

Oralia Diacrónica del Español (ODE)

Descripción del sistema de codificación

Autor:

Gael Vaamonde

Última actualización:

28/10/2025

Cómo citar este trabajo:

Vaamonde, Gael (2024). *Oralia Diacrónica del Español (ODE). Descripción del sistema de codificación*. Zenodo. <https://doi.org/10.5281/zenodo.14060223>.

Oralia Diacrónica del Español (ODE). Descripción del sistema de codificación © 2024 by Gael Vaamonde is licensed under Creative Commons Attribution ShareAlike 4.0 International.



1. INTRODUCCIÓN	4
2. LA CABECERA	6
2.1 Descripción del archivo electrónico	6
2.1.1 Declaración del título	6
2.1.1.1 Título	7
2.1.1.2 Edición y financiación	8
2.1.1.3 Declaración de responsabilidades	10
2.1.2 Declaración de la publicación	11
2.1.3 Descripción de la fuente	12
2.1.3.1 Identificador del manuscrito	12
2.1.3.2 Contenido del manuscrito	15
2.1.3.3 Información adicional	18
2.2 Descripción de la codificación	19
2.2.1 Descripción del proyecto	19
2.2.2 Declaración de la muestra	20
2.2.3 Declaración de la edición	21
2.2.4 Declaración de la clasificación	22
2.3 Descripción del perfil del texto	22
2.3.1 Creación	23
2.3.2 Descripción del contexto	23
2.3.3 Clasificación del texto	24
2.4 Ejemplo completo de cabecera	25
3. LA EDICIÓN FACSIMILAR	28
4. EL TEXTO	29
4.1 La transcripción paleográfica	29
4.1.1 Los elementos estructurales	29
4.1.1.1 Inicio de página	30
4.1.1.2 Inicio de línea	31
4.1.1.3 Inicio de columna	32
4.1.2 Las abreviaturas	33
4.1.3 Texto cancelado	34
4.1.4 Texto añadido	35
4.1.5 Texto resaltado	37
4.1.6 Texto omitido	39
4.1.7 Texto conjeturado	41
4.1.8 Texto superfluo	42
4.1.9 Texto inexacto	42
4.1.10 Texto citado	43
4.1.11 Texto en otra lengua	43
4.1.12 Texto de nueva mano	44
4.1.13 Discurso en estilo directo	44
4.1.14 Firma	45
4.2 El corpus lingüístico	46
4.2.1 Tokenización	46
4.2.2 Normalización ortográfica	48
4.2.3 Lematización y etiquetado morfosintáctico	49
4.2.4 Anotación lingüística adicional	52
5. TEITOK VS. TEI P5	55

1. INTRODUCCIÓN

Oralia Diacrónica del Español (en adelante, ODE) es un corpus diacrónico especializado compuesto por documentación manuscrita inédita producida entre el siglo XVI y el siglo XIX. Actualmente, está conformado por tres tipos textuales: inventarios de bienes, declaraciones de testigos y certificados médicos. La mayor parte de estos documentos procede de la región meridional de la península ibérica, aunque también se están incorporando muestras textuales de otras zonas de España.

ODE no solo reúne una amplia colección de fuentes manuscritas del español clásico y moderno, sino que las ofrece en dos formatos preparados para la búsqueda: el de la edición paleográfica digital y el del corpus lingüísticamente anotado. Esto lo convierte en un recurso electrónico versátil, que puede explorarse como archivo digital de documentos manuscritos o como corpus histórico. Cada documento se almacena en un archivo XML individual, que integra información sobre la edición paleográfica y la anotación lingüística. En esencia, por tanto, ODE se compone de múltiples archivos XML, tantos como documentos conforman este recurso electrónico.

ODE está disponible en acceso abierto desde su página oficial (<http://corpora.ugr.es/ode/>) y se gestiona mediante la plataforma web TEITOK, que permite consultar el corpus y explorar documentos individualmente de forma eficiente. Los diferentes modos de consulta y visualización disponibles en ODE se detallan en Vaamonde (2024)¹. Al consultar un documento, el usuario puede descargar el archivo XML en dos modelos diferentes: el modelo TEITOK (*TEITOK XML*), diseñado para su uso interno en la plataforma, y el modelo TEI P5 (*TEI P5 XML*), basado rigurosamente en las directrices propuestas por la versión más reciente de la [Text Encoding Initiative](#).

Select download format

TEI P5 XML
TEITOK XML

¹ Vaamonde, Gael (2024). *Oralia Diacrónica del Español (ODE). Descripción del sistema de consulta*. Zenodo. <https://doi.org/10.5281/zenodo.14080389>.

El modelo TEITOK es un modelo interno, personalizado, no estandarizado y diseñado fundamentalmente para facilitar el análisis y procesamiento de datos dentro de la plataforma TEITOK. Debido a su naturaleza específica, no es adecuado para el intercambio de datos fuera del proyecto ODE, salvo en contextos relacionados con la creación de corpus históricos desarrollados en esta plataforma. Para resolver este inconveniente, ODE también pone a disposición del usuario la descarga de cada archivo XML en TEI P5, un estándar ampliamente reconocido en el ámbito de las Humanidades Digitales que garantiza la interoperabilidad y la reutilización de datos en diversos entornos. La conversión del modelo TEITOK en el modelo TEI P5 se realiza automáticamente mediante el script *ode2teip5*, desarrollado por Gael Vaamonde.

La presente guía de usuario se enfoca en explicar el modelo TEI P5 utilizado en los documentos de ODE, con el objetivo de facilitar su comprensión y aplicación en proyectos de investigación, tanto dentro como fuera de la plataforma TEITOK.

2. LA CABECERA

Todos los documentos presentan una cabecera destinada a proporcionar información descriptiva de carácter metatextual. Esta información se incluye dentro del elemento [`<teiHeader>`](#), esto es, la cabecera o encabezado TEI. En ODE, la información contenida dentro de este elemento se organiza en tres subapartados:

- [`<fileDesc>`](#) (Descripción del archivo). Contiene una descripción bibliográfica completa del archivo electrónico.
- [`<encodingDesc>`](#) (Descripción de la codificación). Documenta la relación entre el texto electrónico y la fuente de la que deriva.
- [`<profileDesc>`](#) (Descripción del perfil del texto). Proporciona una descripción detallada de los aspectos no bibliográficos del texto.

2.1 Descripción del archivo electrónico

El elemento [`<fileDesc>`](#) contiene una descripción bibliográfica completa del archivo. En ODE, este elemento consta de tres subapartados:

- [`<titleStmt>`](#) (Declaración del título). Agrupa la información referente al título de la obra y a los responsables de su contenido intelectual.
- [`<publicationStmt>`](#) (Declaración de la publicación). Agrupa la información referente a la publicación o distribución del texto electrónico.
- [`<sourceDesc>`](#) (Descripción de la fuente). Proporciona una descripción del texto fuente del que deriva el texto electrónico.

2.1.1 Declaración del título

El elemento [`<titleStmt>`](#) agrupa la información referente al título de una obra y a los responsables de su contenido intelectual. En ODE, este elemento consta de cuatro elementos:

- [`<title>`](#) (Título). Título completo de la obra.
- [`<editor>`](#) (Editor). Responsable de edición de la obra.
- [`<funder>`](#) (Responsable de la financiación). Responsable de la financiación del proyecto.
- [`<respStmt>`](#) (Declaración de responsabilidad). Otros responsables del contenido intelectual de la obra.

2.1.1.1 Título

a. El elemento `<title>` contiene la información referente al título del texto electrónico, que en ODE siempre es la edición digital de un documento manuscrito, producido entre el siglo XVI y el siglo XIX.

b. La redacción del título no sigue un esquema preestablecido; más bien, se trata de texto libre; no obstante, guarda relación con el tipo textual en el que se clasifique el documento (ver [sección 2.2.3](#) y [sección 2.3.3](#)). En el caso de los inventarios de bienes, por ejemplo, el título suele comenzar por expresiones como *inventario*, *relación*, *carta de dote*, *testamento*, etcétera; en el caso de las declaraciones de testigos, por otras como *declaración*, *juicio*, *pleito*, *probanza*, etcétera; y en el caso de los certificados médicos, subtipo especial dentro de las declaraciones de testigos, por otras como *certificado médico*, *declaración de médico*, *declaración de cirujano*, *declaración de maestro sangrador*, *declaración de esencia*, etcétera. Se aportan a continuación algunos ejemplos:

Inventarios de bienes:

```
<title>Inventario de los bienes de doña Luisa Francisca de Olerson  
para su venta a José del Villar</title>
```

```
<title>Carta de pago de dote de Juan Serrano para María Núñez</title>
```

```
<title>Testamento de María Sánchez, mujer de García Sánchez de  
Velasco</title>
```

```
<title>Partición de los bienes de María Gómez Cabeza entre sus tres  
hermanas</title>
```

Declaraciones de testigos:

```
<title>Declaración de Andrés Cardeña de la Guerra, procurador y vecino  
de Málaga, testigo presentado por el solicitador del fiscal en el  
pleito entre Luis Muñoz y Juan Pérez</title>
```

```
<title>Juicio sobre las cuchilladas que un esclavo le dio a otro,  
provocándole la muerte</title>
```

<title>Probanza. El fiscal contra Esteban y Francisco Carvajal, vecinos de Vélez Málaga, presos, sobre la tenencia de armas prohibidas</title>

<title>Pleito incoado de oficio por la justicia de Atarfe sobre averiguación de cómo el niño de nueve años Antonio Triana, vecino de dicha villa, cayó a un pozo y se ahogó</title>

Certificaciones médicas:

<title>Certificado médico de Julián Román, sobre una herida en la cabeza</title>

<title>Declaración del cirujano Isidro Moraita por el reconocimiento de las heridas que tenían los hermanos Melchor y Dionisio Otañez</title>

<title>Declaración de esencia del sangrador Diego Lozano por la cura de las heridas de Francisco de Castro</title>

c. La amplia mayoría de textos incluidos en ODE son documentos originales. No obstante, también se admite un número limitado de documentos que son copias o traslados de originales (ver [sección 2.1.3.2i](#) y s.). En estos casos, el título generalmente comienza con expresiones como *copia de*, *traslado de*, *copia coetánea de* o *traslado coetáneo de*, según corresponda: Por ejemplo:

<title>Traslado de testimonio de presentación de demanda del Dr. Arias contra el regimiento de la villa de Medina del Campo por impago de unos servicios médicos prestados contenido en un pleito de 1543-1546</title>

<title>Copia coetánea de declaración de Bartolomé Hernández, esclavo del obispo D. Diego de la Calzada, sobre las heridas propinadas a Juana de Herrera, hija de Martín de Herrera, presentada en el pleito de 1600-1602 que sigue este contra Cristóbal de Hornazo, alguacil de la villa, por negligencia al dejar escapar de la cárcel a dicho esclavo</title>

2.1.1.2 Edición y financiación

a. El corpus ODE está conformado es el resultado de diversos proyectos de investigación financiados por organismos autonómicos y nacionales. En ODE, los elementos [<editor>](#) y [<funder>](#) se utilizan para detallar esta información.

b. El elemento `<editor>` contiene información relacionada con el nombre del proyecto de investigación en el que se ha desarrollado la edición digital del texto. Además, contiene un atributo `@xml:id` que sirve para asignar un identificador único a cada proyecto de investigación. Actualmente, existen cinco opciones: (ver [Tabla 1](#)):

Tabla 1. Relación entre el elemento `<editor>` y el atributo `@xml:id` en ODE.

Nombre del proyecto de investigación <code><editor></code>	Identificador único <code>@xml:id</code>
HISPATESD: Hispanae Testium Depositiones. Las declaraciones de testigo en la historia de la lengua española. 1492-1833	HISPATESD
ALEA XVIII: Atlas lingüístico y etnográfico de Andalucía, siglo XVIII. Patrimonio documental y Humanidades Digitales	ALEA18
ALEA oriental-XVIII. Atlas Lingüístico y Etnográfico de Andalucía oriental, s. XVIII. Patrimonio documental y Humanidades Digitales	ALEAO18
ALEA oriental-XIX. Atlas Lingüístico y Etnográfico de Andalucía oriental, s. XIX. Patrimonio documental y Humanidades Digitales	ALEAO19
VIVE. Vita Verborum. Los peritajes de las Chancillerías castellanas en la historia del español (1650 - 1833)	VIVE

c. El elemento `<funder>` contiene información referente al nombre del organismo que financia el proyecto de investigación. Actualmente, se contemplan cinco posibilidades, una por cada proyecto de investigación mencionado anteriormente (ver [Tabla 2](#)).

Tabla 2. Contenido de los elementos `<editor>` y `<funder>` en ODE.

