Premodern Geographical Description: Data Retrieval and Identification

Masoumeh Seydi
University of Leipzig
Leipzig, Saxony
m.seydi@uni-leipzig.de

Maxim Romanov
University of Vienna
Vienna, Vienna
romanov.maxim@gmail.com

Chiara Palladino
University of Leipzig
Leipzig, Saxony
chiara.palladino@dh.uni-leipzig.de

ABSTRACT

Geographical and spatial descriptions in the premodern world are structurally different from the modern era, where spatial understanding is based on cartographic navigation. This paper presents an experimental process to tag, retrieve, and identify geographical information as described in premodern primary sources, together with the issues and possible solutions. The proposed method defines specific categories of geographical information and a markdown system to mark these categories in the source. Having tagged the data, we extract it and geographical locations and their connections are identified through a heuristic approach: the extracted geographical entities are initially aligned with existing geographical references and secondary sources. String similarity approaches might provide fuzzy identifications which need to be verified and disambiguated. In this paper, we describe the process of annotation and extraction of geographical descriptions, experiment some toponyms matching metrics, report the results, and offer possible solutions to handle disambiguation through the existing contextual information in the source. The process is applied to two different datasets, proposed as test cases: a classical Arabic geographical text and a Roman itinerary.

CCS CONCEPTS

• **Information systems → Information extraction**; **Multilingual and cross-lingual retrieval**;

KEYWORDS

Premodern Geography, OpenITI mARkdown, Digital Humanities, Spatial Humanities, Medieval Arabic Geography, Roman Empire, Modeling, Itineraries

ACM Reference Format:
Masoumeh Seydi, Maxim Romanov, and Chiara Palladino. 2017. Premodern Geographical Description: Data Retrieval and Identification. In *GIR'17: 11th Workshop on Geographic Information Retrieval, November 30-December 1, 2017, Heidelberg, Germany.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3155902.3155911

1 INTRODUCTION

It has long been acknowledged that, in premodern societies up to the cartographic revolution of the 16th century, the ways of navigating and orienting through the landscape were not served by diagrammatic representations, such as maps, but were essentially conveyed through geographical storytelling [8, 16, 27, 35]. Consequently, the linguistic encoding of such spatial descriptions was specifically functional to navigation: it was a "system of shared knowledge" produced by a society in order to navigate through space, and it followed specific linguistic and expressive criteria. This has two important consequences:

- Textual sources are the main repositories of information about spatial understanding in premodern societies.
- The recognized systematicity of their linguistic encoding of space provides a test-case for automatic and semi-automatic methods of extraction of semantically meaningful patterns functional for the narrative of space.

The purpose of this paper is to propose an explorative method to extract and make use of such meaningful linguistic patterns, in order to introduce proper modelings and create visual representations of them for further studies and inquiries of a source. This process is especially complex, in view of its necessity to deal with the specificities of ancient languages and their ways of expressing toponyms and spatial indications. Inspired by [30], our approach provides a semi-automated process by which one can combine quantitative and qualitative studies, distant and close reading (see Fig. 1). It means one will be able to iteratively analyze premodern geographical sources and model the geographical descriptions to gain a better understanding of them, disambiguate imprecise perceptions, and reach a consensus on interpreting the source—when possible. Fig. 2 depicts this idea as a procedure in which each step can be connected to the previous ones in order to make possible modifications and/or corrections. We may need to return to the source in each iteration where it is required. This point is crucial since a fully automated method does not necessarily require close attention to the sources in the whole process. The automated part of identifying toponyms using the state-of-the-art string similarity approaches can be engaged into this process after semi-automatic iterative steps.

Our method consists of the following efforts as a workflow:

- Annotating relevant information in the source
- Extracting the places in order to accurately locate them by associating them to real-world spatial entities
- Using other conceptual and spatial information retrieved from the source to resolve ambiguous locating of the places

GIR'17, November 30-December 1, 2017, Heidelberg, Germany
© 2017 Association for Computing Machinery.
ACM ISBN 978-1-4503-5338-0/17/11…$15.00
https://doi.org/10.1145/3155902.3155911

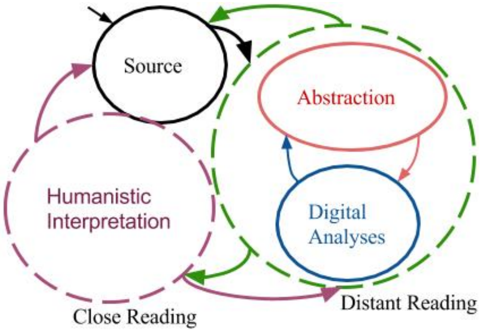


Figure 1: Iterative algorithmic analyses of a text (based on Fig. 1 from [30]

- Associating the places to a series of characteristics according to the source, the chronological period, the historical/cultural context, and so on.

