

Extracción y uso de las etimologías del Wikcionario en inglés  
específicamente  
para MS Office Word

y  
para Libre Office Writer

Víctor Fresco Barbeito

Salamanca, sábado 11 de noviembre de 2023



**X Jornadas Anuales  
Salamanca 2023**  
Cultura libre y patrimonio abierto  
10 - 12 de noviembre



**VNIVERSIDAD  
D SALAMANCA**



# CONTENIDOS

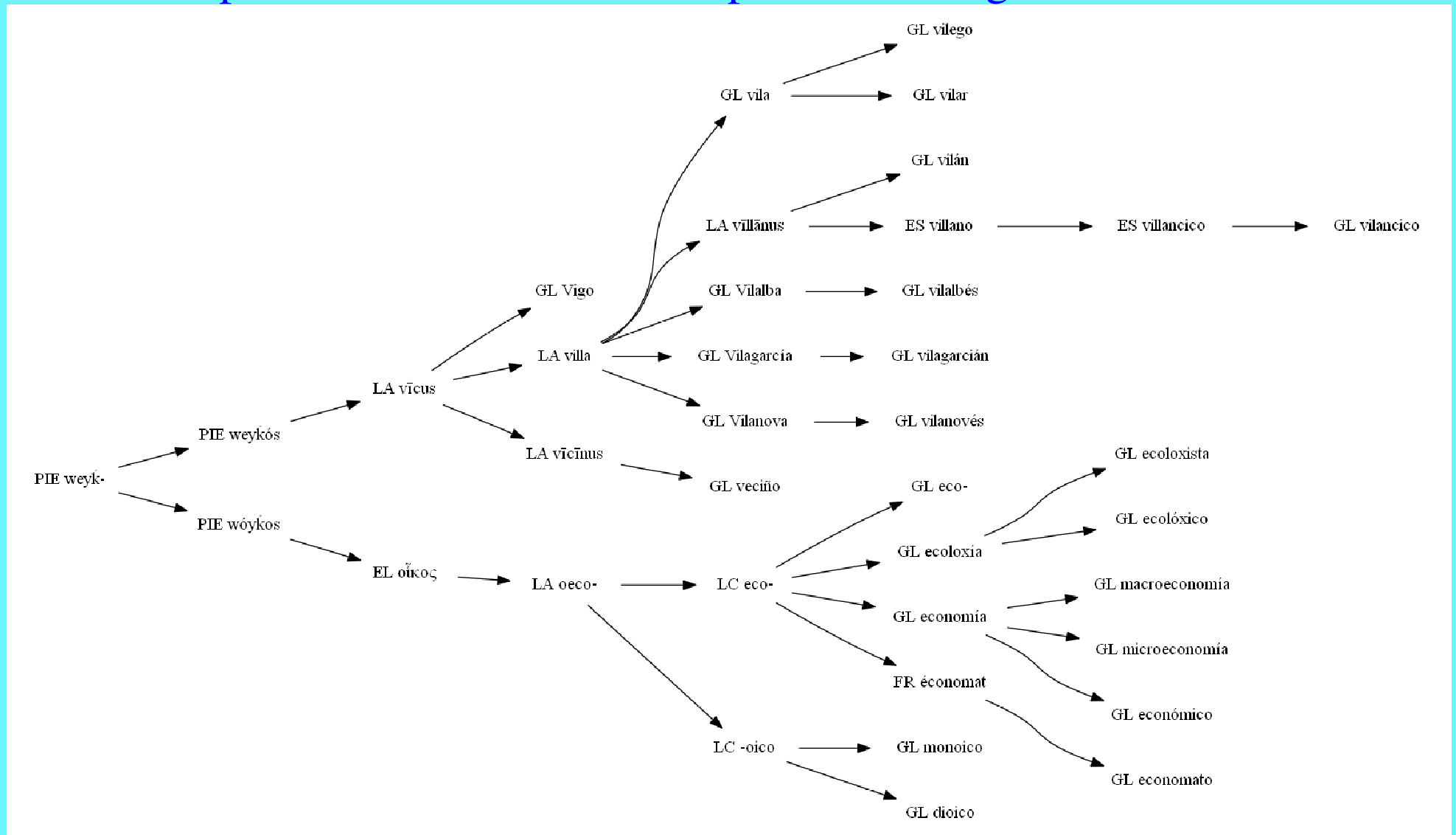
(aquí)	1
OBJETIVO	2
ANTECEDENTES	4
EXPORTACIÓN	10
SELECCIÓN	14
LO EXPORTADO:XML	17
PROCESAMIENTO <1>	24
PROCESAMIENTO ==2==	37
PROCESAMIENTO ===3===	48
PRECAUCIONES	56
OBJECIONES	57
APÉNDICE para MSOW	58
APÉNDICE para OLOW	61
PARA LA GENTE DE MACROS (MSOW)	65

/65

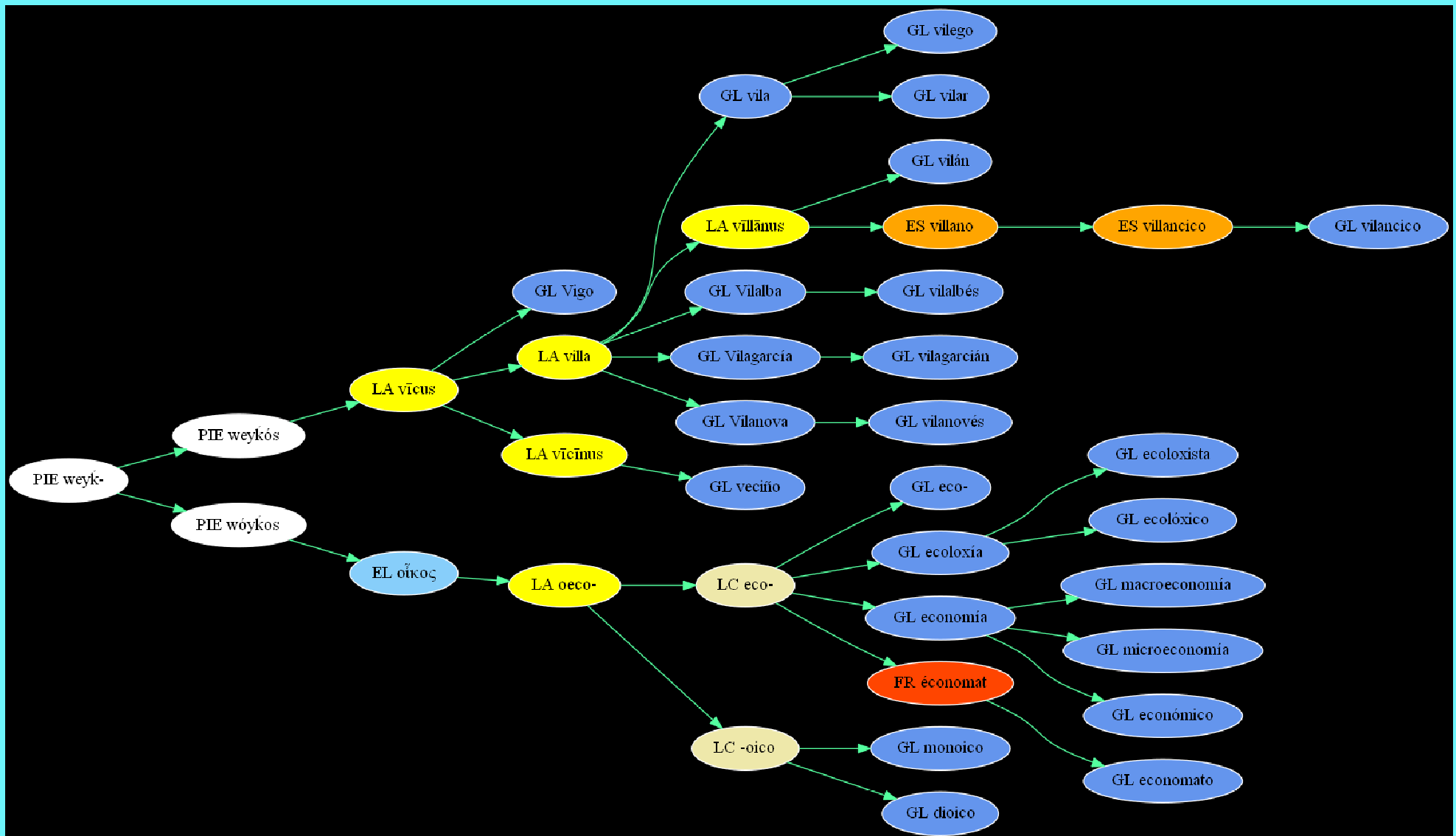
# OBJETIVO

Imaginemos que queremos extraer las etimologías de en.wikt y relacionarlas.

Al representar sus derivaciones podríamos llegar a obtener esto:



Lo cual, si utilizásemos códigos de color identificativos y mejorásemos un poco su estética daría lo siguiente:



# ANTECEDENTES




[https://github.com/esterpantaleo/viz\\_galicia/blob/master/viz\\_galicia\\_etymtree.png](https://github.com/esterpantaleo/viz_galicia/blob/master/viz_galicia_etymtree.png)  
[https://esterpantaleo.github.io/viz\\_galicia/](https://esterpantaleo.github.io/viz_galicia/)

y para este ejemplo aprovechamos una herramienta,

# resultado de una beca Individual Engagement Grant de la Wikimedia Foundation en 2016, llamada *etytree*.

← → ↻ [esterpantaleo.github.io](https://esterpantaleo.github.io)

 **Ester Pantaleo** [github.com/esterpantaleo](https://github.com/esterpantaleo)

I'm a researcher at the University of Bari, Italy, in the Department of Physics within the Medical Physics Group led by prof. Roberto Bellotti. I hold a PhD in Physics from the University of Bari, Italy. During my career I have done research in interdisciplinary fields like **Genomics**, **Econometrics**, **Natural Language Processing**, **Image Analysis**. I am also very much interested in the communication of technical and scientific content through writing and/or interactive visualizations.

