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
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
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/]
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.

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 use, 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:

https://digilex.hypotheses.org/
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
Customizing oXygen XML for Pure Lexicographic Pleasure

**Topic:** This guideline is about customizing oXygen XML.

**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**

**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. `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.

   ![Screenshot of oXygen XML with a tag selection dialog box]

   - **Specify the tag:** `gramGrp`
   - **Options:**
     - `OK`
   - **Text:** `ερ, <gramGrp n.s. δλεξίκον and γράμματα; a harmless drudge, that busies itself in literary matters. Co`

[GitHub Repository](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!