

COSUIZA EN TEITOK

Andrea Escobar Castillo

Última actualización: 2021-12-16

Contents

| | |
|--|-----------|
| Informaciones generales | 5 |
| Introducción | 7 |
| La codificación de textos | 8 |
| COSUIZA | 13 |
| Biblioteca de Ginebra | 13 |
| Biblioteca cantonal y universitaria de Lausana | 13 |
| CHARTA | 15 |
| TEITOK | 17 |
| Edición de textos en COSUIZA/Teitok | 19 |
| Guía de etiquetado | 19 |
| Automatización en la plataforma | 39 |
| Videotutorial | 61 |

Informaciones generales

¿A quién está dirigido este documento?

A todos los colaboradores del corpus COSUIZA que vayan a realizar ediciones digitales en nuestra página web.

Objetivo

Este documento tiene por objetivo servir como guía de trabajo en todas las etapas de creación de ediciones digitales con la herramienta Teitok en el servidor del COSUIZA.

Requisitos

Para hacer uso de este tutorial no se necesita tener ningún conocimiento previo, pero es necesario tener:

1. Una cuenta de usuario o de administrador en la página del COSUIZA.
2. Un editor de código fuente. Recomendamos Visual Studio Code por su versatilidad y por ser una herramienta gratuita compatible con Mac, Linux y Microsoft.

¿Cómo está estructurado este tutorial?

El primer capítulo se ofrece une breve introducción a la codificación de textos, su importancia, fundamentos y herramientas. Este apartado resulta imprescindible para el principiante en estas cuestiones.

El segundo capítulo presenta el corpus documental que motiva el proyecto en el cual se enmarca este documento.

El tercer capitulo expone muy brevemente el proyecto y el macrocorpus del cual hace parte el COSUIZA.

En el cuarto capítulo se presenta la herramienta que nos permite crear, anotar y publicar nuestro corpus documental.

El quinto capítulo es un tutorial que enseña cómo proceder en todas las etapas de la edición de textos en el COSUIZA.

Contacto: andrea.escobarcastillo@unil.ch

Este documento está en construcción

Introducción

En la era digital, los datos textuales se producen en cantidades incommensurables, sean estos de origen digital o no. Diversas disciplinas fundamentan sus investigaciones en datos textuales. En este sentido, las humanidades no han experimentado un cambio, puesto que desde siempre, los datos textuales, o simplemente, el texto, han sido la materia prima de los estudiosos de esta disciplina. Sin embargo, pese a conservar su materia de estudio intacta, la llegada de las computadoras presentó, inevitablemente, desafíos y posibilidades en cuanto a la colecta, constitución y análisis de datos, o más bien dicho de los corpus de estudio. Un corpus lingüístico, que es el tipo de corpus del que nos ocuparemos en este documento, requiere del trabajo de filólogos en la colecta, transcripción y edición de datos y del uso de ciertas herramientas informáticas destinadas a la presentación de los documentos editados.

En la imagen debajo vemos la reproducción de un documento que hace parte del corpus COSUIZA:

Fonds Carlos de Goyeneche y Silvela, marquis de Balbueno, IS 5318/1/3/23

¿Cómo podemos hacer de este documento una fuente de datos susceptibles de almacenarse, transportarse y hacer parte de un corpus en el cual puedan realizarse búsquedas específicas?, pero además, ¿cómo podemos transcribir y presentar fielmente su estructura? Más allá de la etapa evidente de la captura de la imagen, nos debemos plantear la cuestión de cómo extraer el texto que contiene el facsímil. Esto lo hacemos transcribiendo, pero para ello debemos determinar las herramientas que nos van a permitir producir esta transcripción. Existen programas, que muy frecuentemente exigen una suscripción para usarlos, y permiten transcribir documentos manuscritos de manera automática, por ejemplo, Textract de la compañía Amazon. Pero esta etapa constituye solamente una parte del proceso de preparación de un corpus lingüístico digital. En la actualidad, las transcripciones de textos manuscritos se deben realizar teniendo en consideración que estas serán publicadas empleando herramientas informáticas, lo que supone, desde luego, que estén codificadas de modo que puedan ser legibles por las computadoras.

La codificación de textos

En rigor, el texto que leemos en este documento está codificado, de otro modo el computador, que solamente “habla” binario, no podría presentar los caracteres de nuestra lengua. Para la computadora, la palabra *español* equivale a la secuencia \x65\x73\x70\x61\xc3\xb1\x6f\x6c en el sistema de codificación UTF-8. Esta codificación es muy cercana a la lengua de la computadora y no es fácilmente legible por los seres humanos (probablemente algunas personas estarán en desacuerdo con esta afirmación). Esta codificación es la que permite obtener texto plano o texto sin formato. Ahora bien, sabemos que las características textuales y estructurales de los documentos provenientes de las humanidades son diversas. Volvamos al documento presentado más arriba: podemos observar columnas, títulos subrayados, signos, etcétera. Una alternativa sería, simplemente, transcribir este documento en un procesador de texto como Word y aplicar algunos estilos tipográficos. Sin embargo, esta opción es restrictiva, no solamente por el hecho de que Word es un programa que exige una licencia, sino que, además, este no asegura la interoperabilidad imprescindible para el intercambio de datos. Pero por sobre todo, los estilos tipográficos no incluyen información semántica. Somos los seres humanos quienes tenemos la capacidad de interpretar el contenido semántico de los estilos tipográficos.

XML

Una solución al problema de la codificación de textos fue propuesta hace bastante tiempo con la creación del metalenguaje de marcado XML, sigla de eXtensible Markup Language. XML es un estándar que permite la estructuración y transmisión de información a través de un sistema de etiquetas que se extienden en una estructura de subordinación. La cantidad de etiquetas que componen un documento es variable y no es forzosamente permanente, característica que explica el adjetivo eXtensible. Las ventajas que se pueden encontrar haciendo uso de XML como medio para estructurar datos son múltiples. Primeramente, XML no es solamente legible por las computadoras, sino también por los seres humanos. Es una herramienta que existe desde hace más de cuarenta años, razón por la cual podemos estar seguros de que no caerá repentinamente en desuso. Es particularmente adecuada para la estructuración de datos textuales que no se adaptan a un modelo tabular.

Un archivo en formato XML se estructura como vemos en el cuadro debajo. Un elemento raíz (root) contiene todos los elementos y subelementos. La cantidad de niveles no tiene restricción más allá de los evidentes problemas de legibilidad que conlleva un documento con excesivos niveles de anidación.

```
<raíz>
  <elemento>
    <subelemento>...Datos...</subelemento>
```

```
</elemento>
</raíz>
```

Un elemento puede contener atributos, cuya utilidad es aportar información descriptiva sobre el elemento en el que aparece, dicho de otro modo, contiene datos sobre los metadatos. El valor de un atributo debe ir siempre entre comillas como vemos en el cuadro a continuación:

```
<elemento id="20">...Datos...</elemento>
```

Estos principios constituyen el fundamento de la TEI (Text encoding initiative) dado que esta es una aplicación del metalenguaje de marcado XML.

TEI

TEI es una sigla que hace referencia al consorcio de la Iniciativa para la codificación de textos que nació a finales de la década de los ochenta con el fin de establecer una metodología para la codificación informática de textos de carácter humanístico. La TEI representa, en la actualidad, el proyecto más longevo y exitoso de las humanidades digitales. El consorcio TEI desarrolló un sistema de etiquetado, basado en XML, que permite agregar al texto plano marcas o anotaciones con información relativa a la estructura y contenido semántico de un texto. Dicho de otro modo, TEI puede entenderse como una colección de etiquetas con relaciones jerárquicas. Esta colección de etiquetas está orientada a representar textos procedentes de las humanidades: obras no literarias y literarias de todo género, manuscritas o impresas. Para cada documento, y en función de sus características textuales, TEI concibe diferentes módulos. Por ejemplo, para la descripción de manuscritos podemos recurrir a la etiqueta `<foliation>` que nos permite describir la numeración de los folios de un manuscrito. Las directrices o, mejor dicho, recomendaciones (puesto que no han sido concebidas como una normativa), pueden encontrarse en la página del consorcio.¹ Supongamos que formamos parte de un proyecto en el cual se ha decidido codificar diversas ediciones de *La Araucana*. Comenzaremos la codificación por la primera plana que vemos a continuación:

Una manera de transcribir esta primera plana sería como vemos en el cuadro debajo (ignoramos todos los elementos tipográficos para los cuales existe una gran cantidad de elementos previstos por la TEI):

```
<front>
  <titlePage>
    <docTitle>
      <titlePart type="main">LA ARAVCAna <choice><orig>ARAVCAna</orig><reg>Araucana</reg></choice>
```

¹Directrices TEI

```

        </titlePart>
    </docTitle>
    <byline>de don<docAuthor>
    <choice><orig>Alonio</orig><reg>Alonso</reg>
    de Ercilla y <choice><orig>çuñiga</orig><reg>Zúñiga</reg>.
    </docAuthor></byline>
    <figure>
        <graphic url="https://es.wikipedia.org/wiki/Literatura_colonial_de_Chile#/media/
        <figDesc>Grabado del Ángel Gabriel</figDesc>
    </figure>
    <docImprint>
        EN <pubPlace>SALAMANCA</pubPlace>
        En cafa de <name>Domingo de Portonarijs</name>, Impreffor de fu Catholica Magefta
    </docImprint>
    <imprimatur>Con privilegio de Caftilla, y de Aragon. A cofta de Vicente, y Simon d
        marauedis el pliego.</imprimatur>
    </titlePage>
</front>
```

Esta brevíssima transcripción nos permite constatar ciertos puntos antes mencionados en este tutorial:

Un elemento engloba a todo el resto de los elementos, en este caso es el elemento `<front>` cuya función es contener toda la información paratextual de un documento.

Cada elemento tiene una etiqueta de apertura y una de cierre. La barra oblicua / indica el cierre de una etiqueta.

Se puede, facultativamente, agregar atributos a los elementos. En este caso el elemento `<titlePart>` tiene un atributo para especificar el título principal de un texto. En ciertos casos, podríamos encontrar un subtítulo explicativo o un título alternativo precedido por la conjunción disyuntiva *o*, en cuyos casos el valor del atributo `type` debe ser "desc" y "alt" respectivamente.

Podemos constatar la finalidad para cual se ha concebido el sistema de codificación TEI en XML: la descripción detallada del texto.

Sin embargo, esta transcripción se encuentra fuera de contexto. El elemento `<front>` no es el elemento raíz de un documento codificado en TEI. El elemento raíz de un documento en TEI es `<TEI>`. Este siempre lleva dos elementos hijos, `<teiHeader>` para los metadatos y `<text>` para el texto en sí.

TEI ha publicado cinco versiones de sus directrices. La más reciente es la quinta propuesta (P5), una publicación de más de mil páginas y casi seiscientas etiquetas. Esta información nos permite deducir que un texto puede ser exhaustivamente descrito en TEI. Ahora bien, esta exhaustividad dependerá de los requerimientos de cada proyecto de edición. En cambio, TEI exige una

cantidad mínima de información en el encabezado (`<teiHeader>`) para que un texto se considere codificado en TEI.

```
<?xml version="1.0" encoding="UTF-8"?>
<TEI>
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title>Título</title>
      </titleStmt>
      <publicationStmt>
      </publicationStmt>
      <sourceDesc>
      </sourceDesc>
    </fileDesc>
  </teiHeader>
  <text>
    Texto
  </text>
</TEI>
```

Este documento contiene la información mínima requerida de una codificación en TEI. La primera línea es lo que comúnmente se conoce como la declaración XML que tiene como función proveer la versión de XML que se utiliza y el estándar de codificación de caracteres. El `<teiHeader>` puede contener cinco elementos hijos que describen el contexto de publicación de un documento. Pero solamente el elemento `<fileDesc>` es obligatorio, este debe proveer la información relativa a la edición electrónica en al menos tres etiquetas enumeradas a continuación:

`<titleStmt>`: proporciona la información relativa al título del documento y sus autores o responsables.

`<publicationStmt>`: reúne la información acerca de la publicación del documento electrónico. Puede ser descrito simplemente en prosa en un elemento de párrafo `<p>` o se puede recurrir a elementos específicos como `<pubPlace>` o `<Publisher>`.

`<sourceDesc>`: contiene la información referente al documento que sirve de fuente para la edición electrónica. Al igual que el elemento precedente, este puede ser descrito en prosa o con elementos propios.

Como se ha mencionado previamente, la exhaustividad de la descripción con ayuda del estándar de marcación TEI depende de los requisitos de cada proyecto. En nuestro caso, haremos el examen detallado de la cabecera que utilizaremos para cada documento del COSUIZA en el último capítulo.

COSUIZA

COSUIZA es un corpus de textos hispánicos conservados en bibliotecas y archivos suizos que fue iniciado en 2013 por miembros de la Sección de español de la Universidad de Lausana. El objetivo general que se plantearon los integrantes de la sección en ese entonces fue el de realizar un inventario para disponer de una visión general de todos los manuscritos en español custodiados por la Confederación suiza. El objetivo siguiente fue realizar las ediciones de estos documentos siguiendo los criterios propuestos por la red § CHARTA. Los manuscritos escritos en español están repartidos por todo el país y su naturaleza es muy diversa, pero hay dos instituciones que sobresalen por la riqueza tanto cualitativa como cuantitativa de la documentación que conservan.