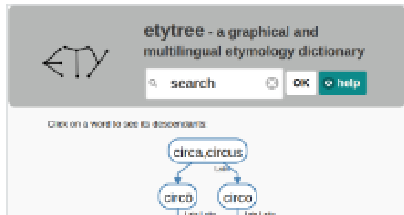
In this website I'm collecting some projects I have developed as a researcher at the Universities of Bari, Palermo, Cambridge and Chicago, or as a freelancer for the Adler Planetarium, Chicago, the Wikimedia Foundation, the Italian Government and the Agronomic Institute CIHEAM IAMB. Among other tools I developed [etytree](#) an interactive tool for the visualization of the etymology of words with an associated RDF database, funded by an IEG grant of the Wikimedia Foundation. If you have comments on my work please do not hesitate to contact me at [esterpantaleo at gmail dot com](mailto:esterpantaleo@gmail.com).

[tools](#) [visualizations](#) [research](#)

## tools

### etytree

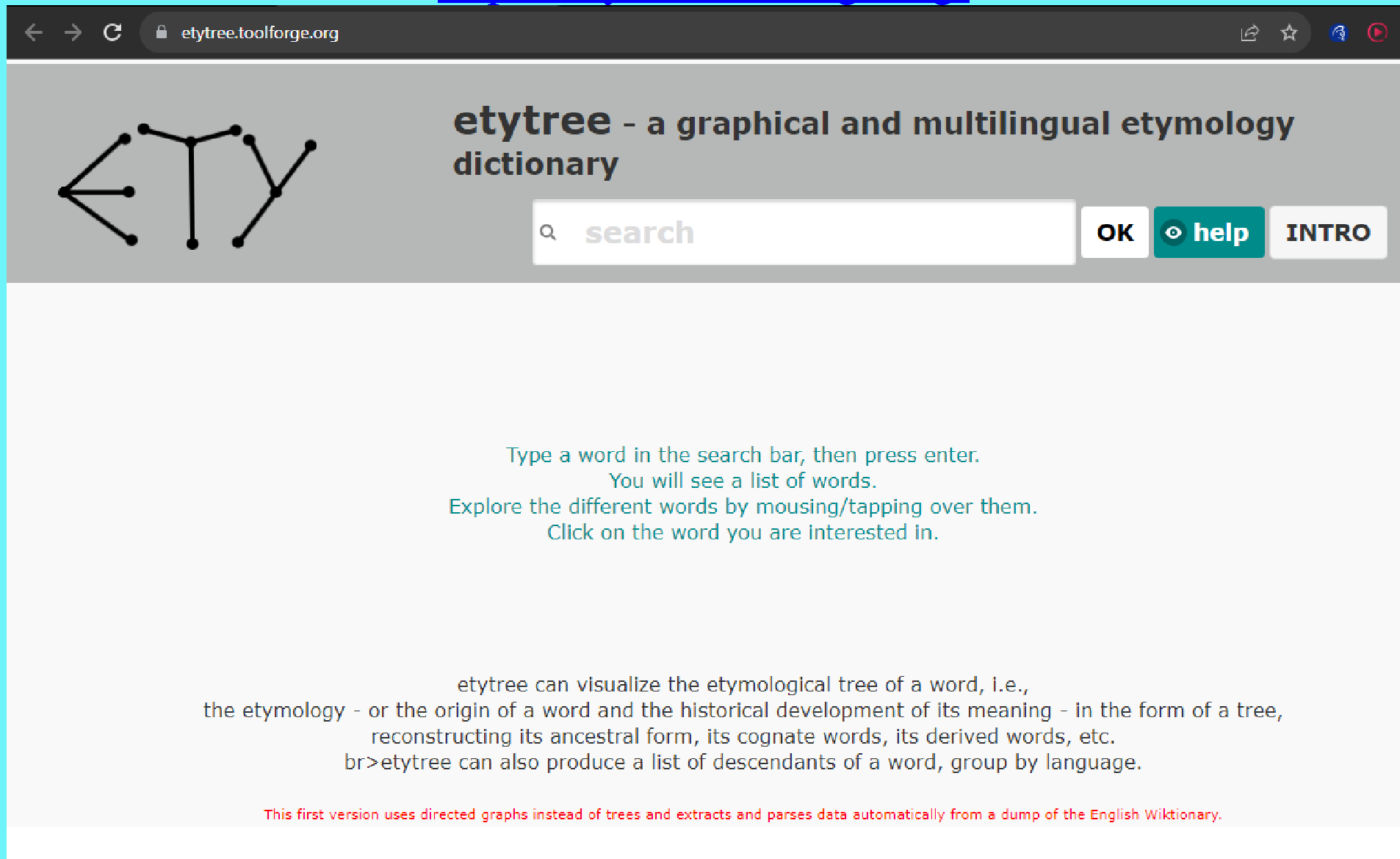
[Website](#) · [GitHub](#) · [Bitbucket](#) · [grant](#) · [demo](#)



<https://esterpantaleo.github.io/>

Esta herramienta es de código abierto:

<https://etytree.toolforge.org/>



etytree - a graphical and multilingual etymology dictionary

search OK help INTRO

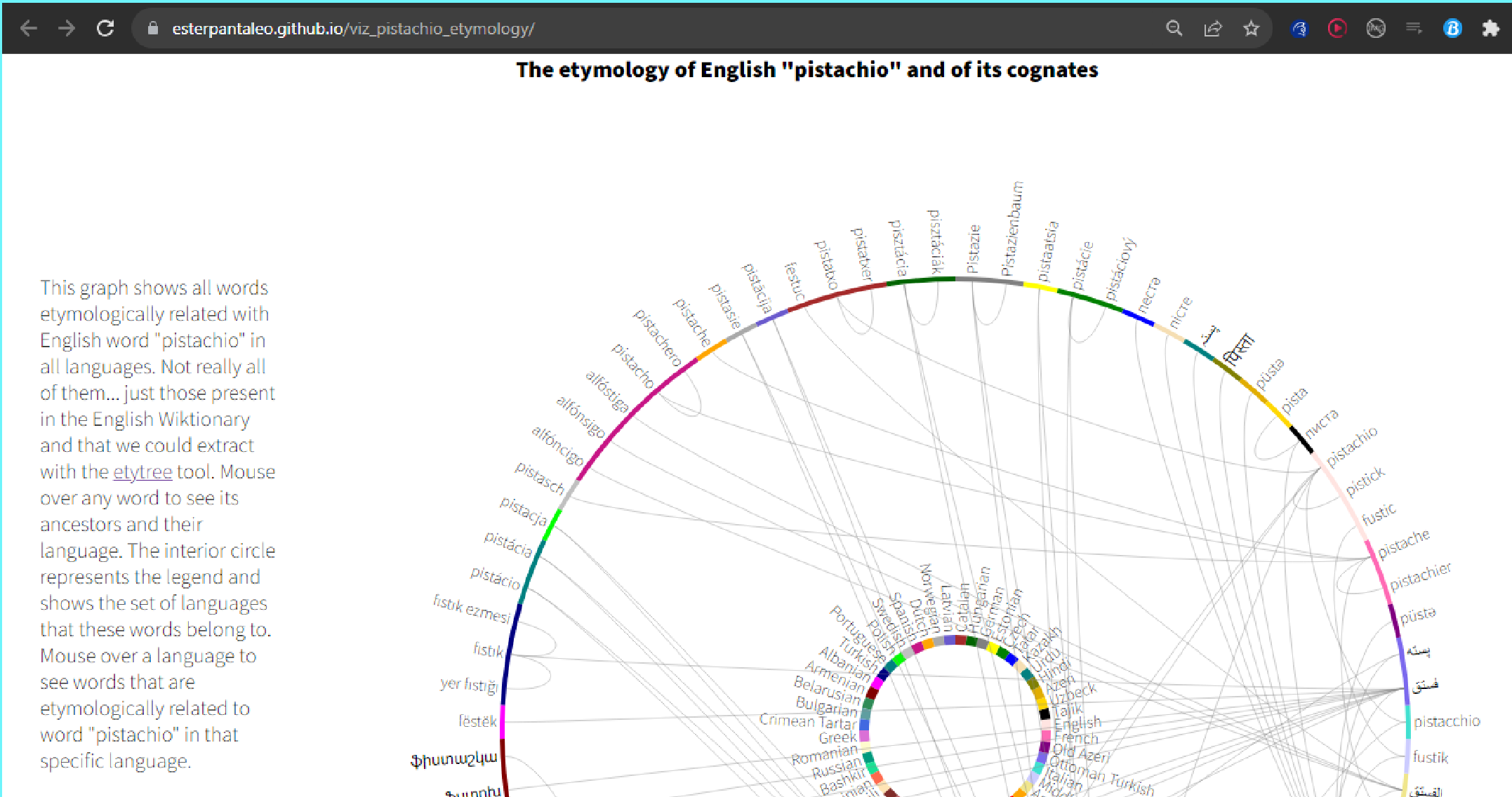
Type a word in the search bar, then press enter.  
You will see a list of words.  
Explore the different words by mousing/tapping over them.  
Click on the word you are interested in.

etytree can visualize the etymological tree of a word, i.e., the etymology - or the origin of a word and the historical development of its meaning - in the form of a tree, reconstructing its ancestral form, its cognate words, its derived words, etc.  
br>etytree can also produce a list of descendants of a word, group by language.

This first version uses directed graphs instead of trees and extracts and parses data automatically from a dump of the English Wiktionary.

(también en WM labs: <https://tools.wmflabs.org/etytree/>, que redirecciona al anterior)

# Tiene sus derivadas como esta:



[https://esterpantaleo.github.io/viz\\_pistachio\\_etymology/](https://esterpantaleo.github.io/viz_pistachio_etymology/)





Esto, sin embargo, plantea dos problemas:

- 1) tenemos una dependencia externa (y no sólo por la red)
- 2) no tenemos control de los datos y de su formato para otros usos

(y 3)



Pues bien, lo que pretendo es darles a ustedes mismos la autonomía de poder extraer la etimología y hacer lo que les plazca con ella.

(OBJETIVO SECUNDARIO)

Este ejercicio teórico-práctico vale para cualquier wiki, para cualquier formato, para cualquier texto, para cualquier sección.

## **EXPORTACIÓN:**

En donde se trata de cómo escojo los datos y cómo los obtengo.

