cost

EUROPEAN COOPERATION IN SCIENCE AND TECHNOLOGY

EUROPEAN
NETWORK OF
e-LEXICOGRAPHY

WG 2 “Retro-digitized dictionaries”

Progress Report

WORKING GROUP 2 IN NUMBERS

- 50 members
  - Gender Balance: 22 female, 28 male, i.e. about 44 % of members are female
  - Early Career Investigators: 18 ESRs, i.e. about 36 % of members are ESRs
- Internationality: 20 countries (Austria (2), Belgium (1), Croatia (1), Denmark (3), Finland (3), France (4), Germany (6), Greece (2), Hungary (1), Israel (1), Italy (1), Macedonia (1), Netherlands (2), Poland (3), Portugal (4), Romania (4), Serbia (5), Slovakia (2), Switzerland (3), United Kingdom (1))

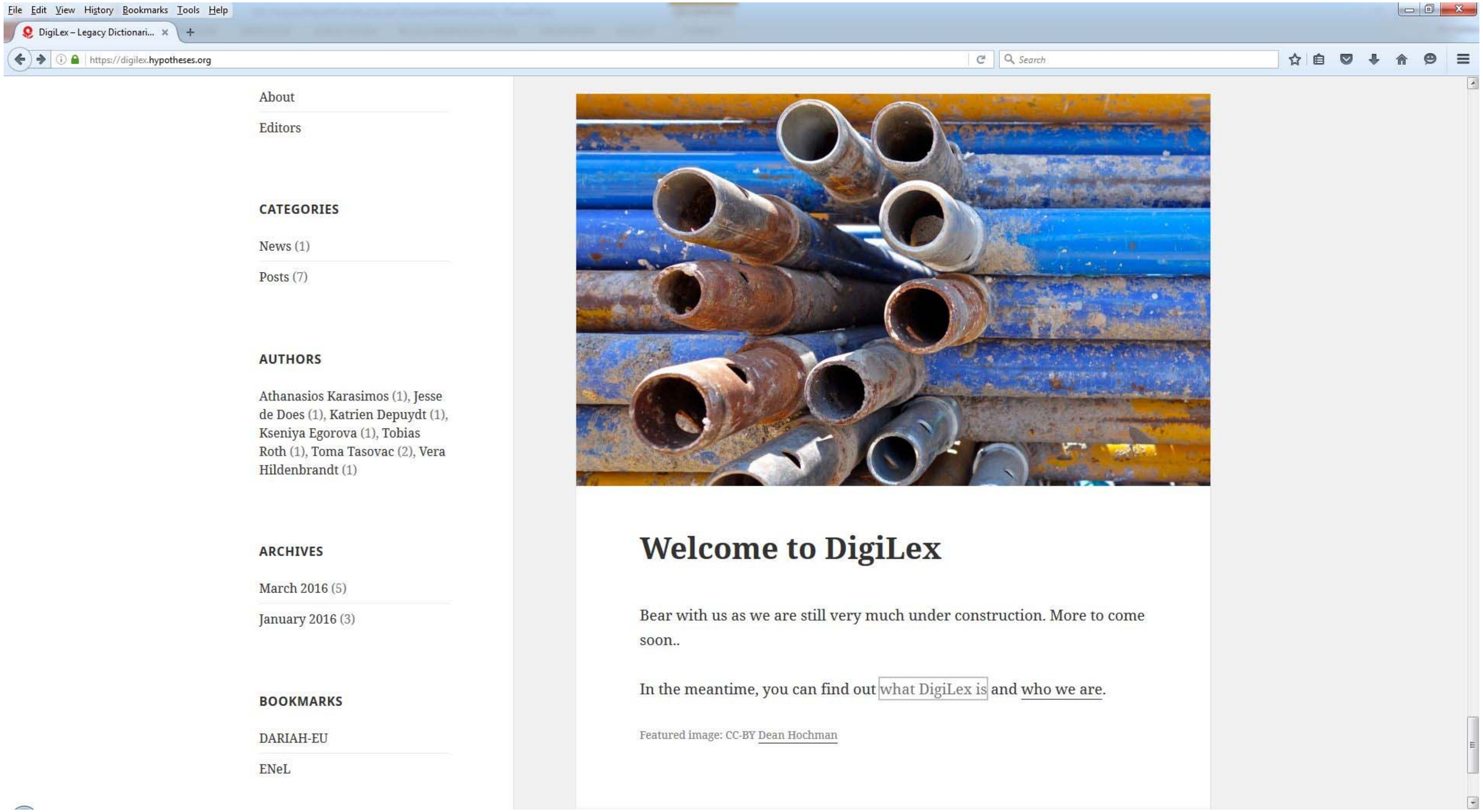
OBJECTIVES OF WORKING GROUP 2

WG 2 will set up guidelines and standards for turning paper dictionaries into a digital format and develop common models in the field of e-lexicography for retro-digitised paper dictionaries already online or planning to go online.

To reach this goal WG 2 will

1. establish an overview of existing retro-digitised dictionaries and dictionaries which should be retro-digitised
2. create guidelines defining standard workflows for the digitisation of dictionaries
3. establish best practices for the encoding of information and the description of relevant information categories for paper dictionaries
4. establish best practices for dictionary enrichment and linking

CONCRETE STEPS – BLOG *DIGILEX. LEGACY DICTIONARIES RELOADED*


File Edit View History Bookmarks Tools Help
DigiLex – Legacy Dictionari... ×
https://digilex.hypotheses.org
Search

About

Editors

CATEGORIES

News (1)

Posts (7)

AUTHORS

Athanasios Karasimos (1), Jesse de Does (1), Katrien Depuydt (1), Kseniya Egorova (1), Tobias Roth (1), Toma Tasovac (2), Vera Hildenbrandt (1)

ARCHIVES

March 2016 (5)

January 2016 (3)

BOOKMARKS

DARIAH-EU

ENeL

Welcome to DigiLex

Bear with us as we are still very much under construction. More to come soon..

In the meantime, you can find out what DigiLex is and who we are.

