A Digital Classics (Sunoikisis) (DH.DC)

Project – 2015

Created by Agata Barcik,

Pierre Motylewicz & Eugen Rein

Supervised by Dr. Monica Berti

Contents[1]

[1] All links to online resources were last checked on the 21st of September 2015.

1. Introduction

The Koios-project is a result of the Sunoikisis DC 2015[2] course held at the Alexander von Humboldt Chair of Digital Humanities at the University of Leipzig and led by Dr. Monica Berti. For further information, please visit the Sunoikisis DC 2015 GitHub and Google+ pages.

"In Greek Mythology, **Coeus** (Ancient Greek: Κοῖος"query, questioning") was one of the Titans […] the embodiment of the celestial axis around which the heavens revolve. The etymology of Coeus' name provided several scholars the theory that Coeus was also the Titan god of intellect, who represented the inquisitive mind"[3]

Therefore Κοῖος seemed a fitting name for a project that aimed to provide a tool for scholarly analyzing, querying and enriching classical texts on the basis of place names.

The goal of Koios was the creation of a program that is able to:

1. Provide identification of place names in Classical texts
2. Go beyond the language barrier and to work with other editions of the same texts
3. Annotate place names
4. Provide a preliminary disambiguation via links/references to external resources
5. Identify of other editions and translations of the same text via comparing annotated place names
6. Create an alignment with other editions/translations
7. Reevaluate links and references based on other editions and translations

The topic of Sunoikisis DC 2015 was focused on the two historical periods of ancient Greek history known as the "Pentekontaetia" and the "Peloponnesian War". The main source for this period is the Athenian author Thucydides. We therefore started our research project with different editions of the "History of the Peloponnesian War". Perseus und the Open Philology Project already provided us with several XML files of Thucydides' work. Soon we realized there are several tools and projects produced by the community of digital humanists which could provide data and guidelines for our work and we had to evaluate them in the process of building our program.

Due to the limits of the available resources, which still need a lot of work in order to be fully machine readable, we achieved some results during the first four steps. We hope the outcomes of our project and even the problems we encountered might provide insight for further ventures.

2. Team

This project is based on teamwork between humanists and computer scientists:

Agata Barcik is doing her master in Computer Science and her bachelor in Linguistics at the University of Leipzig. She is interested in Natural Language Processing.

Pierre Motylewicz is a Theology and Oriental Studies major at the University of Halle-Wittenberg with a focus on Medieval Latin, Greek and Armenian sources and an interest in all digital aspects.

Eugen Rein is a student of computer science at the University of Leipzig. He achieved his bachelor's degree at the Technical University of Berlin specializing in artificial intelligence. At the University Leipzig he is improving his knowledge within this field while pursuing his master's degree.

[2] http://www.dh.uni-leipzig.de/wo/courses/summer-semester-20142015/module-digital-classics-sunoikisis/
[3] https://en.wikipedia.org/wiki/Coeus

3. The available data: stepping stones and stumbling blocks – Pierre Motylewicz

Perseus[4] provided us with twelve XML files[5] of Thucydides' "History of the Peloponnesian War": one French file, three German, six English and two Greek files. The French edition was discarded for our project because none of us possessed any expertise in French.

The three German files are actually two editions. The first was produced by Adolf Wahrmund (1864), it was incredible hard to read because it is poorly structured, with tags all over the place and contains uncountable OCR mistakes. The second German edition created by Theodor Braun (1917) consists of two books (and therefore in this case two XML files). It is far more user friendly, but it still contains quite a lot of OCR errors. In spite of that, the files of this edition were good enough. (The XML files suggest the editions were originally digitalized by HATHI Trust[6] yet the versions provided by HATHI Trust contain completely different OCR errors than the Perseus files.)

The six English files are actually three editions. The last three files contain the same texts as the other three but are updated to (more or less) EpiDoc[7] standards. The editors of the first two versions are Thomas Hobbes (1843) and Benjamin Jowett (1881). The third edition dated 1910 has no mention of the editor or translator and is probably the work of Richard Crawley (based on the year of print and comparison of a few paragraphs with online available PDF versions, but that would indicate the XML document has a wrong title, "The Peloponnesian War" instead of "The History of the Peloponnesian War“). This edition contains as an additional resource placeName tags with key attributes referring to either Perseus or the Getty Thesaurus[8].

The Greek edition is also available in two files, the newer version was updated to EpiDoc standards. It was originally printed by Oxford University Press in 1942.

At this point we had selected three different XML editions we were going to work with. The English version including the placeName tags was chosen as our blue print and data source and we wanted to apply the extracted place names to the German edition from Braun and to the Greek edition.

The first step was extracting the place names and the connected data from the English edition. The second step was adding the German and Greek name variants to the extracted data and the third was enriching the German and Greek texts with the place name data. Besides some pitfalls contained in the different structure of the files (see the following table), the first step was largely successful from a technical point of view. It still took several tries though, mostly because the files contained additional and partially somewhat random undocumented data and structures (unexpected additionally encoded information like lyric parts in the English version or quotes in the Greek text) we had to identify and take into account for our program – something that would become even a bigger problem during the third step (for technical aspects see chapter 4 and 7).