Como herramienta básica vamos a usar una herramienta de la propia wiki:  
<https://en.wiktionary.org/wiki/Special:Export>

Special page

## Special pages

Contents [hide]

- 1 Maintenance reports
- 2 Lists of pages
- 3 Account management
- 4 Users and rights
- 5 Recent changes and logs
- 6 Media reports and uploads
- 7 Data and tools
- 8 Redirecting special pages
- 9 High use pages
- 10 Page tools
- 11 Spam tools
- 12 Other special pages

Maintenance reports

- Broken redirects
- Dead-end pages
- Double redirects
- Lint errors
- Long pages
- Oldest pages
- Short pages
- Uncategorized categories
- Uncategorized files
- Uncategorized pages
- Uncategorized templates
- Unused categories

Page tools

- Book
- Cite This Page
- Compare pages
- Export pages
- URL Shortener
- What links here

Legend

- Normal special pages.
- Restricted special pages.

Wiktionary  
The free dictionary

Main Page  
Community portal  
Preferences  
Requested entries  
Recent changes  
Random entry  
Help  
Glossary  
Donations  
Contact us

Tools

Upload file  
Special pages  
Printable version  
Get shortened URL

Languages

If you have time, leave us a note.

[https://es.wiktionary.org/wiki/Especial:PáginasEspeciales#Herramientas\\_de\\_páginas](https://es.wiktionary.org/wiki/Especial:PáginasEspeciales#Herramientas_de_páginas)  
(en realidad #mw-specialpagesgroup-pagetools)

Lo cual nos lleva a:

<https://en.wiktionary.org/wiki/Special:Export>

The screenshot shows the English Wiktionary Special:Export page. The browser address bar displays [en.wiktionary.org/wiki/Special:Export](https://en.wiktionary.org/wiki/Special:Export). The page title is "Export pages". The main content area contains the following text:

You can export the text and editing history of a particular page or set of pages wrapped in some XML. This can be imported into another wiki using MediaWiki via the [import page](#).

To export pages, enter the titles in the text box below, one title per line, and select whether you want the current revision as well as all old revisions, with the page history lines, or the current revision with the info about the last edit.

In the latter case you can also use a link, for example `Special:Export/Wiktionary:Main Page` for the page "Wiktionary:Main Page".

Below the text, there are two sections for adding pages:

- Add pages from category:** A text input field followed by an "Add" button.
- Add pages manually:** A large text area for manual entry.

At the bottom, there are three checkboxes:

- Include only the current revision, not the full history
- Include templates
- Save as file

An "Export" button is located below the checkboxes. The footer of the page includes links for Privacy policy, About Wiktionary, Disclaimers, Code of Conduct, Developers, Statistics, Cookie statement, and Mobile view. It also features the Wikimedia Project logo and the "Powered by MediaWiki" logo.

<https://es.wiktionary.org/wiki/Especial:Exportar>

Si nos fijamos, podremos ver "Añadir páginas desde la categoría".  
Con esa opción podremos listar las entradas de una categoría que las agrupe.

En general, esta clasificación está bastante lograda en  
todos los proyectos y todas las lenguas.

Pero:

¿que pasa si queremos las páginas de una categoría y de sus subcategorías?

¿que pasa si NO queremos todas las páginas de una categoría?

¿que pasa si la categoría es enorme?

(y nos colapsa o ralentiza la descarga,

o queremos tener cierta consideración de la cortesía para con el servidor)

# SELECCIÓN

Pues que ahora tenemos un problema: ¿con qué rellenar las páginas manualmente?  
Si no hubiese solución, no estaríamos aquí. Está otra vez en WM labs:

The screenshot shows the PetScan tool interface on [petscan.wmflabs.org](https://petscan.wmflabs.org/). The browser address bar shows the URL. The page has a navigation menu with 'PetScan', 'Manual', and 'Issues'. A language selector is set to 'English'. There are links for 'Examples' and 'List'. Below the navigation, there are tabs for 'Categories', 'Page properties', 'Templates&links', 'Other sources', 'Wikidata', and 'Output'. The 'Categories' tab is active. The form fields are: 'Language' (en), 'Project' (wikipedia), 'Depth' (0), 'Categories' (empty), 'Combination' (Intersection selected), and 'Negative categories' (empty). A 'Do it!' button is located at the bottom left.

<https://petscan.wmflabs.org/>

PetScan consiste en una herramienta para extraer datos, para elegir los títulos de las páginas con ciertas condiciones.

Los criterios de selección son múltiples:

- lengua, proyecto y espacio
- tamaño
- redirección
- contenido
- WikiData

con flexibilidad hasta para hacer consultas SQL

Podemos escoger el formato de exportación en el que queremos los resultados de la consulta.

A nosotros no nos interesa más que la lista sencilla para llevarla a Export, es decir, con formato Output: Plain text

Y entonces copiamos la lista y la pegamos en Export, le damos al botón y obtenemos un fichero XML, que abriremos con nuestro editor de texto.



Podríamos abrirlo con el bloc de notas,  
pero podemos hacerlo directamente con el editor con el que trabajaremos:

Para MS Word:

*File > Open > [Files of type: All files]*

*Fichero > Abrir > [Ficheros de tipo: Todos]*

Buscamos nuestro XML y abrimos.

Entonces preguntará el formato en el que tenemos el fichero.

La conversión será *Unicode (UTF-8)* [están alfabeticamente]

Para Libre Office Write:

*Archivo > Abrir*

Buscamos nuestro XML y abrimos.

[a mí ya ni me pide tipo de archivo ni formato de conversión]

(parece que OOLOW da problemas cuando se abre directamente, por no preguntar:  
si se abre con un editor simple -Bloc de notas- y copia-pegamos el contenido, ya no)

También da problemas debido a los colores (que vamos a usar después)

así que seleccionamos todo y lo pasamos a rojo.

# Y ahora, antes de seguir, vamos a observar detenidamente lo que hemos obtenido:

```
<mediawiki
xmlns="http://www.mediawiki.org/xml/export-0.10/"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.mediawiki.org/xml/export-0.10/http://www.mediawiki.org/xml/export-0.10.xsd"
version="0.10" xml:lang="en">
```

## <siteinfo>

```
<sitename>Wiktionary</sitename>
<dbname>en-wiktionary</dbname>
<base>https://en.wiktionary.org/wiki/Wiktionary:Main_Page</base>
<generator>MediaWiki 1.42.0-wmf.3</generator>
<case-sensitive</case>

<namespaces>
<namespace key="2" case="case-sensitive">Media</namespace>
<namespace key="1" case="first-letter">Special</namespace>
<namespace key="0" case="case-sensitive" />
<namespace key="1" case="case-sensitive">Talk</namespace>
<namespace key="2" case="first-letter">User</namespace>
<namespace key="3" case="first-letter">User talk</namespace>
<namespace key="4" case="case-sensitive">Wiktionary</namespace>
<namespace key="5" case="case-sensitive">Wiktionary talk</namespace>
<namespace key="6" case="case-sensitive">File</namespace>
<namespace key="7" case="case-sensitive">File talk</namespace>
<namespace key="8" case="first-letter">MediaWiki</namespace>
<namespace key="9" case="first-letter">MediaWiki talk</namespace>
<namespace key="10" case="case-sensitive">Template</namespace>
<namespace key="11" case="case-sensitive">Template talk</namespace>
<namespace key="12" case="case-sensitive">Help</namespace>
<namespace key="13" case="case-sensitive">Help talk</namespace>
<namespace key="14" case="case-sensitive">Category</namespace>
<namespace key="15" case="case-sensitive">Category talk</namespace>
<namespace key="90" case="case-sensitive">Thread</namespace>
<namespace key="91" case="case-sensitive">Thread talk</namespace>
<namespace key="92" case="case-sensitive">Summary</namespace>
<namespace key="93" case="case-sensitive">Summary talk</namespace>
<namespace key="100" case="case-sensitive">Appendix</namespace>
<namespace key="101" case="case-sensitive">Appendix talk</namespace>
<namespace key="102" case="case-sensitive">Concordance</namespace>
<namespace key="103" case="case-sensitive">Concordance talk</namespace>
<namespace key="106" case="case-sensitive">Rhymes</namespace>
<namespace key="107" case="case-sensitive">Rhymes talk</namespace>
<namespace key="108" case="case-sensitive">Transwiki</namespace>
<namespace key="109" case="case-sensitive">Transwiki talk</namespace>
<namespace key="110" case="case-sensitive">Thesaurus</namespace>
<namespace key="111" case="case-sensitive">Thesaurus talk</namespace>
<namespace key="114" case="case-sensitive">Citation</namespace>
<namespace key="115" case="case-sensitive">Citations talk</namespace>
<namespace key="116" case="case-sensitive">Sign gloss</namespace>
<namespace key="117" case="case-sensitive">Sign gloss talk</namespace>
<namespace key="118" case="case-sensitive">Reconstruction</namespace>
<namespace key="119" case="case-sensitive">Reconstruction talk</namespace>
<namespace key="710" case="case-sensitive">TimedText</namespace>
<namespace key="711" case="case-sensitive">TimedText talk</namespace>
<namespace key="828" case="case-sensitive">Module</namespace>
<namespace key="829" case="case-sensitive">Module talk</namespace>
<namespace key="2300" case="case-sensitive">Gadget</namespace>
<namespace key="2301" case="case-sensitive">Gadget talk</namespace>
<namespace key="2302" case="case-sensitive">Gadget definition</namespace>
<namespace key="2303" case="case-sensitive">Gadget definition
talk</namespace>
```

```
</namespaces>
```

```
</siteinfo>
```