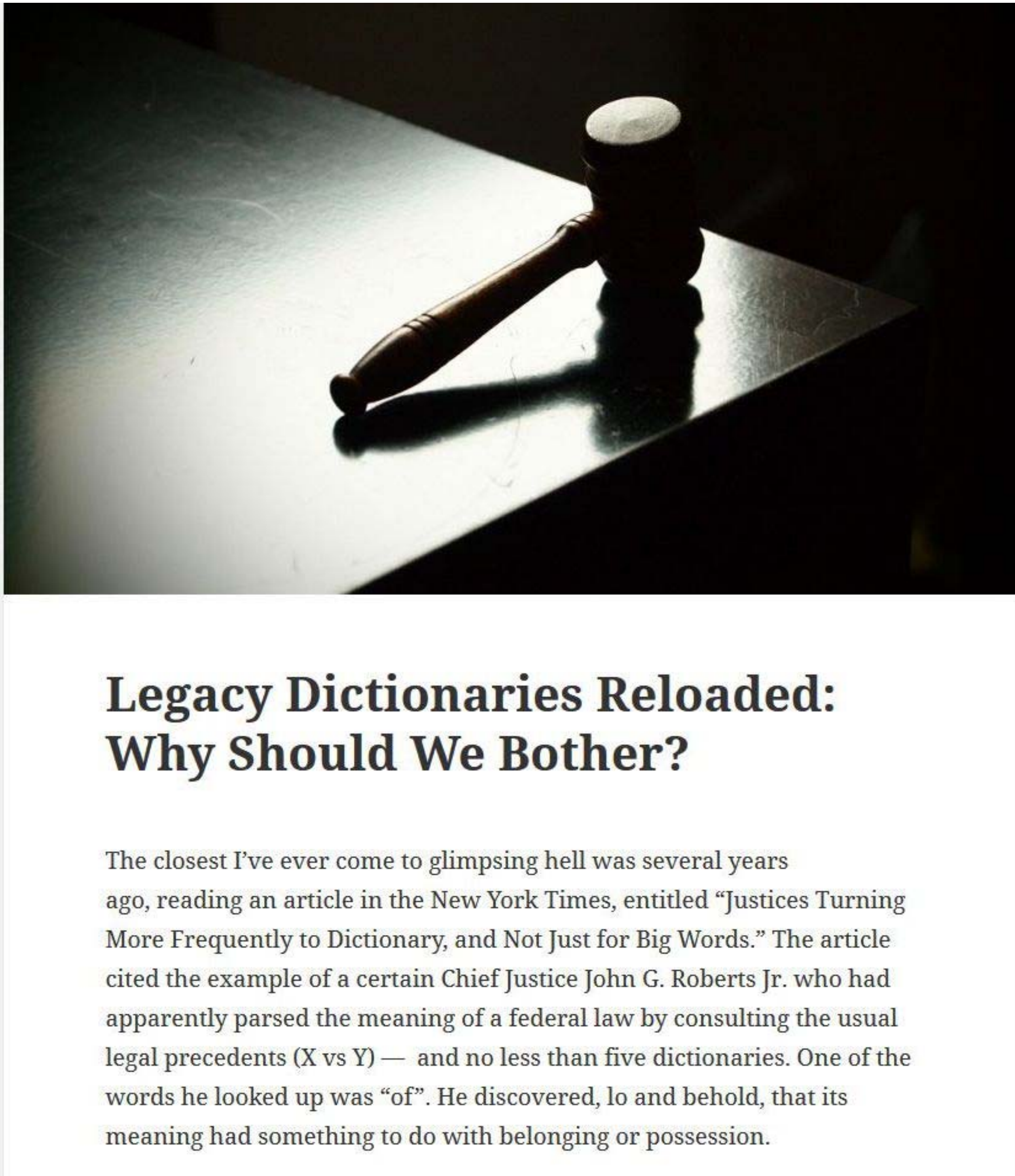
Featured image: CC-BY Dean Hochman

CONCRETE STEPS – BLOG *DIGILEX*

The blog posts reflect upon the central topics of WG 2 by

- describing the workflows for the digitisation of different dictionaries, stressing challenges, problems and solutions
- tackling encoding problems and thus exposing encoding standards
- dealing with issues of dictionary enrichment and dictionary linking.

https://digilex.hypotheses.org/



Legacy Dictionaries Reloaded: Why Should We Bother?