Then, the data are prepared for further usage in state-of-the-art geospatial analysis tools and technologies in order to enrich existing information on locations, reachability, travel and space divisions in the corresponding area and to provide material for studying premodern geography. More clearly, this can be seen as a process to convert geographical narrative in natural language, consisting of complicated and detailed descriptions in a text, to a machine-actionable abstraction for distant reading analysis (for further details see [30]). Such abstraction is then used to produce relevant analysis and visualizations of data that can help us improve the previous steps, correct mistakes, obtain new interpretation of descriptions, and disambiguate fuzzy interpretations of a text.

Tagging the geographical data is mostly limited to tag simple structures that can be done through *Name Entity Recognition*. Tagging of the semantic and complex geographical description requires engaging NLP and machine learning approaches as [25] proposes and considering language-dependent characteristics. These approaches are still limited to specific patterns and is designed for fully automated approaches which is not our main goal here.

Toponym identification and resolution is the first step after extracting the data. Focusing on toponym resolution, [21] proposes an annotation approach and discuses all the relevant issues which might happen in toponym matching in various cases of geography. In this regard, string similarity metrics are widely used and discussed ([2, 11]). State-of-the-art approaches also use machine learning methods relying on very small datasets, often considering only place names in Romance and Germanic languages, or Romanized toponyms ([23, 24]) while [31, 32] experiments an approach with a huge data set in various languages.

Many other works have designed specific algorithms for matching toponyms, often leveraging some form of canonical representation for toponyms ([10, 20]). In order to match gazetteer records, various studies have also combined heuristically different metrics and computed over particular attributes of toponyms such as names, types, geospatial footprints ([22, 37]). Our dataset is very small, and does not have canonical form of toponyms as it happens normally in premodern sources. There is a limited number of gazetteer records to find the referents and what we report here as a toponym matching process is a finalizing step of our experiment on the proposed iterative process following a different perspective of studying premodern geographical sources and data retrieval. Toponym resolution is not our focus here and we consider the related works on it as future work for this part of our research when there will have more data gathered though the current procedure.

2 GEOGRAPHICAL DATA IN PREMODERN SOURCES

Premodern geographical sources contain various types of constructs functional to the description of the landscape for navigation. These constructs can be generalized on linguistic and expressive basis, independently from the scope and cultural context of the source. The most important of them can be classified as follows:

- Names of places (toponyms)
- Space segmentation: geographical or administrative
- Route connections: forming a relation between places, expressed generally with distance estimates; often other conditions, such as orientation and direction specifications, can appear as additional context. This type of information can be part of a complete route path or itinerary, or part of a network (for this distinction, see 2.2).
- Geographical orientation: contextualized indications of directions, movements or placements in the landscape

Obviously, some of these constructs will be statistically more relevant according to the purpose and viewpoint of the source: for example, comprehensive geographies are more concerned with hierarchical data, functional to the definition of national boundaries, whereas route descriptions and distance estimates will be much more relevant in travelogues.

Two of these categories are especially relevant for the purpose of modeling: (a) hierarchical data describing administrative divisions; (b) route sections and connections. In this paper, we will focus on these two typologies as an attempt to develop and apply a workflow on completely different premodern sources of classical Arabic and Roman geographies. The Arabic source is *Aḥsan al-taqāsīm fī maʿrifat al-aqālīm* ("The best division for the knowledge of the provinces") [6, 7], a comprehensive Islamic world geography written by al-Muqaddasī in the 10th century, covering North Africa (including Iberia), Egypt, The Arabian Peninsula, Greater Syria, Iraq and Upper Mesopotamia, as well as eight non-Arab provinces including Iran and Afghanistan. Its geographical description is structured through multi-level hierarchy of administrative divisions, from the level of major provinces down to settlements, connected through routes and distances. For the Roman world, we choose the so-called *Antonine Itinerary* as a test-case for literary geography [13]: this itinerary is a compilative product about the geography of the Roman Empire, produced between the 3rd and the 4th century AC (the maritime part is probably even later), conceptually divided into a land route and a maritime route, where the places of the empire are connected through the main pattern of the Roman provinces and the road network; it represents a comprehensive geographical picture of the Roman Empire and its travel connections, whose scope is more systematic and cultural than practical [9, 29]. The *Itinerarium Provinciarum*, or land route, follows a pattern from the Strait of Gibraltar, through North Africa,