<div type="textpart" subtype="chapter" n="63"> <div type="textpart" subtype="section" n="1"> <p>Returning from the pursuit, Aristeus perceived the defeat of the rest of the army.</p> <p>Being at a loss which of the two risks to choose, whether to go to Olynthus or to Potidaea, he at last determined to draw his men into as	<div type="textpart" subtype="chapter" n="63"> <div type="textpart" subtype="section" n="1"> <p><milestone ed="P" unit="para"/>ἐπαναχωρῶν δὲ ὁ Ἀριστεὺς ἀπὸ τῆς διώξεως, ὡς ὁρᾷ τὸ ἄλλο στράτευμα ἡσσημένον, ἠπόρησε μὲν ὁποτέρωσε διακινδυνεύσῃ χωρήσας, ἢ ἐπὶ τῆς Ὀλύνθου ἢ ἐς τὴν Ποτείδαιαν: ἔδοξε δ᾽ οὖν ξυναγαγόντι τοὺς μεθ᾽ αὑτοῦ	<div type="textpart" subtype="chapter" n="63"> <p>Aristeus, der die Verfolgung aufgegeben hatte, als er sein<lb/> übriges Heer geschlagen sah, war anfangs zweifelhaft, wohin <lb/> er sich wenden sollte, um sich aus der Gefahr zu ziehen, ob <lb/> nach Olynth oder nach Potidäa, entschloß sich dann aber, <lb/> seine Truppen möglichst eng zusammenzuziehen und sich mit <lb/>

[4] http://www.perseus.tufts.edu/hopper/
[5] https://github.com/PerseusDL/canonical-greekLit/tree/master/data/tlg0003/tlg001
[6] https://www.hathitrust.org/digital_library
[7] http://www.stoa.org/epidoc/gl/latest/
[8] http://www.getty.edu/research/tools/vocabularies/tgn/index.html

small a space as possible, and force his way with a run into Potidaea.</p> <p>Not without difficulty, through a storm of missiles, he passed along by the breakwater through the sea, and brought off most of his men safe, though a few were lost.</p>	ὡς ἐς ἐλάχιστον χωρίον δρόμῳ βιάσασθαι ἐς τὴν Ποτείδαιαν, καὶ παρῆλθε παρὰ τὴν χηλὴν διὰ τῆς θαλάσσης βαλλόμενός τε καὶ χαλεπῶς, ὀλίγους μέν τινας ἀποβαλών, τοὺς δὲ πλείους σώσας.</p>	ihnen im Sturmschritt nach Potidäa durchzuschlagen. Auch <lb/> gelang es ihm, vom Hasendamme her durchs Wasser um die <lb/> Stadt herumzukommen und, wenn auch beschossen und nur <lb/> mit Mühe und nicht ganz ohne Verluste, das meiste in Sicher--<lb/> heit zu bringen. <pb n="39"/> [...] </p>

After completing the extraction we had a list of 233 English place names that revealed some problems when we started working with them:

- same place, same identifier, different name: Achaia/Achaea
- Same place, different identifier, same name:
  key="perseus,Corcyra City“ – key="perseus, Corcyra City“
  key="perseus,Naupaktos" – key="Naupaktos"
- different place, same identifier, same name: Chalcis (Aitolia/Eubioa)
- different place, mixed up identifiers, same name: Argos (key="perseus,Amphilochian Argos"), Argos (key="perseus, Argos")
- lots of places were not tagged
- strings wrongly tagged as places and entirely wrong identifiers
  The <placeName key="tgn,1045322">Mede</placeName> defeated, great Pausanias raised this monument, that Phoebus might be praised.
  This identifier belongs to a modern village in northern Italy.

In spite of these problems, we could still use the list we got in some ways.

First we wanted to add the German versions of all the place names in this list. We also wanted to use additional modern languages (Polish, Armenian and Russian) for evaluation purposes. Initially we used the geonames.org API[9] to query ever place one by one with a Python script. The translations were incomplete and very odd for some uncommon places. The second step was to query the Wikipedia API[10]. Finally in the last step we merged both results from Geonames.org and Wikipedia. In some cases we had to decide which translation was the better one. Sometimes we had to remove additional disambiguation information for the places we got from Wikipedia.[11]

While we didn't have the time to evaluate all the hits we got in detail, the pure numbers are nevertheless interesting:

Language	German	Russian	Polish	Armenian
Articles in total	1853000	1253000	1133000	185357
corresponding Articles	202	181	169	61
Ratio (corresponding /total)	10,9^-5	14,4^-5	14,9^-5	32,9^-5

We couldn't possibly expect to find the same number of articles corresponding to our English list of place names in the other languages. When we looked at ever smaller versions of Wikipedia the number of relevant articles however didn't decline nearly as fast as the overall number of articles.

[9] http://api.geonames.org
[10] https://en.wikipedia.org/w/api.php
[11] This task was undertaken by Eugen Rein.

The reason why number of articles that agree with our list of place names becomes an ever bigger part of the databases when the overall number of articles in a version of Wikipedia gets smaller is yet to be determined. The first ideas we had was that these well represented names belong to a certain core of articles which is either founded on globally available geographical data or on common historical significance – a first test for these ideas would probably consist of including data from a few East and Central Asian Wikipedia versions under the assumption that the history of Greece is probably less relevant for Japan than it is in a European context (while Armenia is not part of Europe geographically, its history was always closely connected). A similar ratio in Asian versions of Wikipedia would therefore point towards globally available data while a smaller ratio would indicate historical significance as determining factor.

It is still obvious however that Wikipedia is a good starting point if you look for modern translations of place names, but only a first step especially when handling rarer languages.

Getting the Greek names was a different task. We had to add the ancient Greek lemma manually, yet that wasn't quite enough because we hoped to automatically identify place names in our texts at a later point. Instructing a computer to look not only for the nominative form, but also for inflected forms isn't that hard when there are only slight changes like the "s" and the end of German forms in genitive case. Greek is way trickier because the identification of forms depends not only on the case but also on:

– neighboring words (for example enclitics)
– the place in a sentences (acute>grave)
– different times and dialects
– the encoding of certain parts of the Greek language
– OCR problems: even more than with other Indo European languages there is chance that OCR produces results that seem to be correct or at least unproblematic to the human eye but result in problems for automated processes for example the mix up between certain dots and lines

Luckily we were allowed to test a tool called "Lexicon Formarum Graecarum Treebankearum" (LeFGreaT)[12] that is based on the efforts of Dr. Guiseppe G. A. Celano and Tariq Yousef (for further information on the tool and the process of extracting the data see chapter 7).

