En el marco del proyecto no fue posible realizar un marcado manual posterior.

### 2.2.2. De OCR a XML/TEI

El proyecto DEM se caracterizó por una gran diversidad de fuentes, lo que afectó directamente a los textos que debían procesarse en el proyecto: colecciones de documentos, textos legales, poemas, un drama y otros géneros debían recopilarse de manera uniforme. Esta heterogeneidad dificultó considerablemente la creación de un corpus de textos coherente basado en TEI. A esto se sumaba que las ediciones utilizadas como base también diferían mucho entre sí en cuanto a formato y convenciones; así, por ejemplo, el significado de los corchetes podía variar de una edición a otra. Dado que no se disponía de tiempo para leer detenidamente todos los prólogos ni todas las ediciones especificaban claramente sus convenciones editoriales, en tales casos hubo que actuar de forma pragmática.

El objetivo principal del procesamiento manual de los textos OCR y de las ediciones de la *Biblioteca Virtual Miguel de Cervantes* fue —al igual que con los textos del HSMS y de ADMYTE— obtener el texto continuo medieval. Por ello no se tuvieron en cuenta las introducciones de los editores, los glosarios, las traducciones, los comentarios ni los índices. Además se aplicaron las siguientes reglas:

**Elementos de texto no incluidos**

- Notas a pie de página y aparatos críticos
- Encabezados que, a juzgar, no formaban parte del original medieval y estaban escritos en español moderno
- Texto en escritura árabe y hebrea
- Metadatos al inicio de los documentos (datación, resumen, descripción del original)
- Datos de foliación así como el símbolo "/" utilizado en las ediciones para marcar los límites de folio[14]

**Procesamiento de signos y convenciones editoriales**

- Los paréntesis redondos y los corchetes se conservaron, incluido su contenido; los comentarios editoriales como "[sic]" se eliminaron.
- Las omisiones se representaron como en la fuente, por ejemplo con "...".

**Normalizaciones a nivel de carácter**

- s en lugar de ſ
- Las letras voladas se bajaron a la línea.
- Las cursivas no se reprodujeron.

**Estructura del texto y maquetación**

- Las estructuras de texto más importantes (capítulos, estrofas, párrafos, saltos de página, etc.) se marcaron con TEI.
- Los saltos de columna y de línea no se tuvieron en cuenta al marcar las estructuras de texto.

[14] Una excepción son los textos t0134 (https://purl.uni-rostock.de/demel/t0134) y t0052 (https://purl.uni-rostock.de/demel/t0052), ya que las fuentes correspondientes b0589 (https://purl.uni-rostock.de/demel/b0589) y b1396 (https://purl.uni-rostock.de/demel/b1396) se citan según su foliación.

Las excepciones a estas reglas se indican en el elemento <encodingDesc> del <teiHeader> del documento correspondiente. A continuación se describen en detalle los demás elementos TEI utilizados.

# 3. Convenciones de anotación TEI

El corpus consta de 200 archivos XML (uno por texto). Se emplea una selección reducida de elementos TEI adaptada a las necesidades del proyecto que se define en un documento ODD (archivo demel_text.odd) y un esquema RELAX NG generado a partir de él (archivo demel_text.rng).

**Encabezado**

El encabezado TEI (<teiHeader>) se divide en los siguientes elementos:

- El elemento <fileDesc> (descripción bibliográfica del archivo) contiene los subelementos <titleStmt> (título y responsabilidades)[15], <editionStmt> (datos sobre la edición), <extent> (extensión), <publicationStmt> (datos de publicación) y <sourceDesc> (descripción del original). Dentro de <sourceDesc>, los datos bibliográficos de la edición que sirvió de base para el texto se registran con <bibl type="analog-source">. Los datos se extraen de la base de datos del DEMel. En <bibl type="original-source"> se remite al texto medieval subyacente y se indica su datación con <date>.
- El elemento <profileDesc> recoge la clasificación del contenido del texto mediante <textClass>, <keywords> y <term>. El atributo @corresp remite a la taxonomía del DEMel.
- La información sobre la codificación se recoge en <encodingDesc>, incluyendo <listPrefixDef> con <prefixDef> para definir los prefijos URI para la taxonomía y las documentaciones, así como <editorialDecl> para la documentación de las decisiones editoriales.
- Los cambios en el archivo se documentan en <revisionDesc> con <listChange> y <change>.

**Texto continuo**

Dentro de <text> no se distingue entre <body>, <front> y <back>, ya que en la mayoría de los casos se omitieron las partes de la edición relevantes para <front> y <back>, como las introducciones y los apéndices. Por ello, todo se incluye bajo <body>. Los textos más extensos o las colecciones de documentos se subdividen en capítulos o documentos mediante . Si estaban numerados en la edición, la numeración se transfirió al atributo @n, empleando números arábigos o romanos según la fuente. A nivel de párrafo, <p> recoge el texto continuo. En textos bíblicos como t0033[16] se utiliza <ab> para los versículos bíblicos si la estructura de la edición original lo permitía. Para los versos se emplean <lg> (grupo de versos) y <l> (verso). En la única obra dramática, t0006[17], el diálogo de los personajes se marca con <sp> y el personaje que habla con <speaker>. Para marcar los títulos se utiliza <head>; para los saltos de página, el elemento vacío <pb/>, en el que el número de página se expresa con el atributo @n. Si una palabra estaba separada por un salto de página, se añadió el atributo @break="no" y se eliminó el guion precedente. Asimismo, se eliminaron sistemáticamente los espacios antes y después de <pb/>.