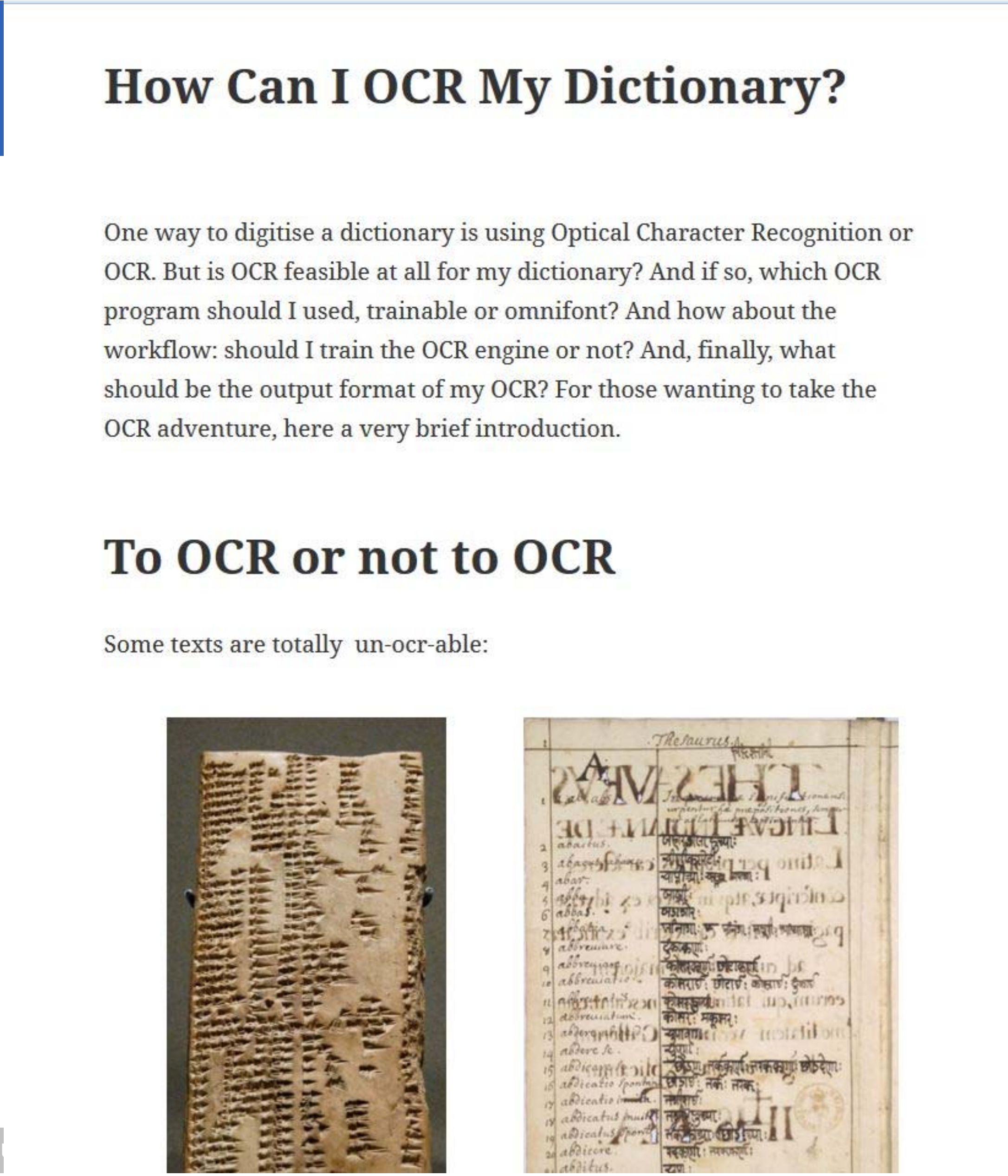
The closest I've ever come to glimpsing hell was several years ago, reading an article in the New York Times, entitled "Justices Turning More Frequently to Dictionary, and Not Just for Big Words." The article cited the example of a certain Chief Justice John G. Roberts Jr. who had apparently parsed the meaning of a federal law by consulting the usual legal precedents (X vs Y) — and no less than five dictionaries. One of the words he looked up was "of". He discovered, lo and behold, that its meaning had something to do with belonging or possession.

CONCRETE STEPS – BLOG *DIGILEX*

The blog will

- enhance the visibility of WG 2 and thus will contribute to the dissemination of the network’s activities
- integrate the members in the working processes of WG 2 and thus help to achieve the defined goals of WG 2.