## <page>

```
<title>unus</title>
<ns>0</ns>
<id>26021</id>
<revision>
<id>76007577</id>
<parentid>71992627</parentid>
<timestamp>2023-09-06T20:48:43Z</timestamp>
<contributor>
<username>AutoDooz</username>
<id>3732454</id>
</contributor>
<minor />
<comment>{"LatinName"/> converted bare quote to templates/comment-
<model>wikitext</model>
<format>text/x-wiki</format>
```

```
<text bytes="3409"
```

```
xml:space="preserve">===Latin===
```

```
[[number box|1]]
```

```
===Alternative forms===
* Symbol: "1"
```

```
===Etymology===
```

```
From [[der|]a]ite-ol[oino]s]], from [[inh|]a]ite-pro[oino]s]], from [[trh|]a]ine-pro[oino]s]]one, single]].
```

```
Cognates include [[cog|]re]o]], [[cog|]sa]re]]-th[ē]ka]], [[cog|]cu]m]]ma]], [[cog|]ga]ben]], and [[cog|]an]]]] [[cog|]one]] and [[m]em]]]].
```

```
===Pronunciation===
```

```
* [[|a-IPA|ec]-yes|unus]]
* [[audio|]a-]s-|unus.ogg|Audio (Classical)]
```

```
===Adjective===
```

```
[[|a-adj|unus&#2D;us&#2D;]]
```

```
# [[|one]], [[|single]]
```

```
#: [[|uc|]a-]]|a-adj]] "unum"tr: [[|unanimously]], [[|universally]], [[|widely]]];]
# [[|alone]]
```

```
===Numeral===
```

```
[[|a-num-adj|unus&#2D;us&#2D;type=card]]
```

```
# [[|c|]n|]a|cardinal numbers]] [[|one]], ]
```

```
#* [[|Q|]a|Ovid|Metamorphoses|6644|thru|645|quote=sans illi adfatu vel "unum" vulnus eni]]
ngulum ferro Philomela resolvit]]ans: Sufficient was this "one" wound to kill, but Philomela
also cut open the throat]]
```

```
#* [[|Q|]a|Jeron|Eulogium|Nehemias|12|quote=et venit Anani "unus" de fratribus meis ipse et
vir ex Juda et interrogavit eos de Iudaeis qui remanserant et supererant de captivitate et de
Hiemal]]em]]ans and Hanani came, "one" of my brethren, he and certain men of Judah, and
asked them concerning the Jews that had escaped, which were left of the captivity, and
concerning Jerusalem]]
```

```
#* "ōnē" — [[w:|Boethius|Boethius]], "[[|s|]c|a|Commentarium in librum Aristotelis Peri
h]om]e]n]e]s|pr]i]m]e]s|editions|Commentarium in librum Aristotelis Peri h]om]e]n]e]s|pr]i]m]e]s|
editions]]", Book 1, section 5]]
#*: [[|q|]u]o|]a|]n|] s]u]m]m]u]m|]g]n]u]r|"unum"r]at]i]o]n]u]m|]a]l]i]e|]s]u]n]t|]s]i]g]n]i]f]i]c]a]t]i]o]n]e|"un]e",|]a]l]i]e|]c]o]n]j]u]n]c]t]i]o]n]e|]
[translation: "In summary therefore, "of one" theme others are (by signification) "one", "sem]
with corrections."]]
```

```
===Usage notes===
```

```
The plural forms are only used with pluralia tantum. For more information see
[[Appendix:Latin cardinal numbers#unus|Appendix:Latin cardinal numbers]]
```

```
===Declension===
[[|a-ades|]unus&#2D;us&#2D;]]
* Sg.gen. "un]i", sg.dat. "un]o", "un]ae" appear in earlier writers.
```

```
===Derived terms===
```

```
[[|col-top|2]]
* [[|]a|]ad]u]m]]
* [[|]a|]u]n]i]c]u]s]]
* [[|]a|]u]n]i]o]]
* [[|]a|]u]n]i]t]a]t]i]s]]
* [[|]a|]u]n]i]t]u]s]]
* [[|]a|]u]n]i]t]u]s]]
* [[|]a|]u]n]i]t]u]s]]
[[|col-bottom|]]
```

```
=== Related terms ===
* [[|]a|]n]o]]
* [[|]a|]l]i]u]s]]
```

```
=== Article ===
[[|head|]a|]a|]r]t]i]c]l]e|]h]e]a]d|=unus]]
```

```
# [[|]b|]a|]M]e]d]i]e]v]a]l|]L]a]t]i]n]] [[|]a|]], [[|]a]n]]
```

```
=== Declension ===
[[|]a-ades|]unus&#2D;us&#2D;us&#2D;g&#2D;]]
```

```
=== Derived terms ===
* [[|]a|]a]l]i]c]u]s]] [[|]q|]V]a]l]g]a]r|]L]a]t]i]n]]
* [[|]a|]n]e]c]u]s]]
```

```
=== References ===
```

```
* [[|R|]L&#2A]m]p;S]]
* [[|R|]e]l]e]m]e]n]t]a]r]y|]L]e]w]i]s]]
* [[|R|]e]l]i]g]i]o]u]s|]t]e]x]t]u]s]]:|]U]N]I]T]A]T]I]O]]
* [[|R|]G]a]t]f]i]o]]
* [[|R|]M&#2A]m]p;A]]
* [[|R|]N]L]W]]
```

```
=== Yalan ===
```

```
=== Noun ===
[[|head|]a|]n]o]]
```

```
# [[|famine]]</text>
```

```
<sha 1>gdjaerc8nnl28peavka2cu88i|hw6<</sha 1>
</revision>
```

```
</page>
```

```
<page>
```

```
<title>tot</title>
```

```
<ns>0</ns>
<id>53952</id>
<revision>
<id>76433538</id>
<parentid>76307599</parentid>
<timestamp>2023-10-21T12:56:19Z</timestamp>
<contributor>
<username>AutoDooz</username>
<id>3732454</id>
</contributor>
<minor />
<comment>{"English:Eymology 1|Derived terms"/> promoted all child sections to
siblings, {"English:Eymology 1|Noun"/> adopted English:Eymology 1|Derived terms,
English:Eymology 1|Translations/comment-
<model>wikitext</model>
<format>text/x-wiki</format>
```

```
<text bytes="50538"
xml:space="preserve">{{also|Appendix:Va
```

```
riations of "tot"}}
==English==
```

# que traducido quiere decir:

```
<mediawiki
BLABLABLABLE BLABLABLABLE
BLABLABLABLE BLABLABLABLE
BLABLABLABLE BLABLABLABLE
BLABLABLABLE BLABLABLABLE
version="0.10" xml:lang="en">
```

```
<siteinfo>
BLABLABLABLE BLABLABLABLE
BLABLABLABLE BLABLABLABLE
BLABLABLABLE BLABLABLABLE
  <namespaces>
<namespaces BLABLABLABLE>
<namespaces BLABLABLABLE>
<namespaces BLABLABLABLE>
<namespaces BLABLABLABLE>
  </namespaces>
</siteinfo>
```

```
<page>
<title>unus</title>
```

```
<BLABLABLABLE BLABLABLABLE>
<BLABLABLABLE BLABLABLABLE>
<BLABLABLABLE BLABLABLABLE>
<BLABLABLABLE BLABLABLABLE>
```

```
<text bytes="3409"
xml:space="preserve">
```

```
==Latin==
BLABLABLABLE BLABLABLABLE
BLABLABLABLE BLABLABLABLE
BLABLABLABLE BLABLABLABLE
BLABLABLABLE BLABLABLABLE
```

```
==Polish==
BLABLABLABLE BLABLABLABLE
BLABLABLABLE BLABLABLABLE
BLABLABLABLE BLABLABLABLE
BLABLABLABLE BLABLABLABLE
```

```
==Spanish==
BLABLABLABLE BLABLABLABLE
BLABLABLABLE BLABLABLABLE
BLABLABLABLE BLABLABLABLE
BLABLABLABLE BLABLABLABLE
```

```
</text>
<sha1>jj4a1c8n1d28peav1a2cu88ihw6</sha1>
</revision>
</page>
```

```
<page>
<title>tot</title>
```

```
<BLABLABLABLE BLABLABLABLE>
<BLABLABLABLE BLABLABLABLE>
<BLABLABLABLE BLABLABLABLE>
<BLABLABLABLE BLABLABLABLE>
```

```
<text bytes="50538"
xml:space="preserve">
```

```
{{also|Appendix: Variations of "tot"}}
==English==
BLABLABLABLE BLABLABLABLE
BLABLABLABLE BLABLABLABLE
BLABLABLABLE BLABLABLABLE
BLABLABLABLE BLABLABLABLE
```

```
==Latin==
BLABLABLABLE BLABLABLABLE
BLABLABLABLE BLABLABLABLE
BLABLABLABLE BLABLABLABLE
BLABLABLABLE BLABLABLABLE
```

```
==Portuguese==
BLABLABLABLE BLABLABLABLE
BLABLABLABLE BLABLABLABLE
BLABLABLABLE BLABLABLABLE
BLABLABLABLE BLABLABLABLE
```

```
</text>
<sha1>jj4a1c8n1d28peav1a2cu88ihw6</sha1>
</revision>
</page> </mediawiki>
```

es decir

<etiqueta inicial, que no importa>  
<etiqueta de información SITEINFO y NAMESPACES, que no importa>  
< cierre de información SITEINFO >

[ <etiqueta de página para cada entrada PAGE>  
<entre ellas, etiqueta de título> título <cierre de título>  
<etiqueta de texto> WIKITEXTO <cierre de texto>  
<cierre de pagina PAGE>

[ <etiqueta de página para cada entrada PAGE>  
<entre ellas, etiqueta de título> título <cierre de título>  
<etiqueta de texto> WIKITEXTO <cierre de texto>  
<cierre de pagina PAGE>

<etiqueta final, que no importa>

¿QUÉ NOS INTERESA? Lo verde y lo verde

A título meramente informativo:  
como funciona el formato XML (y el HTML):

- una etiqueta es algo que va entre "<" y ">", como "[[ ]]" en wikicódigo
- cada etiqueta tiene que cerrarse (con excepciones), como los paréntesis
- se cierran igual que se abren poniendo una barra "/", y sería como ")"
- cada par de etiquetas tiene que tener contenido interno:  
    <abro> **BLABLABLA** <cierro>  
    o en realidad  
    <nombre> **BLABLABLA** </nombre>
- la información de una etiqueta puede ser otra (se pueden anidar) u otras
- .... pero se tienen que cerrar las más internas primero

Vamos, como los paréntesis en las operaciones matemáticas o en la ortografía.