Nombre del proyecto (o subcorpus) <code><editor></code>	Financiación <code><funder></code>
HISPATESD: Hispanae Testium Depositiones. Las declaraciones de testigo en la historia de la lengua española. 1492-1833	MINECO/AEI/FEDER/UE: FFI2017-83400-P
ALEA XVIII: Atlas lingüístico y etnográfico de Andalucía, siglo XVIII. Patrimonio documental y Humanidades Digitales	Junta de Andalucía, Consejería de Transformación Económica, Industria, Comercio y Universidades/FEDER: P18-FR-695
ALEA oriental-XVIII. Atlas Lingüístico y Etnográfico de Andalucía oriental, s. XVIII. Patrimonio documental y Humanidades Digitales	Junta de Andalucía, Consejería de Transformación Económica, Industria, Comercio y Universidades/FEDER. Proyectos I+D+i del Programa Operativo FEDER 2020: A-HUM-116-UGR20
ALEA oriental-XIX. Atlas Lingüístico y Etnográfico de Andalucía oriental, s. XIX. Patrimonio documental y Humanidades Digitales	Proyectos de Investigación Aplicada del Plan Propio de Investigación y Transferencia de la Universidad de Granada, 2023, financiados por el Programa operativo FEDER Andalucía 2021-2027: C-HUM-038-UGR23
VIVE. Vita Verborum. Los peritajes de las Chancillerías castellanas en la historia del español (1650 - 1833)	Ministerio de Ciencia e Innovación. Proyectos de Generación de Conocimiento 2022. PID2022-136256NB-I00

El elemento `<funder>` contiene los atributos `@from` y `@to`, que sirven para informar sobre el año de inicio y fin de la financiación. La correspondencia entre entidad financiadora e intervalo temporal de recoge en la [Tabla 3](#):

Tabla 3. Contenido de los elementos `<editor>` y `<funder>` en ODE

Financiación <code><funder></code>	Intervalo temporal <code>@from - @to</code>
MINECO/AEI/FEDER/UE: FFI2017-83400-P	2018 - 2022
Junta de Andalucía, Consejería de Transformación Económica, Industria, Comercio y Universidades/FEDER: P18-FR-695	2020 - 2023
Junta de Andalucía, Consejería de Transformación Económica, Industria, Comercio y Universidades/FEDER. Proyectos I+D+i del Programa Operativo FEDER 2020: A-HUM-116-UGR20	2021 - 2023
Proyectos de Investigación Aplicada del Plan Propio de Investigación y Transferencia de la Universidad de Granada, 2023, financiados por el Programa operativo FEDER Andalucía 2021-2027: C-HUM-038-UGR23	2024 - 2026
Ministerio de Ciencia e Innovación. Proyectos de Generación de Conocimiento 2022. PID2022-136256NB-I00	2023 - 2027

d. Se aportan a continuación un par de ejemplos de codificación de estos dos elementos en ODE:

```
<editor xml:id="HISPATESD">HISPATESD: Hispanae Testium Depositiones.
Las declaraciones de testigo en la historia de la lengua española.
1492-1833</editor>
<funder from="2018" to="2022">MINECO/AEI/FEDER/UE: FFI2017-83400-P,
2018-2021</funder>
```

```
<editor xml:id="VIVE">VIVE. Vita Verborum. Los peritajes de las
Chancillerías castellanas en la historia del español (1650 -
1833)</editor>
<funder from="2023" to="2027">Ministerio de Ciencia e Innovación.
Proyectos de Generación de Conocimiento 2022. PID2022-136256NB-
```

2.1.1.3 Declaración de responsabilidades

a. El elemento `<respStmt>` proporciona información referente a la responsabilidad intelectual de la edición digital. Consta del elemento `<resp>`, para designar la naturaleza de la responsabilidad, y del elemento `<name>`, para designar el nombre del responsable correspondiente.

b. Para cada documento XML, se distinguen cuatro tipos de responsabilidad, que se corresponden con cuatro tareas realizadas de forma secuencial: la transcripción del texto, la revisión manual de la normalización ortográfica, la revisión manual de la anotación lingüística y la revisión general del resultado

final. Estas tareas se enumeran en el mismo orden a través del atributo **@n** dentro del elemento **<respStmt>**. Por ejemplo:

```
<respStmt n="1">
  <resp>Transcripción</resp>
  <name>Diego Antonio Reinaldos Miñarro</name>
</respStmt>
<respStmt n="2">
  <resp>Normalización</resp>
  <name>Diego Antonio Reinaldos Miñarro</name>
</respStmt>
<respStmt n="3">
  <resp>Anotación lingüística</resp>
  <name>Miguel Calderón Campos</name>
</respStmt>
<respStmt n="4">
  <resp>Revisión</resp>
  <name>Inmaculada González Sopeña</name>
</respStmt>
```

c. Las tareas de normalización ortográfica y anotación lingüística se realizan mediante herramientas de procesamiento automático del texto incluidas en la plataforma TEITOK (ver [sección 4.2.2](#) y [sección 4.2.3](#)). Los nombres recogidos en este apartado se refieren a los responsables de la revisión manual del resultado automático generado por estas herramientas.

2.1.2 Declaración de la publicación

El elemento **<publicacionStmt>** agrupa la información referente a la publicación o distribución de la edición digital. En ODE, este elemento consta de tres elementos:

- **<publisher>** (Editorial). Proporciona el nombre del responsable de la publicación.
- **<pubPlace>** (Lugar de la publicación). Contiene el nombre del lugar de publicación.
- **<distributor>** (Distribuidor). Proporciona el nombre del responsable de la distribución.

El contenido de estos tres elementos es invariable en ODE. El responsable de la publicación es siempre la *Universidad de Granada*; el lugar de publicación, *Granada*, y el responsable de distribución, el grupo de investigación *DiLEs* (*Diacronía de la Lengua Española*). Por tanto, el elemento **<publicacionStmt>** se codifica del modo siguiente para todos los documentos del corpus:

```
<publicationStmt>
  <publisher>UGR, Universidad de Granada</publisher>
  <pubPlace>Granada</pubPlace>
  <distributor>HUM-278. Grupo DiLEs: Diacronía de la Lengua
    Española. UGR-Junta de Andalucía</distributor>
</publicationStmt>
```

2.1.3 Descripción de la fuente

El elemento [<sourceDesc>](#) agrupa información relativa al texto fuente, que en el caso de ODE será siempre un texto manuscrito inédito. Por tanto, toda la información aquí recogida se integra en un elemento [<msDesc>](#), destinado a la descripción de fuentes manuscritas. Y, en ODE, el elemento [<msDesc>](#) consta a su vez de tres elementos:

- [<msIdentifier>](#) (Identificador del manuscrito). Contiene la información necesaria para identificar el manuscrito que se examina.
- [<msContents>](#) (Contenido del manuscrito). Describe el contenido intelectual del manuscrito o parte del manuscrito.
- [<additional>](#) (Distribuidor). Agrupa información adicional sobre el manuscrito.

2.1.3.1 Identificador del manuscrito

a. El elemento [<msIdentifier>](#) permite una identificación precisa del manuscrito. La documentación almacenada en ODE suele localizarse en fondos notariales y judiciales de archivos históricos. Teniendo esto en cuenta, la información de identificación del manuscrito se estructura en cinco elementos, que representan distintos niveles de especificidad sobre su localización:

- [<country>](#) (País). Contiene el país donde se conserva el manuscrito.
- [<settlement>](#) (Ciudad). Contiene la ciudad se conserva el manuscrito.
- [<institution>](#) (Institución). Contiene el archivo histórico donde se conserva el manuscrito.
- [<repository>](#) (Fondo). Contiene el fondo documental donde se conserva el manuscrito.
- [<idno>](#) (Número identificativo). Contiene información para identificar el manuscrito a nivel inferior al fondo documental.

b. El elemento [<country>](#) contiene el nombre del país en donde se conserva el manuscrito. Actualmente, todos los archivos históricos consultados se encuentran en España.

c. El elemento [<settlement>](#) contiene el nombre de la ciudad donde se conserva el manuscrito y el elemento [<institution>](#) contiene el nombre del archivo histórico. Hasta la fecha, se han visitado más de una veintena de archivos históricos — municipales, provinciales y estatales— en diversas localidades de la geografía española. Se recoge a continuación una lista actualizada y ordenada alfabéticamente (ver [Tabla 4](#)):

Tabla 4. Lista de archivos históricos consultados: contenido del elemento [<institution>](#).

Nombre del archivo histórico
Archivo General de Simancas
Archivo Histórico Municipal de Baeza
Archivo Histórico Municipal de Loja
Archivo Histórico Municipal de Lorca
Archivo Histórico Provincial de Almería
Archivo Histórico Provincial de Badajoz
Archivo Histórico Provincial de Burgos
Archivo Histórico Provincial de Cáceres
Archivo Histórico Provincial de Cádiz
Archivo Histórico Provincial de Córdoba
Archivo Histórico Provincial de Huelva
Archivo Histórico Provincial de Jaén
Archivo Histórico Provincial de Málaga
Archivo Histórico Provincial de Sevilla
Archivo Histórico de Protocolos de Granada
Archivo Histórico de Protocolos de Madrid
Archivo Municipal de Huéscar
Archivo Municipal de Puerto Real
Archivo Municipal de Vera
Archivo Municipal e Histórico de Protocolos Notariales de Guadix
Archivo de la Real Chancillería de Granada
Archivo de la Real Chancillería de Valladolid
Archivo del Patronato de la Alhambra y Generalife

d. El elemento [<repository>](#) contiene el nombre del fondo documental en donde se conserva el manuscrito. Generalmente, los fondos consultados para la construcción del corpus ODE suelen tener denominaciones relacionadas con protocolos notariales, escribanías, probanzas, pleitos civiles o pleitos criminales.

e. El elemento [<idno>](#) contiene información sobre la ubicación del manuscrito en niveles de organización inferiores al fondo documental, es decir, en unidades como cajas, legajos, expedientes, etcétera. No obstante, la información relativa a la foliación del manuscrito, o del fragmento del manuscrito transcrito, se recoge dentro del elemento [<locus>](#) (ver [sección 2.1.3.2d](#) y ss.).

f. El contenido del elemento [<idno>](#) sigue en ODE un esquema predeterminado, compuesto por un código único, asignado internamente a cada archivo, seguido de la información correspondiente a la ubicación del manuscrito. Esta última información está tomada generalmente de la signatura usada por el propio archivo. A continuación, se presentan las correspondencias entre el nombre de cada archivo y el código asignado en ODE (ver [Tabla 5](#)).

Tabla 5. Relación entre los elementos [<institution>](#) e [<idno>](#) en ODE.

Nombre del archivo <institution>	Código interno <idno>
Archivo General de Simancas	AGS
Archivo Histórico Municipal de Baeza	AHMB
Archivo Histórico Municipal de Loja	AHMLo
Archivo Histórico Municipal de Lorca	AHML
Archivo Histórico Provincial de Almería	AHPAL
Archivo Histórico Provincial de Badajoz	AHPB
Archivo Histórico Provincial de Burgos	AHPBU
Archivo Histórico Provincial de Cáceres	AHPCC
Archivo Histórico Provincial de Cádiz	AHPC
Archivo Histórico Provincial de Córdoba	AHPCo
Archivo Histórico Provincial de Huelva	AHPH
Archivo Histórico Provincial de Jaén	AHPJ
Archivo Histórico Provincial de Málaga	AHPMa
Archivo Histórico Provincial de Sevilla	AHPSe
Archivo Histórico de Protocolos de Granada	AHPGr
Archivo Histórico de Protocolos de Madrid	AHPM
Archivo Municipal de Huéscar	AMH
Archivo Municipal de Puerto Real	AMPR
Archivo Municipal de Vera	AMV
Archivo Municipal e Histórico de Protocolos Notariales de Guadix	AMHPG
Archivo de la Real Chancillería de Granada	ARCHGR
Archivo de la Real Chancillería de Valladolid	ARCHV
Archivo del Patronato de la Alhambra y Generalife	APAG

g. Opcionalmente, el elemento [<idno>](#) incluye un atributo [@source](#), cuyo valor contiene la dirección electrónica correspondiente a la descripción de la unidad documental en la base de datos [PARES](#) (Portal de Archivos Españoles).

h. Considerando lo expuesto, se presentan a continuación algunos ejemplos de codificación del elemento [<msIdentifier>](#) en ODE:

```
<msIdentifier>
  <country>España</country>
  <settlement>Valladolid</settlement>
  <institution>Archivo de la Real Chancillería de
  Valladolid</institution>
  <repository>Sala de lo criminal</repository>
  <idno>
    source="https://pares.mcu.es/ParesBusquedas20/catalogo/descrip
    tion/534500?nm">ARCHV PCR 0033/0003</idno>
</msIdentifier>
```

```
<msIdentifier>
  <country>España</country>
  <settlement>Vera</settlement>
  <institution>Archivo Municipal de Vera</institution>
  <repository>Legajos judiciales</repository>
  <idno>AMV 01-24</idno>
</msIdentifier>
```

```
<msIdentifier>
  <country>España</country>
  <settlement>Baeza</settlement>
  <institution>Archivo Histórico Municipal de
  Baeza</institution>
  <repository>Protocolos notariales</repository>
  <idno>AHMB sala 3/estante 1/número 1</idno>
</msIdentifier>
```

2.1.3.2 Contenido del manuscrito

a. El elemento [<msContents>](#) proporciona información referente al contenido intelectual del manuscrito. Consta en ODE de dos elementos:

- [<summary>](#) (Resumen). Contiene un resumen del contenido del manuscrito.
- [<msItem>](#) (Elemento del manuscrito). Describe una obra individual dentro del contenido intelectual del manuscrito.

b. El elemento [<summary>](#) contiene una breve descripción del contenido del manuscrito. En ODE, su propósito es ampliar la información del título, proporcionando un nivel de detalle algo mayor. La extensión del contenido es variable, como muestran los dos ejemplos siguientes:

```
<summary>Probanza. Juan Ramírez, escribano público de la ciudad de
Loja, en el pleito de denuncia contra Juan Sedano, arrendador de
las penas del campo de dicha ciudad, por haber hecho daños y cortes
en los montes.</summary>
```

`<summary>`D. Matías Rodríguez, natural y vecino de Cádiz, hijo de D. Juan de Dios Rodríguez y D^a Manuela Jerónima Guerrero, vecinos y naturales de la misma ciudad, otorga recibo de los bienes dotales concertados en su matrimonio con D^a Ignacia Fernández de Otáñez, hija de D. Miguel Fernández de Otáñez y D^a Petronila de Gálvez y Montenegro, naturales y vecinos de Cádiz, que ella había recibido a cuenta de sus legítimas en prendas de oro, plata, ropa, alhajas y dinero que D. Juan Fernández de Otáñez, vecino de México, había remitido a sus padres en la última flota de Indias. Por su parte, D. Matías le da en arras 280 pesos escudos asignados en unas casas situadas al final de la calle de la Horca de los Franceses de Cádiz, esquina con el molino del Viento, heredadas de su madre.`</summary>`

c. El elemento `<summary>` no se ha completado sistemáticamente en ODE. En los casos en que no se ha incluido esta información, el elemento aparece vacío.