Biblioteca de Ginebra

La BGE aloja en su acervo documental una colección de manuscritos provenientes del condado de Altamira. La colección Favre, llamada así para honorar a quien fue el donador de los manuscritos en el año 1907, el historiador Edouard Favre, contiene epistolarios, documentos reales y también privados relativos a la casa de Altamira. En 1914 el conservador de manuscritos de la BGE, Léopold Micheli, publicó un inventario que actualmente se encuentra disponible en versión electrónica.

Biblioteca cantonal y universitaria de Lausana

La BCUL conserva desde 1997 el fondo Carlos de Goyeneche y Silvela, marqués de Balbueno, que ofrece una serie de documentos manuscritos de genealogía y también de tipo notarial: apeos, arrendamientos y títulos, entre otros. Para el lector que quiera ir más allá de esta presentación lacónica del fondo Carlos de Goyeneche y Silvela, puede consultar una exposición prolífica e informativa en Castillo Lluch y Diez del Corral Areta (2015)² o también en la página del Servicio de manuscritos de la Biblioteca cantonal y universitaria de Lausana.

²Castillo Lluch y Diez del Corral Areta (2015)

Para una descripción detallada de los fondos mencionados aquí y otros conservados en Suiza se puede consultar Castillo Lluch y Diez del Corral Areta (2018).³

³Castillo Lluch y Diez del Corral Areta (2018)

CHARTA

Como se ha mencionado en la sección precedente, los documentos que hacen parte del COSUIZA son transcritos y editados conforme a los criterios establecidos por CHARTA.⁴ Esta sigla tiene una doble significación, por una parte hace referencia a una red de investigación internacional que nació en 2005 y por otra da el nombre al Corpus Hispánico y Americano en la Red: Textos Antiguos. En primer lugar, el grupo estableció un estándar de edición aplicable a documentos de vertiente archivística con el propósito de uniformar las transcripciones de este tipo de documentos que suelen tener una alta dosis de subjetividad. En segundo lugar, se planteó la idea de constituir un corpus diacrónico (siglo XII al siglo XIX) de textos españoles e hispanoamericanos editados con los mencionados criterios y presentarlos en tres versiones: paleográfica, crítica y facsimilar. Las posibilidades ofrecidas por bases de datos como el corpus CHARTA han sido de enorme utilidad para las diferentes disciplinas que se interesan en la documentación histórica y no literaria. La creación y evolución de este corpus (y de sus subcorpora) ha sido de vital importancia para el estudio histórico de la lengua. Sin embargo, la abundancia de datos textuales siempre plantea dificultades si las herramientas que utilizamos para el procesamiento de estos datos presentan limitaciones. No cabe duda que el acceso a grandes cantidades de datos textuales ha modificado el proceder metodológico de los lingüistas. Sus exigencias han evolucionado y desde hace un tiempo ya no es suficiente contar con datos disponibles en un mismo sitio en la red, ahora se hace necesario contar con herramientas capaces de procesar y manipular los datos textuales de modo que, por ejemplo, un fenómeno lingüístico pueda ser estudiado y analizado con ayuda de las computadoras y ya no solamente con la calma y perseverancia del lingüista tradicional. En este contexto herramientas digitales como el estándar de etiquetado TEI y la plataforma Teitok permiten la convergencia de la actividadecdótica con la lingüística. Veremos en § TEITOK las posibilidades que ofrece esta plataforma en este sentido. También es necesario señalar que para llegar a construir fuentes de datos textuales fiables es indispensable el trabajo colaborativo de filólogos rigurosos dispuestos a revisar y corregir los datos que se ponen a disposición de la comunidad científica en estos corpus digitales. A

⁴Para una exposición del proyecto ver Areta and Aizpuru [2014]. Para acceder a la página del proyecto y del corpus ir a <https://www.corpuscharta.es/>

este respecto, Teitok ofrece un entorno que facilita este trabajo colectivo.

TEITOK

Teitok es una plataforma web desarrollada por Maarten Janssen [Janssen, 2016] en el contexto del proyecto Post Scriptum de la Universidad de Lisboa en 2014.⁵ Esencialmente, Teitok permite visualizar y procesar datos textuales codificados en TEI/XML. Es una herramienta versátil que ofrece además, diferentes módulos en función del corpus para el cual se utiliza. Comúnmente entendemos por corpus una colección de textos. En Teitok, esta colección de textos se presenta como una serie de archivos en formato TEI/XML. La particularidad de Teitok, y su diferencia con el estándar de marcación TEI, es que en el segmento <text> se aplica un sistema de tokenización que permite llevar a cabo el etiquetado morfosintáctico y la lematización de manera automática con herramientas de lingüística computacional almacenadas en el servidor de la plataforma. Esto se traduce en el reemplazo de la etiqueta <w> del estándar TEI, utilizada para marcar lingüísticamente las palabras, por la etiqueta <tok>. Para el resto de los elementos textuales se conservan las anotaciones de TEI (<pb> para el inicio de página, <lb> para el cambio de línea, <cb> para el inicio de columna, entre otros que veremos en el siguiente capítulo).

A continuación vemos el resultado de la anotación en TEI y luego con el modelo de Teitok:

```
1) TEI
<w lemma="haber" pos="VAIP2PO"
    xml:id="A19883-003-a-0180">
    <choice>
        <orig>habeis</orig>
        <reg>habéis</reg>
    </choice>
</w>
```

```
1) Teitok
<tok id="w-102" nform="habéis" lemma="haber" pos="VAIP2PO">
habeis
</tok>
```

⁵Ver Vaamonde [2015]

Conforme al sistema de tokenización, la normalización, la anotación lingüística y la lematización se inscriben en un solo nodo por medio de los atributos, lo que permite la modificación en la plataforma de manera muy intuitiva.⁶ Siguiendo la línea de desarrollo de los corpus electrónicos, resultado de la vinculación de la lingüística de corpus con las herramientas informáticas, la plataforma permite trabajar directamente en línea, lo que resulta en la utilización crítica del corpus por medio de su corrección continua [Kabatek, 2016]. Ya veremos en § Edición de textos en COSUIZA/Teitok las posibilidades que ofrece la plataforma web en cuanto a estas labores de edición.

Cabe señalar que la red CHARTA se encuentra haciendo la migración de su Corpus 2.0 a una versión 3.0 implementando Teitok. Este proyecto se puso en marcha en 2020 y, desde entonces, ha integrado al entorno de Teitok una serie de instrumentos para adaptar los criterios de edición CHARTA a las posibilidades y requerimientos de Teitok y TEI.

⁶Para un modelo de anotación cabalmente conforme a TEI, el lector puede consultar el libro *Edición digital de documentos antiguos: marcación XML-TEI basada en los criterios CHARTA*. [Martínez et al., 2020]

Edición de textos en COSUIZA/Teitok

En este capítulo presentaremos una guía detallada de los procedimientos de edición en la página del COSUIZA. Este capítulo se divide en tres apartados. En el primero se hace una presentación de las etiquetas que componen la cabecera de todos los documentos del corpus y un repertorio de las etiquetas aplicables al texto. El segundo expone las posibilidades de automatización que ofrece Teitok en el proceso de edición. Finalmente, el tercer apartado, ofrece una serie de videotutoriales que exponen el proceso de creación y edición de un documento en la página del COSUIZA.

Guía de etiquetado

Metadatos

Como se ha indicado previamente, los metadatos se presentan dentro del elemento `<teiHeader>`. Esta sección puede ser muy suscinta o muy extensa dependiendo de los requerimientos del proyecto en el que se inscribe la edición. El `<teiHeader>` del COSUIZA recoge los datos necesarios según los criterios de la red CHARTA y las exigencias de TEI. Este elemento que, de ahora en adelante, llamaremos cabecera se compone de cuatro elementos hijos que detallamos a continuación:

1. `<fileDesc>`
2. `<encodingDesc>`
3. `<profileDesc>`
4. `<revisionDesc>`

1. <fileDesc>

Contiene los elementos que proporcionan información bibliográfica del documento electrónico.

1.1 <titleStmt> El enunciado de título ofrece la información relativa al título del documento electrónico y las personas u organismos responsables de su realización.

1.1.1 <title> Todos los documentos del COSUIZA llevan por título una secuencia alfanumérica como se ilustra en el ejemplo debajo:

```
<title>COSUIZA-0001</title>
```

1.1.2 <funder> Ofrece la información sobre la institución o individuos que financian el proyecto. Para todos los documentos del COSUIZA el contenido de este elemento es el mismo:

```
<funder>Universidad de Lausana</funder>
```

1.1.3 <respStmt> El enunciado de responsabilidad provee información sobre las personas que han participado en la producción del documento electrónico, también se precisa el rol de cada una. Se deben distinguir tres roles de acuerdo a los criterios de edición CHARTA y uno específico a la producción del documento electrónico. Cada uno de estos enunciados de responsabilidad contienen un elemento <resp> con un atributo para especificar si se trata del transcriptor, los revisores o el encargado de producir el documento XML y, un elemento <name> para identificar a la persona encargada de la función correspondiente. Se deben completar solamente los nombres que correspondan en la etiqueta <name>, el resto es invariable.

```
<respStmt>
    <resp resp="transcriber">Transcriptor</resp>
    <name>Nombre del transcriptor</name>
</respStmt>
<respStmt>
    <resp resp="reviewer1">Revisor 1</resp>
    <name>Nombre del revisor 1</name>
</respStmt>
<respStmt>
    <resp resp="reviewer2">Revisor 2</resp>
    <name>Nombre del revisor 2</name>
```

```

</respStmt>
<respStmt>
    <resp resp="XMLconverter">Conversor a XML</resp>
    <name>Nombre del conversor a XML</name>
</respStmt>

```

1.2 <publicationStmt> Proporciona los datos relativos a la publicación y distribución del documento electrónico.

1.2.1 <publisher> Contiene la información sobre la organización responsable de la publicación del documento. En nuestra cabecera encontramos dos elementos `<orgName>` para proporcionar el nombre del grupo GRAFILA (Grupo de Análisis Filológico de Lausana) —acrónimo del grupo de estudio miembro de la red CHARTA— y el de la UNIL. El tipo de institución se detalla con ayuda del atributo `type`. Para el COSUIZA estos elementos y sus atributos son inmutables:

```

<publisher>
    <orgName type="group">GRAFILA</orgName>
    <orgName type="institution">Universidad de Lausana</orgName>
</publisher>

```

1.2.2 <pubPlace> Contiene el nombre del lugar de publicación del documento electrónico. Para todos los documentos del corpus este elemento tiene el mismo contenido:

```
<pubPlace>Lausana</pubPlace>
```

1.2.3 <idno> Este elemento aparece dos veces en la cabecera y está destinado a identificar el corpus y el número del documento dentro del corpus. Estas informaciones son de utilidad para el macrocorpus CHARTA en el proceso de integración de los documentos de todos sus subcorpora. El contenido de la primera etiqueta es invariable a diferencia del número identificador:

```

<idno type="corpus">COSUIZA</idno>
<idno type="corpus-num">Número identificador</idno>

```

1.2.4 <distributor> Recoge la información sobre el o los organismos responsables de la distribución del documento. En nuestra cabecera permanece invariable:

```
<distributor>Universidad de Lausana</distributor>
```

1.2.5 <availability> Proporciona la información relativa a la disponibilidad o restricción de publicación de un documento. Dada la naturaleza de los documentos que hacen parte del COSUIZA, en la mayoría de los casos el contenido de esta etiqueta será el mismo. Ahora bien, si se tratara de un documento sobre el cual se aplican ciertos derechos o reservas deberá ser señalado en esta sección. Para todo el resto de los documentos, este elemento se presenta dentro de una etiqueta *párrafo* (<p>) como se muestra a continuación:

```
<availability>
    <p>Licencia CREATIVE COMMONS</p>
</availability>
```

1.3 <sourceDesc> Suministra la información sobre el documento original que sirve de fuente para la edición digital. En virtud de que todos los documentos originales son manuscritos, vamos a utilizar la etiqueta <msDesc> concebida para describir documentos de este tipo.

1.3.1 <msDesc> 1.3.1.1 <msIdentifier> Este elemento contiene todas las etiquetas destinadas a identificar la ubicación exacta de un manuscrito. Vamos a indicar el país, el cantón, la ciudad, el nombre del archivo y el identificador dentro del archivo. El único elemento que no varía es el país como podemos ver debajo:

```
<msIdentifier>
    <country>Suiza</country>
    <region type="canton">Cantón en donde se conserva el documento</region>
    <settlement>Ciudad en donde se conserva el documento</settlement>
    <repository>Archivo en donde se conserva el documento</repository>
    <idno>Identificador dentro del archivo</idno>
</msIdentifier>
```

En caso de que el documento electrónico tenga como fuente el fragmento de un manuscrito se deben indicar los folios (f o ff.) dentro de la etiqueta <idno>.