https://digilex.hypotheses.org/



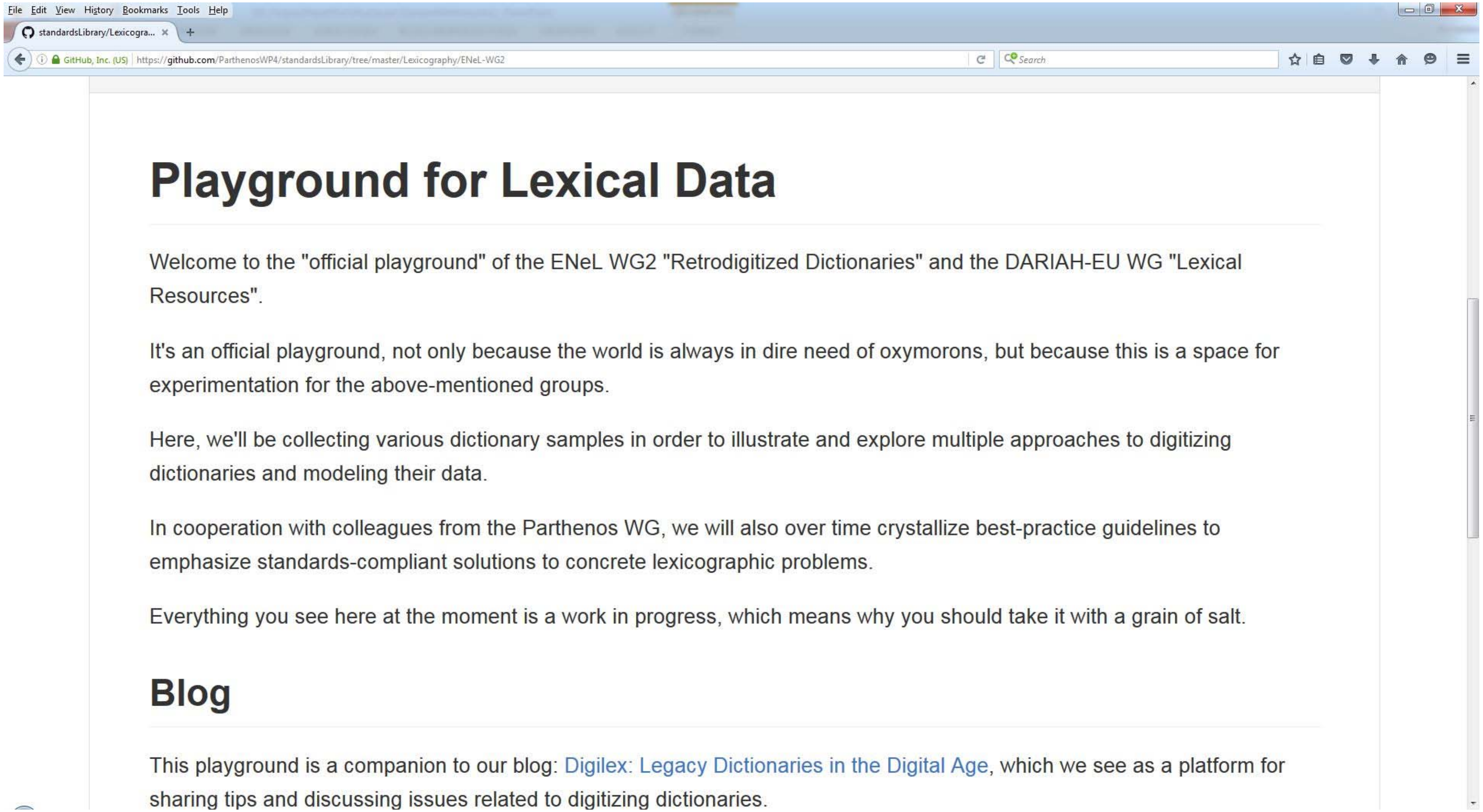
How Can I OCR My Dictionary?

One way to digitise a dictionary is using Optical Character Recognition or OCR. But is OCR feasible at all for my dictionary? And if so, which OCR program should I used, trainable or omnifont? And how about the workflow: should I train the OCR engine or not? And, finally, what should be the output format of my OCR? For those wanting to take the OCR adventure, here a very brief introduction.

To OCR or not to OCR

Some texts are totally un-ocr-able:

CONCRETE STEPS: PLAYGROUND ON GITHUB



File Edit View History Bookmarks Tools Help
standardsLibrary/Lexicogra...
GitHub, Inc. (US) https://github.com/ParthenosWP4/standardsLibrary/tree/master/Lexicography/ENeL-WG2
Search

Playground for Lexical Data

Welcome to the "official playground" of the ENeL WG2 "Retrodigitized Dictionaries" and the DARIAH-EU WG "Lexical Resources".

It's an official playground, not only because the world is always in dire need of oxymorons, but because this is a space for experimentation for the above-mentioned groups.

Here, we'll be collecting various dictionary samples in order to illustrate and explore multiple approaches to digitizing dictionaries and modeling their data.

