Workflow Linking Syriac Geographic Data

This document describes the workflow we have developed throughout the Pelagios-funded work group project Linking Syriac Geographic Data. This project, with collaborators of the Vrije Universiteit Amsterdam and Syriaca, an LOD project on named entities for the Syriac world, builds on and is developed largely in conjunction with a CLARIAH research pilot LinkSyr. The general goal of this project is to provide tools to progress from a non-digitised Syriac text to a digital version which is LOD-ready. In particular, this results in the creation of OCR tools, a morphological analyser, and to provide a Linked-Data plugin to provide annotations to these texts from online sources. Within the Pelagios work group, we look more concretely at the difficulties of linking concrete texts to Linked Data resources. We document these difficulties, so that work towards a general workflow of adding new texts to the LOD cloud. Our documentation focuses on the Book of the Laws of the Countries, a second century text written by Bardaisan, which is famous for its description of the large variety of peoples and cultures of the then known world. Needless to say, this text contains many named entities, which in turn provide a salient basis for discussing its reconciliation with LOD resources.

At the centre of our attention is the Syriaca database, a collection of named entities for the Syriac world. Within the context of this project, we limit ourselves to the place names, as they were already added to Pelagios. Since the database was developed by use of several English and French dictionaries and encyclopaedias, which do not always contain a Syriac reference to the entity under scrutiny, their database often has no Syriac word given for an entry. On the other hand, we have discovered that many varieties in attested names exist, which are also not yet available in the database. This is usually due to the high variation throughout concrete text concerning named entity spelling. Therefore, we propose our system as a linkage tool between a corpus of texts and the database, which delivers suggestions of potential named entities, which can then in a manual step be checked and, if correct, added to the database. This step does not only enrich the database at hand, but also keeps the links to the original texts, hence providing a concordance of attestations of named entities and their variations. The entries of Syriaca contain both Syriac entries, and the rendition in other important languages throughout time. These include, among others, Hebrew, Greek, Latin, but also English. It is the latter language which is necessary for the Recogito tool to work. For this reason, we have developed an additional step where the named entities, which are matched to the Syriaca database, are given the English counterpart of their name, as they are entered into Pelagios.
In trying to link the BLC text to the Syriaca entries, we have come across a threefold division of named entities to expect: the ones readily found in Syriaca, the ones for which we find a variant to the form found in Syriaca, and finally the ones which are not yet in Syriaca.

Concerning the first variety of matching against terms already found in the Syriaca database, we have found the SpaCy package of Python to provide salient results. Their Named Entities module allows, next to machine learning methods which are data are not yet enough elaborated for, also possibilities to perform rule-based matching. Nevertheless, since we want to simultaneously extract named entities, and discover additional salient terms to enrich the Syriaca database, this package does not completely suit our needs. Hence, for the second variety, we had to take refuge to other possibilities.

Using the fuzzywuzzy package, we allow to perform fuzzy matching on the data in the text. This is necessary, because our research has pointed out that many varieties of name attestations are available in the text, often variants of named entities which are readily available in the Syriaca database. As in most languages, place names give rise to variation of writing to refer to these places. Syriac, being a semitic language, has most variation in the plene or defective writing, which refers respectively to writing with or without matres lectionis, the so-called reading mothers, consonants used as constraints to determine the reading of vowels. These matres lectionis are not necessarily written, hence the formation of a source of written variation. An clear example of this

variation from the Book of the Laws of the Countries (BLC) would be the gentilics for Britain. The standard form is ܒܪܛܢܝ (brytwny), although a form using additional matres lectionis, ܒܪܝܛܘܢܝ (byrwtwny) is also read. In these words, y and w are used as reading mothers respectively.
Other, less obvious variation from a syriacist point of view attested in BLC, is the variation in phonologically close consonants, which provide a variation on the common and expected root. An example from our text would include two references for Egypt, respectively the location name (ܐܓܦܛܐ, 'gptws) and the derived gentilic (ܐܓܒܝ, 'gbty). Here, the P and B consonants are interchangeably, which is not common in Syriac. As we will point out in the next section as well, we believe that at this stage we do not need to circumvent false positives, in order to find the most balanced results. We think it is necessary that we attract as many variants as possible, because at the moment it is difficult to estimate how many variants in general text we are likely to encounter. The program for this is called Term_matching.py

Concerning the third variety, Hannes Vlaardingerbroek of the LinkSyr project has developed a morphological analyzer, which allows to devise rules to discern high likelihood candidates to be a place name. For example, if we find a noun (which we up to this point have not yet identified as a place name, which is preceded by a preposition, and not followed by a suffix, and which is also not matched to the SEDRA database. Because Syriaca and SEDRA are linked, we can use the fact that if a term is not found in SEDRA, it is likely a named entity which we have not yet taken up in the Syriaca database. Of course, this process will produce some false positives, which in this case does not really form a problem. We want to not let any possible named entities slip, and allow to select the good ones by manual selection. Work on this entry database will require more textual data to be trained on, which is in the process of being digitalized. In the future, this will allow machine learning algorithms to be trained on the recognition of named entities, as is provided in the SpaCy package we have used for this workflow.

Other reconciliatory issues we have as yet left unaddressed, such as the problem that place names can also have parallel, unrelated meanings. A clear example from the Book of the Laws of the Countries would include ܣܢܝ, being the standard term referring to China, but which also refers to a Syriac saint. Another example from our text would be ܦܪܣܝܐ or ܦܪܣܝ, the gentilic name referring to the Persian identity, which nevertheless is also the root for words in the semantic field of laying bare and nakedness.