1.3.1.2 <msContents> Esta etiqueta nos permite describir los contenidos del manuscrito. Incluirímos el resumen que hace parte de todas las cabeceras de las ediciones CHARTA, además señalaremos si se trata de una copia o de un original en el valor del atributo **class** del elemento <msItem>, el nombre del scriptor en caso de conocer su nombre y, añadiremos su tipo en el atributo **type** con el valor que corresponda:

```
<msContents>
  <summary>Resumen del contenido</summary>
  <msItem class="original/copia">
    <editor>
      <persName type="escribano/notario">Nombre del scriptor</persName>
    </editor>
  </msItem>
</msContents>
```

1.3.1.3 <physDesc> Proporciona la información respecto de las características físicas del documento fuente. Este elemento contiene tres etiquetas mayores:

- <objectDesc> En donde incluiremos los elementos relativos a las medidas del documento y su estado de conservación. En la mayoría de los casos los atributos de los elementos son invariables.
- <handDesc> En donde identificaremos y describiremos la o las manos presentes en el manuscrito. En nuestro ejemplo vamos a suponer que existen dos manos en el manuscrito, una de ellas será la preponderante (major) y, la otra, la secundaria (minor).
- <bindingDesc> En donde podemos describir la encuadernación.

```
<physDesc>
  <objectDesc form="ms">
    <supportDesc material="papel">
      <extent>
        <note>Descripción en prosa del manuscrito</note>
        <dimensions scope="all" unit="mm">
          <height></height>
          <width></width>
        </dimensions>
      </extent>
      <condition><p>Estado de conservación</p></condition>
    </supportDesc>
  </objectDesc>
  <handDesc hands="2">
    <handNote resp="#h1" scope="major"><p>Descripción de la mano 1</p></handNote>
    <handNote resp="#h2" scope="major"><p>Descripción de la mano 2</p></handNote>
  </handDesc>
  <bindingDesc>
    <binding><p>Descripción en prosa de la encuadernación</p></binding>
  </bindingDesc>
</physDesc>
```

1.3.1.4 <history> Etiqueta que contiene la información sobre la historia del manuscrito. Usaremos la etiqueta <origin> para incluir los elementos

específicos a la fecha y locación originales del manuscrito. Todos los atributos son invariables, con la excepción del atributo `when`.

```
<history>
<origin>
  <origDate when="1688-11-12" type="explicit">1688 noviembre 12</origDate>
  <origPlace type="explicit">
    <placeName type="settlement">Valladolid</placeName>
    <region type="region">Valladolid</region>
    <country>España</country>
    <geo>41.651981 -4.728561</geo>
  </origPlace>
</origin>
</history>
```

El atributo `when` debe tener como valor la fecha en formato `aaaa-mm-dd`. El contenido del elemento `<geo>` permite la geovisualización de los documentos en la plataforma Teitok. Para lograr la correcta distribución de los documentos en el mapa, es imprescindible estandardizar las coordenadas geográficas de las diferentes localidades representadas en el mapa. Para acceder al repertorio de coordenadas del COSUIZA pinchar aquí. En los casos en donde solo se sepa el país de origen del documento, el repertorio de coordenadas se encuentra en este enlace. En el supuesto de que una localidad o país no se hallaren en el repertorio se debe contactar al administrador de la página del COSUIZA.

2. `<encodingDesc>`

Elemento que permite describir los fundamentos editoriales sobre los cuales se ha realizado la edición del documento.

2.1 `<projectDesc>` Ofrece las razones por las cuales se ha realizado la codificación del documento. Se trata de una descripción somera dentro de un elemento *párrafo* (`<p>`). El contenido será el mismo para todos los documentos del COSUIZA:

```
<projectDesc>
  <p>Edición electrónica preparada para la investigación CHARTA-TEI</p>
</projectDesc>
```

2.2 `<editorialDecl>` Contiene los detalles de los principios editoriales aplicados en la codificación del documento. Su contenido se precisa en un elemento *párrafo* (`<p>`) y es idéntico en todos los documentos del COSUIZA:

```
<editorialDecl>
  <p>Este documento sigue los criterios de edición CHARTA adaptados para el estándar internacional</p>
</editorialDecl>
```

2.3 <classDecl> En este elemento se incluye la tipología documental propuesta en los criterios CHARTA.⁷ Este elemento es invariable y debe estar presente en todos las ediciones electrónicas, puesto que sirve como fuente de la categoría de documento que se señalará posteriormente en la cabecera. Dada la extensión de esta clasificación, se expone solamente un extracto a continuación:

```
<classDecl>
  <taxonomy xml:id="Tip-CH">
    <bibl>Tipología CHARTA, propuesta de octubre de 2013</bibl>
    <category xml:id="tex-leg">
      <category xml:id="tex-leg-ord">
        <catDesc>Ordenanzas</catDesc>
      </category>
      <category xml:id="tex-leg-fue">
        <catDesc>Fueros</catDesc>
      </category>
      <category xml:id="tex-leg-pri">
        <catDesc>Privilegios</catDesc>
      </category>
      <category xml:id="tex-leg-cpue">
        <catDesc>Cartas pueblas</catDesc>
      </category>
      <category xml:id="tex-leg-pra">
        <catDesc>Pragmáticas</catDesc>
      </category>
    </category>
    <category xml:id="car-com">
      <category xml:id="car-com-con">
        <catDesc>Contratos</catDesc>
      </category>
      <category xml:id="car-com-com">
        <catDesc>Compraventas</catDesc>
      </category>
      <category xml:id="car-com-ces">
        <catDesc>Cesiones</catDesc>
      </category>
      <category xml:id="car-com-don">
        <catDesc>Donaciones</catDesc>
      </category>
    </category>
  </taxonomy>
</classDecl>
```

⁷Ver tipología documental en el enlace

```

<category xml:id="car-com-conc">
    <catDesc>Conciertos</catDesc>
</category>
<category xml:id="car-com-acu">
    <catDesc>Acuerdos</catDesc>
</category>
<category xml:id="car-com-per">
    <catDesc>Permutas</catDesc>
</category>
<category xml:id="car-com-tru">
    <catDesc>Truques</catDesc>
</category>
<category xml:id="car-com-pac">
    <catDesc>Pactos</catDesc>
</category>
<category xml:id="car-com-int">
    <catDesc>Intercambios de bienes</catDesc>
</category>
<category xml:id="car-com-ccen">
    <catDesc>Cartas de censo</catDesc>
</category>
</category>
</taxonomy>
</classDecl>

```

3. <profileDesc>

Elemento que se utiliza para indicar informaciones no bibliográficas del documento.

3.1 <langUsage> Esta etiqueta contiene las lenguas utilizadas en el texto. Cada lengua se señala en un elemento `<language>` que incluye el atributo `ident` para identificar el código de cada lengua. Este código proviene de la norma ISO-639-1. La mayoría de los documentos del COSUIZA están escritos en castellano y para expresarlo usaremos el código `es`. En el ejemplo debajo vamos a suponer estamos frente a un documento en el cual se utilizan dos lenguas, una principal (*major*) y una secundaria (*minor*), la extensión de la utilización de cada una en el manuscrito se expresa en el atributo `scope` que permite expresar una noción de medida:

```

<langUsage type="hybrid">
    <language ident="es" scope="major">castellano</language>
    <language ident="fr" scope="minor">francés</language>
</langUsage>

```

Cuando codifiquemos un documento con más de una lengua debemos precisarlo en el atributo `type` del elemento `<langUsage>` con el valor `hybrid`. No incluiremos este atributo en los documentos monolingües.

3.2 <textClass> Proporciona la información relativa a la clasificación del documento de acuerdo a una tipología o esquema establecidos. En esta sección recogeremos los identificadores correspondientes de la taxonomía CHARTA presentada previamente. También vamos a incluir las palabras clave propias de la cabecera CHARTA. Supongamos que estamos codificando un testamento, primero buscaremos la categoría Testamento en la taxonomía declarada más arriba y vamos a copiar el valor de su atributo `xml:id` que es: `tes-inv-tes`. También vamos a incluir la supracategoría que contiene a Testamento como vemos en la figura debajo:

```

<category xml:id="tes-inv"> Supracategoría
  <category xml:id="tes-inv-inv">
    | <catDesc>Inventarios</catDesc>
  </category>
  <category xml:id="tes-inv-lis">
    | <catDesc>Listados</catDesc>
  </category>
  <category xml:id="tes-inv-alm">
    | <catDesc>Almonedas</catDesc>
  </category>
  <category xml:id="tes-inv-tes"> Subcategoria
    | <catDesc>Testamentos</catDesc>
  </category>
  <category xml:id="tes-inv-cod">
    | <catDesc>Codicilos</catDesc>
  </category>
  <category xml:id="tes-inv-mtes">
    | <catDesc>Mandas testamentarias</catDesc>
  </category>
  <category xml:id="tes-inv-cue">
    | <catDesc>Cuentas</catDesc>
  </category>
  <category xml:id="tes-inv-cpob">
    | <catDesc>Censos de población</catDesc>
  </category>
  <category xml:id="tes-inv-des">
    | <catDesc>Deslindes y amojonamientos</catDesc>
  </category>
</category>

```

Se indica en primera posición la supracategoría y, luego, la subcategoría, cada una en un elemento `<catRef>` y en el atributo `target` se debe escribir el identificador de la categoría antecedido por un signo almohadilla o numeral. La codificación de esta sección de un documento testamentario quedaría como vemos debajo:

```

<textClass>
  <catRef target="#tes-inv"/>
  <catRef target="#tes-inv-tes"/>
  <keywords>

```

```

<term>Palabra clave 1</term>
<term>Palabra clave 2</term>
<term>Palabra clave 3</term>
<term>etc</term>
</keywords>
</textClass>
```

3.3 <particDesc> Este elemento permite describir los participantes o interlocutores presentes en un texto. Esta sección solo debe utilizarse para ciertas categorías documentales como la correspondencia o, documentos reales en donde existen diversos roles y podemos diferenciar la autoría de la delegación y la escritura. Para cada rol o papel usaremos una etiqueta <person> con un atributo **rol** cuyo valor va a depender de la categoría documental. En el ejemplo debajo vemos los roles que debemos especificar, en el caso de que sea posible identificar todos los roles, en un documento real.

```

<particDesc>
  <person role="author">
    <persName>Autor</persName>
  </person>
  <person role="ordered">
    <persName>Iussor</persName>
  </person>
  <person role="scriptor">
    <persName>Scriptor</persName>
  </person>
  <person role="signed">
    <persName>Firmante</persName>
  </person>
</particDesc>
```

En el caso de estar frente a un documento epistolar, se debe utilizar la secuencia a continuación, solamente si es posible identificar estos roles:

```

<particDesc>
  <person role="sent">
    <persName>Emisor</persName>
  </person>
  <person role="received">
    <persName>Destinatario</persName>
  </person>
</particDesc>
```

3.4 <correspDesc> Ofrece datos relativos a un documento epistolar. Esta sección debe ser utilizada solamente en la codificación de cartas. Se detalla el lugar de origen y destino además de sus coordenadas respectivas.

```
<correspDesc>
  <correspAction type="sent">
    <settlement>
      <placeName>Origen</placeName>
      <geo>Coordenadas</geo>
    </settlement>
  </correspAction>
  <correspAction type="received">
    <settlement>
      <placeName>Origen</placeName>
      <geo>Coordenadas</geo>
    </settlement>
  </correspAction>
</correspDesc>
```

4. <revisionDesc>

Este elemento permite registrar las modificaciones que se han hecho en el documento electrónico. Cada intervención se detalla en un elemento `<change>` con un atributo `when` para precisar la fecha en que ha tenido lugar la modificación. En el COSUIZA detallaremos dos modificaciones principales: la creación del documento XML y la revisión de la presentación crítica. El atributo `when` debe indicar la fecha de la modificación y se puede agregar el atributo `who` para indicar las iniciales del colaborador que las ha llevado a cabo.

```
<revisionDesc>
  <change when="aaa-mm-dd" who="MCL">Creación del archivo</change>
  <change when="aaa-mm-dd" who= "ECDA">Revisión final de la presentación crítica</change>
</revisionDesc>
```

Texto

`<text>` es el elemento que contiene los datos textuales en un documento TEI. Hemos señalado que en Teitok la cabecera se ajusta a los criterios de la TEI, sin embargo, en lo que respecta al texto tiene su propio sistema basado en la *tokenización* de las palabras y de los signos de puntuación. En este sentido, en este tutorial nos limitaremos a presentar solamente los elementos editoriales que CHARTA ha adaptado para su migración a Teitok.

1. Elementos codicológicos

1.1 Numeración de hoja, columna y línea

1.1.1 Inicio de página CHARTA

CHARTA-TEITOK

{h 1r}

<pb n = “1r” facs = “COSUIZA-0014-1r.jpeg” id = “e-1”/>

{h 1v}

<pb n = “1v” facs = “COSUIZA-0014-1v.jpeg” id = “e-23”/>

La etiqueta <pb> indica el inicio de página. Es pertinente recordar que conforme a los criterios CHARTA todos los textos editados, aunque sean fragmentos de un manuscrito, comienzan por la hoja 1. De modo que el atributo **n** (*número*), siempre comienza por el número 1. También consta de un atributo **facs** para precisar el nombre del archivo facsimilar. Para denominar los archivos de los facsímiles se utiliza el nombre del archivo XML seguido de un guión, la hoja y cara correspondiente: COSUIZA-XXXX-1r. El atributo **id** adjudica un identificador único al elemento <pb>, este identificador es de la forma **e-nº** y, si bien no es aleatorio, no corresponde necesariamente al número de hoja. Veremos en la sección práctica de este documento, que la plataforma de Teitok se encarga de adjudicar este identificador.