`<summary/>`

d. El elemento `<msItem>` se utiliza para codificar dos tipos de información: (i) la foliación que ocupa el fragmento transcrito dentro del manuscrito, y (ii) la autenticidad del fragmento transcrito, esto es, si corresponde a un original o a una copia posterior. La información relativa a la foliación se codifica mediante el elemento `<locus>` y la información relativa a la autenticidad del documento se codifica mediante el elemento `<filiation>`. Por tanto, el elemento `<msItem>` consta de dos elementos en ODE, en este orden:

- `<locus>` (Paginación). Define la secuencia de folios del fragmento transcrito dentro del manuscrito.
- `<filiation>` (Filiación). Contiene información referente a la filiación del fragmento transcrito, esto es, a su posible relación con otros testimonios del mismo texto.

e. Por regla general, el elemento `<locus>` señala el intervalo que abarca el fragmento transcrito dentro de un manuscrito, especificando el folio inicial y final que ocupa e incluyendo la orientación (recto o vuelto) en cada caso. Por ejemplo, si un fragmento abarca desde el folio 124 recto hasta el folio 126 vuelto, se representará del modo siguiente:

`<locus>124r-126v</locus>`

f. Si se trata de varios intervalos no consecutivos, se indican todos, separando cada intervalo por coma:

`<locus>323r-323v, 324v-325v, 326v-327r, 329r-329v, 330v</locus>`

g. Si el fragmento transcrito carece de numeración, pero esta puede deducirse a partir de la numeración de la unidad documental, el elemento `<locus>` incluye un atributo `@cert` con el valor "high". Esta misma estrategia se aplica en el atributo `@n` del elemento `<pb>` (ver [sección 4.1.1.1c](#)).

```
<locus cert="high">5r-5v, 7r-7v</locus>
```

h. Si no hay foliación explícita, el contenido del elemento `<locus>` será la abreviatura "sf" (sin foliación). Esta misma estrategia se aplica en el atributo `@n` del elemento `<pb>` (ver [sección 4.1.1.1d](#)).

```
<locus>sf</locus>
```

i. El elemento `<filiation>` contiene un atributo `@type` que puede tomar dos valores —"original" o "copy"—, según si el fragmento transcrito es un original o una copia posterior. La mayoría de los documentos en el corpus ODE son originales. Sin embargo, algunos pleitos consultados en los archivos históricos incluyen documentación que es resultado de un proceso de transmisión documental, de tal forma que las partes litigantes aportan como prueba inventarios o testamentos que son copias de originales de años anteriores. El uso del elemento `<filiation>` en ODE es precisamente para identificar estos casos, absolutamente excepcionales dentro del corpus.

j. En el caso de tratarse de una copia, el elemento `<filiation>` incluye dos elementos: `<origDate>` y `<date>`. El primero informa del año de creación del texto original, mientras que el segundo informa del año de creación de la copia. El siguiente ejemplo corresponde a una copia de 1706 de un inventario escrito originalmente en 1690:

```
<filiation type="copy">
  <origDate>1690</origDate>
  <date>1706</date>
</filiation>
```

k. En el caso de tratarse de un original, el elemento `<filiation>` se deja vacío:

```
<filiation type="original"/>
```

l. La fecha de creación del texto se recoge, en cualquier caso, dentro del apartado [<setting>](#) (ver [sección 2.3.2d](#)).

2.1.3.3 Información adicional

a. El elemento [<additional>](#) agrupa información adicional sobre el manuscrito. En ODE consta de dos elementos:

- [<adminInfo>](#) (Información administrativa). Contiene información relativa a la gestión y a la disponibilidad del manuscrito.
- [<surrogates>](#). (Reproducción). Contiene información relativa a posibles representaciones del manuscrito en diferentes formatos.

b. El elemento [<adminInfo>](#) incluye un elemento [<availability>](#), utilizado en ODE para especificar los permisos de uso de las imágenes facsimilares. La licencia de uso de estas imágenes, almacenadas en el corpus ODE, depende de la autorización de la institución correspondiente, es decir, del archivo histórico que custodia el manuscrito (ver [sección 2.1.3.1c](#)). Por tanto, el contenido de [<availability>](#) consiste en un párrafo prácticamente invariable, en el que solo cambia el nombre de la institución correspondiente en cada caso. Por ejemplo:

```
<adminInfo>
  <availability>
    <p>Para hacer uso de las imágenes es necesaria la
    autorización del Archivo Municipal de Vera</p>
  </availability>
</adminInfo>
```

```
<adminInfo>
  <availability>
    <p>Para hacer uso de las imágenes es necesaria la
    autorización del Archivo de la Real Chancillería de Granada</p>
  </availability>
</adminInfo>
```

c. El elemento [<surrogates>](#) se utiliza en ODE para especificar el formato de las reproducciones facsimilares. Su estructura en XML incluye un elemento [<bibl>](#), que contiene a su vez un elemento [<title>](#) con el atributo [@type](#) y el valor "gmd" (*general material designation*). Puesto que todas las ediciones facsimilares de ODE están guardadas en formato JPEG, el contenido de este elemento [<title>](#) es invariable. Por ejemplo:

```
<surrogates>
  <bibl>
    <title type="gmd">facsimil digital guardado en formato
      JPEG</title>
  </bibl>
</surrogates>
```

2.2 Descripción de la codificación

El elemento [<encodingDesc>](#) está diseñado para documentar la relación entre el texto electrónico y la fuente de la que deriva. En ODE, se utiliza para describir cuatro tipos de información: (i) las características principales del corpus, (ii), el subcorpus al que pertenece el documento transcrito, (iii) las herramientas empleadas para llevar a cabo las diversas tareas de procesamiento automático del texto, y (iv) la clasificación utilizada para asignar el tipo textual a cada documento transcrito. La información general del corpus se incluye en el elemento [<projectDesc>](#), el subcorpus en que se integra el documento se incluye en el elemento [<samplingDecl>](#), la información sobre el uso de herramientas de procesamiento automático se incluye en el elemento [<editorialDecl>](#) y la información relativa a la clasificación textual se incluye en el elemento [<classDecl>](#). Por tanto, el elemento [<encodingDesc>](#) consta de cuatro elementos en ODE, en este orden:

- [<projectDesc>](#) (Descripción del proyecto). Describe el propósito para el que un archivo electrónico ha sido codificado.
- [<samplingDecl>](#) (Declaración de la muestra). Contiene información sobre la creación de una selección de textos.
- [<editorialDecl>](#) (Declaración de la edición). Proporciona detalles de principios editoriales y prácticas aplicadas en la codificación de un texto.
- [<classDecl>](#) (Declaraciones de clasificación). Contiene una o más taxonomías que definen cualquier código usado en algún punto del texto.

2.2.1 Descripción del proyecto

El elemento [<projectDesc>](#) contiene la descripción general del corpus ODE. Se trata, por tanto, de un elemento con contenido invariable que adopta la forma siguiente:

```
<projectDesc>
  <p><ref target=" http://corpora.ugr.es/ode/">Oralia diacrónica
    del español (ODE)</ref> es un corpus de inventarios de bienes,
    declaraciones de testigos en juicios penales y certificaciones
    de barberos y cirujanos de los siglos XVI a XIX. Actualmente,
    ofrece documentación de todas las provincias andaluzas, para
    contribuir a la elaboración de un ALEA de carácter histórico,
    y de algunas regiones peninsulares, que sirven de corpus de
    control de la documentación andaluza</p>
</projectDesc>
```

2.2.2 Declaración de la muestra

a. Actualmente, ODE está constituido por tres subcorpus, diferenciados de acuerdo con un criterio geográfico. El elemento [<samplingDecl>](#) se utiliza para proporcionar información relativa al subcorpus al que pertenece cada documento. Los tres subcorpus que constituyen ODE en la actualidad son los siguientes:

- Extremadura - CORTENEX
- Andalucía
- Madrid y otros

b. El elemento [<samplingDecl>](#) contiene un atributo [@xml:id](#) que puede tomar tres valores —"EXT", "AND" y "MAD"—, según el subcorpus al que pertenezca el documento transcrito.

c. El subcorpus de Extremadura se denomina CORTENEX (Corpus de textos notariales extremeños) y está siendo compilado por Inmaculada González Sopena. El subcorpus de Andalucía es resultado del trabajo realizado en diferentes proyectos de investigación hasta la actualidad, fundamentalmente CORDEREGR (Corpus Diacrónico del Español del Reino de Granada), ALEA XVIII, ALEA oriental-XVIII y ALEA oriental-XIX. El tercer subcorpus es fruto del trabajo realizado en distintos archivos históricos fuera de Andalucía, con el objetivo de ir conformando un subcorpus de control que sirva de referencia comparativa respecto a la documentación andaluza. Actualmente, está formado, en su mayoría, por documentación procedente de Madrid, con una representación significativa de Burgos y Valladolid, y una presencia menor de otros fondos documentales del centro y norte peninsular, como Toledo, León, Soria, Álava, Cantabria, Palencia o Zamora, entre otros.

d. Se presentan a continuación algunos ejemplos de codificación del elemento `<samplingDecl>` en ODE:

```
<samplingDecl xml:id="EXT">
  <p>Este texto forma parte de la muestra de documentos de ODE
  producidos en Extremadura e integra el subcorpus CORTENEX (Corpus de
  textos notariales extremeños), compilado por <name>Inmaculada
  González Sopeña</name></p>
</samplingDecl>
```

```
<samplingDecl xml:id="AND">
  <p>Este texto forma parte de la muestra de documentos de ODE
  producidos en Andalucía</p>
</samplingDecl>
```

```
<samplingDecl xml:id="MAD">
  <p>Este texto forma parte de la muestra de documentos de ODE
  producidos en Madrid y otras provincias del centro y norte
  peninsular</p>
</samplingDecl>
```

2.2.3 Declaración de la edición

El elemento `<editorialDecl>` tiene como propósito enumerar las diferentes tareas de procesamiento automático aplicadas al texto, así como las herramientas empleadas para llevar a cabo cada una de ellas. Se trata de un elemento con contenido invariable que se presenta de la forma siguiente:

```
<editorialDecl>
  <p>La tokenización se ha realizado con el script
  <term>xmltokenize.pl</term>, desarrollado por Maarten
  Janssen</p>
  <p>La normalización se ha realizado con el script
  <term>nformtreat.pl</term>, desarrollado por Maarten
  Janssen</p>
  <p>La lematización y el etiquetado morfosintáctico se han
  realizado con el anotador <term>NeoTag</term>, desarrollado por
  Maarten Janssen</p>
  <p>La anotación de información fonética, etimológica y
  semántica se ha realizado con el script <term>ltags.pl</term>,
  desarrollado por Gael Vaamonde</p>
  <p>La conversión de la codificación XML al estándar TEI P5 se
  ha realizado con el script <term>ode2teip5.pl</term>,
  desarrollado por Gael Vaamonde</p>
</editorialDecl>
```

2.2.4 Declaración de la clasificación

a. El elemento [<classDecl>](#) contiene un elemento [<taxonomy>](#) que recoge la taxonomía empleada en ODE para clasificar los documentos en función del tipo textual. Esta taxonomía comprende actualmente cuatro categorías:

- inventarios de bienes
- certificados médicos
- declaraciones de testigos
- otros

b. Cada categoría de la taxonomía se define mediante un elemento [<category>](#), al que se asigna un identificador único: "inv" para inventarios, "cer" para certificados médicos, "dec" para declaraciones de testigos y "oth" para otros. Finalmente, el elemento [<catDesc>](#) describe brevemente cada categoría. Según lo expuesto, elemento [<classDecl>](#) es de contenido invariable y adopta la siguiente estructura XML en ODE:

```
<classDecl>
  <taxonomy>
    <category xml:id="inv">
      <catDesc>inventarios de bienes</catDesc>
    </category>
    <category xml:id="cer">
      <catDesc>certificados médicos</catDesc>
    </category>
    <category xml:id="dec">
      <catDesc>declaraciones de testigos</catDesc>
    </category>
    <category xml:id="oth">
      <catDesc>otros</catDesc>
    </category>
  </taxonomy>
</classDecl>
```

c. La categoría específica asignada a cada documento se indica dentro del elemento [<textClass>](#) (ver [sección 2.3.3](#)).

2.3 Descripción del perfil del texto

El elemento [<profileDesc>](#) agrupa aspectos no bibliográficos del texto. En ODE este elemento consta de cuatro elementos:

- [<creation>](#) (Creación). Contiene información sobre la creación del texto.
- [<settingDesc>](#) (Descripción del contexto). Describe el marco contextual en el que tiene lugar una interacción lingüística.
- [<textClass>](#) (Clasificación del texto). Agrupa información que describe la temática de un texto de acuerdo con un esquema predefinido.

2.3.1 Creación

a. En ODE, el elemento [<creation>](#) se emplea para registrar el nombre del escribano que redactó el manuscrito. Para ello, incluye un elemento [<name>](#) que contiene dicho nombre. Por ejemplo:

```
<creation>
  <name>Diego Hernández Cárdenas</name>
</creation>
```

b. Este elemento no se ha completado sistemáticamente. En los casos en que no se ha incluido esta información, el elemento [<name>](#) aparece vacío:

```
<creation>
  <name/>
</creation>
```

2.3.2 Descripción del contexto

a. El elemento [<settingDesc>](#) incluye un elemento [<setting>](#), que recoge información referente al lugar y la fecha de creación del manuscrito.

b. La codificación de la información relativa al lugar de creación del manuscrito es de particular importancia en ODE, ya que este corpus ha sido diseñado específicamente para llevar a cabo estudios de dialectología histórica. Esta información se incorpora mediante tres estrategias complementarias, todas ellas incluidas dentro del elemento [<placeName>](#). En primer lugar, este elemento incluye tres valores separados por comas, que representan el país, la provincia y la ciudad donde se creó el manuscrito. En segundo lugar, estos mismos tres valores se distribuyen en elementos [<name>](#) individuales. Cada elemento [<name>](#) tiene un atributo [@n](#) que asigna una numeración correlativa (1, 2, 3) y un atributo [@type](#) que indica el nivel de organización territorial correspondiente

(*country*, *province*, *city*). En tercer lugar, se añade un elemento [<geo>](#), que contiene las coordenadas geográficas correspondientes al nivel de organización más específico (*city*). Estas coordenadas incluyen la latitud y la longitud, en ese orden y separadas por espacio. Por ejemplo:

```
<placeName>
  España, Granada, Loja
  <name n="1" type="country">España</name>
  <name n="2" type="province">Granada</name>
  <name n="3" type="city">Loja</name>
  <geo>37.1664839 -4.1496374</geo>
</placeName>
```

c. En caso de que el nombre de la provincia y el de la ciudad coincidan, se repite el valor coincidente, de manera que la información siempre se presente en tres partes:

```
<placeName>
  España, Granada, Granada
  <name n="1" type="country">España</name>
  <name n="2" type="province">Granada</name>
  <name n="3" type="city">Granada</name>
  <geo>37.3210180 -4.0111230</geo>
</placeName>
```

d. La información relativa a la fecha de creación del manuscrito se especifica de dos maneras: año y siglo. Esta información se distribuye en sendos elementos [<date>](#), que utilizan un atributo [@n](#) para ser numerados correlativamente y un atributo [@type](#) con uno de dos valores posibles, "year" o "century", según corresponda. Por ejemplo:

```
<date n="1" type="year">1690</date>
<date n="2" type="century">XVII</date>
```

e. En caso de tratarse de una copia producida en años posteriores a la creación del testimonio original, la fecha aquí recogida es siempre la del testimonio original. En cualquier caso, ambas fechas, la de la copia y la del original, se recogen dentro del elemento [<filiation>](#) (ver [sección 2.1.3.2i](#) y ss.).

2.3.3 Clasificación del texto

a. El elemento [<textClass>](#) se utiliza para clasificar el texto según la taxonomía descrita en el elemento [<classDecl>](#) (ver [sección 2.2.3](#)). Esta clasificación se codifica dentro del elemento [<catRef>](#), específicamente en el valor del atributo [@target](#).

b. Los posibles valores para el atributo **@target**, conforme a la taxonomía utilizada en ODE, son "#inv", "#cer", "#dec" y "#oth". La almohadilla (#) se utiliza para señalar referencias a identificadores de elementos dentro del mismo documento (ver [Tabla 6](#)).