The Lexicon is built on different data sets:

- 1) http://nlp.perseus.tufts.edu/syntax/treebank/greek.html
- 2) The New Testament and Herodotus from PROIEL
- 3) Some Aesop and Apollodorus
- https://github.com/PerseusDL/treebank_data/tree/master/AGDT2/corpus/stableannotations

We were hoping to get some variants you'd expect to see in a real text for our Greek place names. We used 177 Greek names derived from the list with the 233 English place names (not all Greek names could be found or identified with absolute certainty) and exactly 100 of those names were attested in the LeFGreaT with a frequency ranging from one to well over hundred hits.

With those attested forms there was roughly a one in five chance to get the expected grammar table forms (one form per case except vocative) but there are examples with up to thirteen attested forms. These additional forms were partially dialect forms, partially midsentence forms (changing of accents, elision, …) but were also created through different encoding or writing (with our without diaresis, …) with only an insignificant number caused by mistakes in the underlining treebanks (wrong number, case or gender). While some lemmas also contained adverb and adjective forms others didn't without any obvious reasons and thus probably originated in different treebanking guidelines for certain corpora of the Lexicon.

While the Lexicon gave us some valuable information, it is still at a stage that isn't sufficient for the things we had in mind, but it might become an irreplaceable resource for similar projects – or might already be one for projects working on texts with a closer relation to the sources the Lexicon was built on.

[12] http://tariq-yousef.com/GT/

As already mentioned it was easier working with the German text because it wasn't that hard to search the names we already had on our list semi automatically, while looking at the same time for additional place names in the XML file. But we had to realize we weren't that sure what exactly is considered a place name.

TEI[13] provides us with some tags that have a close relation to the <place name> tag itself: <rs>, <name>, <place>, <geogName>, <geogFeat> but there are many options to handle those and not many were clearly spelled out, so we were looking for guidelines with a narrower focus. The EpiDoc guidelines state: "References to places (such as settlements, regions, nations), peoples (such as ethnic groups, civic populations) and other geographical features (such as seas, mountains, rivers) can be identified in the transcribed text using the elements <placeName> (for mostly political entities: cities, nations, etc.) or <geogName> (for physical features: rivers, mountains, etc.). Further information about the kind of geographical entity being named may be specified using a type attribute, for example identifying ethnic names as opposed to the place names themselves; or subcategorising geographic names as the type of physical entity being names. Names may be linked using the attribute nymRef, which will contain a URL or other URI pointing to the standard form of this name (nominative singular; normalized spelling) in a local XML table or database or online onomasticon. More usefully the names may be linked using the ref attribute to a place identifier, via a pointer to a local place database or online gazetteer such as Pleiades." This was way more specific and seemed applicable, even though it seemed somewhat at odds with the TEI guidelines when it comes to ethnic names and it pointed us to Pleiades.

The Gazetteer Pleiades[14] is part of the Pelagios[15] Network which provides its own tagging principles:

- Tag names but not categories: e.g. Mount **Aetna** (NOT **Mount Aetna**)
- Don't tag articles or prepositions UNLESS the toponym cannot be properly understood without it. e.g. the Brigantes, OR al-Andaluz, (NOT the Brigantes, al-Andaluz)
- Don't tag place references inserted by modern translators or editors, such as footnotes, clarifications or summary titles.
- Descriptive place names (including those containing prepositions or definite articles) are OK : e.g. ad aras, the Stone Tower
- Tag multiple parts of a single name as one place reference, but alternative names as multiple references: e.g. Segida also surnamed Augurina, BUT Troy, also called Ilion.
- Ethnoi SHOULD be tagged where they are used as shorthand for the territory they occupy. They SHOULD NOT be tagged where they are being referred to as an agent in a narrative. e.g. Beyond the River Ister are the Marcomanni BUT the Marcomanni crossed the Ister.

While these principles are partially hard to bring in line with the TEI examples for geogNames, geogFeats and for the definitions of what actually is part of a name, they are concrete for most parts. The principles for ethnoi on the other hand are ambiguous. This lead not only to inconsistent tagging within the team, but also to different results over time.

The other problem was the disambiguation process, while Pleiades is pretty much the best tool to get references to place names in historical texts, it's still far away from being perfect. The main issues we encountered are

- Stability: while the site is available nearly constantly from time to time it's excruciatingly slow
- Lack of a Manual: While it's easy enough to start working with Pleiades at first, it is hard work to figure out how to make valuable contributions – its openness to contributions is one of the big perks of Pleiades – and to find and use some of neat features Pleiades provides is difficult.
- Because it is a community sourced project you need a login to make contributions and your contributions have to be approved. While this guarantees for some quality, it might also take quite a while till you get feedback and see your additions online.