# Volviendo al tema: ¿Qué nos interesa? Lo verde y lo verde

```
<mediawiki
BLABLABLABLE BLABLABLABLE
BLABLABLABLE BLABLABLABLE
BLABLABLABLE BLABLABLABLE
BLABLABLABLE BLABLABLABLE
version="0.10" xml:lang="en">
```

```
<siteinfo>
BLABLABLABLE BLABLABLABLE
BLABLABLABLE BLABLABLABLE
BLABLABLABLE BLABLABLABLE
  <namespaces>
<namespaces BLABLABLABLE>
<namespaces BLABLABLABLE>
<namespaces BLABLABLABLE>
<namespaces BLABLABLABLE>
  </namespaces>
</siteinfo>
```

```
<page>
<title>unus</title>
```

```
<BLABLABLABLE BLABLABLABLE>
<BLABLABLABLE BLABLABLABLE>
<BLABLABLABLE BLABLABLABLE>
<BLABLABLABLE BLABLABLABLE>
```

```
<text bytes="3409"
xml:space="preserve">
```

```
==Latin==
BLABLABLABLE BLABLABLABLE
BLABLABLABLE BLABLABLABLE
BLABLABLABLE BLABLABLABLE
BLABLABLABLE BLABLABLABLE
```

```
==Polish==
BLABLABLABLE BLABLABLABLE
BLABLABLABLE BLABLABLABLE
BLABLABLABLE BLABLABLABLE
BLABLABLABLE BLABLABLABLE
```

```
==Spanish==
BLABLABLABLE BLABLABLABLE
BLABLABLABLE BLABLABLABLE
BLABLABLABLE BLABLABLABLE
BLABLABLABLE BLABLABLABLE
```

```
</text>
<sha1>jj4a1c8n1d28peav1a2cu88ihw6</sha1>
</revision>
</page>
```

```
<page>
<title>tot</title>
```

```
<BLABLABLABLE BLABLABLABLE>
<BLABLABLABLE BLABLABLABLE>
<BLABLABLABLE BLABLABLABLE>
<BLABLABLABLE BLABLABLABLE>
```

```
<text bytes="50538"
xml:space="preserve">
```

```
{{also|Appendix: Variations of "tot"}}
==English==
BLABLABLABLE BLABLABLABLE
BLABLABLABLE BLABLABLABLE
BLABLABLABLE BLABLABLABLE
BLABLABLABLE BLABLABLABLE
```

```
==Latin==
BLABLABLABLE BLABLABLABLE
BLABLABLABLE BLABLABLABLE
BLABLABLABLE BLABLABLABLE
BLABLABLABLE BLABLABLABLE
```

```
==Portuguese==
BLABLABLABLE BLABLABLABLE
BLABLABLABLE BLABLABLABLE
BLABLABLABLE BLABLABLABLE
BLABLABLABLE BLABLABLABLE
```

```
</text>
<sha1>jj4a1c8n1d28peav1a2cu88ihw6</sha1>
</revision>
</page> </mediawiki>
```

# ¿Qué NO nos interesa? Lo azul (siempre al principio), lo rojo (principio y fin), lo naranja y lo marrón

```
<mediawiki
BLABLABLABLA BLABLABLABLA
BLABLABLABLA BLABLABLABLA
BLABLABLABLA BLABLABLABLA
BLABLABLABLA BLABLABLABLA
version="0.10" xml:lang="en">
```

```
<siteinfo>
BLABLABLABLA BLABLABLABLA
BLABLABLABLA BLABLABLABLA
BLABLABLABLA BLABLABLABLA
  <namespaces>
<namespaces BLABLABLABLA>
<namespaces BLABLABLABLA>
<namespaces BLABLABLABLA>
<namespaces BLABLABLABLA>
  </namespaces>
</siteinfo>
```

```
<page>
  <title>unus</title>

<BLABLABLABLA BLABLABLABLA>
<BLABLABLABLA BLABLABLABLA>
<BLABLABLABLA BLABLABLABLA>
<BLABLABLABLA BLABLABLABLA>
```

```
  <text bytes="3409"
xml:space="preserve">
```

```
==Latin==
BLABLABLABLA BLABLABLABLA
BLABLABLABLA BLABLABLABLA
BLABLABLABLA BLABLABLABLA
BLABLABLABLA BLABLABLABLA
```

```
==Polish==
BLABLABLABLA BLABLABLABLA
BLABLABLABLA BLABLABLABLA
BLABLABLABLA BLABLABLABLA
BLABLABLABLA BLABLABLABLA
```

```
==Spanish==
BLABLABLABLA BLABLABLABLA
BLABLABLABLA BLABLABLABLA
BLABLABLABLA BLABLABLABLA
BLABLABLABLA BLABLABLABLA
```

```
</text>
  <sha1>jjdaec8nml28peavla2cu88ihww6</sha1>
  </revision>
</page>
```

```
<page>
  <title>tot</title>
```

```
<BLABLABLABLA BLABLABLABLA>
<BLABLABLABLA BLABLABLABLA>
<BLABLABLABLA BLABLABLABLA>
<BLABLABLABLA BLABLABLABLA>
```

```
  <text bytes="50538"
xml:space="preserve">
```

```
{{also|Appendix: Variations of "tot"}}
==English==
BLABLABLABLA BLABLABLABLA
BLABLABLABLA BLABLABLABLA
BLABLABLABLA BLABLABLABLA
BLABLABLABLA BLABLABLABLA
```

```
==Latin==
BLABLABLABLA BLABLABLABLA
BLABLABLABLA BLABLABLABLA
BLABLABLABLA BLABLABLABLA
BLABLABLABLA BLABLABLABLA
```

```
==Portuguese==
BLABLABLABLA BLABLABLABLA
BLABLABLABLA BLABLABLABLA
BLABLABLABLA BLABLABLABLA
BLABLABLABLA BLABLABLABLA
```

```
</text>
  <sha1>jjdaec8nml28peavla2cu88ihww6</sha1>
  </revision>
</page> </mediawiki>
```

## ¿Cual va a ser nuestra estrategia?

Deshacernos de lo que no nos interesa:

- Lo azul: al estar siempre al principio, se puede quitar a mano
  - Lo **rojo**: al estar siempre al principio y fin se puede quitar a mano
  - Lo **naranja**:
  - Lo **marrón**:
- } tienen en común ser etiquetas

En realidad, todo lo que NO nos interesa son:

- etiquetas
- contenido de etiquetas excepto dos:

`<title> .... </title>`

`<text> .... </text>`

Vamos a separar lo que nos interesa marcándolo de colores.



## **PRIMER PROCESAMIENTO:**

Donde se dan cuenta de las etiquetas y se obtiene wikicódigo puro y duro.

Antes que nada, hay unos espacios al principio de los primeros párrafos muy molestos y que después nos pueden complicar las instrucciones (\*).

Así que vamos a quitarlos. El procedimiento sería el siguiente:

Para MS Word:

"^p " (sin comillas; es decir ^p seguido de espacio, en minúscula)

( ^ se escribe con MAY+el acento a la derecha de la tecla P, dos veces y borra una )

Reemplazar con:

^p

※ NO Use wildcards/comodines (si acabamos de encender el programa)  
(Sin formato de reemplazamiento)

y repetimos *Reemplazar todo* muchas veces, ocho o diez por lo menos.  
(hasta que ponga que el número de veces que lo encuentra es sólo 0 ó 1)

\* (estimo que esta sustitución no es imprescindible, pero por si acaso)

El procedimiento no es igual para los dos programas:

Para Libre Office Write:

"\n " (sin comillas; es decir \n seguido de espacio)

( \i se escribe con AltGr y la tecla a la izquierda de l )

Reemplazar con:

\n

√ SÍ Expresiones regulares

Sin formato de reemplazamiento

y necesita una segunda instrucción

"^ " (sin comillas; es decir ^ seguido de espacio)

Reemplazar con:

(NADA)

√ SÍ Expresiones regulares

Sin formato de reemplazamiento

y ya está.

Vamos a hacer una *prueba* algo más elaborada.

Queremos deshacernos de:

las etiquetas <namespace> de apertura,  
de su información y

de su contraetiqueta </namespace> de cierre.

Están al principio del documento. Nosotros pinchamos el cursor al principio.

En vez de destruirlas, vamos primero a ponerlas en rojo.

Para MS Word:

`\<namespace(*)\</namespace\>`

(no ponemos nada en Reemplazar con)  SÍ Use wildcards/comodines

Format Repl. : Font > Font > Font Color : violeta

Para Libre Office Write:

`<namespace(.*)</namespace>`

`$0`

SÍ Expresiones regulares

Formato Reempz : Efectos tipográficos > Color de letra : violeta

(y va a quedar una etiqueta suelta sin localizar, por no estar cerrada)

y Reemplazar todo

Ahora vamos a deshacernos realmente de ello.  
Pinchamos en la parte Reemplazar y quitamos contenidos y formatos.  
El resto, igual.

Están al principio del documento. Nosotros pinchamos el cursor al principio.  
En vez de destruirlas, vamos primero a ponerlas en rojo.

Para MS Word:

`\<namespace(*)\</namespace\>`

(no ponemos nada en Reemplazar con) ✓ Use wildcards/comodines

Para Libre Office Write:

`<namespace(.*)</namespace>`

(nada en Reempz)

✓ Expresiones regulares

(y va a quedar una etiqueta suelta sin localizar, por no estar cerrada)

y Reemplazar todo

*(esto que acabamos de hacer no es imprescindible, sólo para practicar)*

Y así podríamos ir etiqueta por etiqueta,  
eliminando una por una.

<sitename> </sitename>

<dbname> </dbname>

<base> </base>

<generator> </generator>

<case> </case>

<ns> </ns>

<id> </id>

<parentid> </parentid>

<timestamp> </timestamp>

<username> </username>

<comment> </comment>

<model> </model>

<format> </format>

Pero no. Nos importa quitarlas, pero voy a dejarlas aparcadas por un rato.

Sin embargo, hay seis etiquetas concretas que nos interesan especialmente. La primera es la del **título**, porque esa información no está disponible en otro lado.

Su formato es:

```
<title> BLABLABLA </title>
```

Se da la casualidad de que en las wikis normalmente no se utiliza la sección de nivel máximo, con sólo un "=", así que vamos a aprovecharlo y transformar nuestro título en una sección máxima.

**IMPORTANTE:**

Vamos también a ponerla en verde (o algún color no usado) para identificarla mejor (y considerarla después).

El procedimiento sería el siguiente:

Para MS Word:

\<title\>(\*)\</title\>

Reemplazar con:

=\1=

√ Use wildcards/comodines

Format Repl. : Font > Font > Font Color : verde

Para Libre Office Write:

<title>(.\*</title>

Reemplazar con:

=\$1=

√ Expresiones regulares

Formato Reempz : Efectos tipográficos > Color de letra : verde



Las otras cinco son cinco pares de etiquetas anidadas,  
que en su contenido interno incluyen la información que nos importa  
(título y wikicódigo).