Tabla 6. Relación entre los elementos [<catRef>](#) y [<category>](#) en ODE.

Categoría asignada al documento <catRef>	Categoría referenciada en la taxonomía <category>
<code><catRef target="#inv"/></code>	<code><category xml:id="inv"> <catDesc>inventarios de bienes</catDesc> </category></code>
<code><catRef target="#cer"/></code>	<code><category xml:id="cer"> <catDesc>certificados médicos</catDesc> </category></code>
<code><catRef target="#dec"/></code>	<code><category xml:id="dec"> <catDesc>declaraciones de testigos</catDesc> </category></code>
<code><catRef target="#oth"/></code>	<code><category xml:id="oth"> <catDesc>otros</catDesc> </category></code>

c. De acuerdo con lo mencionado anteriormente, en el siguiente ejemplo el valor "#inv" hace referencia al elemento que contiene el identificador único "inv", es decir, al elemento `<category xml:id="inv">` incluido en la taxonomía:

```
<textClass>
  <catRef target="#inv"/>
</textClass>
```

2.4 Ejemplo completo de cabecera

Se recoge a continuación un ejemplo completo de cabecera ([<teiHeader>](#)). Corresponde al archivo electrónico CA170812525.xml, que contiene un inventario de bienes producido en Cádiz en el año 1708. La edición digital del documento está disponible para su consulta en la página web de ODE, específicamente en [este enlace](#).

```
<teiHeader>
  <fileDesc>
    <titleStmt>
      <title>Inventario de los bienes del fallecido D. Lorenzo Hernández de Menda,
sargento mayor</title>
      <editor xml:id="ALEA18">ALEA XVIII: Atlas lingüístico y etnográfico de Andalucía,
siglo XVIII. Patrimonio documental y Humanidades Digitales</editor>
      <funder from="2020" to="2023">Junta de Andalucía/FEDER: P18-FR-695</funder>
      <respStmt n="1">
        <resp>Transcripción</resp>
        <name>Diego Reinaldos Miñarro</name>
      </respStmt>
      <respStmt n="2">
        <resp>Normalización</resp>
        <name>Pilar Arrabal Rodríguez</name>
      </respStmt>
      <respStmt n="3">
        <resp>Anotación lingüística</resp>
        <name>Pilar Arrabal Rodríguez</name>
      </respStmt>
      <respStmt n="4">
        <resp>Revisión</resp>
        <name>Gael Vaamonde</name>
      </respStmt>
    </titleStmt>
    <publicationStmt>
      <publisher>UGR, Universidad de Granada</publisher>
      <pubPlace>Granada</pubPlace>
      <distributor>HUM-278. Grupo DiLEs: Diacronía de la Lengua Española. UGR-Junta
de Andalucía</distributor>
    </publicationStmt>
    <sourceDesc>
      <msDesc>
        <msIdentifier>
          <country>España</country>
          <settlement>Cádiz</settlement>
          <institution>Archivo Histórico Provincial de Cádiz</institution>
          <repository>Protocolos notariales</repository>
          <idno>AHPC 1820</idno>
        </msIdentifier>
        <msContents>
          <summary>D. Francisco de Aguilera, sargento mayor del regimiento del
marqués de Alcántara, y D. Alonso Berral, teniente del regimiento de Baena, albaceas
testamentarios de D. Lorenzo Hernández de Menda, fallecido, realizan inventario y
aprecio de los bienes dejados en su testamento, para proceder a venta y remate en
almoneda pública, ante Juan Miguel Bermúdez Luna, escribano público</summary>
          <msItem>
            <locus>134r-141r</locus>
            <filiation type="original"/>
          </msItem>
        </msContents>
      </msDesc>
      <additional>
        <adminInfo>
          <availability>
            <p>Para hacer uso de las imágenes es necesaria la autorización del
Archivo Histórico Provincial de Cádiz</p>
          </availability>
        </adminInfo>
        <surrogates>
          <bibl>
            <title type="gmd">facsimil digital guardado en formato JPEG</title>
          </bibl>
        </surrogates>
      </additional>
    </msDesc>
  </sourceDesc>
</fileDesc>
```

```
<encodingDesc>
  <projectDesc>
    <p><ref target="http://corpora.ugr.es/ode/">Oralia diacrónica del español
(ODE)</ref> es un corpus de inventarios de bienes, declaraciones de testigos en
juicios penales y certificaciones de barberos y cirujanos de los siglos XVI a XIX.
Actualmente, ofrece documentación de todas las provincias andaluzas, para contribuir
a la elaboración de un ALEA de carácter histórico, y de algunas regiones
peninsulares, que sirven de corpus de control de la documentación andaluza</p>
  </projectDesc>
  <samplingDesc xml:id="AND">
    <p>Este texto forma parte de la muestra de documentos de ODE producidos en
Andalucía</p>
  </samplingDesc>
  <editorialDecl>
    <p>La tokenización se ha realizado con el script <term>xmltokenize.pl</term>,
desarrollado por Maarten Janssen</p>
    <p>La normalización se ha realizado con el script <term>nfmtreat.pl</term>,
desarrollado por Maarten Janssen</p>
    <p>La lematización y el etiquetado morfosintáctico se han realizado con el
anotador <term>NeoTag</term>, desarrollado por Maarten Janssen</p>
    <p>La anotación de información fonética, etimológica y semántica se ha
realizado con el script <term>ltags.pl</term>, desarrollado por Gael Vaamonde</p>
    <p>La conversión de la codificación XML al estándar TEI P5 se ha realizado con
el script <term>ode2teip5.pl</term>, desarrollado por Gael Vaamonde</p>
  </editorialDecl>
  <classDecl>
    <taxonomy>
      <category xml:id="inv">
        <catDesc>inventarios de bienes</catDesc>
      </category>
      <category xml:id="cer">
        <catDesc>certificados médicos</catDesc>
      </category>
      <category xml:id="dec">
        <catDesc>declaraciones de testigos</catDesc>
      </category>
      <category xml:id="oth">
        <catDesc>otros</catDesc>
      </category>
    </taxonomy>
  </classDecl>
</encodingDesc>
<profileDesc>
  <creation>
    <name>Juan Miguel Bermúdez de Luna</name>
  </creation>
  <settingDesc>
    <setting>
      <placeName>España, Cádiz, Cádiz
      <name n="1" type="country">España</name>
      <name n="2" type="province">Cádiz</name>
      <name n="3" type="city">Cádiz</name>
      <geo>36.5297438 -6.2928976</geo>
    </placeName>
    <date n="1" type="year">1708</date>
    <date n="2" type="century">XVIII</date>
  </setting>
</settingDesc>
  <textClass>
    <catRef target="#inv"/>
  </textClass>
</profileDesc>
</teiHeader>
```

3. LA EDICIÓN FACSIMILAR

a. El elemento [<facsimile>](#) contiene el conjunto de imágenes facsimilares correspondientes al texto transcrito. En ODE, cada imagen corresponde siempre a una página y se guarda en formato JPEG. La lista de imágenes dentro de [<facsimile>](#) se codifica utilizando un elemento [<graphic>](#) por cada imagen existente. El atributo [@url](#) dentro del elemento [<graphic>](#) almacena el nombre del archivo JPEG correspondiente. Por ejemplo

```
<facsimile>
  <graphic url="GR_0358.jpg"/>
  <graphic url="GR_0359.jpg"/>
  <graphic url="GR_0360.jpg"/>
  <graphic url="GR_0361.jpg"/>
</facsimile>
```

b. Los valores capturados en el atributo [@url](#), esto es, los nombres de los archivos JPEG correspondientes a cada imagen facsimilar, permiten vincular cada página con su imagen correspondiente. Esta vinculación se establece mediante el atributo [@facs](#) dentro del elemento [<pb>](#), que señala el inicio de cada página del manuscrito (ver [sección 4.1.1.1](#)).

4. EL TEXTO

En ODE, todo documento consta de un único texto incluido en el elemento `<text>`. Este elemento contiene obligatoriamente un elemento `<body>`, que incluye la codificación relacionada con el contenido textual del manuscrito. Esta codificación se realiza teniendo en cuenta dos perspectivas de análisis diferentes, que se aplican conjuntamente dentro del archivo XML:

- La transcripción paleográfica
- El corpus lingüístico

4.1 La transcripción paleográfica

ODE ofrece una edición digital paleográfica de los textos manuscritos, en la que se respetan las abreviaturas, que han sido debidamente desarrolladas, así como los cambios de línea y de párrafo, los tachones, las lagunas textuales y las adiciones fuera de línea, entre otros aspectos del texto fuente.

Únicamente se ha normalizado la delimitación de palabras respecto al texto original, ya que las dificultades para determinar con precisión dónde comienza y termina cada palabra ortográfica podrían generar inconsistencias indeseadas. Por lo tanto, dos o más palabras consecutivas que en el manuscrito original están escritas —o parecen estar escritas— sin espacios en blanco se separan con espacio en la edición digital; del mismo modo, una única palabra que en el manuscrito original aparece partida —o parece estar partida— por uno o más espacios en blanco se transcribe sin espacios en la edición digital.

En esta sección se explican los criterios de transcripción adoptados en ODE para la edición paleográfica digital de los textos.

4.1.1 Los elementos estructurales

Los elementos estructurales sirven para subdividir el texto en bloques. En ODE se hace uso de tres elementos estructurales, todos ellos vacíos de contenido:

- `<pb>` (Inicio de página). Marca el inicio de una nueva página de un texto.
- `<lb>` (Inicio de línea). Marca el inicio de una nueva línea (topográfica).
- `<cb>` (Inicio de columna). Marca el inicio de una nueva columna.

Los dos primeros elementos —[<pb>](#) y [<lb>](#)— son de uso muy frecuente en ODE, ya que se encuentran en todas las transcripciones; en cambio, el tercero —[<cb>](#)— aparece de forma muy esporádica, dado que las páginas de los documentos transcritos en ODE rara vez están divididas en columnas.

4.1.1.1 Inicio de página

a. El elemento [<pb>](#) marca el inicio de página. En ODE incluye siempre al menos tres atributos: [@n](#), que especifica el número de página y su orientación (recto o vuelto); [@break](#), que indica si el inicio de página marca el final de la palabra ortográfica adyacente, y [@facs](#), que proporciona el nombre del archivo JPEG que contiene la imagen, permitiendo así vincular cada página con su correspondiente imagen facsimilar:

```
<pb n="436r" facs="JA_0902.JPG" break="no"/>
```

b. El atributo [@n](#) indica el número de página. Nótese que los textos transcritos en ODE están numerados conforme a la foliación de la unidad documental a la que pertenecen, que suele ser un pleito o un documento notarial. Por ello, la numeración de los distintos elementos [<pb>](#) en el texto transcrito refleja esta foliación: 1 recto, 1 vuelto, y así sucesivamente. De este modo, el valor del atributo [@n](#) suele estar compuesto por uno o más dígitos, seguidos de las letras *r* (recto) o *v* (vuelto), según corresponda.

c. Puede suceder que el texto transcrito, o alguna de sus páginas, no tenga numeración, aunque esta puede inferirse a partir de la numeración de la unidad documental. En tales casos, el elemento [<pb>](#) puede incluir opcionalmente un cuarto atributo, [@cert](#), que indica el grado de certeza de la numeración. Dado que en estos casos la numeración de cada página es siempre deducible de la unidad documental, el valor de este atributo es siempre "high". Esta misma estrategia se aplica dentro del elemento [<locus>](#) (ver [sección 2.1.3.2g](#)).

```
<pb n="6r" facs="IMG_7085.JPG" cert="high" break="no"/>
```

d. En caso de que la unidad documental en su conjunto no tenga numeración, el valor del atributo será "sf" (sin foliación). Esta misma estrategia se aplica dentro del elemento [<locus>](#) (ver [sección 2.1.3.2h](#)).

```
<pb n="sf" facs="CC_6836.JPG" break="no"/>
```

e. El atributo [@facs](#) permite vincular cada página con su correspondiente imagen facsimilar. El valor de este atributo es el nombre del archivo JPEG que contiene la imagen, incluyendo la extensión del archivo (*.jpg*). La lista completa de las imágenes facsimilares para cada documento transcrito se recoge dentro del elemento `<facsimile>` (ver [sección 3](#)).

f. El atributo [@break](#) señala si el inicio de página coincide con el final de una palabra o si, por el contrario, se encuentra en mitad de una palabra. En el primer caso, el atributo toma el valor "yes", y en el segundo caso, el valor es "no". Así, en el primer ejemplo recogido a continuación el inicio de página se sitúa entre las palabras *qual* y *en*, lo que significa que el atributo [@break](#) adopta el valor "yes" (es decir, el inicio de página marca el final de la palabra ortográfica adyacente, que es *qual*). En cambio, en el segundo ejemplo el inicio de página se encuentra a mitad de la palabra *ratificaron*, por lo que el atributo [@break](#) toma el valor "no":

```
el qual <pb n="76v" facs="IMG_0257.jpg" break="yes"/> en el día
```

```
ratifica<pb n="77r" facs="IMG_0258.jpg" break="no"/>ron
```

4.1.1.2 Inicio de línea

a. El elemento `<lb>` marca el inicio de línea. En ODE incluye siempre dos atributos: [@n](#), que señala el número de línea, y [@break](#), que indica si el inicio de línea marca el final de la palabra ortográfica adyacente:

```
<lb n="35" break="yes"/>
```

b. El atributo [@n](#) proporciona información sobre el número de línea en el texto. Este atributo comienza con el valor "1" en la primera línea y aumenta de manera correlativa, con independencia del número de páginas, hasta llegar al final del texto transcrito. Por ejemplo, un documento constituido por 30 líneas distribuidas en dos páginas y con 15 líneas por página tendría una estructura similar a la siguiente:

```
<pb n="1r" facs="IMG_0001.jpg" break="yes"/>
<lb n="1" break="yes"/> <!-- inicio de la primera línea -->
<lb n="2" break="yes"/> <!-- inicio de la segunda línea -->
<!-- continúa la numeración correlativa de líneas -->
<lb n="15" break="yes"/>
<pb n="1v" facs="IMG_0002.jpg" break="yes"/>
<lb n="16" break="yes"/>
<lb n="17" break="yes"/>
<!-- continúa la numeración correlativa de líneas -->
<lb n="30" break="yes"/> <!-- inicio de la última línea -->
```

c. El atributo **@break** señala si el inicio de línea coincide con el final de una palabra o si, por el contrario, se encuentra en mitad de una palabra. En el primer caso, el atributo toma el valor "yes", y en el segundo caso, el valor es "no". Así, en el primer ejemplo recogido a continuación el inicio de línea se sitúa entre las palabras *en* y *veinte*, lo que significa que el atributo **@break** adopta el valor "yes" (es decir, el inicio de línea marca el final de la palabra ortográfica adyacente, que es *en*); en cambio, en el segundo ejemplo el inicio de línea se encuentra a mitad de la palabra *veinte*, por lo que el atributo **@break** toma el valor "no":

```
en <lb n="264" break="yes"/> veinte reales
```

```
ochocientos ve<lb n="12" break="no"/>inte reales
```