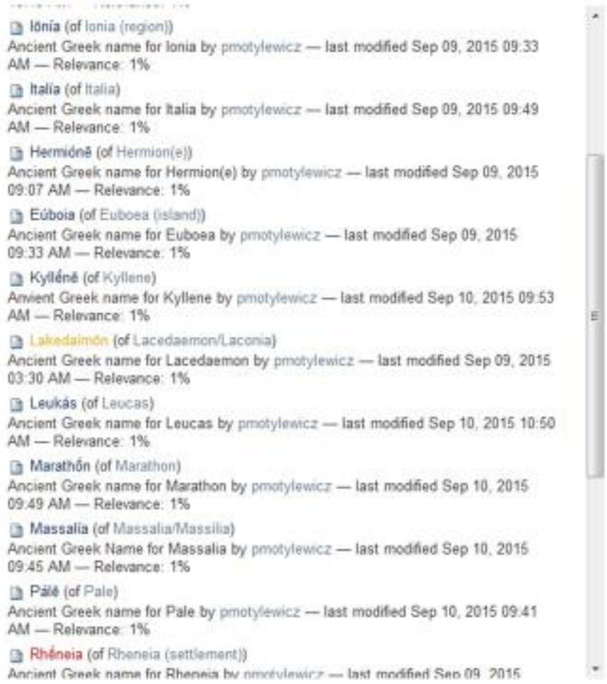
[13] http://www.tei-c.org/index.xml

[14] http://pleiades.stoa.org/

[15] http://pelagios-project.blogspot.co.uk/

While there is right now a good chance to find most places you are looking for especially in Europe and the Mediterranean area and some parts of Asia, it often takes some time to find what you are looking for. Most places are only available under their English name and only under one possible English name: the modern form, the name derived from the Greek form or the name derived from the Latin form – so you might need to give it a few tries. On the other hand, it is sometimes hard to identify the correct version of something looking for – if possible at all, because there are often several entries with the same name and location, for an island/city/territory in different time periods without any or only little explanation.

That said, the potential of Pleiades is amazing and we are happy we could make a few small contributions by adding Ancient Greek forms, links and some explanations to a lot of places we found in our extracted list of place names from Thucydides.



Iōnía (of Ionia (region))
Ancient Greek name for Ionia by pmotylewicz — last modified Sep 09, 2015 09:33 AM — Relevance: 1%

Italía (of Italia)
Ancient Greek name for Italia by pmotylewicz — last modified Sep 09, 2015 09:49 AM — Relevance: 1%

Hermiónē (of Hermion(e))
Ancient Greek name for Hermion(e) by pmotylewicz — last modified Sep 09, 2015 09:07 AM — Relevance: 1%

Eúboia (of Euboea (island))
Ancient Greek name for Euboea by pmotylewicz — last modified Sep 09, 2015 09:33 AM — Relevance: 1%

Kyllḗnē (of Kyllene)
Anwent Greek name for Kyllene by pmotylewicz — last modified Sep 10, 2015 09:53 AM — Relevance: 1%

Lakedaímōn (of Lacedaemon/Laconia)
Ancient Greek name for Lacedaemon by pmotylewicz — last modified Sep 09, 2015 03:30 AM — Relevance: 1%

Leukás (of Leucas)
Ancient Greek name for Leucas by pmotylewicz — last modified Sep 10, 2015 10:50 AM — Relevance: 1%

Marathṓn (of Marathon)
Ancient Greek name for Marathon by pmotylewicz — last modified Sep 10, 2015 09:49 AM — Relevance: 1%

Massalia (of Massalia/Massilia)
Ancient Greek Name for Massalia by pmotylewicz — last modified Sep 10, 2015 09:45 AM — Relevance: 1%

Pálē (of Pale)
Ancient Greek name for Pale by pmotylewicz — last modified Sep 10, 2015 09:41 AM — Relevance: 1%

Rhḗneia (of Rheneia (settlement))
Ancient Greek name for Rheneia by pmotylewicz — last modified Sep 09, 2015

+
−
Powered by Leaflet and Mapbox. Map base by AWMC, 2014 (cc-by-nc).

4. Querying the files – *Agata Barcik*

The main aim is to use the already exisiting Greek corpus in **Perseus** and extract – just as in the English version of the corpus – the named entities.

In the English version named entities are linked and get a unique number. All of the files are available in XML format so it is easier to process them and extract the information you need.

As already mentioned there are XML files available for download. First we need to find out what kind of information and how it is marked in the English version of the corpus. This is a really important issue, because every name has an assigned number, which is unique, so it would be easy to connect the same information in two different languages.

```
<div2 type="chapter" n="2"><p>


  <milestone unit="section" n="1"/>

              <seg>

                <milestone unit="para" ed="P"/>For instance, it is evident that the
country

                now called <placeName key="tgn,1000074">Hellas</placeName> had
in ancient times no settled population; on the contrary, migrations were of
```

CCode snippet 1

In the example above you can see how the text is built. The XML structure contains the main tags like **div2** with chapter's information, then the milestones and finally placeNames. In the first step we extracted all these information using XQuery.

```
xquery version "3.0" encoding "utf-8";


declare variable $y := doc("/Users/macbook/Desktop/thucEng150526.xml");

let $j := $y//ancestor::text




for $o in $j//placeName
```

Code snippet 2

The extracted list looks like the one below:

```
<placeName key="tgn,1000074">Hellas</placeName>

<placeName key="tgn,7001399">Thessaly</placeName>

<placeName key="tgn,7002683">Boeotia</placeName>

<placeName key="tgn,7017076">Peloponnese</placeName>
```

```
<placeName key="tgn,7002735">Arcadia</placeName>
…
```

The list contains – after sorting out the repeated forms – 233 unique placeNames. After that, all the places where manually translated into Greek. There are four cases, so every English form was assigned four Greek forms.