1.1.2 Inicio de columna CHARTA

CHARTA-TEITOK

{a}

<cb n = “a” id = “e-9”/>

{h 1ra}

<pb n = “1r” facs = “COSUIZA-0014-1r.jpeg” id = “e-1”/> <cb n = “a” id = “e-9”/>

El inicio de columna se indica con la etiqueta <cb>. Consta de un atributo **n** cuyo valor es la letra de la columna y un atributo **id**. En la segunda línea de la tabla tenemos un inicio de página que coincide con el inicio de una columna, coincidencia muy frecuente en los manuscritos. Como vemos en la marcación CHARTA, página y columna se señalan en una misma secuencia entre llaves. Sin embargo, en la marcación CHARTA-TEITOK el inicio de página y el de columna se indican con etiquetas independientes no anidadas.

1.1.3 Inicio de línea CHARTA

CHARTA-TEITOK

{1}

<lb n = “1” id = “e-1”/>

Del mismo modo que los elementos precedentes, el cambio de línea contiene dos atributos: **n** y **id**, y el identificador puede no coincidir con el número de línea.

1.1.4 Cambio de línea en acotaciones marginales CHARTA

CHARTA-TEITOK

|

<lb id = “e-4”/>

El cambio de línea en el texto marginal se anota con un elemento <lb>, sin embargo, carece de atributo **n**.

1.2 Deterioro del original CHARTA

CHARTA-TEITOK

[***]

<gap reason = “ilegible”/>

*

<gap reason = “ilegible” extent = “1 char”/>

**

<gap reason = “ilegible” extent = “2 chars”/>

<gap reason = “ilegible” extent = “3 chars”/>

<gap reason = “ilegible” extent = “4 chars”/>

<gap reason = “ilegible” extent = “5 chars”/>

<gap reason = “ilegible” extent = “6 or more chars”/>

[roto]

<gap reason = “roto”/>

[doblez]
<gap reason = “doblez”/>
[mancha]
<gap reason = “mancha”/>

Valga recordar que el uso de corchetes en la marcación CHARTA tiene lugar solamente cuando no se sabe el número exacto de caracteres ilegibles. En CHARTA-TEITOK se puede indicar hasta un máximo de cinco letras ilegibles, para cualquier cantidad superior se debe dar el valor **6 or more chars** al atributo **extent**. Otro cambio que ha tenido lugar en los criterios CHARTA-TEITOK en relación a los criterios CHARTA es que no se puede emplear la marcación relativa a la cantidad de letras ilegibles y la razón del deterioro al mismo tiempo. El transcriptor debe escoger la información que sea pertinente en cada caso.

1.3 Signos o elementos especiales

1.3.1 Firma CHARTA

CHARTA-TEITOK

[firma: Alfonso de Fonseca]

<signed> <tok id = “w-1”> Alfonso</tok> <tok id = “w-2”>de</tok> <tok id = “w-3”>Fonseca</tok></signed>

La etiqueta `<signed>` se utiliza para marcar las firmas. Debe contener el texto de la firma debidamente *tokenizado*.

CHARTA

CHARTA-TEITOK

[firma mano 2: Alfonso de Fonseca]

<signed hand = “#h2”> <tok id = “w-1”> Alfonso</tok> <tok id = “w-2”>de</tok> <tok id = “w-3”>Fonseca</tok></signed>

Se diferencia la firma hecha por otra mano con el atributo **hand**, cuyo valor debe estar debidamente mencionado dentro del elemento `<handDesc>` en los metadatos.

CHARTA

CHARTA-TEITOK

[firma en ar][firma en he]

<signed xml:lang = “ar”/> <signed xml:lang = “he”/>

Se usa el atributo `xml:lang` para marcar una firma en un alfabeto distinto al latino. Nótese que entre corchetes en la marcación CHARTA se aconseja utilizar el código ISO-639-1.

1.3.2 Rúbrica CHARTA

CHARTA-TEITOK

[rúbrica]

`<figure type = “rúbrica”/>`

[rúbrica: A]

`<figure type = “rúbrica”><tok id = “w-1”>A</tok> </figure>.`

La rúbrica se designa con la etiqueta `<figure>`. Al igual que en la firma, el texto que contiene debe estar *tokenizado*.

1.3.3 Signos CHARTA

CHARTA-TEITOK

[sello][crismón][cruz][signo][quirógrafo]

`<figure type = “cruz”/>`

Se emplea la etiqueta `<figure>` para los signos especiales. En el ejemplo hemos usado solamente el signo de la cruz, no obstante, se debe dar el valor que corresponda en cada caso al atributo `type`.

1.3.4 Impreso CHARTA

CHARTA-TEITOK

[impreso: texto]

`<hi rend = “impreso”> <tok id = “w-1”>texto</tok></hi>`

Los pasajes impresos se marcan con la etiqueta `<hi>` (*highlighted*). Este elemento se utiliza para destacar fragmentos que se diferencian de su contexto. El atributo `rend(rendition)` sirve para indicar cómo se representa un elemento en el manuscrito. En este contexto, el valor de este atributo es invariable. Como se ha señalado previamente, el texto debe estar *tokenizado*.

1.4 Intervenciones en el texto

1.4.1 Tachado, raspado, cancelado CHARTA

CHARTA-TEITOK

[tachado]

<del type = “tachado”/>

[tachado: texto]

<del type = “tachado”> <tok id = “w-1”>texto</tok>

[tachado mano 2: texto]

<del type = “tachado” hand = “#h2”> <tok id = “w-1”>texto</tok>

Usamos la etiqueta `` (*deletion*) para marcar un fragmento suprimido. Utilizaremos la misma secuencia para otras intervenciones como *raspado* y *cancelado*. Para ello, se debe completar el valor del atributo `type` con la intervención que corresponda.

1.4.2 Sobre escrito CHARTA

CHARTA-TEITOK

mu[sobrescrito: l+g]er

<tok id = “w-1”>mu<subst><del type = “sobrescrito”>l<add type = “sobrescrito”>g</add></subst>er</tok>

Puesto que esta es una intervención que tiene lugar dentro de una palabra o un `<tok>`, integramos los elementos propios a la intervención en el interior de esta etiqueta. Utilizamos `<subst>` para indicar la supresión y adición de elementos en el texto. Para indicar la supresión empleamos la etiqueta `` cuyo contenido es, en este caso, la letra *l* y, para el texto sobrescrito, la letra *g*, usamos `<add>`. Ambas deben contar con un atributo `type` para declarar el tipo de intervención.

1.4.3 Sobre raspado CHARTA

CHARTA-TEITOK

[sobre raspado: muger]

<subst><del type = “sobre-raspado”/><add type = “sobre-raspado”>muger</add></subst>

Utilizamos los mismos elementos que en el caso precedente, sin embargo, dado que es muy improbable que podamos restituir el texto raspado, la etiqueta `` no lleva contenido. El texto, en este caso *muger*, debe ir correctamente *tokenizado*, cosa que no hacemos en este ejemplo para no sobrecargarlo visualmente. Si el sobre raspado se hallase al interior de una palabra, la intervención debe indicarse al interior de la etiqueta `<tok>` como podemos apreciar en el ejemplo del sobrescrito.

1.4.4 Interlineado CHARTA

CHARTA-TEITOK

[interlineado: texto]

```
<add place = "interlineado">texto</add>
```

Para el interlineado utilizamos la etiqueta `<add>` que consta de un atributo `place`cuyo valor es invariable. Si el interlineado fuese obra de otra mano, esto se indica agregando el atributo `hand` con el valor “#h2” o el que corresponda a la mano debidamente declarada en los metadatos. Este mismo principio se puede aplicar para todo el resto de las intervenciones hechas por otra mano en el texto.

1.4.5 Margen CHARTA

CHARTA-TEITOK

[margen: texto]

```
<add place = "margen">texto</add>
```

[margen mano 2: texto]

```
<add hand = "#h2" place = "margen">texto</add>
```

Usamos las mismas etiquetas que en el caso del interlineado, sin embargo, el valor del atributo `place`cambia. Opcionalmente se puede añadir el lado en el cual tiene lugar la intervención al margen. En cuyo caso el atributo `place`llevaría por valor: margen izquierdo, margen derecho, margen superior o margen inferior. Como en el resto de los casos previamente señalados y, como se puede ver en la tabla, también se puede precisar un cambio de mano.

Para agregar el texto de un sobre se puede utilizar el mismo marcado que se usa para el margen. Se debe escribir *sobre* en el valor del atributo `place`

1.4.6 Encabezamiento y título CHARTA

CHARTA-TEITOK

[encabezamiento: texto]

```
<div type = "encabezamiento">texto</div>
```

[título: texto]

```
<div type = "título">texto</div>
```

Se emplea la etiqueta `<div>` para el encabezamiento y el título. Este elemento se utiliza para englobar párrafos que hacen parte de una unidad. Se añade el atributo `type`para detallar el tipo de subdivisión.

1.4.7 Blanco CHARTA

CHARTA-TEITOK

[blanco]

<gap reason = “blanco”>texto</gap>

Este elemento se marca con las mismas etiquetas que en los casos de deterioro del original, en cambio, el valor del atributo **reason** es *blanco*.

1.5 Intervenciones en el texto por parte del editor Las conjeturas del editor se pueden clasificar de acuerdo al grado de certeza con el que se realizan.

1.5.1 Conjetura cierta CHARTA

CHARTA-TEITOK

[ilegible: texto]

<supplied reason = “ilegible”>texto</supplied>

[roto: texto]

<supplied reason = “roto”>texto</supplied>

[doblez: texto]

<supplied reason = “doblez”>texto</supplied>

[mancha: texto]

<supplied reason = “mancha”>texto</supplied>

Se utiliza el elemento <supplied> para agregar una conjetura con certeza. La razón se añade en el valor del atributo **reason**.

1.5.2 Conjetura incierta CHARTA

CHARTA-TEITOK

[ilegible: texto]

<unclear reason = “ilegible”>texto</unclear>

[roto: texto]

<unclear reason = “roto”>texto</unclear>

[doblez: texto]

<unclear reason = “doblez”>texto</unclear>

[mancha: texto]

<unclear reason = “mancha”>texto</unclear>

La conjetura con menos certeza se trata de manera similar a la precedente, sin embargo, se utiliza la etiqueta `<unclear>`. Se debe recordar que todo texto al interior de estas etiquetas debe estar *tokenizado*.

2. Transcripción paleográfica y presentación crítica

2.1 Transcripción paleográfica CHARTA

CHARTA-TEITOK

villa

`<tok id = "w-1">villa</tok>`

Para marcar las palabras empleamos la etiqueta `<tok>`, la cual siempre debe constar de un atributo `id` para su identificador de la forma "`w-nº`". El contenido de esta etiqueta es la transcripción paleográfica de la palabra, en nuestro ejemplo: *villa*. Sabemos que esta palabra no sufrirá ningún cambio en la presentación crítica, a menos que esté precedida de un punto y tengamos que poner la *v* en mayúscula. En este ejemplo supondremos que *villa* no requiere de ninguna modificación en la presentación crítica. En todos los casos similares la etiqueta `<tok>` no lleva otro atributo que el de `id`.

2.2 Presentación crítica CHARTA

CHARTA-TEITOK

uilla

`<tok id = "w-1" nform = "villa">uilla</tok>`

En este caso debemos cambiar la grafía *u* por la *v* ya que la primera se presenta con valor consonántico, para ello, agregamos el atributo `nform` para escribir la forma que deseamos para nuestra presentación crítica.

2.3 Abreviaturas CHARTA

CHARTA-TEITOK

escriu

`<tok id = "w-1" form = "escrui" fform = "escriuano" nform = "escrivano">escriu<ex>ano</ex>`

Para indicar la presencia de una abreviatura, debemos hacerlo por medio del elemento `<ex>`. Dado que es un fenómeno que ocurre al interior de una palabra, su etiqueta debe situarse al interior de la etiqueta `<tok>`. Ante la presencia de una abreviatura se deben agregar dos atributos a esta etiqueta. En primer lugar se añade el atributo `form` cuyo valor debe ser la forma escrita, sin desarrollar la abreviatura ni incluir el elemento `<ex>`. Este atributo se usa en todos los casos

en donde encontramos otra etiqueta al interior de una etiqueta `<tok>`. En segundo lugar, se incorpora el atributo `fform` cuyo valor debe ser la palabra con su abreviatura expandida. Finalmente, como el ejemplo requiere la modificación de la grafía *u* por *v*, al igual que en el ejemplo precedente, añadimos un atributo `nform`.

Recordemos que cuando se agregan atributos a las etiquetas, estos deben tener un valor, de otro modo no los añadiremos. En el caso susodicho, la abreviatura exige la existencia de los atributos `formy` `fform`, pero no de `nform`.

2.4 Unión y separación de palabras CHARTA

CHARTA-TEITOK

dela

```
<tok id = "w-1" nform = "de la">dela<dtok id = "d-1-1" form = "de"/>
<dtok id = "d-1-2" form = "la"/></tok>
```

Se trata la unión de palabras con la adición del elemento `<dtok>` el cual no lleva contenido, sin embargo, debe imperativamente llevar un atributo `form`. También debe constar de un identificador que incluye el número identificador del `<tok>` que lo contiene. En los casos que se requiera, la etiqueta `<dtok>` debe llevar los atributos `fformy` `nform`, no obstante, esto no descarta la necesidad de también incluir el atributo `nform` en el `<tok>` principal.