4.1.1.3 Inicio de columna

a. El elemento **<cb>** marca el inicio de una nueva columna de texto en una página dividida en columnas. Aunque esta disposición es poco común en los documentos de ODE, existen algunos casos en los que se utiliza, por lo que se incluye este elemento estructural en esta guía. Al igual que el elemento **<lb>**, el elemento **<cb>** incluye dos atributos: **@n**, que señala el número de columna, y **@break**, que indica si el inicio de columna marca el final de la palabra ortográfica adyacente. Por ejemplo, una página constituida por 20 líneas distribuidas en dos columnas y 10 líneas por columna tendría una estructura similar a la siguiente:


```
<pb n="1r" facs="IMG_0001.jpg" break="yes"/>
<cb n="1" break="yes"/> <!-- inicio de la primera columna -->
<lb n="1" break="yes"/> <!-- inicio de la primera línea -->
<lb n="2" break="yes"/> <!-- inicio de la segunda línea -->
<!-- continúa la numeración correlativa de líneas -->
<lb n="10" break="yes"/>
<cb n="2" break="yes"/> <!-- inicio de la segunda columna -->
<lb n="11" break="yes"/>
<lb n="12" break="yes"/>
<!-- continúa la numeración correlativa de líneas -->
<lb n="2'" break="yes"/> <!-- inicio de la última línea -->
```

b. El atributo **@break** señala si el inicio de columna coincide con el final de una palabra o si, por el contrario, se encuentra en mitad de una palabra. En el primer caso, el atributo toma el valor "yes", y en el segundo caso, el valor es "no". Así, en el primer ejemplo recogido a continuación el inicio de columna se sitúa entre las palabras *doña* y *Manuela*, lo que significa que el atributo **@break** adopta el valor "yes" (es decir, el inicio de columna marca el final de la palabra ortográfica adyacente, que es *doña*); en cambio, en el segundo ejemplo el inicio de columna se encuentra a mitad de la palabra *Manuela*, por lo que el atributo **@break** toma el valor "no":

doña <cb n="45" break="yes"/> Manuela

doña Ma<cb n="45" break="no"/>nuela

4.1.2 Las abreviaturas

a. En ODE se marcan y se desarrollan todas las abreviaturas que aparecen en el cuerpo del texto. Para la marcación de abreviaturas se utiliza el elemento **<abbr>** y para la forma desarrollada correspondiente se utiliza el elemento **<expan>**. Puesto que se trata de dos alternativas de transcripción que remiten a un mismo fragmento de texto, ambos elementos se incluyen dentro de un elemento **<choice>**. Por ejemplo, la forma *vo* como abreviatura de *vecino* se marca del modo siguiente:

```
<choice>
  <abbr>vo</abbr>
  <expan>vecino</expan>
</choice>
```

b. El desarrollo de la abreviatura no implica necesariamente su normalización ortográfica. Por ejemplo, la forma *vzo* como abreviatura de *vezino* se marca del modo siguiente:

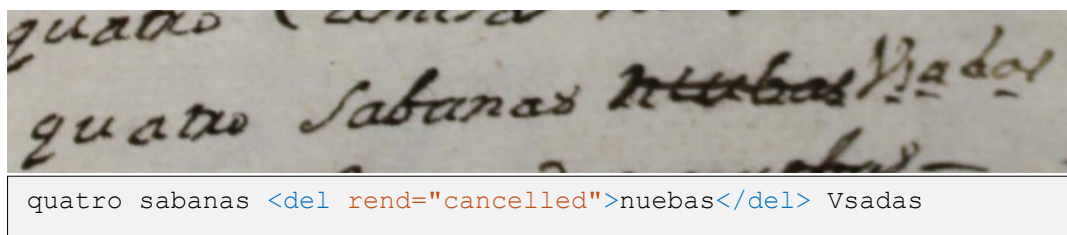
```
<choice>
  <abbr>vzo</abbr>
  <expn>vezino</expn>
</choice>
```

c. La estructura de marcado de las abreviaturas se incluye dentro del elemento `<w>`, que contiene la información referente a cada palabra del texto (ver [sección 4.2.1](#)).

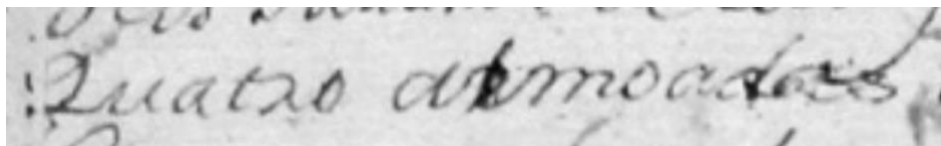
4.1.3 Texto cancelado

a. El elemento `` se utiliza para marcar un segmento de texto que ha sido cancelado, pero cuyo contenido se puede leer. Este elemento incluye siempre un atributo `@rend`, que señala cómo está representada la cancelación en el texto fuente. En ODE, el atributo `@rend` solo puede tomar uno de dos valores posibles: "cancelled" y "overwritten".

b. El valor "cancelled" se usa para marcar segmentos de texto que han sido tachados o cancelados mediante algún tipo de rayado. Un caso ilustrativo es la palabra *nuebas* en el ejemplo siguiente:



c. El valor "overwritten" se usa para marcar segmentos de texto que fueron reemplazados por contenido nuevo, escrito encima del anterior. En estos casos, el elemento `` se combina siempre con el elemento `<add>`, que indica el contenido añadido (ver [sección 4.1.4](#)). Un ejemplo ilustrativo es la *r* de la palabra *armoadas*, sobre la que se superpuso una *l*, como se observa a continuación:



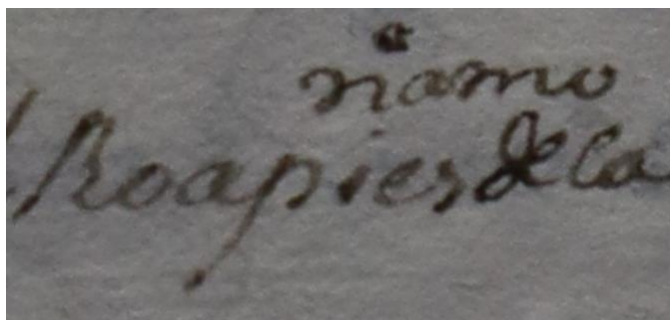
```
quatro  
a<del rend="overwritten">r</del><add place="inline">r</add>moadas
```

4.1.4 Texto añadido

a. El elemento `<add>` se utiliza para marcar un segmento de texto que ha sido añadido. Este elemento incluye siempre un atributo `@place`, que señala en qué parte del manuscrito se ha añadido el segmento. En ODE, el atributo `@place` puede tomar uno de siete valores posibles:

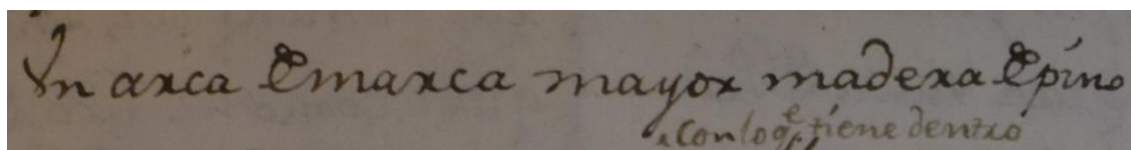
- "above": encima de la línea
- "below": debajo de la línea
- "left": en el margen izquierdo
- "right": en el margen derecho
- "top": en el margen superior
- "bottom": en el margen inferior
- "inline": en la línea de escritura

b. El valor "above" se usa para marcar segmentos de texto añadidos por encima de la línea; por ejemplo, el segmento *ñamo* de la palabra *cáñamo*, que se observa a continuación:



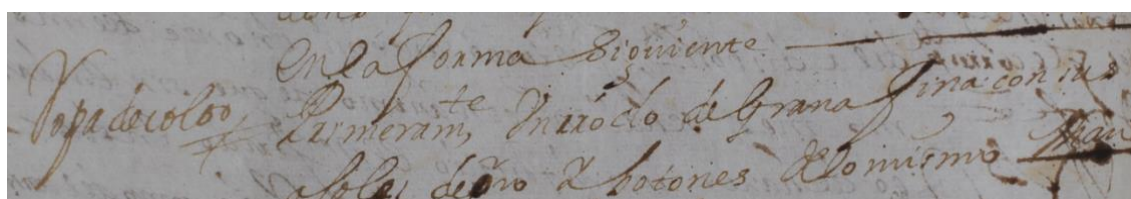
```
Roapies de ca<add place="above">ñamo</add>
```

c. El valor "below" se usa para marcar segmentos de texto añadidos por debajo de la línea: por ejemplo, el segmento *con lo que tiene dentro*, que se observa a continuación:



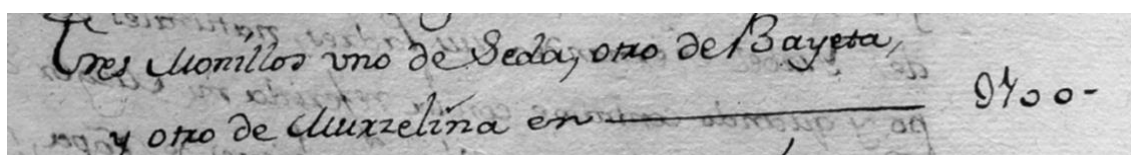
Vn arca de marca mayor madera de pino `<add place="below">con lo que tiene dentro</add>`

d. El valor "left" se usa para marcar segmentos de texto añadidos en el margen izquierdo de la página: por ejemplo, el segmento *ropa de color*, que se observa a continuación:



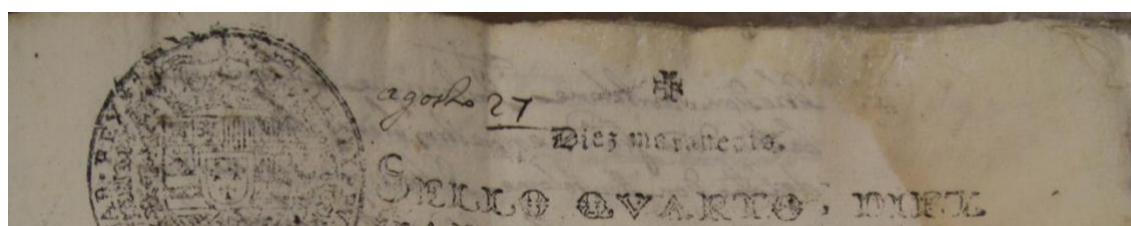
`<add place="left">ropa de color</add>`

e. El valor "right" se usa para marcar segmentos de texto añadidos en el margen derecho de la página. En ODE, estos casos suelen ser precios referidos a tasaciones de objetos: por ejemplo, el segmento *d100*, que se observa a continuación:



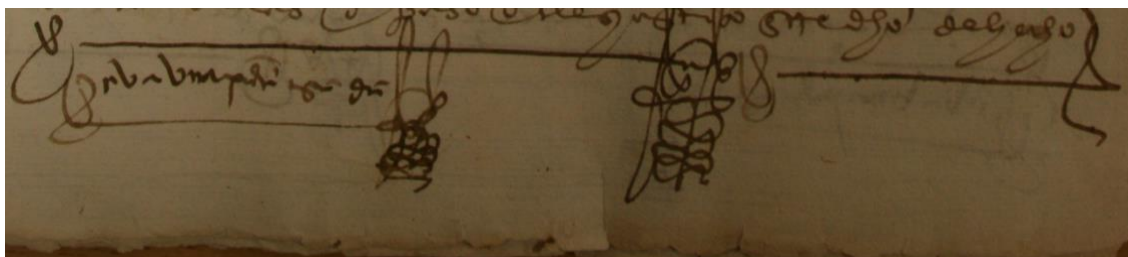
`<add place="right">d100</add>`

f. El valor "top" se usa para marcar segmentos de texto añadidos en el margen superior de la página: por ejemplo, el segmento *agosto 27*, que se observa a continuación:



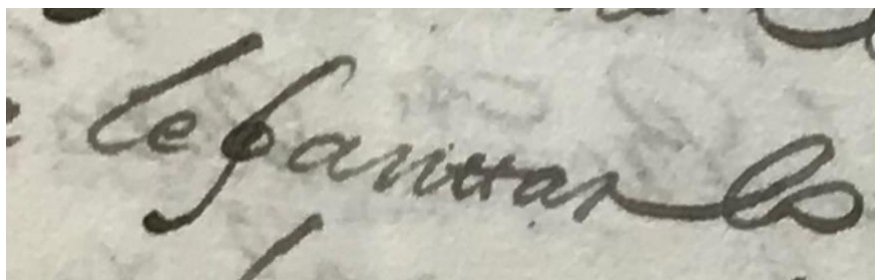
`<add place="top">agosto 27</add>`

g. El valor "bottom" se usa para marcar segmentos de texto añadidos en el margen inferior de la página: por ejemplo, el segmento *va vna pat testada*, que se observa a continuación



```
<add place="bottom">va vna pat testada</add>
```

h. El valor "inline" se utiliza para señalar segmentos añadidos directamente en la línea del texto. En ODE, este valor se reserva específicamente para marcar casos de segmentos de texto sobrescritos, en donde el elemento `<add>` se utiliza en combinación con el elemento `` (ver [sección 4.1.3](#)). El contenido reemplazado se marca con `` y el contenido nuevo con `<add>`: por ejemplo, en la palabra *lebanttarlo*, la *b* se escribió encima de una *p* original, como se observa a continuación:



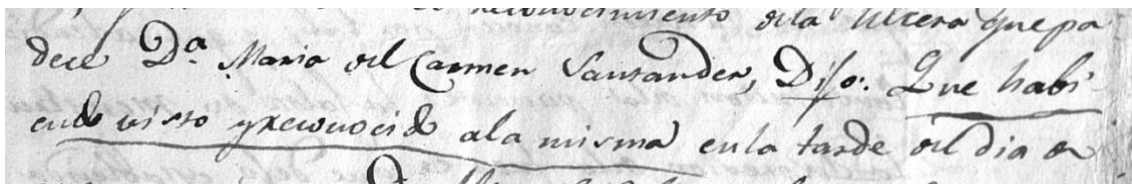
```
le<del rend="overwritten">p</del><add place="inline">p</del>anttarlo
```