4.1 Data sets and first evaluations

The first data set is based on the 10 frequent forms, which were manually translated and the right different word forms were assigned. (See listing below)

```
<placeNames>
<placeName ref="http://pleiades.stoa.org/places/579885">Ἀθῆναι</placeName>
<placeName ref="http://pleiades.stoa.org/places/579885">Ἀθηνῶν</placeName>
<placeName ref="http://pleiades.stoa.org/places/579885">Ἀθῆναις</placeName>
<placeName ref="http://pleiades.stoa.org/places/579885">Ἀθῆνας</placeName>
<placeName ref="http://pleiades.stoa.org/places/579885">Ἀθῆνάι</placeName>
<placeName ref="http://pleiades.stoa.org/places/579885">Ἀθῆνάις</placeName>
<placeName ref="http://pleiades.stoa.org/places/579885">Ἀθῆνάς</placeName>
<placeName ref="http://pleiades.stoa.org/places/1001896">Ἑλλάς</placeName>
<placeName ref="http://pleiades.stoa.org/places/1001896">Ἑλλάδος</placeName>
<placeName ref="http://pleiades.stoa.org/places/1001896">Ἑλλάδι</placeName>
<placeName ref="http://pleiades.stoa.org/places/1001896">Ἑλλάδα</placeName>
<placeName ref="http://pleiades.stoa.org/places/540689">Βοιωτία</placeName>
...
</placeNames>
```

All these word forms were searched by using the following query:

```
xquery version "1.0";
declare variable $original := doc("/Users/macbook/Desktop/Sunokisis/perseus-grc2-1.xml");
declare variable $places := doc("/Users/macbook/Desktop/Sunokisis/samplePlaces.xml");
copy $string3 := $original
modify
(
for $y in $string3//p
for $z in $places//placeName
  where $y[contains(.,$z)] and $z
  return ( insert node (<placeName  ref="{data($z/@ref)}">{string($z)}</placeName>)  into $y)
```

```
 )
return $string3
```

In 70 chapters there were 337 place names.

```
xquery version "1.0";
let $places := doc("/Users/macbook/Desktop/Sunokisis/insertedPlaces.xml")
for $y in $places//div//div[@subtype="chapter"]
return (count($y//placeName), "&#10;")
```

The same kind of comparison resulted in the German version: 105 places were found. It's not a representative result since the file was badly structured and just a part of it was analyzed.

The XQuery would be a perfect solution, if the lists of places are already done and the file contains just the right structure – every word element besides being a part of a chapter or a subsection is a single element concerning the XML structure.

That was unfortunately the case in this project. So first the data had to be completely restructured starting with the tokenization, which meant treating each element of a sentence as a token. This could help querying the data and the comparison could contain the token tag and an element from the list of place names. Since it was not the case, the results were inserted at the end of the subsection.

5. Recogito as an alternative – *Agata Barcik*

The second approach was based on working with Recogito. Recogito is a kind of framework for automatic data annotation, on the webpage you can already see the annotated data.

Using Recogito seemed to be a good alternative to our semi-automatic solution. Theoretically the idea is easy, because for your own purpose you “just” need to install it, make it run and then upload your text with some place names. Because the installation steps were very bad documented on their main web page, we struggled a lot to make it run.   (See the link below)

https://docs.google.com/document/d/1z_L39-KuMcYv-o356ELlIX8gv1hsGgejB_CjSIsYl9E/edit

After the fully successful installation process, the conversion of the XML-JSON format failed, so it was impossible to upload the data and test it.

Summing up: All these two steps would have gone much better, if we had concentrated at the file format in the first place. Being not that familiar with the right XML structure causes a lot of problems.

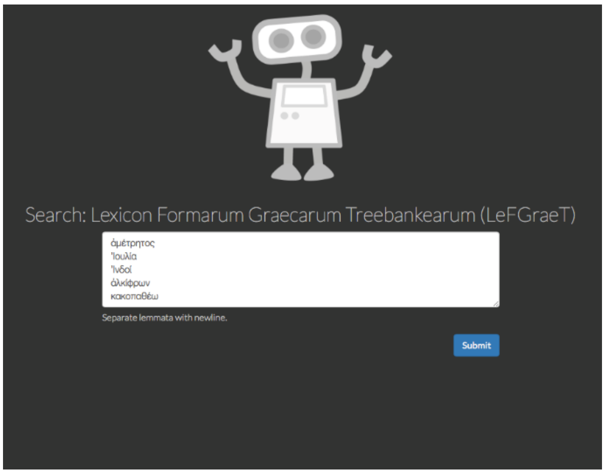
6. Querying Lexicon Formarum Graecarum Treebankearum (LeFGraeT) – *Eugen Rein*

After examining a part of Thucydides' work we had many Greek words. In order to produce better results during manual and automatic named entity recognition our idea was to use the Greek lemmas of the words we had already found.

Treebanking is a process that helped us to get the Greek lemma form, but it would have been a far too long procedure to get us the data we needed. For this reason Dr. Giuseppe G.A. Celano provided us with a tool called "Lexicon Formarum Graecarum Treebankearum" (LeFGreaT)[16] that he developed together with Tariq Yousef. With this online software one can search for a Greek word and get one or possibly multiple Greek lemma forms of this word. If you wanted to get the lemma for many Greek words, you would have to pick the first one from your list, search for it in LeFGreaT, copy the results and repeat this three steps for every single word in your remaining list. This would have been a tedious work for a list that consisted of 177 words.

We had to take a shortcut to create a more convenient way to get the lemma forms. So we created an online tool that can search for a list of words on LeFGreaT.[17] This software is mainly made up of two sites which are written with the programming language PHP. On the first site one can enter a list of Greek words into a text-field. Every Greek word has to be separated by a newline, independently of the newline character/sequence your operating system uses.[18] Internally the system will replace Windows and old OS X newlines by a UNIX newline character (ASCII code: 0x0A). In the next step it splits this modified list using the UNIX newline character to create an array of words. For every word in the array a request is sent to LeFGreaT. The relevant parts of the response are extracted using some significant delimiters found in the content of LeFGreaT. After the data is available to be shown the second site is loaded. See pictures 1 and 2 to get a better understanding of the procedure mentioned above. Picture 1 shows the first site with a list of Greek words. In picture 2 you can see the contents loaded from LeFGreaT.