CHARTA

CHARTA-TEITOK

juris prudencia

```
<tok id = "w-1" nform = "jurisprudencia">juris prudencia</tok>
```

La separación irregular de palabras se conserva en el contenido de la etiqueta `<tok>` y se normaliza en el atributo `nform`.

2.5 Ruptura de palabra a final de renglón CHARTA

CHARTA-TEITOK

razo{2}nes

```
<tok id = "w-1" form = "razones">razo<lb n = "2" id = "e-9"/>nes</tok>
```

Se introduce la etiqueta de salto de línea `<lb>` al interior del *token* en el que ocurre la ruptura. Recordemos que todo *token* que lleve otra etiqueta al interior debe añadir el atributo `form` con la palabra en su forma de transcripción paleográfica.

2.6 Cambio de lengua CHARTA

CHARTA-TEITOK

[en.: fish]

```
<foreign xml:lang = “en”> <tok id = “w-1”>fish</tok></foreign>
```

Se trata el cambio de lengua con la etiqueta `<foreign>`, que permite indicar la presencia de una lengua distinta a su contexto. Se debe añadir el código ISO-639-1 como valor del atributo `xml:lang`.

2.7 Signos de puntuación CHARTA-TEITOK

```
<tok id = “w-1” nform = “.”>,</tok>
```

Se ha señalado previamente que en Teitok los signos de puntuación son considerados *tokens*, de modo que para editar un signo de puntuación se añade un atributo `nform` con el signo deseado en la presentación crítica.

CHARTA-TEITOK

```
<tok id = “w-1” nform = “–”>,</tok>
```

Para eliminar un signo de puntuación de la presentación crítica, se escriben dos guiones en el atributo `nform`.

CHARTA-TEITOK

```
<tok id = “w-1” nform = “,”><ee/></tok>
```

Si se desea añadir un signo de puntuación únicamente en la presentación crítica, es necesario agregar una etiqueta `<ee/>` al interior del elemento `<tok>` y se escribe en el atributo `nform` el signo deseado.

2.8 Etiquetado gramatical CHARTA-TEITOK

```
<tok id = “w-1” pos = “VMIS3S0” lemma = “venir”>vino</tok>
```

Para indicar la anotación morfosintáctica se añaden dos atributos: `pos`(*part of speech*) para incluir la categoría gramatical y `lemma` para la forma lematizada.

Para obtener más detalles sobre las categorías gramaticales y los códigos a emplear en el atributo `pos` se recomienda visitar la página del etiquetario.

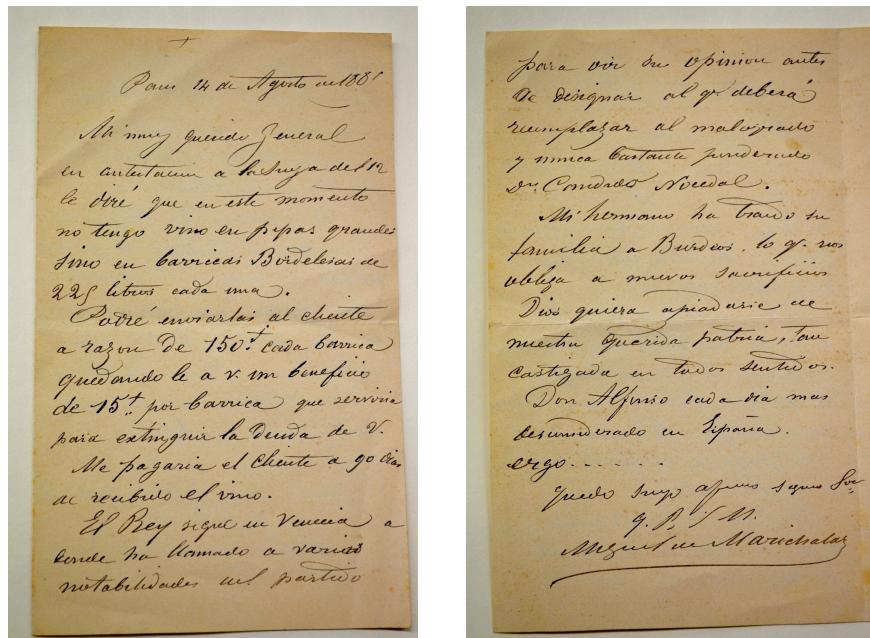
Automatización en la plataforma

En esta sección se presenta una lección con las instrucciones para la edición de textos en la página del COSUIZA. Ahora bien, aunque la sección precedente

pueda parecer abrumadora por la profusión de información, es absolutamente necesario que el lector esté al tanto de la estructura y contenido de los documentos codificados.

En este apartado veremos que todo el proceso de codificación está ampliamente automatizado en la plataforma, sin embargo, las aplicaciones informáticas, aunque son convenientes, no siempre son infalibles.

El documento que codificaremos es una carta que se conserva en la Biblioteca cantonal y universitaria de Friburgo:



Biblioteca cantonal y universitaria de Friburgo, B-126

Acceso

Como se ha mencionado en el primer apartado, para hacer ediciones en el COSUIZA, es necesario tener una cuenta de usuario o administrador. Satisficho este requisito, ir a la página de inicio del COSUIZA. En el menú principal seleccionar la pestaña *login* e ingresar usuario y contraseña.

Creación de archivo XML

Tras identificarse en la plataforma, se abrirá una página con la lista de funciones con privilegios de administrador. Se debe pinchar en la primera opción *create new XML file* como vemos a continuación:

Esta selección nos dirige a una página en donde se debe denominar el archivo, escoger las opciones relativas a los metadatos y agregar su contenido textual.

Create New XML File

XML Filename

XML id (filename):

Initial Metadata

Leave empty 1

Use a template

Paste a TEI/XML file (will keep text content as well)

Use an existing XML file

Initial Content

Leave empty 2

Create as WYSIWYG (or paste from Word, HTML, etc.)

Create from plain text

Create XML File - instead of the methods here, you can also create a new XML starting from a PDF document or Facsimile images 3

Provide more options

You do not have a teiHeader template defined for editing; using such a template allows you to easily edit the metadata in an HTML form. You can create an edit template [here](#)

1. En primer lugar se completa el campo *XML id (filename)* con el nombre del archivo siguiendo el estándar expuesto en § 1.1.1. Para los metadatos, se debe escoger la opción *Leave empty* (posteriormente nos ocuparemos de los metadatos).

Create New XML File

XML Filename

XML id (filename): ←

Initial Metadata

Leave empty ←

Use a template

Paste a TEI/XML file (will keep text content as well)

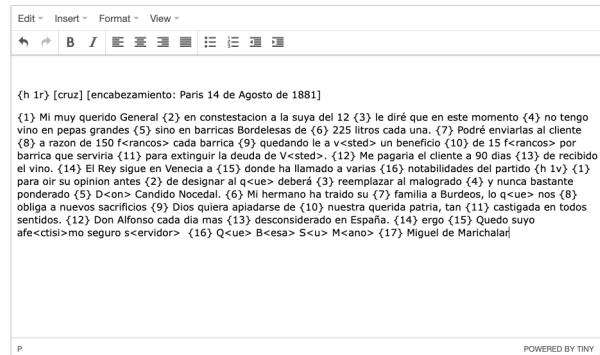
Use an existing XML file

2. Para el contenido textual, pinchamos la opción *Create as WYSIWYG* que nos permite transcribir directamente en el cuadro de texto o pegar el texto del portapapeles. Vamos a pegar la transcripción paleográfica de nuestro documento hecha conforme a los criterios CHARTA.
3. Verificar que no existan errores, que cada corchete de apertura vaya acompañado de un corchete de cierre y que las anotaciones hechas entre corchetes no se encuentren escritas en cursiva. Posteriormente, pinchamos en el botón *Create XML File* en la parte inferior izquierda de la pantalla.

Initial Content

- Leave empty
 Create as WYSIWYG (or paste from Word, HTML, etc.)

Here you can write or paste rich text, which will then be converted to TEI/XML. This conversion keeps only limited typesetting information, and can only be used for the initial creation of the XML file; after the file is in TEI/XML this editor will no longer work. You can switch to fullscreen, and drafts will be stored automatically every 20 seconds.



- Create from plain text

[Create XML File](#) ← odds here, you can also create a new XML starting from a PDF document or [Facsimile Images](#)

Esta acción crea un archivo XML mínimo con el elemento raíz <TEI>, sus dos elementos hijos, <teiHeader> y <text> y, dos elementos <p> que contienen los dos párrafos de nuestra carta. Sin embargo, este archivo XML no contiene ninguna etiqueta previamente presentadas en la guía de etiquetado. Para aplicar esta marcación vamos a pinchar en la opción **TP (CHARTA 3.0)** en el apartado *Admin options* como vemos en la imagen debajo:

COSUIZA-0013xml

[edit header data](#) • [view teiHeader](#)
 This XML has not been tokenized yet, and only the text is shown below. To edit, click [here](#).
To tokenize the text and start editing token-level attributes, select the tokenization link from the bottom of the page

{h lr} [cruz] [encabezamiento: Paris 14 de Agosto de 1881]
 {l} Mi muy querido General {2} en constestacion a la suya del 12 {3} le diré que en este momento {4} no tengo vino en pepas grandes {5} sino en barricas Bordelenses de {6} 225 litros cada una. {7} Podré enviarlas al cliente {8} a razon de 150 francos> cada barrica {9} quedando le a <usted> un beneficio {10} de 15 francos> por barrica que serviría {11} para extinguir la deuda de V<sted>. {12} Me pagaría el cliente a 90 días {13} de recibido el vino. {14} El Rey sigue en Venecia a {15} donde ha llamado a varias {16} notabilidades del partido {h 1v} {1} para oír su opinion antes {2} de designar al q<ue> deberá {3} reemplazar al malogrado {4} y nunca bastante ponderado {5} D<on> Candido Nocedal. {6} Mi hermano ha traído su {7} familia a Burdeos, lo q<ue> nos {8} obliga a nuevos sacrificios {9} Dios quiera apiadarse de {10} nuestra querida patria, tan {11} castigada en todos sentidos. {12} Don Alfonso cada dia mas {13} desconsiderado en España. {14} ergo {15} Quedo suyo afe<ctisi>mo seguro s<ervidor> {16} Q<ue> B<esa> S<u> M<ano> {17} Miguel de Marichalar

[Descargar XML](#) • [Descargar texto](#)

Admin options
 Custom actions:

[TP \(CHARTA 3.0\)](#) ←

[Tokenize the text](#) (will introduce token nodes into the XML)

Esta etapa permite marcar elementos textuales como los cambios de línea, de página, los signos y las firmas, entre otros. En la imagen debajo vemos un

mensaje subrayado, este mensaje nos indica que la secuencia de comandos se aplicó correctamente. De no hacerlo, es necesario reiniciar el proceso desde la creación del archivo XML. Si el script **TP (CHARTA 3.O)** se ha aplicado correctamente, pinchar en el enlace indicado en la parte inferior de la imagen a continuación:

Script Done

Script successfully executed. Result:

Document converted to TEI/XML (CHARTA 3.0)

- Click [here](#) to return to the XML file 

Tokenización del texto

La aplicación del script **TP (CHARTA 3.O)** nos permite visualizar los elementos que ya han sido marcados. Si hacemos clic en el botón **<pb>** podemos ver los cambios de página y, el botón **<lb>** nos permite visualizar las líneas. Para poder modificar cada palabra y poder aplicar el etiquetado morfosintáctico, debemos aplicar otro programa especial para dar a cada palabra una etiqueta **<tok>**. Para ello, pincharemos en el enlace **Tokenize the text** en la parte inferior de la página. Esta acción nos permite hacer clic en cualquiera de las palabras del texto para poder editarla.

COSUIZA-0013xml

[edit header data](#) • [view teiHeader](#)

Opciones de visualización

Mostrar: [Formato](#) <[pb>](#) <[lb>](#) 

This XML has not been tokenized yet, and only the text is shown below. To edit, click [here](#).
 To tokenize the text and start editing token-level attributes, select the tokenization link from the bottom of the page

[tr]

[cruz]

[encabezamiento: Paris 14 de Agosto de 1881]

[1] Mi muy querido General [2] en contestacion a la suya del 12 [3] le diré que en este momento [4] no tengo vino en pepas grandes [5] sino en barricas Bordelessas de [6] 225 litros cada una. [7] Podré enviarlas al cliente [8] a razon de 150 francos cada barrica [9] quedando le a usted un beneficio [10] de 15 francos por barrica que serviría [11] para extinguir la deuda de Vsted. [12] Me pagaría el cliente a 90 días [13] de recibido el vino. [14] El Rey sigue en Venecia a [15] donde ha llamado a varias [16] nobilidades del partido [17] [18] para oír su opinión antes [2] de designar al que deberá [3] reemplazar al malogrado [4] y nunca bastante ponderado [5] Don Cándido Nocedal. [6] Mi hermano ha traído su [7] familia a Burdeos, lo que nos [8] obliga a nuevos sacrificios [9] Dios quiera apiadarse de [10] nuestra querida patria, tan [11] castigada en todos sentidos. [12] Don Alfonso cada día mas [13] desconsiderado en España. [14] ergo [15] Quedo suyo afectísimo seguro servidor [16] Que Besa Su Mano [17] Miguel de Marichalar

[Descargar XML](#) • [Descargar texto](#)

Admin options

Custom actions:

[Tokenize the text](#) (will introduce token nodes into the XML) 

[Recover a previous version of this file](#)
 Last change to this file: 14 Dec 2021

Edición de token

En la imagen debajo hemos hecho clic en la palabra *Bordelessas* para poder cambiar la *B* inicial por *b*. Esto abre una página en donde podemos agregar cualquiera de las formas posibles de un *token*. En la imagen debajo hemos completado el campo **nform Critical form** con la forma de la presentación crítica. Pinchar en **Save** en la parte inferior izquierda de la página para volver a la visualización del documento.