4.1.5 Texto resaltado

a. El elemento `<hi>` se utiliza para marcar un segmento de texto que aparece destacado de algún modo. Este elemento incluye siempre un atributo `@rend` que indica cómo se presenta el resaltado en el texto original. En ODE, el atributo `@rend` puede tomar uno de dos valores posibles:

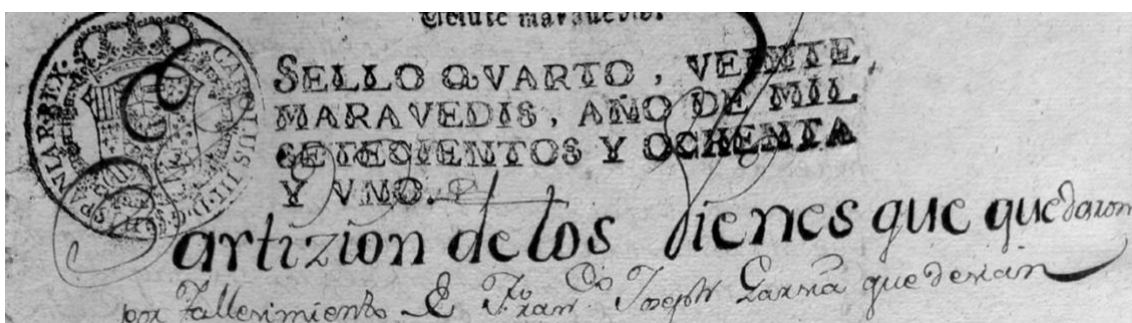
- "underlined": texto subrayado
- "bold": texto en negrita

b. El valor "underlined" se usa para marcar segmentos de texto subrayados: por ejemplo, el segmento *Dijo que habiendo visto y reconocido a la misma* que se observa a continuación:



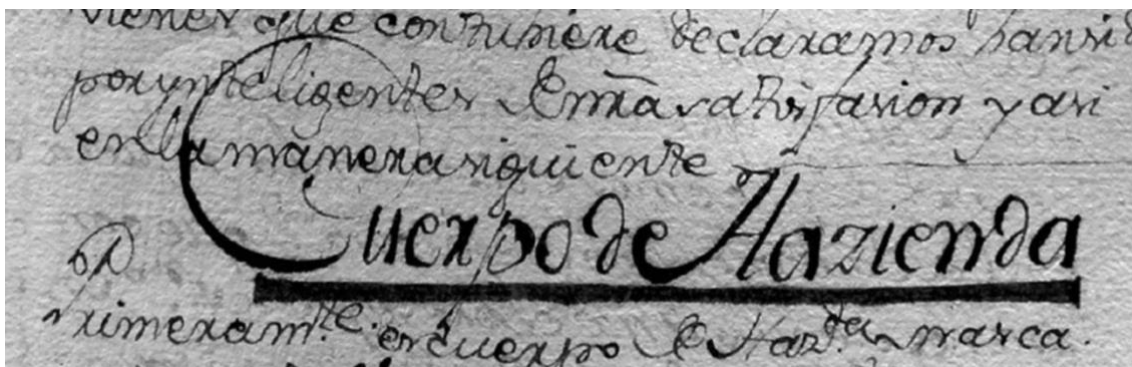
```
<hi rend="underlined">Dijo que habi<lb n="11" break="no"/>endo visto  
y reconocido a la misma</hi>
```

c. El valor "bold" se usa para marcar segmentos de texto que presentan tinta más intensa: por ejemplo, el segmento *Partizion de los Vienes que quedaron* que se observa a continuación:



```
<hi rend="bold">Partizion de los Vienes que quedaron</hi>
```

d. Los valores "underlined" y "bold" pueden aparecer simultáneamente dentro del atributo **@rend**. En este caso, los valores deben ir separados por espacio. Es lo que sucede con el segmento *Cuerpo de Hazienda* que se observa a continuación:



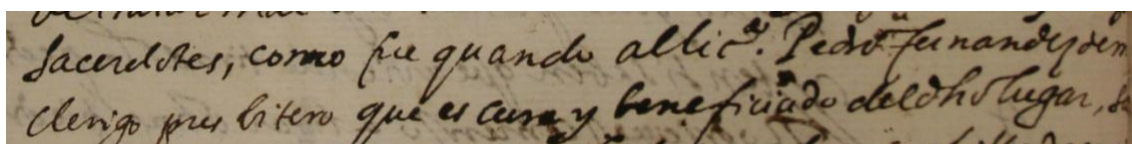
```
<hi rend="bold underlined">Cuerpo de Hazienda</hi>
```

4.1.6 Texto omitido

a. El elemento `<gap>` se utiliza para marcar un segmento de texto que ha sido omitido en la transcripción, bien por el editor —esto es, por quien transcribe el texto en lenguaje XML—, bien por el propio escribano. Este elemento, que por definición es siempre un elemento vacío, incluye un atributo `@reason` que indica la razón de la omisión. En ODE, el atributo `@reason` puede tomar uno de tres valores posibles:

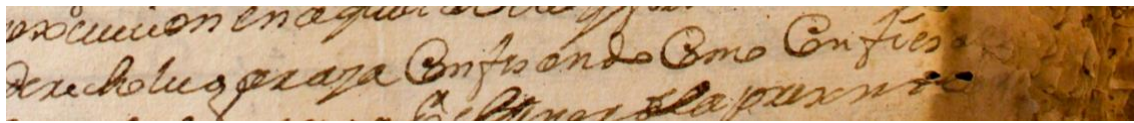
- "illegible": texto ilegible, que el editor no puede transcribir
- "editorial": texto irrelevante, que el editor decide no transcribir
- "omitted": texto omitido por el propio escribano en el texto fuente

b. El valor "illegible" abarca diversas causas que impiden la correcta lectura del texto, tales como daños en el manuscrito, grafías difíciles de interpretar o fragmentos inaccesibles debido a la encuadernación, entre otras. Por ejemplo, el segmento de texto posterior a *Pedro fernandez de* es ilegible debido a la costura del documento, como se observa a continuación:



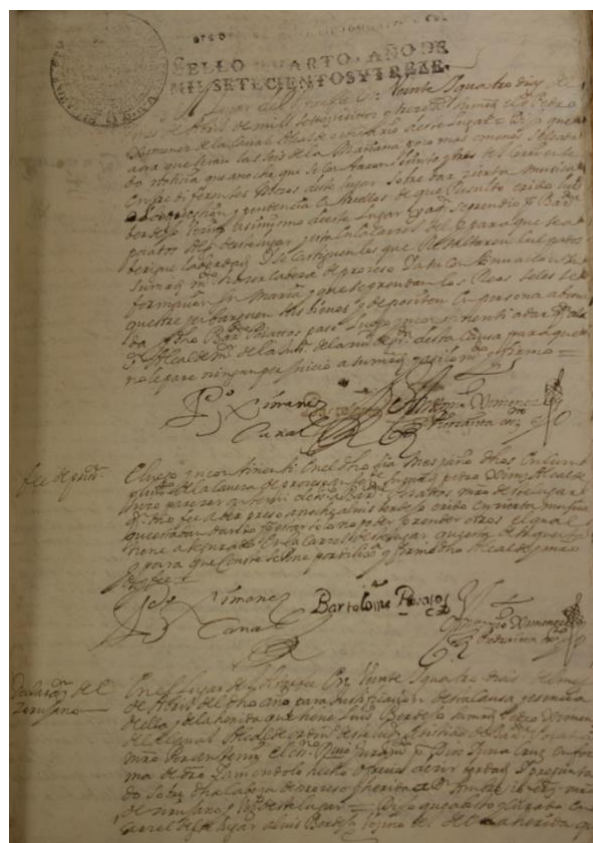
```
Pedro fernandez de <gap reason="illegible"/>
```

De modo análogo, el segmento de texto posterior a *confesando como confiesa* es ilegible debido al deterioro del papel, como se observa a continuación:



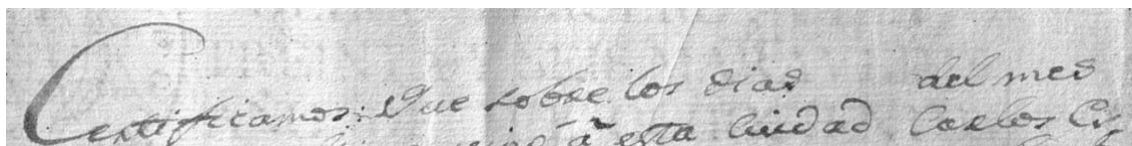
confesando como confiesa <gap reason="illegible"/>

c. El valor "editorial" se utiliza para omitir fragmentos de texto que el editor decide excluir de la transcripción, bien por tratarse de contenido anterior o posterior al discurso de interés para el corpus, bien por ser repetitivo en relación con lo ya transcrito. Por ejemplo, en el último párrafo del siguiente documento comienza un certificado médico, que constituye el verdadero objeto de interés, de modo que en la transcripción de esa página se omite todo el contenido previo al inicio del certificado:



<gap reason="editorial"/>
<add place="left">Declarazon del zirujano</add> En el lugar del Atarfee,
en veinte y quatro dias del mes de abril del dho año,

d. El valor "omitted" se utiliza para marcar lagunas textuales producidas por el propio escribano en el texto fuente: por ejemplo, la que se observa entre *días* y *del* en el segmento *los días del mes*, a continuación:



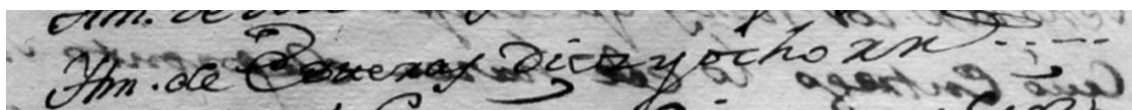
Certificamos que sobre los días `<gap reason="omitted"/>` del mes

4.1.7 Texto conjeturado

a. El elemento `<supplied>` se utiliza para marcar un segmento de texto conjeturado por el editor, cuando el texto original presenta dificultades que impiden una lectura clara. Este elemento incluye un atributo `@cert` que indica el grado de certeza del editor con respecto a la interpretación del texto. En ODE, el atributo `@cert` puede tomar uno de dos valores posibles:

- "high": alto grado de certeza
- "low": bajo grado de certeza

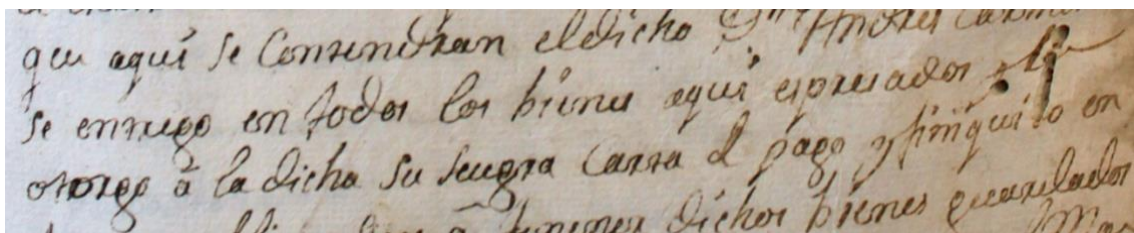
b. Como ejemplo de uso del valor "high" valga el siguiente ejemplo, en donde la tinta del reverso del folio dificulta una lectura clara del texto; no obstante, el editor conjetura, con un alto grado de certeza, que en el manuscrito aparece la palabra *extteras*:



Ytm de `<supplied cert="high">`extteras`</supplied>` diez y ocho rr

c. Como ejemplo de uso del valor "low" valga el siguiente caso. Se trata de un manuscrito parcialmente dañado que impide leer la palabra posterior a la expresión *los bienes aquí espresados* y. El editor conjetura que dicha palabra podría ser el pronombre clítico *le*, aunque también cabe la posibilidad de que sea el clítico *la*. Dado que esta decisión es lingüísticamente relevante —podría tratarse de un caso de leísmo con referente femenino o de laísmo—, el transcriptor conjetura la opción que parece más plausible, pero marca la esa

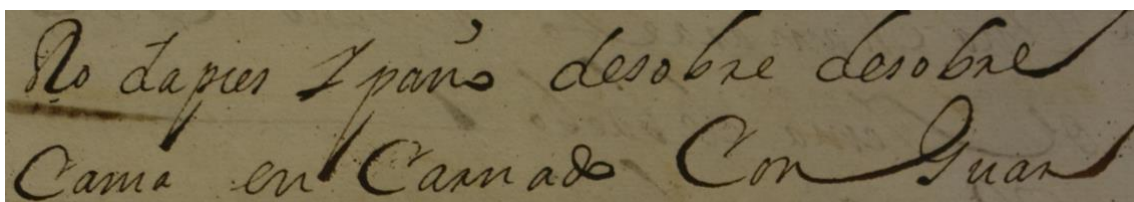
conjetura con un bajo grado de certeza, para que el usuario del corpus la considere con las debidas cautelas:



se entrego en todos los bienes aqui espresados y <sup><supplied
cert="low">le</supplied></sup> **<lb n="62" break="yes"/>** otorgo a la dicha
suegra carta de pago

4.1.8 Texto superfluo

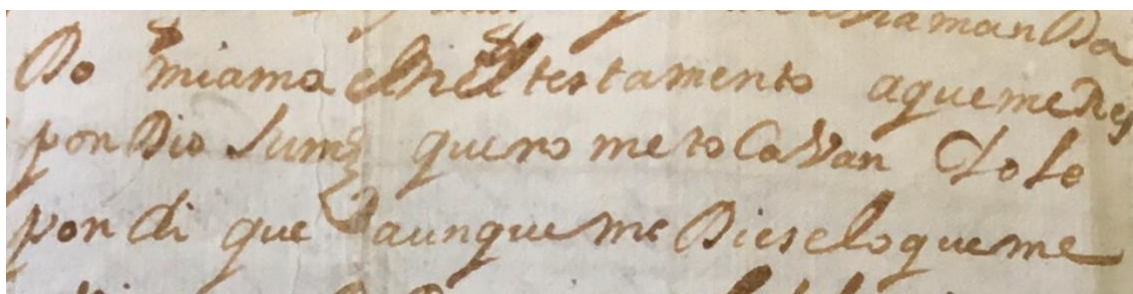
El elemento [<surplus>](#) se utiliza para marcar un segmento que el editor considera superfluo o redundante: por ejemplo, el segmento *de sobre* que se observa a continuación:



Rodapiés y paño ^{<surplus>de sobre</surplus>} de sobre **<lb n="150" break="no"/>** cama encarnado