One of our decisions was to finish this online tool as fast as possible and not to waste too much time thinking about design, but we also wanted it to be conveniently usable on mobile devices like smartphones and tablet computers. For this reason we chose the Bootstrap framework that provides a mobile first approach for the development of websites.[19] With this setup we were able to query the LeFGreaT website in a semiautomatic manner and to get results much faster compared to the manual strategy introduced in the beginning.

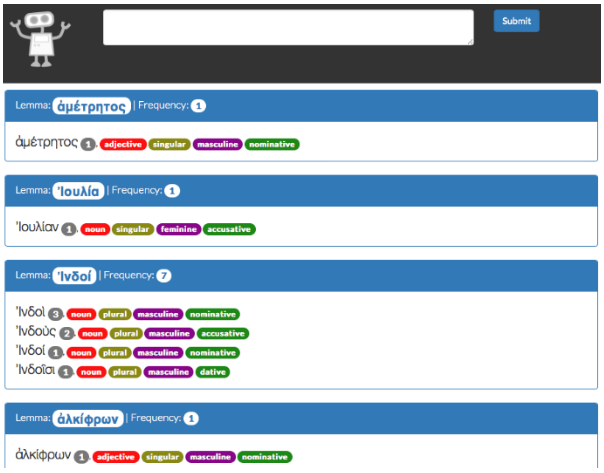


Search: Lexicon Formarum Graecarum Treebankearum (LeFGraeT)

ἀμέτρητος
'Ιουλία
'Ινδοί
ἀλκίφρων
κακοπαθέω

Separate lemmata with newline.

Picture 1: The first site with a list of Greek works in a text-area with every word in one line



Lemma: ἀμέτρητος | Frequency: 1

ἀμέτρητος 1 adjective singular masculine nominative

Lemma: 'Ιουλία | Frequency: 1

'Ιουλίαν 1 noun singular feminine accusative

Lemma: 'Ινδοί | Frequency: 7

'Ινδοὶ 3 noun plural masculine nominative
'Ινδοὺς 2 noun plural masculine accusative
'Ινδοί 1 noun plural masculine nominative
'Ινδοῖσι 1 noun plural masculine dative

Lemma: ἀλκίφρων | Frequency: 1

ἀλκίφρων 1 adjective singular masculine nominative

Picture 2: The second page shows a list of Greek lemma form requested from LeFGreat

[16] http://tariq-yousef.com/GT/
[17] https://github.com/eugenrein/sunoikisislemma/
[18] https://en.wikipedia.org/wiki/Newline
[19] http://getbootstrap.com

Another program that we developed is called just like the name of our project "Koios".[20] We will describe Koios in the following chapter.

7. Automatic named entity recognition with Koios – *Eugen Rein*

The main idea behind Koios is to have an online program that takes a plain text or a TEI XML file and automatically finds and visualizes place names. Additionally it possibly should enrich the XML file by TEI place name tags[21] with a reference to the corresponding Pleiades page.[22] A plain text file should be converted to a valid TEI XML file with place name tags. A user should be able to download these XML files later from Koios.

In order to recognize place names we used those one manually extracted ones from the German and English Thucydides' work mentioned earlier. To every place name we looked for a matching Pleiades page and used the URI to establish a relation between place name and Pleiades. So if you search for Athens on Pleiades for example, you get a list of many matching results. We chose the result that we thought was the best matching one. In the Athens example we used http://pleiades.stoa.org/places/579885 as the URI.

We used a hash table[23] as a data structure to be able to access data efficiently. The place name is used as a key and the value is the URI. With this approach we can access the data in linear time without the need to traverse through the whole table to find the place name and the Pleiades URI.

For the search of place names within the uploaded file we had to ignore all additional information of the XML structure. So the search only works on the main text content inside the body tag. The program picks the first element (key and value – place name and URI) from the hash table. Then it searches within the document for one matching word or many matching words if the place name consists of more than one word. If a match occurs the named entity is replaced by a HTML snippet. We first thought of implementing a fuzzy search, but we quickly abandoned the idea quickly in the first version of Koios due to time constraints.

This HTML snippet contains a map that is provided by Google Maps and a clickable link to the Pleiades page of the named entity. Pleiades also has a Keyhole Markup Language (KML) data export of the geographic data shown on the page. Google Maps has a build in functionality to load KML data from URLs which makes the work with the map simple and convenient. We used this data to visualize the place name on the map which appears likes a popup below a recognized named entity on the Koios website. (see picture 7)

The whole system was created with the programming language PHP. It is built on top of the Yii2 framework[24] which implements the rapid development approach (RAD).[25] It encourages the use of model-view-controller (MVC) pattern for a tidy coding structure and provides beneficial tools and software libraries like Bootstrap. With Bootstrap it was easy to have Koios be usable both on desktop computers and mobile devices (responsiveness). Besides PHP Javascript was used in some places of the code. The upload of files for example uses Javascript – AJAX to be precise – to send data from the client to the server.

At the time of this documentation Koios is in the alpha version. The Koios online programm consists of two sub-pages that a user can access. On the first page – the entry page – a user is required to upload a TEI XML file. (see picture 3) This file is checked against some validation rules. The validation progress analyses the file size and file type. The file size is limited in Koios to 10 MB. All files that are bigger than this limit are rejected. This limit also depends on the server settings which can decrease the allowed file sizes further (but never increase it). All file types that don't appear to be plain text or XML files are also rejected. The browser displays an error message accordingly. (see pictures 5, 6) If all validation rules are passed without failing, the file is transferred to an internal upload folder on the server and the browser redirects the user to the second page.