El proceso es igual para las tres primeras  
y en los dos programas:

*Para MS Word:*

*ETIQUETA*

*Reemplazar con:*

*(NADA)*

*√ Use wildcards/comodines*

*Sin formato Repl*

*Para Libre Office Write:*

*ETIQUETA*

*Reemplazar con:*

*(NADA)*

*√ Expresiones regulares*

*Sin formato Reempz*

Las etiquetas son:

ETIQUETA	MS Word <i>Encontrar</i>	LO Writer <i>Buscar</i>	Reemp. <i>con</i>	comodín expr. regul.
+page	\<page\>	<page>	(NADA)	√
-page	\</page\>	</page>	(NADA)	√
+revision	\<revision\>	<revision>	(NADA)	√
-revision	\</revision\>	</revision>	(NADA)	√
+contributor	\<contributor\>	<contributor>	(NADA)	√
-contributor	\</contributor\>	</contributor>	(NADA)	√
minor	\<minor^\>	<minor/>	(NADA)	√

Se puede copiar y pegar de la tabla, pero no copien nada después del ">".

Metí una más ( <minor/>, donde ^ son dos barras / y \),  
que lleva la barra al final, no al principio  
y no tiene cierre.

No es importante, pero ya que estamos....

Y ya sólo nos quedan dos importantes.

Son diferentes porque tienen contenido dentro de la etiqueta,  
y tenemos que eliminar también ese contenido.

Para eso usamos un código especial para "cualquier tipo de contenido".

ETIQUETA	MS Word <i>Encontrar</i>	LO Writer <i>Buscar</i>	Reemp. <i>con</i>	comodín expr. regul.
+mediawiki	\<mediawiki*\>	<mediawiki.*>	(NADA)	√
-mediawiki	\</mediawiki\>	</mediawiki>	(NADA)	√
+text	\<text*\>	<text.*>	(NADA)	√
-text	\</text\>	</text>	(NADA)	√

(nótese el punto para LO Writer)

Las dos primeras se pueden hacer a mano. Y ya casi está....

Las etiquetas que dejamos aparcadas (p. 26) son todas de apertura y cierre.

Todas tienen contenido que no nos interesa.

Todas están en el mismo párrafo.

Y (por eso) las eliminamos todas con un sólo reemplazamiento.

Un poco complicado, pero muy efectivo (recuerda: el  $\wedge$  son dos barras / y  $\backslash$ ):

Para MS Word:

$\backslash\langle(*)\backslash\rangle*\backslash\langle\wedge 1\backslash\rangle$

Reemplazar con:

(nada)

✓ Use wildcards/comodines

Sin formato Repl

Para Libre Office Write:

$\langle(.*)\rangle.*\langle\wedge 1\rangle$

Reemplazar con:

(nada)

✓ Expresiones regulares

Sin formato Reempz

En especial:

Para Libre Office Write ADEMÁS deberían quedar bailando

*<siteinfo> </siteinfo>*

y

*<namespaces> </namespaces>*

al principio de todo sin eliminar.

Toca quitarlas a mano (o no, eso creo: no estoy seguro, pero por si acaso).

Ya está. Ya tenemos sólo wikicódigo.

Si copiásemos y pegásemos en la wiki, ya veríamos resultados.

Este es el fin de la primera parte.

## **SEGUNDO PROCESAMIENTO:**

Donde separamos la(s) sección(es) que interesan.

Ahora tenemos un wikicódigo con toda la información de todas las páginas.

Queremos sonsacar una sección (Latín), y de ella una subsección (Etimología).

No nos interesa el inglés.

Non nos interesan las pronunciaciones, definiciones, ejemplos, sinónimos/antónimos, conjugación/declinación, referencias o notas.

Y la información de las secciones va en bloques,  
y cada bloque tiene distintos párrafos.

No nos interesan los párrafos, sólo los bloques de secciones.

Pero para el programa de edición de texto no es fácil editar bloques.

*(se puede hacer, pero se precisa más conocimiento)*

La clave aquí es

- 1) DESTRUIR las separaciones de párrafos preexistentes,
- 2) RECONSTRUIR las secciones como párrafos,
- 3) ESCOGER las que quiera,
- 4) DESCARTAR el resto(y las escojo por colores, por eso marqué los títulos en color)
- y 5) RESTITUIR separaciones.

Antes de nada quiero regularizar un detalle

(se trata de los tipos de salto de párrafo, sé que dan problemas;

no voy a entrar a explicar con mucho detalle porque no compensa pero, en breve: existen dos -se pueden distinguir si se hacen visibles- y quiero todos igual):

Para MS Word:

^l

Reemplazar con:

^p

※ NO Use wildcards/comodines

Sin formato

Para Libre Office Write:

\n

Reemplazar con:

\n

√ SÍ Expresiones regulares

Sin formato



Y otro tema, ya que estamos.  
De la misma manera que quitamos espacios al principio de todo,  
a mí, que soy tiquismiquis y apañado,  
no me gusta ver tantos párrafos en blanco.  
Vamos a quitarlos y compactar todo un poco.

Para MS Word:

^p^p

Reemplazar con:

^p

※ NO Use wildcards/comodines

Sin formato

y repetimos varias veces hasta que haya 0 o 1 resultado, por lo menos cuatro.

Para Libre Office Write:

^\$

Reemplazar con:

(NADA)

※ NO Expresiones regulares

Sin formato

en este caso sin repetir.

## 1) DESTRUIR

Queremos destruir las separaciones entre párrafos, pero no perder su localización.

Vamos a buscar una cadena (una serie de caracteres) que NO exista en mi texto.

Mis favoritas son las combinaciones "|-|-|-|" o "||||", pero voy a escoger "ññññ".

La voy a llamar Mi cadena personal.

Pueden escoger la suya (",,,"), pero no vale cualquiera.

Es preferible (casi obligado) evitar caracteres especiales (depende del programa):

. ^ \$ \* + ? \ [ ( { | [ ] { } < > ( ) - @ ? ! \* \

Voy a usar esa cadena que escogí como separador.

## 1) DESTRUIR

Para MS Word:

^p

Reemplazar con (su cadena personal):

ññññ

※ NO Use wildcards/comodines

Sin formato

Para Libre Office Write:

\$

Reemplazar con (su cadena personal):

ññññ

√ SÍ Expresiones regulares

Sin formato

Ya está, absolutamente todo nuestro texto es un sólo párrafo.  
Como un bloque de jamón cocido. No hay quien le meta el diente.

## 2) RECONSTRUIR

Y ahora voy a cortarlo en bloques por secciones,  
siempre por el inicio de sección, cada vez que empiece cualquier idioma:

Para MS Word (idem cadena):

ññññ==[!=]

Reemplazar con:

^p^&

√ Use wildcards/comodines

Sin formato

Para Libre Office Write (idem cadena):

ññññ==[^=]

Reemplazar con:

\n\$0

√ Expresiones regulares

Formato Reempz: Color automático

### 3) ESCOGER

Escogemos las secciones que nos interesan, la de nuestro idioma concreto:  
(esto se puede repetir si interesase más de un idioma)

Para MS Word (id. cadena):

ññññ==Latin\*^13

Reemplazar con:  
(NADA)

√ Use wildcards/comodines

Formato Repl: Color azul (uno que NO escogiésemos antes)  
(y hay que repetir con espacio por si acaso: ññññ== Latin\*^13 )

Para Libre Office Write (id. cadena, nótese el punto):

ññññ==Latin.\*

Formato Buscar: Color rojo (el de la página 16)

Reemplazar con:  
\$0

√ Expresiones regulares

Formato Reempz: Color azul (uno que NO escogiésemos antes)  
(y hay que repetir con espacio por si acaso: ññññ== Latin.\* )

## 4) DESCARTAR

Eliminamos lo que no nos interesa

Para MS Word:

(NADA)

Formato Buscar: Color Automático (o negro)

Reemplazar con:

(NADA)

※ Use wildcards/comodines (o ya ni te deja)

Formato Repl: Sin formato

Para Libre Office Write:

.\*

Formato Buscar: Color rojo (el de la página 16)

Reemplazar con (cadena personal):

ññññ

√ Expresiones regulares

Formato Reempz: Color azul (o no)

## 5) RESTITUIR

Restituímos (el paso contrario a 1)

Para MS Word (ibid. cadena):

ññññ

Formato Buscar: Sin formato

Reemplazar con:

^p

※ √ Use wildcards/comodines (de esta vez vale cualquiera)

Formato Repl: Sin formato

Para Libre Office Write (ibid. cadena):

ññññ

Formato Buscar: ninguno

Reemplazar con:

\n

√ Expresiones regulares

Formato Reempz: Color rojo (o el de la página 16)

Antes de acabar, si queremos salvar nuestros epígrafes de Latin  
ante procesamientos posteriores, lo vamos a pasar a verde,  
el color en el que está el idioma (P31)

Para MS Word:

==Latin\*^13

Reemplazar con:

(NADA)

√ Use wildcards/comodines

Formato Repl: Color verde

*(y hay que repetir con espacio por si acaso: ññññ== Latin\*^13 )*

Para Libre Office Write:

==Latin.\*

Reemplazar con:

\$0

√ Expresiones regulares

Formato Reempz: Color verde

*(y hay que repetir con espacio por si acaso: ññññ== Latin.\* )*



## **TERCER PROCESAMIENTO:**

Donde separamos la(s) SUBsección(es) que interesan.

Nuestra tercera parte va a ser muy similar a la segunda,  
con algunos cambios menores.