4.1.9 Texto inexacto

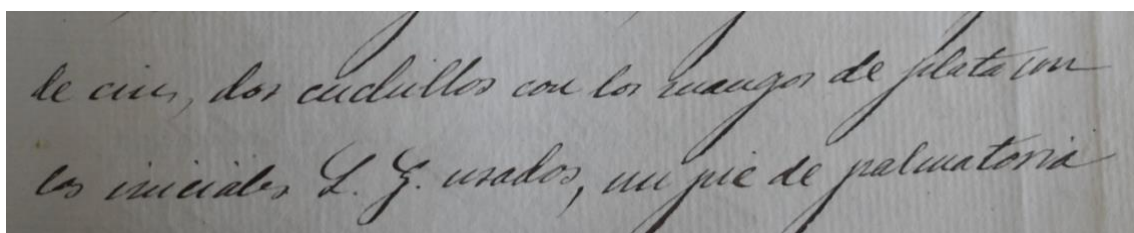
El elemento [<sic>](#) se utiliza para marcar un segmento que el editor considera incorrecto o inexacto: por ejemplo, el segmento *pondí* (en lugar de *respondí*) que se observa a continuación:



a que me Res<lb n="66" break="no"/>pondio su mz que no me tocavan yo
le <lb n="67" break="yes"/> <sic>pondi</sic> que aunque me Diese lo que me

4.1.10 Texto citado

El elemento [<q>](#) se utiliza para marcar texto que se cita de cualquier otra fuente escrita, como sucede con la expresión L. G. en el siguiente ejemplo:



dos cuchillos con los mangos de plata <lb n="63" break="yes"/> las
iniciales <q>L. G.</q> usados

4.1.11 Texto en otra lengua

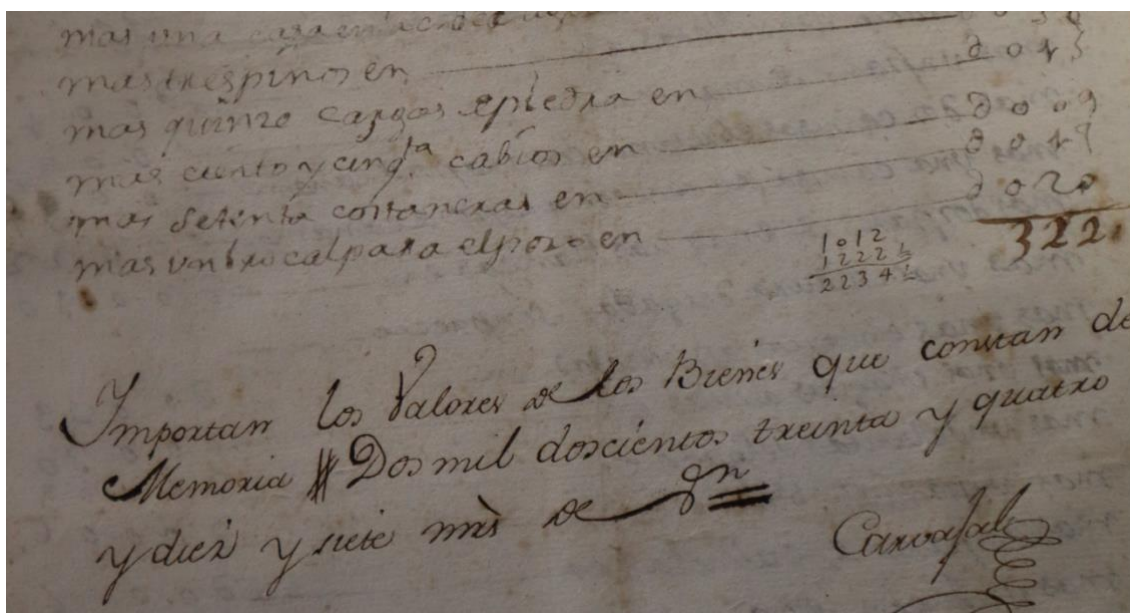
El elemento [<foreign>](#) se utiliza para marcar un segmento de texto en un idioma diferente al resto del documento, es decir, que no está en español. Este elemento incluye el atributo [@xml:lang](#), que especifica la lengua en la que está escrito dicho segmento. En ODE, los casos marcados con [<foreign>](#) corresponden generalmente a expresiones en latín:



cada vno de por si y por el todo <lb n="66" break="yes"/> <foreign
xml:lang="lang">yn solidum</foreign>, Renunziando como expresante

4.1.12 Texto de nueva mano

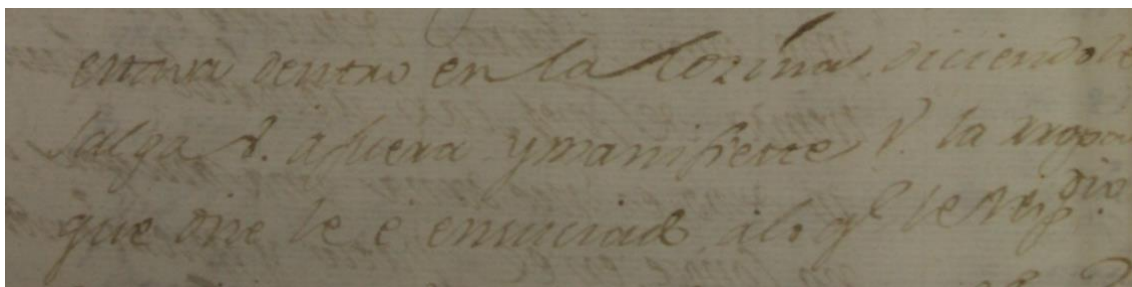
El elemento [<handShift>](#) se utiliza para marcar el inicio de un segmento de texto atribuible a otra mano. Por ejemplo, en el siguiente documento —un inventario de bienes—, la sección final incluye un fragmento adicional escrito por otra mano: el que comienza con *ymportan los Valores de los bienes*.



```
mas setenta costaneras en d019 <lb n="56" break="yes"/> mas un brocal  
para el pozo en d020 <lb n="57" break="yes"/> <handShift/> ymportan  
los Valores de los Bienes que constan
```

4.1.13 Discurso en estilo directo

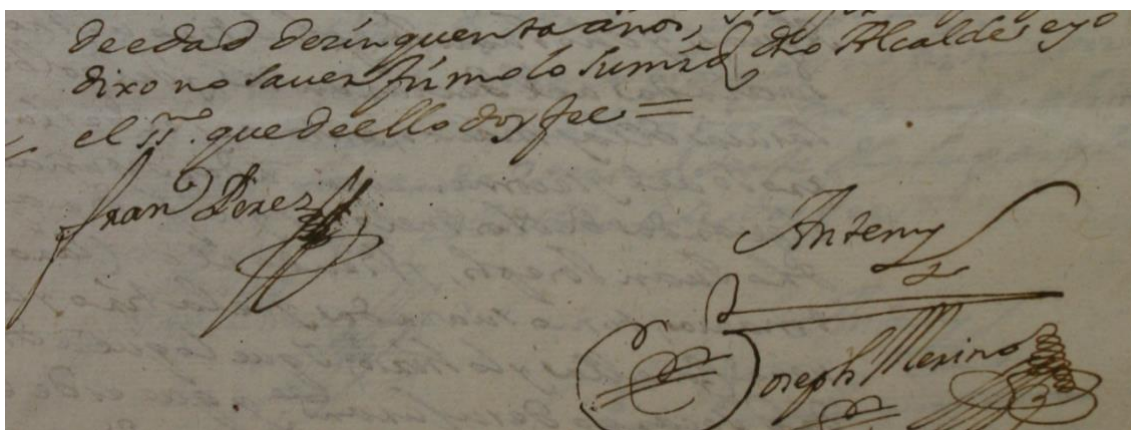
El elemento [<quote>](#) se utiliza para marcar discurso en estilo directo, particularmente, la reproducción exacta de expresiones puestas en boca de declarantes en el transcurso de un juicio. Este tipo de discurso, difícil de documentar en textos históricos y registrado de manera esporádica en las declaraciones de testigos, constituye el subtipo textual de ODE más próximo a la oralidad, lo que le otorga una relevancia especial en el corpus. Tómese como ejemplo la expresión que comienza *salga V afuera* en el siguiente fragmento:



estaba dentro en la cocina diciendole <lb n="19" break="yes"/>
<q>salga V afuera y manifieste V la ropa <lb n="20" break="yes"/> que
dize le e ensuciado</q>

4.1.14 Firma

Las firmas se marcan con el elemento `<seg>`, que va siempre acompañado del atributo `@type` con el valor "signature"; por ejemplo, los segmentos *Franco Perez* y *Joseph Merino* que se observan a continuación:



<seg type="signature">Franco Perez</seg> ante my <lb n="82"
break="yes"/> <seg type="signature">Joseph Merino</seg>

Dado que no presentan interés lingüístico, en ODE no se marcan las rúbricas, garabatos, figuras u otros símbolos sin contenido lingüístico que puedan acompañar a las firmas o aparecer en otras partes del manuscrito.

4.2 El corpus lingüístico

Como se señaló en la introducción, ODE persigue un doble objetivo: por un lado, ofrecer una edición paleográfica digital de cada manuscrito; por otro lado, convertir el contenido textual de estos manuscritos en un corpus anotado. Este apartado se centra en el segundo de estos objetivos. Concretamente, se detalla la estrategia de marcación implementada en ODE para enriquecer las transcripciones paleográficas, codificadas en TEI P5, con el análisis lingüístico generado en la plataforma TEITOK. El resultado se almacena en los mismos archivos electrónicos, convertido también en lenguaje TEI P5. Ese análisis lingüístico abarca los siguientes aspectos:

- Tokenización
- Normalización ortográfica
- Lematización y etiquetado morfosintáctico
- Anotación lingüística adicional

4.2.1 Tokenización

a. La tokenización se refiere al proceso de segmentar un texto para identificar y delimitar las unidades del corpus que van a ser anotadas —los tokens—, habitualmente palabras, números y signos de puntuación. En ODE, este procedimiento se realiza automáticamente mediante el script *xmltokenize*, desarrollado por Maarten Janssen y ejecutado desde la plataforma TEITOK. Por esta razón, los espacios en blanco, que en lenguaje XML suelen ser irrelevantes, adquieren un papel fundamental en ODE, ya que delimitan cada palabra ortográfica y garantizan la eficacia del script. La conversión del resultado de la tokenización a TEI P5 implica la consideración de dos elementos principales:

- `<w>` (Palabra). Representa una palabra gramatical (no necesariamente ortográfica).
- `<pc>` (Puntuación). Contiene un signo de puntuación.

b. El contenido textual del elemento `<w>` contiene la transcripción de la palabra, junto con cualquier tipo de marcación aplicada a esa palabra durante la creación de la edición paleográfica digital. Así, el contenido de `<w>` puede consistir en caracteres alfanuméricos únicamente (por ejemplo, *vezino*) o en una combinación de estos con elementos de marcación en TEI P5, esto es, posibles

adiciones o cancelaciones de grafías dentro de la palabra, cambios de línea en mitad de palabra, desarrollos de abreviaturas, etcétera:

```
<w>vezino</w>
```

```
<w>vez<lb n="32" break="no"/>ino</w>
```

c. Además de su contenido textual, el elemento [<w>](#) incorpora diversos atributos, cada uno asociado a un nivel específico de edición o anotación efectuada dentro del corpus. Al ser la palabra —es decir, el token— la unidad básica del análisis, es en [<w>](#) donde se almacenan estas capas de información. Por ejemplo, el atributo [@norm](#) (*normalized*) contiene la forma de la palabra con grafía normalizada, mientras que [@pos](#) y [@msd](#) capturan su categoría sintáctica (*part of speech*) y sus rasgos morfosintácticos (*morphosyntactic description*). En total, se contemplan cinco atributos diferentes dentro de [<w>](#):

- [@orig](#). Contiene la palabra original, esto es, tal como aparece en el manuscrito.
- [@norm](#). Contiene la palabra con ortografía normalizada.
- [@lemma](#). Contiene el lema de la palabra.
- [@pos](#). Contiene la clase gramatical de palabra.
- [@msd](#). Contiene la descripción morfosintáctica de la palabra.

Los dos primeros atributos —[@orig](#) y [@norm](#)— representan distintos enfoques de edición textual: una edición conservadora que preserva la grafía original del manuscrito frente a una edición con grafía normalizada al español estándar contemporáneo. Los tres últimos atributos —[@lemma](#), [@pos](#) y [@msd](#)— se relacionan con diferentes niveles de anotación gramatical. No obstante, la anotación lingüística presente en el corpus ODE no se restringe a estos tres atributos, sino que incluye otros tipos de información también a nivel de palabra (ver [sección 4.2.4](#)).

d. El elemento [<pc>](#) contiene los signos de puntuación presentes en el texto. En el corpus ODE, los signos de puntuación no se normalizan, por lo que este elemento puede contener hasta cuatro atributos: [@orig](#), [@lemma](#), [@pos](#) y [@msd](#), es decir, todos los que pueden aparecer en [<w>](#), excepto el correspondiente a la normalización ([@norm](#)).

e. El atributo [@orig](#) contiene la forma original de la palabra, desprovista de cualquier marca en lenguaje XML. Si el contenido textual de [<w>](#) se limita a caracteres alfanuméricos, su valor coincidirá con el de [@orig](#):

```
<w orig="vezino">vezino</w>
```

Si, por el contrario, el contenido textual de [<w>](#) incluye código XML relacionado con la marcación de aspectos de la edición paleográfica digital, el valor de [@orig](#) contendrá únicamente los caracteres alfanuméricos correspondientes:

```
<w orig="vezino">vez<lb n="32" break="no"/>ino</w>
```

f. En el caso de las abreviaturas, el contenido textual del elemento [<w>](#) sigue la marcación de TEI P5 especificada para la expansión de abreviaturas (ver [sección 4.1.2](#)), mientras que el valor del atributo [@orig](#) contiene la forma original abreviada, tal como aparece en el manuscrito:

```
<w orig="vo">
  <choice>
    <abbr>vo</abbr>
    <expansion>vecino</expansion>
  </choice>
</w>
```

g. De lo expuesto anteriormente en este apartado se deduce que, aunque cada archivo XML contiene mucha información de diferente naturaleza, los valores del atributo [@orig](#) permiten reconstruir fielmente el texto original del manuscrito, respetando sus peculiaridades gráficas; por otro lado, el contenido textual del elemento [<w>](#) permite reconstruir fielmente la edición paleográfica digital, incluyendo todo su aparato de marcación en TEI P5. Esta estrategia garantiza la preservación de la autenticidad del manuscrito y logra una integración coherente, aunque diferenciada, entre el contenido textual, la edición digital y el corpus anotado.