Edit Token

Filename COSUIZA-0013.xml
Title Without Title

Token value (w-33): Bordelesas

| | |
|---------------------------------|------------|
| pform Transcription (Inner XML) | Bordelesas |
| form Transcribed form | |
| fform Full form | |
| nform Critical form | bordelas |

pos POS tag
lemma Lemma

construcción de etiquetas • lookup

insert tok after: attached / separate • before: attached / separate • insert elm before: paragraph ; linebreak • split in dtoks: 2 ; 3
edit context XML • merge left to w-32 • create mtok left: 1 ; 2
treat similar tokens

[cruz]

Paris 14 de Agosto de 1881
Mi muy querido General en constestacion a la suya del 12 le diré que en este momento no tengo vino en pepas grandes sino en barricas **Bordelesas** de 225 litros cada una. Podré enviarlas al cliente a razon de 150 *francos* cada barrica quedando le a *vsted* un beneficio de 15 *francos* por barrica que serviría para extinguir la deuda de *Vsted*. Me pagaría el cliente a 90 días de recibido el vino. El Rey sigue en Venecia a donde ha llamado a varias notabilidades del partido para oír su opinión antes de designar al que deberá reemplazar al malogrado y nunca bastante ponderado *Don* Cándido Nocedal. Mi hermano ha traído su familia a Burdeos, lo que nos obliga a nuevos sacrificios Dios quiera apiadarse de nuestra querida patria, tan castigada en todos sentidos. Don Alfonso cada dia mas desconsiderado en España. ergo Quedo suyo afe*cito*mo seguro *servidor* Que *Besa Su Mano* Miguel de Marichalar

Save • Cancel • Token Details

Como podemos ver en la imagen a continuación, solo hay que añadir una **nform** a una palabra para que se haga visible el botón que permite visualizar la presentación crítica. Debajo vemos *bordelasas* con *b* inicial como corresponde.

COSUIZA-0013.xml

edit header data • view teiHeader

Opciones de visualización

Texto: Transcripción paleográfica | Forma transcripta | Forma expandida | **Presentación crítica** - Mostrar: Colores | Formato | <pb> | <lb>

Edit the information about each word of this file by clicking on the word in the text below, or click here to edit the raw XML

[b] Paris 14 de Agosto de 1881
[i] Mi muy querido General [2] en constestacion a la suya del 12 [3] le diré que en este momento [4] no tengo vino en pepas grandes [5] sino en barricas **bordelasas** de [6] 225 litros cada una. [7] Podré enviarlas al cliente [8] a razon de 150 francos cada barrica [9] quedando le a *vsted* un beneficio [10] de 15 francos por barrica que serviría [11] para extinguir la deuda de *Vsted*. [12] Me pagaría el cliente a 90 días [13] de recibido el vino. [14] El Rey sigue en Venecia a [15] donde ha llamado a varias [16] notabilidades del partido [b] [1] para oír su opinión antes [2] de designar al que deberá [3] reemplazar al malogrado [4] y nunca bastante ponderado [5] *Don* Cándido Nocedal. [6] Mi hermano ha traído su [7] familia a Burdeos, lo que nos [8] obliga a nuevos sacrificios [9] Dios quiera apiadarse de [10] nuestra querida patria, tan [11] castigada en todos sentidos. [12] Don Alfonso cada dia mas [13] desconsiderado en España. [14] ergo [15] Quedo suyo afectísimo seguro servidor [16] Que Besa Su Mano [17] Miguel de Marichalar

Descargar XML • Descargar texto

Puntuación

Vamos a añadir dos puntos después del saludo en nuestra carta. Para ello, pinchamos en la palabra *General*, en seguida hacemos clic en **insert tok after: attached**. Esto añadirá un *tok* sin dejar un espacio en blanco antes. La plataforma reconoce esta instrucción, por lo tanto, solamente se debe agregar los dos puntos en el campo **nform Critical form** y hacer en clic en **Save**.

Edit Token

Filename COSUIZA-0013_Lxml
Title Without Title

Token value (w-10): General

| | |
|---------------------------------|---------|
| pform Transcription (Inner XML) | General |
| form Transcribed form | |
| fform Full form | |
| nform Critical form | |

pos POS tag construcción de etiquetas • lookup
 lemma Lemma

insert tok after: attached / separate • before: attached / separate • insert elm before: paragraph ; linebreak • split in dtoks: 2 ; 3
edit context XML: merge left to w-9 • create mtok left: 1 ; 2
treat similar tokens

[cruz]

Paris 14 de Agosto de 1881
 Mi muy querido **General** en constestacion a la suya del 12 le diré que en este momento no tengo vino en pepas grandes sino en barricas Bordelesas de 225 litros cada una. Podré enviarlas al cliente a razon de 150 *francos* cada barrica quedando le a *vuestro* un beneficio de 15 *francos* por barrica que serviría para extinguir la deuda de *Vsted*. Me pagaría el cliente a 90 días de recibido el vino. El Rey sigue en Venecia a donde ha llamado a varias notabilidades del partido para oír su opinion antes de designar al que deberá reemplazar al malogrado y nunca bastante ponderado *Don* Cándido Nocedal. Mi hermano ha traído su familia a Burdeos, lo que nos obliga a nuevos sacrificios Dios quiera apidiarse de nuestra querida patria, tan castigada en todos sentidos. Don Alfonso cada dia mas desconsiderado en España. ergo Quedo suyo afectísimo seguro *servidor* Q. e Besa Su Mano Miguel de Marichalar

Save **Cancel** • **Token Details**

Edit Token

Filename COSUIZA-0013.xml
Title Without Title

Token value (w-170):

| | |
|---------------------------------|--------|
| pform Transcription (Inner XML) | <dtok> |
| form Transcribed form | |
| fform Full form | |
| nform Critical form | |

pos POS tag construcción de etiquetas • lookup
 lemma Lemma

insert tok after: attached / separate • before: attached / separate • insert elm before: paragraph ; linebreak • split in dtoks: 2 ; 3
edit context XML: merge left to w-10 • create mtok left: 1 ; 2
treat similar tokens

[cruz]

Paris 14 de Agosto de 1881
 Mi muy querido General en constestacion a la suya del 12 le diré que en este momento no tengo vino en pepas grandes sino en barricas Bordelesas de 225 litros cada una. Podré enviarlas al cliente a razon de 150 *francos* cada barrica quedando le a *vuestro* un beneficio de 15 *francos* por barrica que serviría para extinguir la deuda de *Vsted*. Me pagaría el cliente a 90 días de recibido el vino. El Rey sigue en Venecia a donde ha llamado a varias notabilidades del partido para oír su opinion antes de designar al que deberá reemplazar al malogrado y nunca bastante ponderado *Don* Cándido Nocedal. Mi hermano ha traído su familia a Burdeos, lo que nos obliga a nuevos sacrificios Dios quiera apidiarse de nuestra querida patria, tan castigada en todos sentidos. Don Alfonso cada dia mas desconsiderado en España. ergo Quedo suyo afectísimo seguro *servidor* Q. e Besa Su Mano Miguel de Marichalar

Save **Cancel** • **Token Details**

Recordemos que para cambiar un signo de puntuación, simplemente se añade el signo deseado en la presentación crítica en el campo **nform Critical form**. Para eliminar un signo de puntuación de la presentación crítica, se agregan dos guiones en el campo **nform Critical form** del token que se desea eliminar.

Unión y separación irregular de palabras

La unión de palabras se trata agregando etiquetas `<dtok>` al interior de la etiqueta `<tok>`. Esta opción se presenta en la ventana de edición de token como vemos en la imagen debajo. Pincharemos en **split in dtoks: 2**.

Edit Token

Filename test.xml
Title Without Title

Token value (w-8): dela

| | | |
|-------|---------------------------|------|
| pform | Transcription (inner XML) | dela |
| form | Transcribed form | |
| fform | Full form | |
| nform | Critical form | |

pos POS tag
lemma Lemma

construcción de etiquetas • lookup

insert tok after: attached / separate • before: attached / separate • insert elm before: paragraph ; linebreak • split in dtoks: 2 ; 3
edit context XML: merge left to w-7 • create mtok left: 1 ; 2
treat similar tokens

Esto es una prueba de unión irregular dela

•

Esto crea los *dtok* con sus campos para las diferentes formas posibles. La forma de la presentación crítica se conserva en el token principal, por lo tanto, separamos las palabras en este campo. Los dos *dtok* deben tener siempre el campo **form** completo. Por último, hacemos clic en **Save**.

Edit Token

Filename test.xml
Title Without Title

Token value (w-8): dela

| | | |
|-------|---------------------------|-------|
| pform | Transcription (inner XML) | dela |
| form | Transcribed form | |
| fform | Full form | |
| nform | Critical form | de la |

pos POS tag
lemma Lemma

construcción de etiquetas • lookup

D-Token (d-8-1)

| | | |
|-------|------------------|----|
| form | Transcribed form | de |
| fform | Full form | |
| nform | Critical form | |
| pos | POS tag | |
| lemma | Lemma | |

D-Token (d-8-2)

| | | |
|-------|------------------|----|
| form | Transcribed form | la |
| fform | Full form | |
| nform | Critical form | |
| pos | POS tag | |
| lemma | Lemma | |

There are DTOK without @form, which will make the dtoks not correctly export to the CQP corpus. Please provide a written form
insert tok after: attached / separate • before: attached / separate • insert elm before: paragraph ; linebreak • add: dtok
edit context XML: merge left to w-7 • create mtok left: 1 ; 2
treat similar tokens

Esto es una prueba de unión irregular dela

•

La separación irregular de palabras se trata con la opción *merge* en la página de edición de token. Primero, se pincha en el segundo o último token de la secuencia que se desea unir.

test.xml

[edit header data](#) • [view teiHeader](#)

Opciones de visualización

Mostrar: <pb>

Edit the information about each word of this file by clicking on the word in the text below, or click [here](#) to edit the raw XML.

[b] [i] Esto es un ejemplo de separación irregular de palabra: juris **prudencia**

[Descargar XML](#) • [Descargar texto](#)

Admin options

Custom actions:

[Recover a previous version of this file](#)
Last change to this file: 14 Dec 2021
[View verticalized version of this text](#)
[\(Pre\)tag this text with POS \(and lemma\)](#)

Luego hacemos clic en *merge*:

Edit Token

Filename test.xml
Title Without Title

Token value (w-12): prudencia

| | |
|--------------------------------|-----------|
| form Transcription (Inner XML) | prudencia |
| form Transcribed form | |
| form Full form | |
| form Critical form | |

pos POS tag construcción de etiquetas • lookup
lemma Lemma

insert tok after: attached / separate • before: attached / separate • insert elm before: paragraph ; linebreak • split in dtoks: 2 ; 3
edit context XML• merge left to w-11 • create mtok left: 1 ; 2
treat similar tokens

Esto es un ejemplo de separación irregular de palabra: juris **prudencia**

[Save](#) • [Cancel](#) • [Token Details](#)

Esta acción abre una página en donde se presenta la visualización del token que resulta de la fusión. Pinchamos en **Save**.

Merging Tokens

In case a token has accidentally been split in two (for instance around a <pb/> or <lb/>) resulting merged token will typically have to be corrected manually, and will start out by It is possible to join tokens that are separate by whitespaces in the original, but bear in mind that this will result in a single token.

Before merge

Token 1: <tok id="w-11">juris</tok>
Token 2: <tok id="w-12">prudencia</tok>

After merge

Raw XML

<tok id="w-11">juris prudencia</tok>

Pre-visualization (token only)

juris prudencia

[Save](#) • [cancel](#) • [edit rawxml](#)

Finalmente, completamos la **nform Critical form** con la palabra en la forma de la presentación crítica y pinchamos en el botón **Save**.

Edit Token

Filename test.xml
Title Without Title

| | |
|-------------------------------------|--|
| Token value (w-11): juris prudencia | |
| pform | Transcription (Inner XML) <input type="text" value="juris prudencia"/> |
| form | Transcribed form <input type="text"/> |
| form | Full form <input type="text"/> |
| nform | Critical form <input type="text" value="jurisprudencia"/> ← |
| pos | POS tag <input type="text"/> |
| lemma | Lemma <input type="text"/> |

construcción de etiquetas • lookup

insert tok after: attached / separate • before: attached / separate • insert elm before: paragraph ; linebreak • split in dtoks: 2 ; 3
edit context XML • merge left to w-10 • create mtok left: 1 ; 2
treat similar tokens

Este es un ejemplo de separación irregular de palabra: juris prudencia

[Save](#) • [cancel](#) • [Token Details](#)

Etiquetado morfosintáctico y lematización

Para aplicar el etiquetado morfosintáctico es indispensable haber realizado las tareas de edición precedentes en todo el documento. Solo entonces podemos aplicar el etiquetado automático. Para ello, debemos pinchar en el enlace **(Pre)tag this text with POS (and lemma)**.