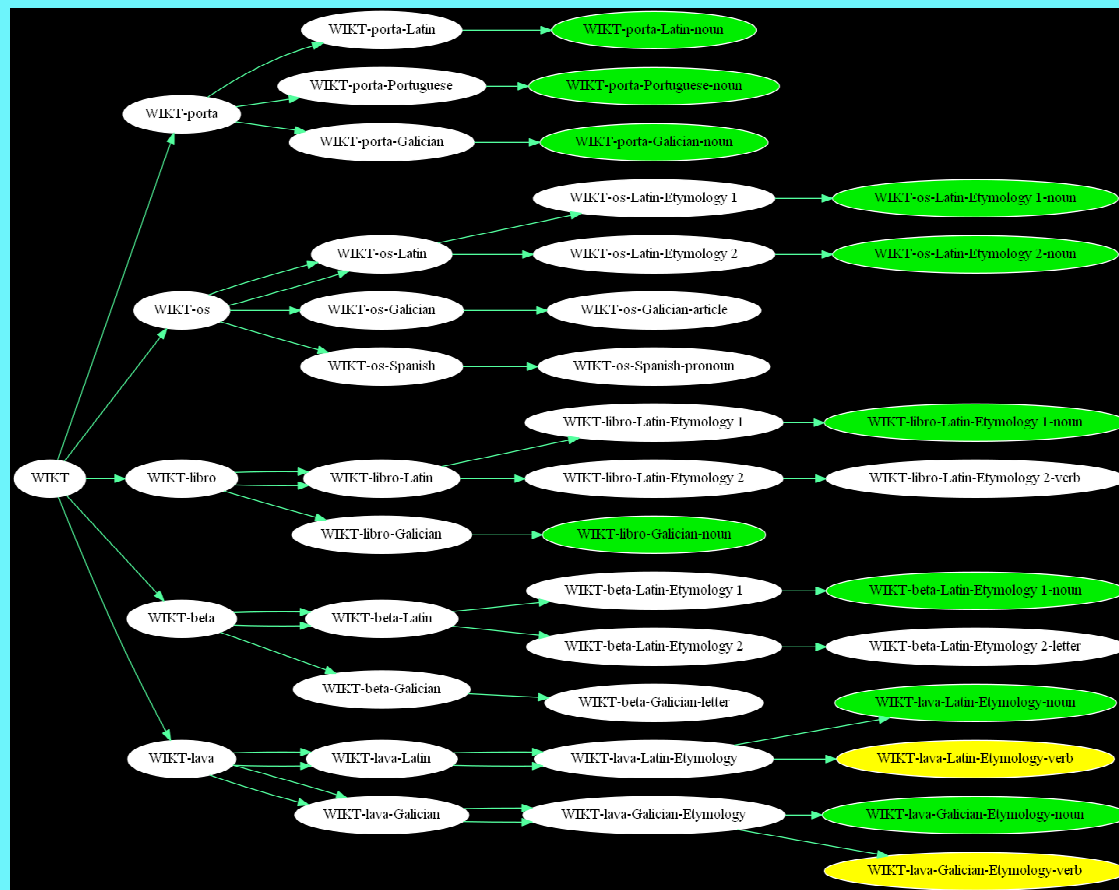
De la misma manera, si hubiese una subsubsección  
también se podría procesar y separar.

Hay que ser consciente de que,  
cada vez que se repita el proceso,  
se está escogiendo lo anidado  
dentro de lo seleccionado previamente.

Después de elegir Latin, no podemos ver los datos de otros idiomas.  
Si se escogió un nivel 2 (==) ya se pierde todo el resto de ese nivel.

Además, hay que conocer la estructura de datos con la que trabajamos.

Por ejemplo (ejemplos escogidos con PetScan y este procedimiento;  
son 543 palabras en 4764 páginas y llevó 13 minutos)



Aquí tenemos casos reales (o casi: adaptados y simplificados) de en.wikt.

Las secciones verdes son substantivos.

Como se ve, no están todas al mismo nivel, porque dependen de la etimología.

Y lo mismo con las amarillas, los verbos. Hay que conocer la estructura.

Pues bueno, volviendo al tema de los reemplazos y substituciones.

Toca otra vez el paso de destrucción de párrafos (42):

de  $\wedge p$  a ññññ

o

de \$ a ññññ

(en teoría los pasos previos de las páginas 39 y 40 no serían necesarios)

La recomendación es repasar el nivel 3 (===),  
por lo tanto el paso de reconstrucción sería:

ññññ===[!]=]

Reemplazar por:

$\wedge p \wedge \&$

✓ y Sin formato

ññññ===[^]=]

Reemplazar por:

$\backslash n \$ 0$

✓ y Formato Reempz: Color automático

Una vez reconstruido podríamos incluir:

- Etymology (dentro de Latin)
- Derived terms (dentro de Latin)
- Related terms (dentro de Latin)
- Descendants (dentro de Latin)

Así que el paso de escoger sería:

ññññ===Etymology\*^13

ññññ===Derived terms\*^13

ññññ===Related terms\*^13

ññññ===Descendants\*^13

y con sus espacios despues de los iguales y sustituir cada uno de ellos por:

(NADA), √ y un nuevo color, Rosa

ññññ===Etymology.\*

ññññ===Derived terms.\*

ññññ===Related terms.\*

ññññ===Descendants.\*

también con sus espacios y formato Color rojo (el de la página 16), cambiarlos por:

\$0, √ y un nuevo color, Naranja

El paso de descartar sería igual que en 45:  
(NADA) y formato Colores Automático y azul  
por: (NADA), ✖ y Sin formato  
. \* y formato Color rojo (el de la página 16)  
por: ññññ, √ y Color azul

La restitución es idéntica:

ññññ, Sin formato  
por: ^p, ✖ √ y Sin formato  
ññññ, Sin formato  
por: \n, √ y Color rojo (o el de la página 16)

Y para salvar lo conseguido (todos en conxunto, brevemente):

===[!=]\*^13  
por: (NADA), √ y Color verde  
===[^=].\*  
por: \$0, √ y Color verde

Llegados a este punto, ya está.

Lo que tenemos ahora es wikicódigo concentrado, depurado, seleccionado:  
pero ni Word ni Writer pueden interpretarlo.

Hay muchas variables internas a la wiki que no se pueden emular.

Podemos (guardar una copia de seguridad si no lo fuimos haciendo y)  
copiar todo nuestro texto, pegarlo en la wiki de donde procede  
y dejar que se procese.

Logicamente yo lo hago en Vista previa  
en la *Sandbox/Zona de pruebas*.

El problema es que una misma página de wiki tiene un máximo de  
Templates, Plantillas, Modelos y módulos de Lua,  
pero hasta llegar a ese punto.

Por lo tanto, habría que ir interpretando por partes.

Yo lo que hago es copiar la parte legible de la wiki  
pegarla en Word/Writer y dejar que procese la siguiente parte.

## PRECAUCIONES Y PROBLEMAS AL PEGAR

Hay (por lo menos dentro de en.wikt) algunos detalles frecuentes que se pueden quitar o modificar:

`{{wikipedia|lang=la}}` y `{{wikipedia}}` (*cajita de remisión a en.wiki*)

`&lt;` y `&gt;` (*que representan < y >*)

y las cajas que empiezan por:

`{{number box|la|`

`{{sensenol|la|`

Cuando pegues, recuerde que hay tablas y cajas colapsables, escondibles.

Para expandirlas todas juntas, en la columna izquierda puede hacer que se muestre según el tema al que corresponda:

declinaciones, conjugaciones, flexión,  
sinónimos, antónimos, relaciones semánticas,  
descendientes, derivadas, relacionadas...

Escoja como considere.



# OBJECCIONES AL MÉTODO

Mayores problemas:

- 1) andar a cambiar de una página a otra, y de un programa a otro
- 2) saber un poco de inglés si se quiere alterar el proceso de exportación
- 3) dividir el total en partes si es demasiado grande
- 4) basarse en la casuística de lo que hay hecho y sus directrices  
(secciones "=" de nivel 1 inexistentes;  
resto de secciones al nivel correcto;  
etiquetas cerradas y anidadas bien\*,....)

Recomendaciones:

- 1) comprobar que las cosas van yendo bien
- 2) guardar copias de seguridad de cada parte del proceso
- 3) seguir la tabla de los apéndices para abreviar

\* (yo ya encontré *Descendants* de nivel 3 y de nivel 6)

# APÉNDICE 1: reemplazamientos para MSOW

acceso a reempz = CTRL + H

C = comodines/wildcards

B = búsqueda; R = reemplazamiento/replace

P = página

C0a = Automático (o negro)

C1v = verde

C2z = azul

C3r = rojo

C4x = naranja

Las comillas son simplemente para indicar/recordar que hay un espacio.

Cuidado si copian y pegan los códigos de la tabla por que pueden capturar un carácter de cierre.

Los formatos en amarillo son imprescindibles.

Primera <parte>

<b>n</b>	<b>C</b>	<b>textoB</b>	<b>textoR</b>	<b>formB</b>	<b>formR</b>	<b>P</b>
1	※	"^p "	^p			25x10
2	√	\<title\>(*)\</title\>	=\1=		C1v	31
3	√	\<page\>				33
4	√	\</page\>				33
5	√	\<revision\>				33
6	√	\</revision\>				33
7	√	\<contributor\>				33
8	√	\</contributor\>				33
9	√	\<minor\>				33
10	√	\<mediawiki*\>				34
11	√	\</mediawiki\>				34
12	√	\<text*\>				34
13	√	\</text\>				34
14	√	\<(*)\>*\<^1\>				35

## Segunda ==parte==

n	C	textoB	textoR	formB	formR	P
15	※√	^l	^p			39
16	※	^p^p	^p			40x4
17	※	^p	ññññ			1)42
18	√	ññññ==[!=]	^p^&			2)43
19a	√	ññññ==Latin*^13			C2z	3)44
19b	√	ññññ== Latin*^13			C2z	3)44
20	※			C0a		4)45
21	※√	ññññ	^p		C0a	5)46
22a	√	==Latin*^13			C1v	47
22b	√	== Latin*^13			C1v	47

16 y 17 pueden tardar bastante en acabar  
(22a-22b) ==[!=]\*^13

Tercera ===parte===

n	C	textoB	textoR	formB	formR	P
23	※	<sup>p</sup>	ññññ			51
24	√	ññññ===[!]=]	<sup>p</sup> &			51
25a	√	ññññ===Etymology* <sup>13</sup>			C4x	52
25b	√	ññññ===Derived terms* <sup>13</sup>			C4x	52
25c	√	ññññ===Related terms* <sup>13</sup>			C4x	52
25d	√	ññññ===Descendants* <sup>13</sup>			C4x	52
25a	√	ññññ=== Etymology* <sup>13</sup>			C4x	52
25b	√	ññññ=== Derived terms* <sup>13</sup>			C4x	52
25c	√	ññññ=== Related terms* <sup>13</sup>			C4x	52
25d	√	ññññ=== Descendants* <sup>13</sup>			C4x	52
26a	※			C0a		53
26b	※			C2z		53
27	※√	ññññ	<sup>p</sup>		C0a	53
28	√	===[!]=]* <sup>13</sup>			C1v	53

## APÉNDICE 2: reemplazos para OOLOW

acceso a reempz = CTRL + H

ER = expresiones regulares

B = búsqueda; R = reemplazo

P = página

C0a = Automático (o negro)

C1v = verde

C2z = azul

C3r = rojo

C4x = naranja

Las comillas son simplemente para indicar/recordar que hay un espacio.