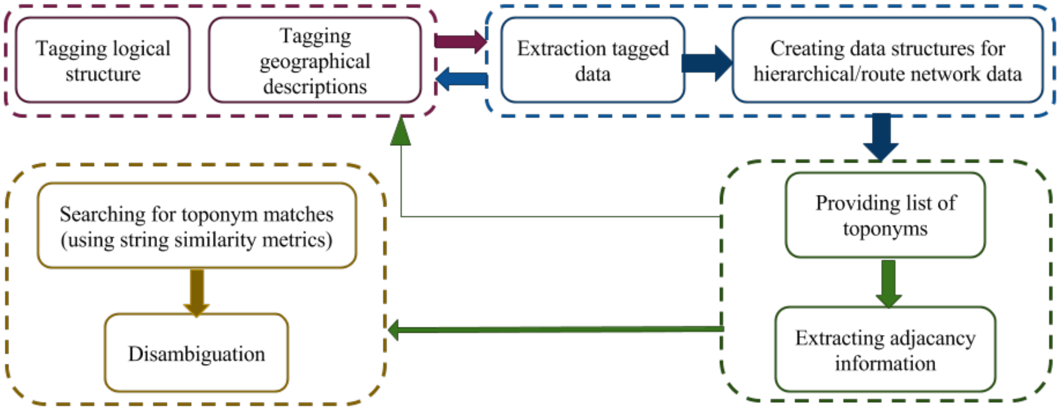

Tagging logical structure
Tagging geographical descriptions
Extraction tagged data
Creating data structures for hierarchical/route network data
Searching for toponym matches (using string similarity metrics)
Disambiguation
Providing list of toponyms
Extracting adjacency information

Figure 2: The iterative process of analyzing geographic sources

to Italy and the Eastern provinces, then back to Gaul, Britain and Spain, whereas the *Itinerarium Maritimum* is more irregular and partial: it traces a maritime route from Greece to Africa and from Rome to Arles, with some additional information about northern sea connections between Gaul and Britain. Structurally, the *Antonine Itinerary* is organized as a list of connections between places, with the indication of the respective distances. The connections follow a ramified, not linear, pattern: from one central point, often an important city, the information provided describes various travel options essentially established through the Roman road network. Because of this structural factor, we treat this itinerary as a network description, analogous to *al-Muqaddasī*’s description, and not as a progressive, narrative travelogue. Although the hierarchical description is not as precise as in *al-Muqaddasī*’s book, there is some hierarchy in the organization of the connections: each section is included under the name of one or two major provinces, and each major route is divided hierarchically into minor segments.

2.1 Hierarchical Data

Hierarchical data in geographical descriptions are based on the systematic division of an empirical territory. Divisions form various non-atomic spatial entities can be seen as higher level categories for atomic locations and places.

Unless demanded by a specific requirement, we do not consider that there is one basic level of hierarchy in all premodern sources, i.e., the area covered by the source as a “macro-region”. What we consider here as hierarchical data are explicit descriptions of grouping places and toponyms under a name of a division of any type (geographical, conceptual or political). Grouping a set of toponyms in micro-regions and forming macro-regions containing micro-regions according to an arbitrary number of hierarchical levels shapes the regional frontier accordingly. The most comprehensive pattern of hierarchical descriptions starts with the macro-regions, or highest level regions, which include smaller regions as subregions, in which more subordinate regions or settlements are included. Each division has its own type as a property. For example, in the following passages (the English translations is also provided below them) al-Muqaddasī describes how “The Peninsula of the Arabs” as a highest-level region/province is divided into four major subregions of a specific type and the first subregion, al-Ḥiğāz, has its capital as well as other types of settlements:

وهذه صورة جزيرة العرب وقد جعلنا اربع كورة جليلة
واربع نواحٍ نفيسة والكور اوّلها الحِجَاز ثم اليَمَن ثم عُمَان ثم هَجَر

“This is the form of the Peninsula of the Arabs. We have divided this region into four extensive provinces, and four large districts. The provinces are al-Ḥiğāz, al-Yaman, ‘Umān, Hajar; the districts al-Ahqāf, al-Ashhār, al-Yamāma, Qurh.” [7]