On the second page the user sees the text of the uploaded file. (see picture 4) In the text there are some words highlighted in blue. These blue highlights are the place names that Koios was able to recognize. With a click on

[20] https://github.com/eugenrein/koios
[21] http://www.tei-c.org/release/doc/tei-p5-doc/de/html/ref-placeName.html
[22] http://pleiades.stoa.org
[23] https://en.wikipedia.org/wiki/Hash_table
[24] http://www.yiiframework.com
[25] https://en.wikipedia.org/wiki/Rapid_application_development

the highlighted place name a small popup opens. (Bootstrap Popover)[26] In the popup there is a map that shows the location of the place name that is encoded in the KML data. (see picture 7)

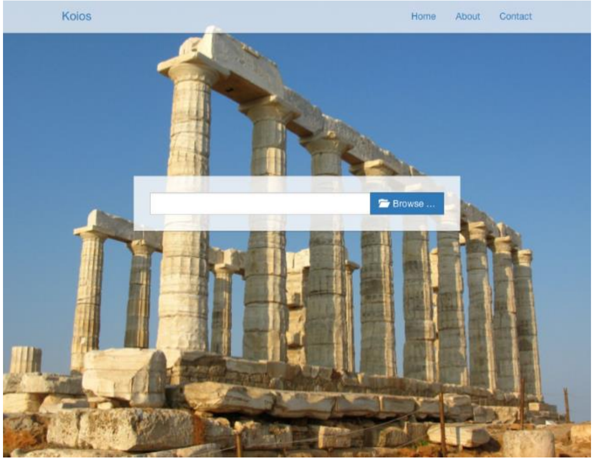
Nevertheless we weren't able to finish Koios the way we planned until the deadline and there were some problems we encountered using Pleiades.

Pleiades does provide KML data, but the data seems to be in a format that Google Maps isn't always able to interpret. Sometimes you see a map when you click on a place name, but there is no pin that shows you the exact location. One answer to this problem might be that some locations are regions that can't be located exactly. Yet there are sometimes regions in the KML data that Google Maps displays with a blue box that encloses an area on the map. Another manifestation of this problem might be that the map sometimes shows the whole world instead of a place located around the Mediterranean area. (see picture 9)

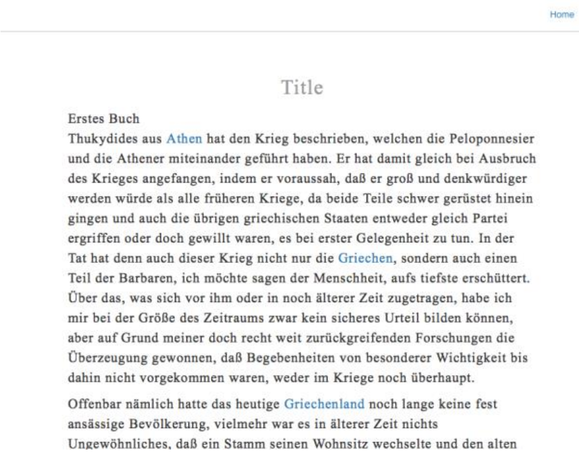
The slow response time while requesting Pleiades might also lead to maps without a pin. Since Google will download the data from Pleiades servers to their own, they will use a timeout to prevent their servers from being blocked by non-responding connections. A solution could be to download all KML data from Pleiades manually and use these on the Koios server.

This is not the only enhancement that has to be implemented before Koios can be published. As for now it is not possible to use plain text files in Koios. We planned to support plain text files, but we didn't manage to implement it in time. The user can only upload TEI XML files. Another issue might be the upload of non TEI files. The behavior is unknown, because Koios relies on an existing body tag inside the XML file. Furthermore a download functionality of the enriched TEI XML files is still missing. The unfinished code can be found in the GitHub repository.[27]

Additionally we were not able to test the accuracy of Koios because of the lack of time. We don't know how good or bad Koios behaves. The fact that we were against implementing a fuzzy search approach means that Koios won't be able to recognize partially unknown named entities. A positive point to our conservative search approach is that the rate of false positives must be very low. This of course is only true if the error rate in the process of manual identification of named entities is also low.



Picture 3: The front page of Koios. A user has to upload a file to get to the second page.



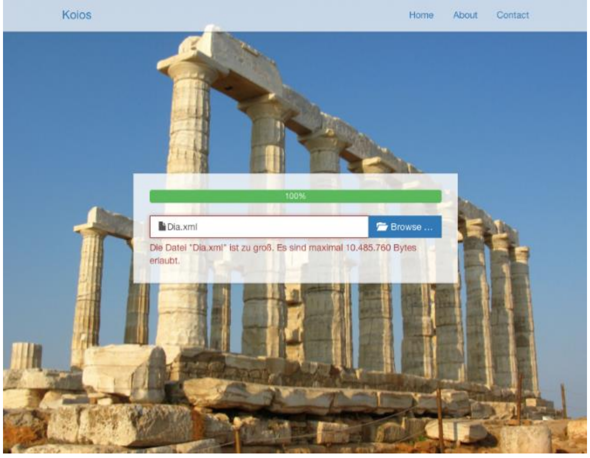
Title

Erstes Buch
Thukydides aus Athen hat den Krieg beschrieben, welchen die Peloponnesier und die Athener miteinander geführt haben. Er hat damit gleich bei Ausbruch des Krieges angefangen, indem er voraussah, daß er groß und denkwürdiger werden würde als alle früheren Kriege, da beide Teile schwer gerüstet hinein gingen und auch die übrigen griechischen Staaten entweder gleich Partei ergriffen oder doch gewillt waren, es bei erster Gelegenheit zu tun. In der Tat hat denn auch dieser Krieg nicht nur die Griechen, sondern auch einen Teil der Barbaren, ich möchte sagen der Menschheit, aufs tiefste erschüttert. Über das, was sich vor ihm oder in noch älterer Zeit zugetragen, habe ich mir bei der Größe des Zeitraums zwar kein sicheres Urteil bilden können, aber auf Grund meiner doch recht weit zurückgreifenden Forschungen die Überzeugung gewonnen, daß Begebenheiten von besonderer Wichtigkeit bis dahin nicht vorgekommen waren, weder im Kriege noch überhaupt.