Cuidado si copian y pegan los códigos de la tabla por que pueden capturar un carácter de cierre.

Los formatos en amarillo son imprescindibles.

Primera <parte>  
(primero, pasar todo a C3r: P16)

n	ER	textoB	textoR	formB	formR	P
1a	√	"\n "	\n			26
1b	√	"^ "				26
2	√	<title>(.*?)</title>	=\$1=		C1v	31
3	√✘	<page>				33
4	√✘	</page>				33
5	√✘	<revision>				33
6	√✘	</revision>				33
7	√✘	<contributor>				33
8	√✘	</contributor>				33
9	√✘	<minor/>				33
10	√	<mediawiki.*>				34
11	√	</mediawiki>				34
12	√	<text.*>				34
13	√	</text>				34
14	√	<(.*?)>.*</1>				35

(3-8) </?(page|revision|contributor)> ; (10-13) </?(mediawiki|text).\*>

Segunda ==parte==  
(tras eliminar las primeras etiquetas a mano, P36)

n	ER	textoB	textoR	formB	formR	P
15*	√	\n	\n			39
16	√	^\$				40
17	√	\$	ññññ			1)42
18	√	ññññ==[^=]	\n\$0		C3r	2)43
19a	√	ññññ==Latin.*	\$0	C3r	C2z	3)44
19b	√	ññññ== Latin.*	\$0	C3r	C2z	3)44
20	√	.*	ññññ	C3r	C2z	4)45
21	√	ññññ	\n		C3r	5)46
22a	√	==Latin.*	\$0		C1v	47
22b	√	== Latin.*	\$0		C1v	47

\*15 : cuidado porque a veces textoB se autocompleta con el espacio de (1a)  
(19-20) ññññ== ?Latin.\* ; (22a+22b) ==[^=].\*



### Tercera parte

n	ER	textoB	textoR	formB	formR	P
23	√	\$	ññññ			51
24	√	ññññ====[^=]	\n\$0		C3r	51
25a	√	ññññ====Etymology.*	\$0	C3r	C4x	52
25b	√	ññññ====Derived terms.*	\$0	C3r	C4x	52
25c	√	ññññ====Related terms.*	\$0	C3r	C4x	52
25d	√	ññññ====Descendants.*	\$0	C3r	C4x	52
25a	√	ññññ==== Etymology.*	\$0	C3r	C4x	52
25b	√	ññññ==== Derived terms.*	\$0	C3r	C4x	52
25c	√	ññññ==== Related terms.*	\$0	C3r	C4x	52
25d	√	ññññ==== Descendants.*	\$0	C3r	C4x	52
26	√	.*	ññññ	C3r	C2z	53
27	√	ññññ	\n		C3r	53
28	√	====[^=].*	\$0		C1v	53

# PARA LA GENTE DE MACROS (MSOW):

Sub WMes()

## PARTE 1

```
Selection.Find.ClearFormatting
Selection.Find.Replacement.ClearFormatting
Selection.Find.Wrap = wdFindContinue
Selection.Find.Forward = True
Selection.Find.Format = False
Selection.Find.MatchCase = False
Selection.Find.MatchWholeWord = False
Selection.Find.MatchWildcards = False
Selection.Find.MatchSoundsLike = False
Selection.Find.MatchAllWordForms = False

Selection.Find.Text = "^p"
Selection.Find.Replacement.Text = "p"
Selection.Find.Execute Replace:=wdReplaceAll
Selection.Find.Execute Replace:=wdReplaceAll
Selection.Find.Execute Replace:=wdReplaceAll
Selection.Find.Execute Replace:=wdReplaceAll
Selection.Find.Execute Replace:=wdReplaceAll
Selection.Find.Execute Replace:=wdReplaceAll
Selection.Find.Execute Replace:=wdReplaceAll
Selection.Find.Execute Replace:=wdReplaceAll
Selection.Find.Execute Replace:=wdReplaceAll
Selection.Find.Execute Replace:=wdReplaceAll
Selection.Find.Replacement.Font.Color = wdColorGreen
Selection.Find.Format = True
Selection.Find.MatchWildcards = True
Selection.Find.Text = "<title>(*)</title>"
Selection.Find.Replacement.Text = "=1="
Selection.Find.Execute Replace:=wdReplaceAll
Selection.Find.Format = False 'FAL TABA
Selection.Find.Text = "<page>"
Selection.Find.Replacement.Text = ""
Selection.Find.Execute Replace:=wdReplaceAll
Selection.Find.Text = "</page>"
Selection.Find.Execute Replace:=wdReplaceAll
Selection.Find.Text = "<revision>"
Selection.Find.Execute Replace:=wdReplaceAll
Selection.Find.Text = "</revision>"
Selection.Find.Execute Replace:=wdReplaceAll
Selection.Find.Text = "<contributor>"
Selection.Find.Execute Replace:=wdReplaceAll
Selection.Find.Text = "</contributor>"
Selection.Find.Execute Replace:=wdReplaceAll
Selection.Find.Text = "<minor>"
Selection.Find.Execute Replace:=wdReplaceAll
Selection.Find.Text = "<mediawiki>"
Selection.Find.Execute Replace:=wdReplaceAll
Selection.Find.Text = "</mediawiki>"
Selection.Find.Execute Replace:=wdReplaceAll
Selection.Find.Text = "<text*>"
Selection.Find.Execute Replace:=wdReplaceAll
```

```
Selection.Find.Text = "</text>"
Selection.Find.Execute Replace:=wdReplaceAll
Selection.Find.Text = "<(*)>*<1>"
Selection.Find.Execute Replace:=wdReplaceAll
PARTE 2
Selection.Find.Text = "A"
Selection.Find.Replacement.Text = "p"
Selection.Find.Execute Replace:=wdReplaceAll
Selection.Find.MatchWildcards = False
Selection.Find.Text = "p^p"
Selection.Find.Replacement.Text = "p"
Selection.Find.Execute Replace:=wdReplaceAll
Selection.Find.Execute Replace:=wdReplaceAll
Selection.Find.Execute Replace:=wdReplaceAll
Selection.Find.Text = "p"
Selection.Find.Replacement.Text = "ñññ"
Selection.Find.Execute Replace:=wdReplaceAll
Selection.Find.MatchWildcards = True
Selection.Find.Text = "ñññ==[!]"
Selection.Find.Replacement.Text = "p^p"
Selection.Find.Execute Replace:=wdReplaceAll
Selection.Find.Format = True
Selection.Find.Replacement.Font.Color = wdColorBlue
Selection.Find.Text = "ñññ==Latin*^13"
Selection.Find.Replacement.Text = ""
Selection.Find.Execute Replace:=wdReplaceAll
Selection.Find.Text = "ñññ== Latin*^13"
Selection.Find.Replacement.Text = ""
Selection.Find.Execute Replace:=wdReplaceAll
Selection.Find.Text = "" 'FAL TABA
Selection.Find.ClearFormatting
Selection.Find.Font.Color = wdColorAutomatic
Selection.Find.Replacement.ClearFormatting
Selection.Find.Execute Replace:=wdReplaceAll
Selection.Find.ClearFormatting
Selection.Find.Replacement.ClearFormatting
Selection.Find.Replacement.Font.Color = wdColorBlue
Selection.Find.Text = "ñññ"
Selection.Find.Replacement.Text = "p"
Selection.Find.Execute Replace:=wdReplaceAll
Selection.Find.ClearFormatting
Selection.Find.Replacement.ClearFormatting
Selection.Find.Replacement.Font.Color = wdColorGreen
Selection.Find.Text = "^13==[!]*^13" 'CAMBIO
Selection.Find.Replacement.Text = "" 'FAL TABA
PARTE 3
Selection.Find.Format = False
Selection.Find.MatchWildcards = False
Selection.Find.Text = "p"
```

```
Selection.Find.Replacement.Text = "ñññ"
Selection.Find.Execute Replace:=wdReplaceAll
Selection.Find.MatchWildcards = True
Selection.Find.Text = "ñññ====[!]"
Selection.Find.Replacement.Text = "p^p"
Selection.Find.Execute Replace:=wdReplaceAll
Selection.Find.Format = True
Selection.Find.ClearFormatting
Selection.Find.Replacement.ClearFormatting
Selection.Find.Replacement.Font.Color = wdColorOrange
Selection.Find.Text = "ñññ===Etymology*^13"
Selection.Find.Replacement.Text = ""
Selection.Find.Execute Replace:=wdReplaceAll
Selection.Find.Text = "ñññ===Derived terms*^13"
Selection.Find.Execute Replace:=wdReplaceAll
Selection.Find.Text = "ñññ===Related terms*^13"
Selection.Find.Execute Replace:=wdReplaceAll
Selection.Find.Text = "ñññ===Descendants*^13"
Selection.Find.Execute Replace:=wdReplaceAll
Selection.Find.Text = "ñññ=== Etymology*^13"
Selection.Find.Execute Replace:=wdReplaceAll
Selection.Find.Text = "ñññ=== Derived terms*^13"
Selection.Find.Execute Replace:=wdReplaceAll
Selection.Find.Text = "ñññ=== Related terms*^13"
Selection.Find.Execute Replace:=wdReplaceAll
Selection.Find.Text = "ñññ=== Descendants*^13"
Selection.Find.Execute Replace:=wdReplaceAll
Selection.Find.ClearFormatting
Selection.Find.Font.Color = wdColorBlue
Selection.Find.Replacement.ClearFormatting
Selection.Find.MatchWildcards = False
Selection.Find.Format = True
Selection.Find.Text = ""
Selection.Find.Replacement.Text = ""
Selection.Find.Execute Replace:=wdReplaceAll
Selection.Find.Font.Color = wdColorAutomatic
Selection.Find.Execute Replace:=wdReplaceAll
Selection.Find.ClearFormatting
Selection.Find.Replacement.ClearFormatting
Selection.Find.Replacement.Font.Color = wdColorAutomatic
Selection.Find.Text = "ñññ"
Selection.Find.Replacement.Text = "p"
Selection.Find.Execute Replace:=wdReplaceAll
Selection.Find.ClearFormatting
Selection.Find.Replacement.ClearFormatting
Selection.Find.Replacement.Font.Color = wdColorGreen
Selection.Find.MatchWildcards = True
Selection.Find.Text = "^13===[!]*^13" 'CAMBIO
Selection.Find.Replacement.Text = ""
Selection.Find.Execute Replace:=wdReplaceAll
End Sub
```