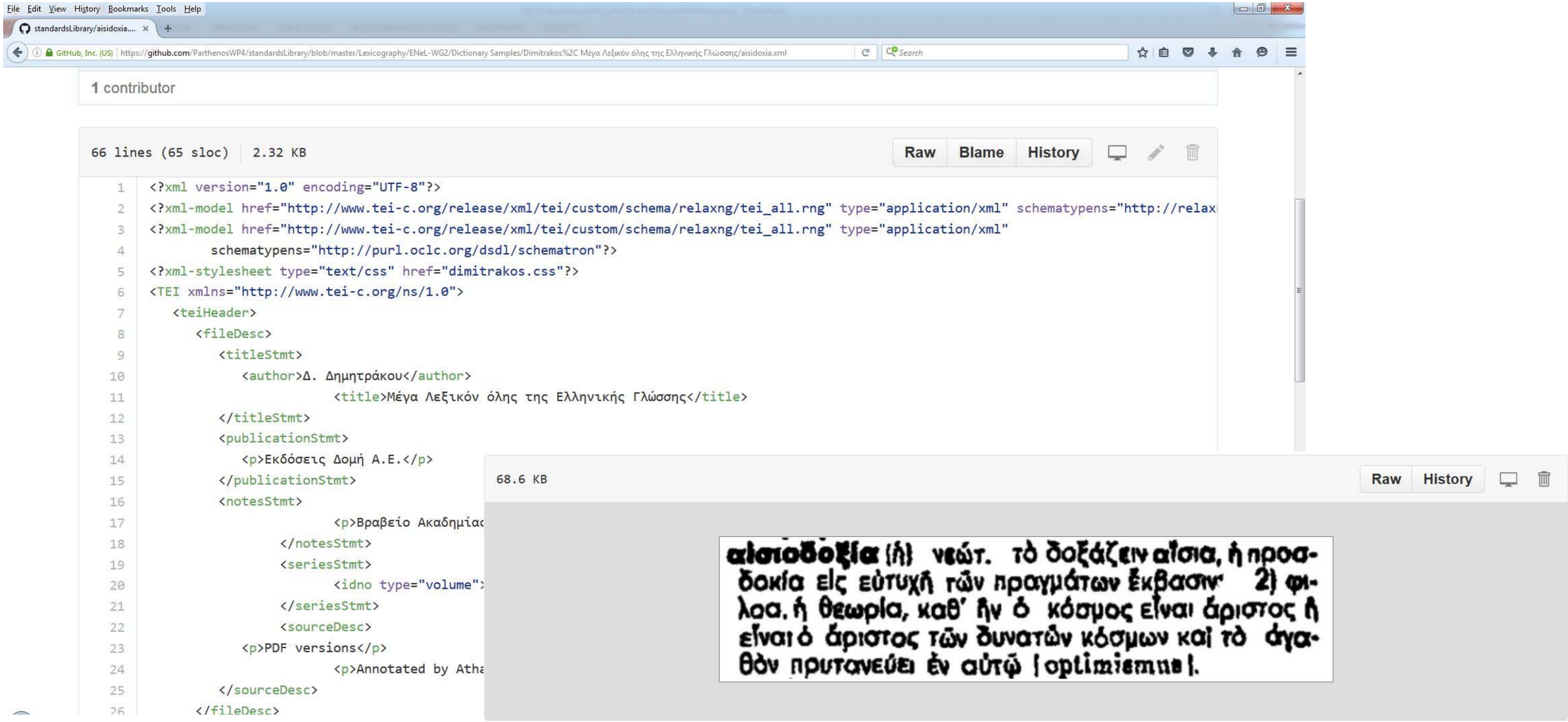
In cooperation with colleagues from the Parthenos WG, we will also over time crystallize best-practice guidelines to emphasize standards-compliant solutions to concrete lexicographic problems.

Everything you see here at the moment is a work in progress, which means why you should take it with a grain of salt.

Blog

This playground is a companion to our blog: Digilex: Legacy Dictionaries in the Digital Age, which we see as a platform for sharing tips and discussing issues related to digitizing dictionaries.