Offenbar nämlich hatte das heutige Griechenland noch lange keine fest ansässige Bevölkerung, vielmehr war es in älterer Zeit nichts Ungewöhnliches, daß ein Stamm seinen Wohnsitz wechselte und den alten

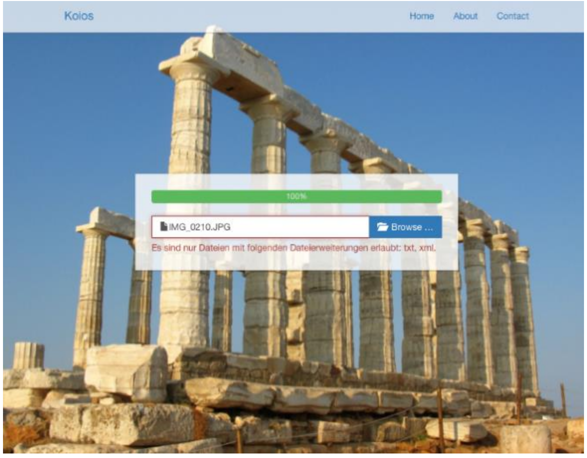
Picture 4: The second page. Highlighted words in blue show recognized place names that can be clicked to open a popup that shows more details

[26] http://getbootstrap.com/javascript/#popovers
[27] https://github.com/eugenrein/koios



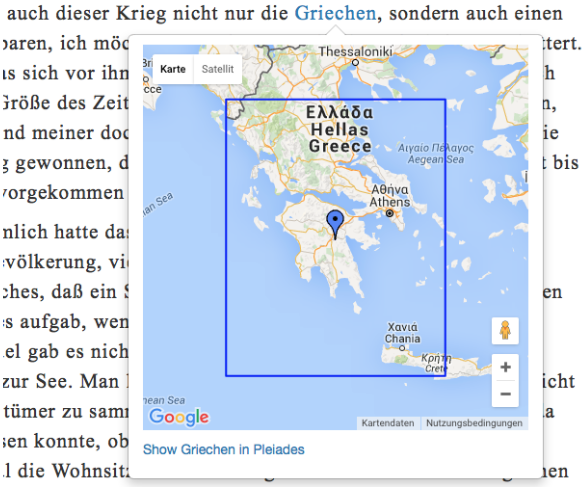
Die Datei "Dia.xml" ist zu groß. Es sind maximal 10.485.760 Bytes erlaubt.

Picture 5: An error message explaining that the uploaded file has exceeded the 10 MB limit



Es sind nur Dateien mit folgenden Dateierweiterungen erlaubt: txt, xml.

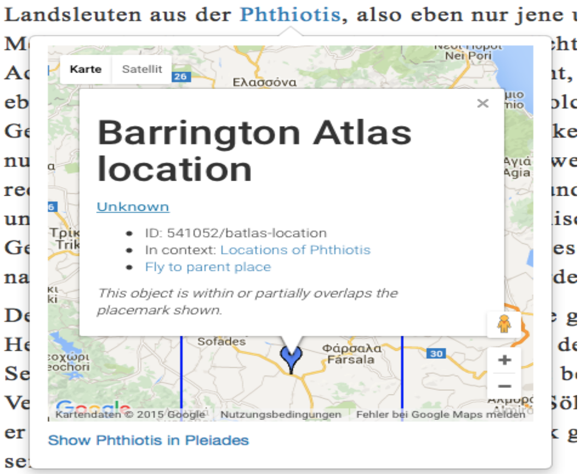
Picture 6: An error message showing that an invalid file was uploaded



auch dieser Krieg nicht nur die Griechen, sondern auch einen

Show Griechen in Pleiades

Picture 7: The popup that appears after a click on a highlighted place name. It shows a region of Greece.



Landsleuten aus der Phthiotis, also eben nur jene

Barrington Atlas location

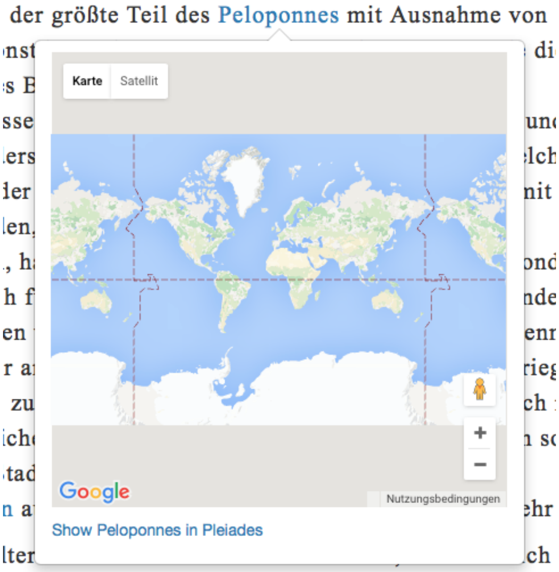
Unknown

- ID: 541052/batlas-location
- In context: Locations of Phthiotis
- Fly to parent place

This object is within or partially overlaps the placemark shown.

Show Phthiotis in Pleiades

Picture 8: You may click on a pin on the map to see more details.



der größte Teil des Peloponnes mit Ausnahme von

Show Peloponnes in Pleiades

Picture 9: An error in the map that occures with Pleiades. The whole world is shown instead of a region or a place.