COSUIZA-0013.xml

[edit header data](#) • [view teiHeader](#)

Opciones de visualización

Texto: [Transcripción paleográfica](#) [Forma transcrita](#) [Forma expandida](#) [Presentación crítica](#) - Mostrar: Colores Formato <pb> <lb>Edit the information about each word of this file by clicking on the word in the text below, or click [here](#) to edit the raw XML

[tr]

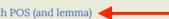
[cruz]

[encabezamiento: Paris 14 de Agosto de 1881]

[5] Mi muy querido General [2] en constestación a la suya del 12 [3] le diré que en este momento [4] no tengo vino en pepas grandes [5] sino en barricas Bordelesas de [6] 225 litros cada una. [7] Podré enviarlas al cliente [8] a razon de 150 francos cada barrica [9] quedando le a *vsted* un beneficio [10] de 15 francos por barrica que serviría [11] para extinguir la deuda de *Vsted*. [12] Me pagaría el cliente a 90 días [13] de recibido el vino. [14] El Rey sigue en Venecia [15] donde ha llamado a varias [16] notabilidades del partido [17] [18] para oír su opinión antes [2] de designar al *que* deberá [3] reemplazar al malogrado [4] y nunca bastante ponderado [5] *Dor* Cándido Nocedal. [6] Mi hermano ha traído su [7] familia a Burdeos, lo *que* nos [8] obliga a nuevos sacrificios [9] Dios quiera apiadarse de [10] nuestra querida patria, tan [11] castigada en todos sentidos. [12] Don Alfonso cada dia mas [13] desconsiderado en España. [14] ergo [15] Quedo suyo afectísimo seguro servidor [16] Que *Besa Su Mano* [17] Miguel de Marichalar

[Descargar XML](#) • [Descargar texto](#)[Admin options](#)[Custom actions:](#)[Recover a previous version of this file](#)

Last change to this file: 14 Dec 2021

[View verticalized version of this text](#)[\(Pre\)tag this text with POS \(and lemma\)](#)

El resultado que se obtenga con la aplicación de este etiquetado automático debe ser cuidadosamente revisado por el editor. Para evitar equivocaciones como la que vemos en la imagen debajo. Al apoyar el cursor sobre la palabra *vino* vemos que la categoría adjudicada automáticamente no es correcta. Hemos obtenido VMIS3SO:

en momento [4] no tengo **vino** en pepas grandes
nte [8] a razo vino
extinguir la Verbo (VMIS3SO)
amado a vari Principal; indicativo; pretérito
rado [4] y nu perfecto; tercera; singular
[8] obliga a nuevos sacrificios [9] DIOS venir
[8] obliga a nuevos sacrificios [9] DIOS quiera

| | |
|---------|--|
| POS tag | Verbo (VMIS3SO) |
| Lema | Principal; indicativo; pretérito perfecto; tercera; singular |

Ahora bien, evidentemente si examinamos el contexto, esta categoría no corresponde al no tratarse de un verbo, sino más bien de un sustantivo. Habremos de modificar este campo directamente en la página de edición del token como vemos en la siguiente imagen:

Edit Token

Filename: COSUIZA-0013_Lxml
Title: Without title

Token value (w-26): vino

| | |
|--------------------------------|------|
| form Transcription (Inner XML) | vino |
| form Transcribed form | |
| fform Full form | |
| nform Critical form | |

POS POS tag → NCMS00 construcción de etiquetas • lookup
lemma Lemma → venir

insert tok after: attached / separate • before: attached / separate • insert elm before: paragraph ; linebreak • split in dtoks: 2 ; 3
edit context XML • merge left to w-25 • create mtok left: 1 ; 2
treat similar tokens

Para mas información sobre el etiquetario pinchar aquí.

Volvemos a la visualización de nuestro documento y la viñeta presenta la información corregida.

[4] no tengo **vino** en pepas grandes [5] sino en barricas [9] vino
 a 90 días [1] POS tag Nombre Común (NCMS00)
 antes [2] de masculino; singular
 su [7] familia Lema vino
 idos. [12] Don Alfonso cada dia mas [13] desconsidera

Esta acción debe repetirse con otros fenómenos que el editor podrá encontrar al revisar el resultado del etiquetado automático. En cuanto a los pronombres atónicos, el editor deberá separar los enclíticos de los verbos utilizando la opción split in dtoks: 2.

[1] Encabezamiento. Paris 14 de Agosto de 18

[1] Mi muy querido General [2] en conste
enviarlas al cliente [8] a razon de 150 frar

enviarlas sig
 POS tag Verbo (VMN0000.PP3FPA00) on C
 Principal; infinitivo] Di
 Lema enviarlas

En la página de edición del token vamos a borrar el campo **pos POS tag** y **lemma Lemma** del token principal y completaremos los campos que correspondan en cada dtok.⁸

Edit Token

Filename COSUIZA-0013.xml
Title COSUIZA-0013

Token value (w-48): enviarlas

| | |
|---------------------------------|--|
| pform Transcription (Inner XML) | <input type="text" value="enviarlas"/> |
| form Transcribed form | <input type="text"/> |
| fform Full form | <input type="text"/> |
| nform Critical form | <input type="text"/> |
| pos POS tag | <input type="text"/>  |
| lemma Lemma | <input type="text"/>  |

D-Token (d-48-1)

| | |
|-----------------------|--|
| form Transcribed form | <input type="text" value="enviar"/>  |
| fform Full form | <input type="text"/> |
| nform Critical form | <input type="text"/> |
| pos POS tag | <input type="text" value="VMN0000"/>  |
| lemma Lemma | <input type="text" value="enviar"/>  |

[delete this dtok](#)

D-Token (d-48-2)

| | |
|-----------------------|---|
| form Transcribed form | <input type="text" value="las"/>  |
| fform Full form | <input type="text"/> |
| nform Critical form | <input type="text"/> |
| pos POS tag | <input type="text" value="PP3FPA00"/>  |
| lemma Lemma | <input type="text" value="lo"/>  |

En la página de visualización del documento podemos ver la información correcta aparecer en la viñeta del token.

⁸ Ante cualquier duda respecto al etiquetado se recomienda encarecidamente consultar el manual concebido para el proyecto de P.S. Post Scriptum (Vaamonde 2018)

| | | |
|------|------------------|-----------------------------|
| a [| enviarlas | [12] |
| > ha | | el] |
| [4] | enviar | Ca |
| cri | Forma transcrita | n |
| ons | POS tag | lo |
| | Lema | enviar |
| | las | |
| | Forma | |
| | transcrita | las |
| | POS tag | Pronombre (PP3FPA00) |
| | | Personal; tercera; |
| | | femenino; plural; acusativo |
| | Lema | lo |

Facsimil

Para adjuntar el facsímil de cada página a nuestro documento vamos a pinchar en el botón que nos permite visualizar los cambios de página (<pb>). Inmediatamente después hacemos clic en el indicador de página [1r].

COSUIZA-0013.xml

[edit header data](#) • [view teiHeader](#)

Opciones de visualización

Texto: Transcripción paleográfica | Forma transcrita | Forma expandida | Presentación crítica - Mostrar: Colores | Formato: <pb> <lb> - Etiquetas: POS tag | Lema

Edit the information about each word of this file by clicking on the word in the text below, or click [here](#) to edit the raw XML

[b]  [cruz]

[encabezamiento: Paris 14 de Agosto de 1881]
 [1] Mi muy querido General [2] en contestacion a la suya del 12 [3] le diré que en este momento [4] no tengo vino en
 pepas grandes [5] sino en barricas Burdelesas de [6] 225 litros cada una. [7] Podré enviarlas al cliente [8] a razon de 150
 francos cada barrica [9] quedando le a *vsted* un beneficio [10] de 15 francos por barrica que serviría [11] para extinguir la
 deuda de *Vsted*. [12] Me pagaría el cliente a 90 dias [13] de recibido el vino. [14] El Rey sigue en Venecia a [15] donde ha
 llamado a varias [16] notabilidades del partido [16] [1] para oír su opinion antes [2] de designar al *que* deberá [3]
 reemplazar al malogrado [4] y nunca bastante ponderado [5] *Don* Candido Nocedal. [6] Mi hermano ha traído su [7]
 familia a Burdeos, lo *que* nos [8] obliga a nuevos sacrificios [9] Dios quiera apiadarse de [10] nuestra querida patria, tan [11]
 castigada en todos sentidos. [12] Don Alfonso cada dia mas [13] desconsiderado en España. [14] ergo [15] Quedo suyo
 afe*cus*imo seguro *servidor* [16] *Que Besa Su Mano* [17] Miguel de Marichalar

En la página de edición que se abrirá vamos a pinchar en el enlace ([see list](#)) como se muestra a continuación:

Edit Element

Structural element (e-1): pb

n Page number 

facsimile image (see list) 

admin Admin-only image 

[Save](#) [Cancel](#)

Se abrirá una nueva pestaña en donde se podrán buscar y adjuntar archivos desde la computadora. Antes de iniciar esta acción se debe denominar el archivo con la secuencia estándar concebida para los facsímiles como se presentó anteriormente en este documento (en 1.1.1). Para ello, se debe pinchar en el botón **Choisir un fichier**, navegar hasta la ubicación del archivo y una vez seleccionado hacer clic en **Save**.

Stored Facsimile images

Add new image Aucun fichier choisi

suggested image filename: Filename-10v.jpg - where Filename is the name of Filenamexml and 10v is the page number
 To insert an image into a <pb> just copy the desired filename from the list below

El archivo se habrá guardado en el directorio destinado a conservar los facsímiles en el servidor del COSUIZA. Ingresamos el nombre del archivo —incluida su extensión— al campo **Facsimile image** y hacemos clic en **Save**.

Edit Element

Structural element (e-1): pb

n Page number facs Facsimile image (see list)
admin Admin-only image

[cruz:

Paris 14 de Agosto de 1881

Mi muy querido General en constestacion a la suya del 12 le diré que en este momento no tengo vino en pepas grandes sino en barricas Bordelenses de 225 litros cada una. Podré enviarlas al cliente a razón de 150 francos cada barrica quedando le a usted un beneficio de 15 francos por barrica que serviría para extinguir la deuda de Vsted. Me pagaría el cliente a 90 días de recibido el vino. El Rey sigue en Venecia a donde ha llamado a varias notabilidades del partido para oír su opinión antes de designar al que deberá reemplazar al malogrado y nunca bastante ponderado Don Candido Nocedal. Mi hermano ha traído su familia a Burdeos, lo que nos obliga a nuevos sacrificios Dios quiera apañarse de nuestra querida patria, tan castigada en todos sentidos. Don Alfonso cada día mas desconsiderado en España. ergo Quedo suyo afectísimo seguro servidor Que Besa Su Mano Miguel de Marichalar]

Terminada esta etapa, podremos ir a la página de visualización de nuestro documento y el botón **Imágenes** nos permitirá visualizar el facsímil.

COSUIZA-0013.xml

[edit header data](#) • [view teiHeader](#)

Opciones de visualización

Texto: Transcripción paleográfica | Forma transcrita | Forma expandida | Presentación crítica - Mostrar: Colores Formato <pb> <db> Imágenes - Etiquetas:
POS tag | Lema

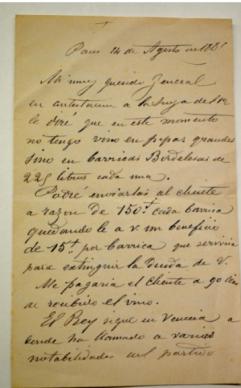
Edit the information about each word of this file by clicking on the word in the text below, or click [here](#) to edit the raw XML.