---

[15] Bajo el campo «autor» se recogen indistintamente autores en sentido estricto, traductores y promotores de los textos medievales.
[16] https://purl.uni-rostock.de/demel/t0033
[17] https://purl.uni-rostock.de/demel/t0006

Las rúbricas de los textos del HSMS y ADMYTE se reprodujeron con `<hi rend="rubric">`. No se prescindió de su marcado, ya que en los originales suelen indicar títulos y aportan así al menos cierta estructura al texto. Las adiciones se etiquetan con `<add>` y las glosas con `<gloss>`.

El elemento `<seg>` permite marcar cualquier segmento de texto. En el corpus, este elemento se utiliza exclusivamente para anotar las documentaciones primarias, concretamente alrededor del 60 % de todas las documentaciones primarias de la base de datos. El ID de la documentación se indica con el atributo `@corresp`, a través del cual, mediante el prefijo "obj" definido en `<prefixDef>`, se puede acceder a la entrada TEI Lex-0[18] correspondiente. Un mismo `<seg>` puede referenciar varios IDs simultáneamente. Cuando la forma documentada consistía en dos palabras no consecutivas, ambas partes se anotaron con su propio elemento `<seg>`. Los dos elementos recibieron cada uno un `@xml:id` y se remiten entre sí mediante `@prev` y `@next`, indicando así su relación.

# 4. Estructura de carpetas y archivos

El archivo ZIP publicado contiene los siguientes elementos:

- un archivo de configuración de Maven (`pom.xml`) para ejecutar la generación de los textos,
- un archivo README,
- los textos generados en formato HTML y TXT,
- las hojas de estilo XSLT para generar los archivos HTML y TXT a partir de los archivos XML/TEI,
- los textos en formato XML/TEI con los esquemas correspondientes.

```
<ROOT>
│     pom.xml
│     README.md
│
├──generated_texts
│     ├──html
│     │        demel_t0001.html
│     │        demel_t0002.html
│     │        ...
│     └──txt
│              demel_t0001.txt
│              demel_t0002.txt
│              ...
│
├──scripts
│     └──xslt
│              demel_tei2html.xsl
│              demel_tei2txt.xsl
│
└──texts
      ├──schema
      │        demel_text.odd
      │        demel_text.rng
      └──tei
               demel_t0001.tei.xml
               demel_t0002.tei.xml
               ....
```

[18] https://lex-0.org

## 4.1. Textos (`texts/`)

Este directorio contiene el núcleo del corpus. Se divide en:

- `tei/`: Almacena los 200 archivos XML de los textos medievales siguiendo el estándar TEI.
- `schema/`: Incluye los archivos de validación RelaxNG (`.rng`) y el archivo ODD (`.odd`). Estos recursos permiten a los usuarios verificar la integridad estructural de los archivos XML.

## 4.2. Lógica de transformación (`scripts/xslt/`)

El directorio `scripts/xslt/` incluye las hojas de estilo XSLT (`.xsl`) utilizadas para automatizar la conversión de los archivos XML/TEI a los formatos HTML y TXT. Las hojas de estilo tienen en cuenta los diferentes tipos de textos del corpus, adaptando la representación visual según se trate de textos en prosa, en verso o dramáticos. De este modo, se pueden generar versiones HTML y TXT actualizadas directamente a partir de los archivos XML/TEI modificados.

## 4.3. Versiones derivadas (`generated_texts/`)

Este directorio contiene las representaciones del corpus en formato HTML para la consulta y lectura, y en formato TXT para el análisis computacional:

- `html/`: Versiones diseñadas para la lectura humana y la consulta mediante navegadores web. Las documentaciones anotadas aparecen subrayadas con una línea punteada y un icono de ojo. En el caso de formas documentadas que consisten en dos palabras no consecutivas se emplea una línea discontinua. Los saltos de página se representan como líneas horizontales con la referencia de página enlazada. Asimismo, cada archivo incorpora metadatos bibliográficos en el encabezado (`citation_title`, `citation_publication_date`, `citation_public_url`, etc.).
- `txt/`: Archivos de texto plano destinados al análisis computacional y la lingüística de corpus. Cada archivo comienza con un bloque de metadatos en formato YAML delimitado por `---`, que incluye los campos `title`, `date`, `purl`, `publisher`, `analog source` y `text encoding`, seguido del texto continuo. Las referencias de página se indican entre corchetes.