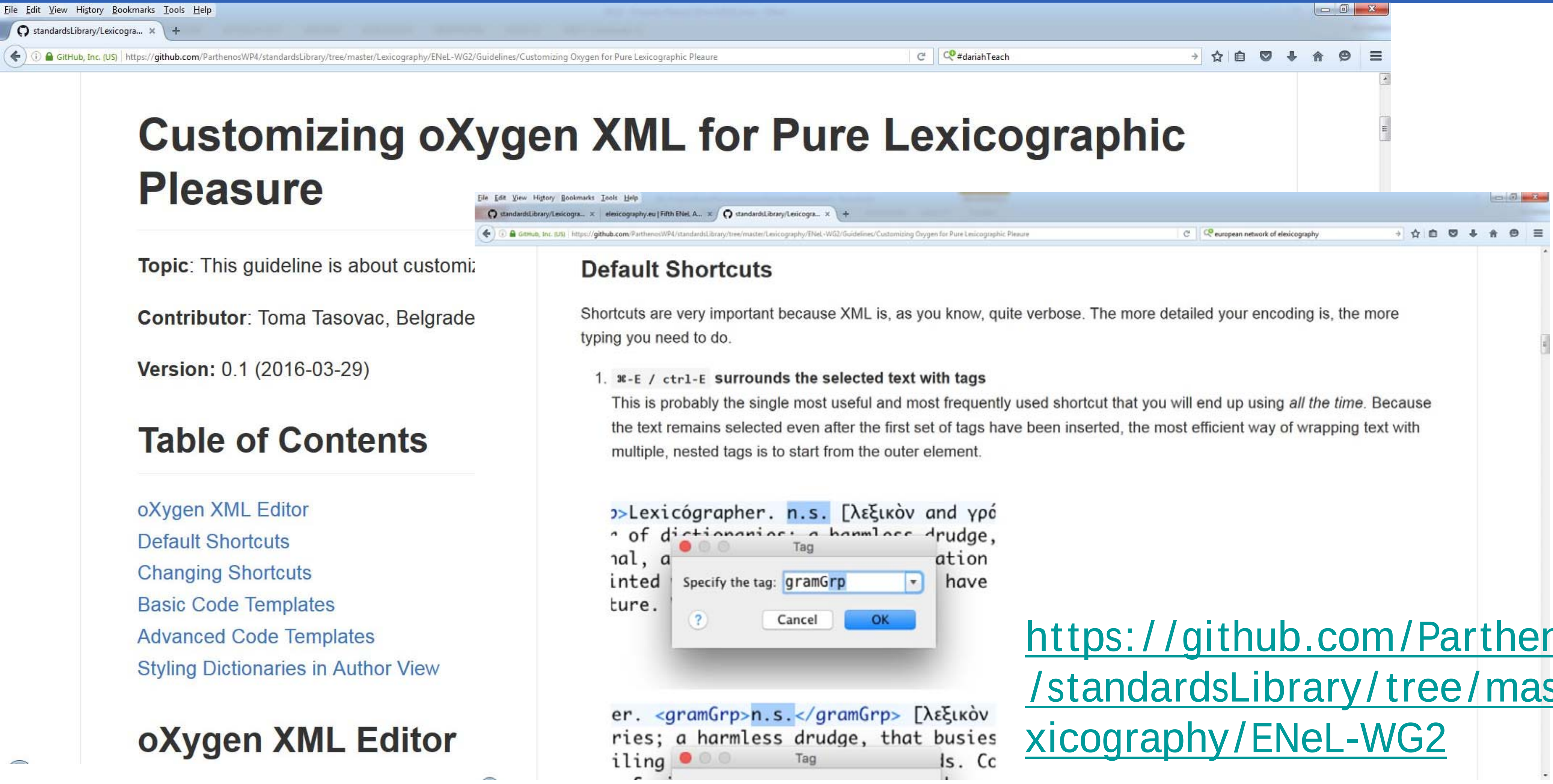
CONCRETE STEPS: COLLECTION OF DICTIONARY SAMPLES


File Edit View History Bookmarks Tools Help
standardsLibrary/aisidoxia....
GitHub, Inc. (US) https://github.com/ParthenosWP4/standardsLibrary/blob/master/Lexicography/ENeL-WG2/Dictionary Samples/Dimitrakos%2C Μέγα Λεξικόν όλης της Ελληνικής Γλώσσης/aisidoxia.xml
Search
1 contributor
66 lines (65 sloc) 2.32 KB
Raw
Blame
History