[r]

[cruz]

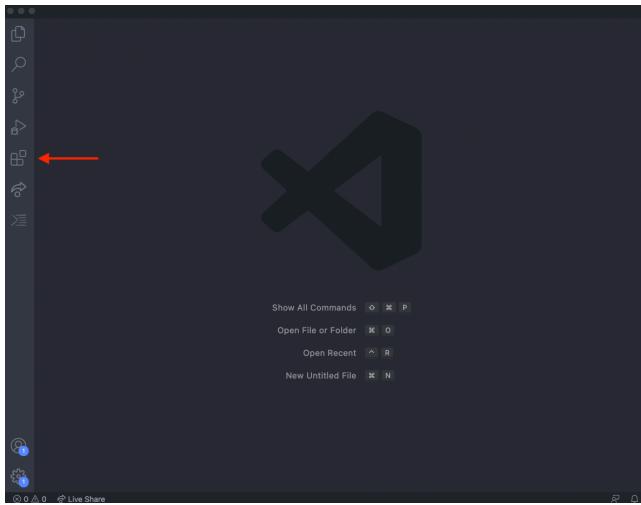
[encabezamiento: Paris 14 de Agosto de 1881]

[1] Mi muy querido General [2] en constestacion a la suya del 12 [3] le diré que en este momento [4] no tengo vino en pepas grandes [5] sino en barricas Bordelenses de [6] 225 litros cada una. [7] Podré enviarlas al cliente [8] a razón de 150 francos cada barrica [9] quedando le a usted un beneficio [10] de 15 francos por barrica que serviría para extinguir la deuda de Vsted. [11] Me pagaría el cliente a 90 días [12] de recibido el vino. [13] El Rey sigue en Venecia a [14] donde ha llamado a varias [15] notabilidades del partido [16] para oír su opinión antes [17] de designar al que deberá [18] reemplazar al malogrado [19] y nunca bastante ponderado [20] Don Candido Nocedal. [21] Mi hermano ha traído su [22] familia a Burdeos, lo que nos [23] obliga a nuevos sacrificios [24] Dios quiera apañarse de [25] nuestra querida patria, tan [26] castigada en todos sentidos. [27] Don Alfonso cada día mas [28] desconsiderado en España. [29] ergo [30] Quedo suyo afectísimo seguro servidor [31] Que Besa Su Mano [32] Miguel de Marichalar

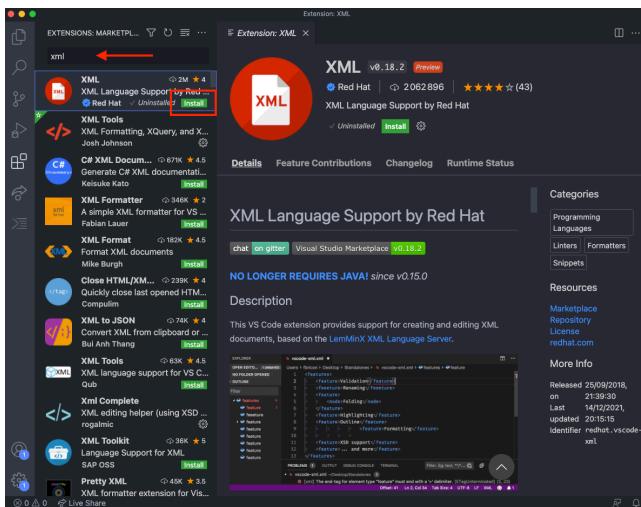


Metadatos

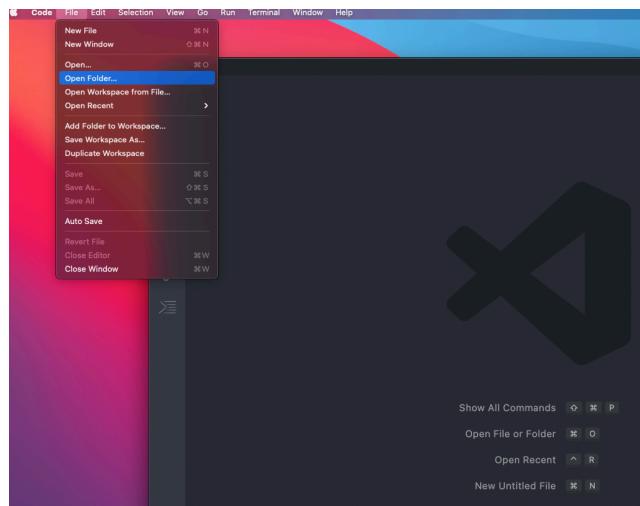
En este apartado vamos a agregar la cabecera de nuestro documento. Comenzaremos instalando el editor de código fuente Visual Studio Code. Tras instalar y abrir el programa, vamos a pinchar en la pestaña de las extensiones como se muestra en la siguiente figura:



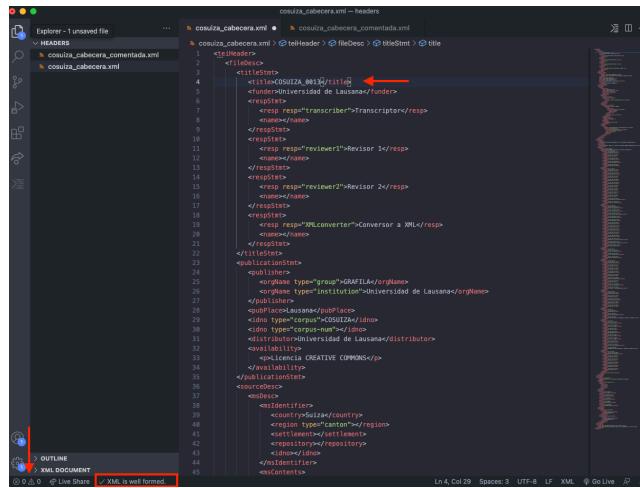
Vamos a escribir *xml* en el cuadro de texto y vamos a pinchar en el botón **install**. Esta extensión nos ayudará a comprobar que nuestra cabecera está bien estructurada.



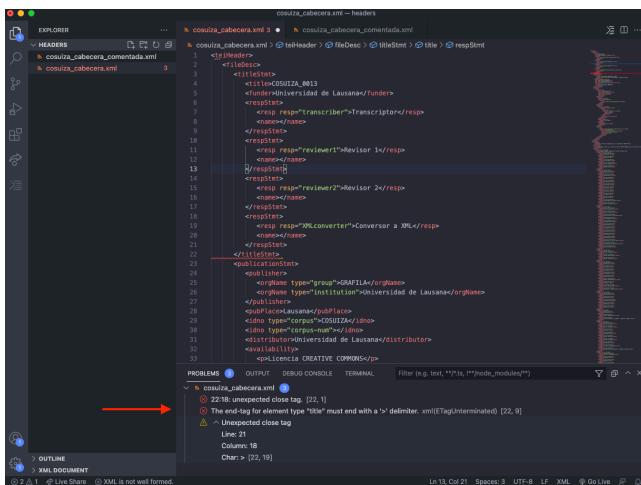
Luego, vamos a abrir nuestra cabecera con la ayuda de este editor de código fuente. Primero, pincha en este enlace. La carpeta que se descargará contiene dos documentos XML, uno es la cabecera que usaremos para completar los datos correspondientes a nuestro documento y el otro es el mismo, sin embargo, todos los campos llevan comentarios para recordar la finalidad de cada etiqueta. Vamos a abrir la carpeta que hemos descargado en nuestro editor de código fuente como vemos en la figura debajo o, simplemente arrastrando la carpeta sobre el icono de Visual Studio Code.



Una vez abierto el archivo XML podemos agregar los datos directamente en nuestro editor de código. En la figura debajo podemos ver que dentro de la etiqueta `<teiHeader>` hemos completado con `COSUIZA-0013`, el nombre de archivo de nuestra carta. En la parte inferior izquierda se indica la cantidad de errores existentes en el archivo y, en la barra inferior, aparece la información respecto a la validez del archivo. En nuestro caso, nuestro archivo es válido.



En el caso de existir errores, estos aparecen señalados en la parte inferior. En la figura debajo, a nuestro elemento `<title>` le falta la etiqueta de cierre.



Después de haber agregado todos los datos pertinentes, vamos a copiar el texto de nuestra cabecera en el portapapeles y vamos a ir a nuestro documento en el COSUIZA.

Vamos a pinchar en la opción que nos permite visualizar el archivo XML sin formato como se muestra en la figura debajo:

The screenshot shows a web-based XML viewer for the file 'COSUIZA-0013.xml'. At the top, there are links for 'edit header data' and 'view teiHeader'. Below that is a section titled 'Opciones de visualización' with various options like 'Transcripción paleográfica', 'Forma transcrita', 'Forma expandida', etc. A prominent red arrow points to the 'View full XML' link at the top right. The main content area shows the XML code with some parts collapsed, indicated by a minus sign icon.

Después pinchamos en el enlace **Switch to full XML including header** como vemos a continuación:

En seguida podremos ver el elemento `<teiHeader>` que se constituye automáticamente al crear un archivo en Teitok, pero que no contiene la información ni la estructura de nuestra cabecera presentada en el apartado de los § metadatos.

COSUIZA-0013_1.xml



```

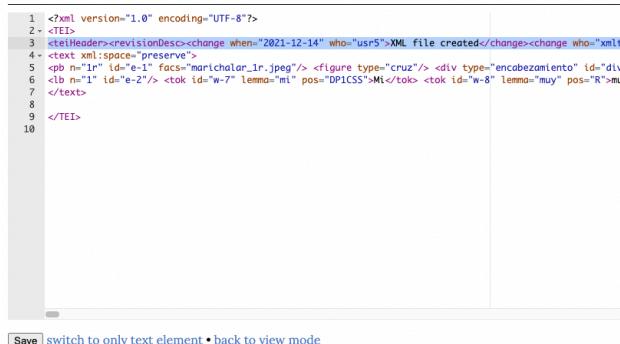
1 <text xml:space="preserve">
2 <pb n="1" id="e-1" facs="marichalar_1r.jpeg"/> <figure type="cruz"/> <div type="encabezamiento" id="div-
3 <lb n="1" id="e-2"/> <tok id="w-7" lemma="mI" pos="DPICSS">Ml</tok> <tok id="w-8" lemma="muy" pos="R">muy
4 </text>

```

Save switch to full XML including header • back to view mode

Vamos a seleccionar el elemento `<teiHeader>` desde su etiqueta de apertura hasta su etiqueta de cierre y la vamos a eliminar. En su lugar vamos a pegar la cabecera que hemos creado en el editor de código fuente.

COSUIZA-0013_1.xml



```

1 <?xml version="1.0" encoding="UTF-8"?>
2 <TEI>
3 <teiHeader><revisionDesc><change when="2021-12-14" who="usr5">XML file created</change><change who="xmit
4 <text xml:space="preserve">
5 <pb n="1r" id="e-1" facs="marichalar_1r.jpeg"/> <figure type="cruz"/> <div type="encabezamiento" id="div-
6 <lb n="1" id="e-2"/> <tok id="w-7" lemma="mI" pos="DPICSS">Ml</tok> <tok id="w-8" lemma="muy" pos="R">muy
7 </text>
8
9 </teiHeader>
10 </TEI>

```

Save switch to only text element • back to view mode

Finalmente pinchamos en **Save**.



```

322 <term>vino</term>
323 <term>barriada</term>
324 <term>Alfonso XII</term>
325 <term>España</term>
326 </keywords>
327 </textClass>
328 <particDesc>
329 <person role="sent" sex="M">
330 <personName>Miguel de Marichalar</personName>
331 </person>
332 </particDesc>
333 </profileDesc>
334 <revisionDesc>
335 <change when="2021-10-24" who="AEC">Creación del archivo</change>
336 <change when="aaa-mm-dd">Revisión final de la presentación crítica</change>
337 </revisionDesc>
338 </text>
339 <text xml:space="preserve">
340 <pb n="1" id="e-1" facs="marichalar_1r.jpeg"/> <figure type="cruz"/> <div type="encabezamiento" id="
341 <lb n="1" id="e-2"/> <tok id="w-7" lemma="mI" pos="DPICSS">Ml</tok> <tok id="w-8" lemma="muy" pos="R">muy
342 </text>
343 </TEI>
344
345

```

Save switch to only text element • back to view mode

Al volver a nuestra página de visualización de nuestro documento podemos ver en la parte superior los metadatos pertinentes.

COSUIZA-0013.xml

Archivo: BCUF
Referencia: B-126
Fecha: 1881 agosto 14
Lugar: París
Resumen: Carta de Miguel de Marchilar dirigida a un general para tratar asuntos relativos a la venta de unas barricas de vino

[edit header data](#) • más metadatos • [view teiHeader](#)

Opciones de visualización

Texto: Transcripción paleográfica [Forma transcrita](#) [Forma expandida](#) [Presentación crítica](#) - Mostrar: Colores Formato <pb> <bp> Imágenes - Etiquetas: POS tag Lema

Edit the information about each word of this file by clicking on the word in the text below, or click [here](#) to edit the raw XML.

[cruz]
[encabezamiento: París 14 de Agosto de 1881]
Mi muy querido General en constestacion a la suya del 12 le diré que en este momento no tengo vino en pepas grandes sino en barricas Bordelesas de 225 litros cada una. Podré enviarlas al cliente a razon de 150 f cada barrica quedando le a v un beneficio de 15 f por barrica que serviría para extinguir la deuda de V. Me pagaría el cliente a 90 días de recibido el vino. El Rey sigue en Venecia a donde ha llamado a varias notabilidades del partido para oír su opinión antes de designar al q deberá reemplazar al malogrado y nunca bastante ponderado D Cándido Nocedal. Mi hermano ha traído su familia a Burdeos, lo q nos obliga a nuevos sacrificios Dios quiera apañárselas de nuestra querida patria, tan castigada en todos sentidos. Don Alfonso cada dia mas desconsiderado en España. ergo Quedo suyo afemo seguro s Q B S M Miguel de Marchilar

[Descargar XML](#) • [Descargar texto](#)

Videotutorial

Acceso

Creación de archivo XML y tokenización

Edición de token

Puntuación

Unión y separación irregular de palabras

Etiquetado morfosintáctico y lematización

Facsímil

Bibliography

Elena Diez Del Corral Areta and Leyre Martín Aizpuru. Sin corpus no hay historia: la red charta como un proyecto de edición común. *Cuadernos de lingüística de El Colegio de México*, 2:287–314, 2014.

Maarten Janssen. Teitok: Text-faithful annotated corpora. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4037–4043, 2016.

Carlota Kabatek, Johannes y de Benito Moreno. *Lingüística de corpus y lingüística histórica iberorrománica*. de Gruyter, 2016.

Carmen Isasi Martínez, Leyre Martín Aizpuru, Santiago Pérez Isasi, Elena Pierazzo, and Paul Spence. *Edición digital de documentos antiguos: marcación XML-TEI basada en los criterios CHARTA*. Universidad de Sevilla, 2020.

Gael Vaamonde. Ps post scriptum: Dos corpus diacrónicos de escritura cotidiana. *Procesamiento del lenguaje natural*, (55):57–64, 2015.