4.2.2 Normalización ortográfica

La normalización consiste en asignar a cada palabra del corpus su forma ortográfica correspondiente según el estándar actual. En ODE, este procedimiento se realiza automáticamente mediante el script *nformtreat.pl*, desarrollado por Maarten Janssen y ejecutado desde la plataforma TEITOK. El

resultado generado por el script es revisado manualmente por el equipo de lingüistas de ODE:

```
<w orig="vezino" norm="vecino">vezino</w>
```

```
<w orig="abia" norm="había">abia</w>
```

```
<w orig="vzo" norm="vecino">
  <choice>
    <abbr>vzo</abbr>
    <expn>vezino</expn>
  </choice>
</w>
```

4.2.3 Lematización y etiquetado morfosintáctico

a. La lematización consiste en la asignación del lema correspondiente a cada token, esto es, la forma que representa al conjunto de variantes morfológicas de una palabra y encabeza la entrada de un diccionario. El etiquetado morfosintáctico consiste en la asignación de una etiqueta con información gramatical a cada token. En ODE, este doble procedimiento se realiza automáticamente mediante el etiquetador morfosintáctico *Neotag*, desarrollado por Maarten Janssen y ejecutado desde la plataforma TEITOK. El resultado generado por *Neotag* es revisado manualmente por el equipo de lingüistas de ODE.

b. El conjunto de etiquetas usado para el etiquetado del corpus ODE ha sido diseñado siguiendo las directrices del proyecto [EAGLES](#) (*Expert Advisory Group on Linguistic Engineering Standards*), uno de los estándares más reconocidos para la anotación lingüística de corpus y usado, por ejemplo, por el anotador [FreeLing](#) o por algunos de los corpus en español contenidos en la herramienta [Sketch Engine](#). Este conjunto de etiquetas sigue una estructura posicional: cada etiqueta está compuesta por una secuencia de símbolos —letras y números—, donde cada símbolo representa un rasgo morfosintáctico específico según su posición en la secuencia. El significado de cada posición se define en función de la categoría principal, indicada por la primera letra de la secuencia y que representa la clase de palabra.

c. Por ejemplo, la forma *vezino* lleva la etiqueta "NCMS000", donde la "N" indica que se trata de un nombre; la "C", que es común; la "M", que es masculino, y la

S, que es singular. De modo análogo, la forma *abia*, se etiqueta VAI13S0, donde la "V" indica que se trata de un verbo; la "A", que es auxiliar; la "I" de la tercera posición, que es indicativo; la "I" de la cuarta posición, que es tiempo imperfecto; el 3, que es tercera persona, y la S, que es singular:

```
<w
  orig="vezino"
  norm="vecino"
  lemma="vecino"
  pos="N"
  msd="NCMS000">vezino</w>
```

```
<w
  orig="abia"
  norm="había"
  lemma="haber"
  pos="V"
  msd="VAI13S0">abia</w>
```

d. Los ceros (0) se utilizan como indicadores de ausencia de información específica para un rasgo particular en una posición determinada. Esto significa que el rasgo correspondiente no está especificado —el rasgo "persona" en una forma verbal de gerundio, por ejemplo— o no aplica en el contexto del análisis morfosintáctico de la palabra etiquetada para la lengua en cuestión —el rasgo "caso" en los sustantivos del español, por ejemplo—. La etiqueta principal para los signos de puntuación es "F", mientras que para los números es "Z". Para una descripción completa de las etiquetas utilizadas en ODE, véase el [etiquetario de ODE](#).

```
<pc
  orig="."
  lemma="."
  pos="F"
  msd="Fp">.</pc>
```

```
<w
  orig="5"
  lemma="5"
  pos="Z"
  msd="Z">5</w>
```

e. Las contracciones y las formas verbales con pronombres enclíticos se consideran tokens complejos, es decir, palabras ortográficas que representan dos o más palabras gramaticales. El etiquetado morfosintáctico de los tokens complejos requiere descomponer la palabra ortográfica en sus unidades

gramaticales constitutivas y asignar a cada una la información lingüística correspondiente. Para representar estos casos en TEI P5, se emplea el elemento `<w>` de manera recursiva. Cada uno de los elementos `<w>` (es decir, palabras gramaticales) anidados dentro de otro `<w>` (palabra ortográfica) lleva un atributo `@part`, que indica el orden de cada palabra gramatical dentro de la secuencia ortográfica: el valor "I" corresponde a la posición inicial, "M" a la posición intermedia y "F" a la posición final:

```
<w orig="del">del
  <w
    part="I"
    orig="de"
    lemma="de"
    pos="S"
    msd="SPS00"/>
  <w
    part="F"
    orig="el"
    lemma="el"
    pos="D"
    msd="DA0MS0"/>
</w>
```

```
<w orig="habiéndoselo">habiéndoselo
  <w
    part="I"
    orig="habiendo"
    lemma="haber"
    pos="V"
    msd="VMN0000"/>
  <w
    part="M"
    orig="se"
    lemma="se"
    pos="P"
    msd="PP3CND00"/>
  <w
    part="F"
    orig="lo"
    lemma="lo"
    pos="P"
    msd="PP3MSA00"/>
</w>
```

f. Si un token complejo requiere normalización ortográfica, esta se aplica tanto al nivel de la palabra ortográfica como al de la palabra gramatical, cuando sea necesario:

```
<w orig="habiéndoselo" norm="habiéndoselo">aviendoselo
  <w
    part="I"
    orig="aviendo"
    norm="habiendo"
    lemma="haber"
    pos="V"
    msd="VMN0000"/>
  <w
    part="M"
    orig="se"
    lemma="se"
    pos="P"
    msd="PP3CND00"/>
  <w
    part="F"
    orig="lo"
    lemma="lo"
    pos="P"
    msd="PP3MSA00"/>
</w>
```

4.2.4 Anotación lingüística adicional

a. El corpus ODE no solo incluye anotación lingüística de índole gramatical —lematización y etiquetado morfosintáctico—, sino que también incorpora otros tipos de anotación lingüística adicional. Actualmente, el corpus está siendo enriquecido con anotaciones de tipo gráfico-fonético, etimológico y semántico. Esta anotación se realiza de forma automática mediante el script *ltags.pl*, desarrollado por Gael Vaamonde y ejecutado desde la plataforma TEITOK. Téngase en cuenta que estas capas de información lingüística todavía se encuentran en fase de revisión, por lo que los resultados deben interpretarse con cautela. Es posible que aún contengan inconsistencias o errores que requieren un análisis más exhaustivo.

b. Todos los tipos de anotaciones adicionales se representan mediante estructuras de rasgos, esto es, utilizando el elemento `<fs>` (*feature structure*). Este elemento incluye siempre un atributo `@type` con el valor "annotation" y presenta una organización interna que consiste en un subelemento `<f>` con un atributo `@name` y, dentro de `<f>`, un subelemento `<symbol>` con el atributo `@value`. En el atributo `@name` se indica el tipo de anotación (por ejemplo, "etymology"), mientras que en el atributo `@value` se indica el valor correspondiente (por ejemplo, "arabism"):

```
<fs type="annotation">
  <f name="etymology">
    <symbol value="arabism"/>
  </f>
</fs>
```

c. Dado que la anotación lingüística adicional implementada en ODE se aplica exclusivamente a nivel de palabra, la estructura de rasgos `<fs>` se incorpora necesariamente dentro del elemento `<w>`. Por ejemplo, la forma *almuadas* se anota de la siguiente manera:

```
<w orig="almuadas"
  norm="almohadas"
  lemma="almohada"
  pos="N"
  msd="NCFS000">almuadas
  <fs type="annotation">
    <f name="etymology">
      <symbol value="arabism"/>
    </f>
  </fs>
</w>
```

De forma análoga, pero ilustrando otros tipos de anotación, las formas *aborresido* y *alcuza* se anotan de la siguiente manera:

```
<w orig="aborresido"
  norm="aborrecido"
  lemma="aborrecer"
  pos="V"
  msd="VMP00SM">aborresido
  <fs type="annotation">
    <f name="phonetics">
      <symbol value="seseo"/>
    </f>
  </fs>
</w>
```

```
<w orig="alcuza"
  lemma="alcuza"
  pos="N"
  msd="NCFS000">alcuza
  <fs type="annotation">
    <f name="semantics">
      <symbol value="cookware"/>
    </f>
  </fs>
</w>
```

d. Una misma palabra puede ser objeto de varias anotaciones. Si la palabra está asociada a dos (o más) valores de un mismo tipo de anotación, estos se separan por guion dentro del mismo atributo @value. Es el caso de la forma *porsolana*, por ejemplo, que se anota del modo siguiente:

```
<w orig="porsolana"
  norm="porcelana"
  lemma="porcelana"
  pos="N"
  msd="NCFS000">porcelana
  <fs type="annotation">
    <f name="phonetics">
      <symbol value="seseo-vowel_system"/>
    </f>
  </fs>
</w>
```

En cambio, si la palabra está asociada a dos (o más) valores referidos a diferentes tipos de anotación, se declaran tantos elementos <f> dentro de <fs> como sean necesarios para dar cuenta de todos los valores asignados a esa palabra. Es el caso de la forma *alcusa*, que se anota del modo siguiente:

```
<w orig="alcusa"
  norm="alcuza"
  lemma="alcuza"
  pos="N"
  msd="NCFS000">alcuza
  <fs type="annotation">
    <f name="phonetics">
      <symbol value="seseo"/>
    </f>
    <f name="semantics">
      <symbol value="cookware"/>
    </f>
  </fs>
</w>
```

5. TEITOK VS. TEI P5

Como se ha explicado en la introducción, los archivos XML de ODE se pueden descargar en dos formatos, correspondientes a dos modelos distintos: el modelo TEITOK, de uso interno en la plataforma, y el formato TEI P5, que sigue estrictamente las directrices de la última versión del estándar TEI y en el que se centra esta guía. A continuación, se presenta unas tablas de correspondencias entre los elementos y atributos de ambos modelos, que permite al usuario identificar las diferencias principales y comprender cómo el formato interno de ODE se adapta a una estructura de marcado conforme a las recomendaciones de la TEI (ver [Tabla 7](#) y [Tabla 8](#)).

Tabla 7. Correspondencias en la cabecera entre el modelo TEITOK y el modelo TEI P5.

TEITOK XML	TEI P5 XML
@id	@xml:id
@lang	@xml:lang
<pre><respStmt> <resp id="transcription"> Gael Vaamonde </resp> </respStmt></pre>	<pre><respStmt n="1"> <resp>Transcripción</resp> <name>Gael Vaamonde</name> </respStmt></pre>
<pre><idno ref="http://..."> ARCHV PCR 2114/0004 </idno></pre>	<pre><idno source="http://..."> ARCHV PCR 2114/0004 </idno></pre>
<pre><msItem class="original"> <p> <locus>422r-444v</locus> </p> </msItem></pre>	<pre><msItem> <locus>422r-444v</locus> <filiation type="original"> </msItem></pre>
<pre><msItem class="copy" when="1706"> <p> <locus>331r-333r</locus> </p> </msItem></pre>	<pre><msItem> <locus>331r-333r</locus> <filiation type="copy"> <origDate>1690</origDate> <date>1706</date> </filiation> </msItem></pre>
<pre><locus>7r-7v</locus></pre>	<pre><locus cert="high">7r-7v</locus></pre>
<pre><locus>[7r-7v]</locus></pre>	<pre><locus cert="low">7r-7v</locus></pre>

<pre><surrogates> <cit> <bibl> facsímil digital guardado en formato JPG </bibl> </cit> </surrogates></pre>	<pre><surrogates> <bibl> <title type="gmd"> facsímil digital guardado en formato JPG </title> </bibl> </surrogates></pre>
<pre><encodingDesc> <classDecl></classDecl> </encodingDesc></pre>	<pre><encodingDesc> <projectDesc></projectDesc> <editorialDecl></editorialDecl> <classDecl></classDecl> </encodingDesc></pre>
<pre><name type="place" subtype="Almería" geo="37.170972 2.839710"> España, Almería, Fiñana </name></pre>	<pre><placeName>España, Almería, Fiñana <name n="1" type="country"> España </name> <name n="2" type="province"> Almería </name> <name n="3" type="city"> Fiñana </name> <geo>37.1709721 -2.8397107</geo> </placeName></pre>
<pre><date when="1789" when-custom="XVIII"> 1789 </date></pre>	<pre><date n="1" type="year">1832</date> <date n="2" type="century">XIX</date></pre>
<pre><catRef target="inv"/></pre>	<pre><catRef target="#inv"/></pre>

Tabla 8. Correspondencias en el texto entre el modelo TEITOK y el modelo TEI P5.

TEITOK XML	TEI P5 XML
<pre><foreign lang="la"></pre>	<pre><foreign xml:lang="la"></pre>
<pre><pb n="[3r]" /></pre>	<pre><pb n="3r" cert="high" /></pre>
<pre>tres <lb id="e-1" /> reales</pre>	<pre>tres <lb n="1" break="yes" /> reales</pre>
<pre>tres re<lb id="e-2" />ales</pre>	<pre>tres re<lb n="2" break="no" />ales</pre>
<pre><tok></pre>	<pre><w> <!-- palabras --> <pc> <!-- signos de puntuación --></pre>
<pre><tok>vecino</tok></pre>	<pre><w orig="vecino">vecino</w></pre>

<pre><tok form="vecino"> ve<lb id="e-2"/>cino </tok></pre>	<pre><w orig="vecino"> ve<lb n="2" break="no"/>cino </w></pre>
<pre><tok fform="vecino"> vo </tok></pre>	<pre><w orig="vo"> <choice> <abbr>vo</abbr> <expan>vecino</expan> </choice> </w></pre>
<pre><tok nform="vecino"> vezino </tok></pre>	<pre><w orig="vezino" norm="vecino"> vezino </w></pre>
<pre><tok lemma="vecino"> vecino </tok></pre>	<pre><w lemma="vecino"> vecino </w></pre>
<pre><tok pos="NCMS000"> vecino </tok></pre>	<pre><w pos="N" msd="NCMS000"> vecino </w></pre>
<pre><tok>del <dtok form="de" lemma="de" pos="SPS00"/> <dtok form="el" lemma="el" pos="DA0MS0"/> </tok></pre>	<pre><w orig="del">del <w part="I" orig="de" lemma="de" pos="S"/> msd="SPS00"/> <w part="F" orig="el" lemma="el" pos="D" msd="DA0MS0"/> </w></pre>
<pre><tok ltags="arabism"> almohada </tok></pre>	<pre><w>almohada <fs type="annotation"> <f name="etymology"> <symbol value="arabism"/> </f> </fs> </w></pre>