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-model href="http://www.tei-c.org/release/xml/tei/custom/schema/relaxng/tei_all.rng" type="application/xml" schematypens="http://relax
<?xml-model href="http://www.tei-c.org/release/xml/tei/custom/schema/relaxng/tei_all.rng" type="application/xml"
        schematypens="http://purl.oclc.org/dsdl/schematron"?>
<?xml-stylesheet type="text/css" href="dimitrakos.css"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0">
   <teiHeader>
      <fileDesc>
         <titleStmt>
            <author>Δ. Δημητράκου</author>
                        <title>Μέγα Λεξικόν όλης της Ελληνικής Γλώσσης</title>
         </titleStmt>
         <publicationStmt>
            <p>Εκδόσεις Δομή Α.Ε.</p>
         </publicationStmt>
         <notesStmt>
                        <p>Βραβείο Ακαδημίας
                 </notesStmt>
                 <seriesStmt>
                        <idno type="volume">
                 </seriesStmt>
                 <sourceDesc>
            <p>PDF versions</p>
                        <p>Annotated by Atha
         </sourceDesc>
      </fileDesc>
```

αἰσιοδοξία (ἡ) νεώτ. τὸ δοξάζειν αἴσια, ἡ προσδοκία εἰς εὐτυχῆ τῶν πραγμάτων ἔκβασιν· 2) φιλοσ. ἡ θεωρία, καθ' ἣν ὁ κόσμος εἶναι ἄριστος ἢ εἶναι ὁ ἄριστος τῶν δυνατῶν κόσμων καὶ τὸ ἀγαθὸν πρυτανεύει ἐν αὐτῷ (optimisme).

CONCRETE STEPS: CREATING AND DELIVERING GUIDELINES



File Edit View History Bookmarks Tools Help
standardsLibrary/Lexicogra...
https://github.com/ParthenosWP4/standardsLibrary/tree/master/Lexicography/ENeL-WG2/Guidelines/Customizing Oxygen for Pure Lexicographic Pleaure
#dariahTeach

Customizing oXygen XML for Pure Lexicographic Pleasure

Topic: This guideline is about customi

Contributor: Toma Tasovac, Belgrade

Version: 0.1 (2016-03-29)

Table of Contents

oXygen XML Editor
Default Shortcuts
Changing Shortcuts
Basic Code Templates
Advanced Code Templates
Styling Dictionaries in Author View

oXygen XML Editor

standardsLibrary/Lexicogra...
elexicography.eu | Fifth ENeL A...
standardsLibrary/Lexicogra...
european network of elexicography

Default Shortcuts

Shortcuts are very important because XML is, as you know, quite verbose. The more detailed your encoding is, the more typing you need to do.

1. `⌘-E / ctrl-E` **surrounds the selected text with tags**
   This is probably the single most useful and most frequently used shortcut that you will end up using *all the time*. Because the text remains selected even after the first set of tags have been inserted, the most efficient way of wrapping text with multiple, nested tags is to start from the outer element.

```
>Lexicógrapher. n.s. [λεξικὸν and γρά
 of dictionaries; a harmless drudge,
ıal, a                            ation
inted                             have
ture.
```

```
er. <gramGrp>n.s.</gramGrp> [λεξικὸν
ries; a harmless drudge, that busies
iling                         ls. Cc
```

https://github.com/ParthenosWP4/standardsLibrary/tree/master/Lexicography/ENeL-WG2

CONCRETE STEPS: COOPERATION WITH #DARIAHTEACH

- #dariahTeach is producing an online platform for the delivery of high-quality Digital Humanities training materials.
- A separate online module on retro-digitizing dictionaries will be completed and published by the summer of 2017. It will be officially co-branded with ENeL.

BARCELONA MEETING

- advanced follow-up to the Lisbon Training School in July 2015
- mixture of short presentations and hands-on sessions
  - offered additional instruction in more advanced topics that were not covered in Lisbon
  - addressed the issue of technical heterogeneity in lexical sources in a joint session with WG 4 chaired by Laurant Romary
  - worked on concrete encoding issues, challenges and problems

NEXT STEPS

Keep doing what we are doing!