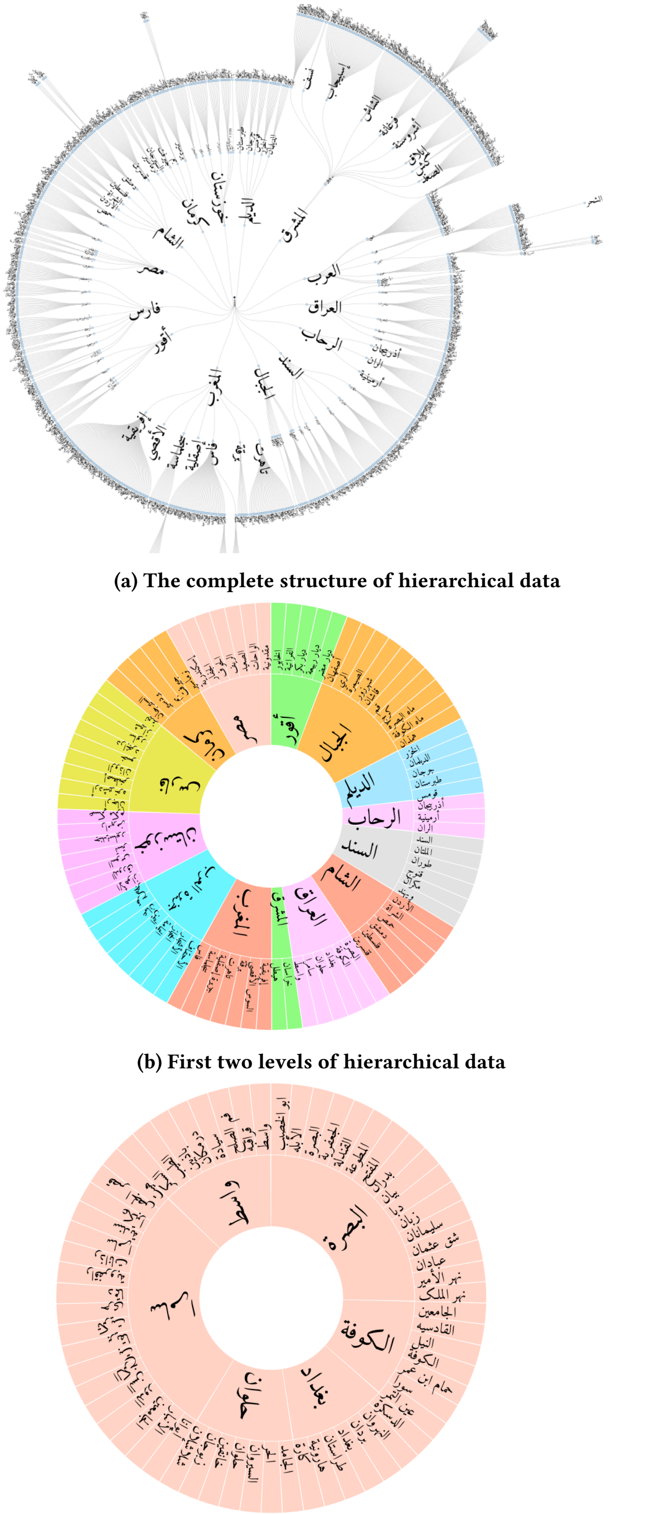
فاما الحجاز فقصبةمكّة ومن مدنها يَثْرِب ويَنْبُع وقُرْح وخَيْبَر
والمَرْوَة والحَوراَء وجُدّة والطّائِف والجار والسُقْيا والعَونِيد
والجُحْفَة والعُشَيرة هذه امّهات، ودونهنّ بَدْر خُلَيص أَمَج الحِجْر
بَدايعقوب السُّوارقيّة الفُرْع السَّيْرة جَبَلَة مَهايع حاذة

"The capital of al-Ḥiğāz is Makka; among its towns are Yathrib, Yanbu, Qurḥ, Khaybar, al-Marwa, al-Hawrā’, Judda, al-Tā’if, al-Jār, al-Suqyā (Yazīd), al-‘Awnid, al-Juhfa, and al-‘Ushayra: these are the larger towns. Lesser towns are Badr, Khulays, Amaj, al-Hijr, Badā Ya‘qūb, al-Suwariqiyyā, al-Fur‘, al-Sayra, Jabala, Mah‘yī, Hādha.” [7]

As shown in Fig. 3a, this pattern forms a hierarchical tree going from the provinces down to the settlements in various levels. Fig. 3b and Fig. 3c depict the zoomed views of the first two levels of this tree, holding provinces together with their subordinate regions and detailed view of the province Iraq with settlements respectively.

2.2 Routes and Connections

Trade routes and connections are a valuable and significant part of premodern spatial descriptions. The description of place connectivity is structurally similar in any geographical narrative (in fact, it appears identical in our both sources). In general, the connections



(a) The complete structure of hierarchical data

(b) First two levels of hierarchical data

(c) The province of Iraq with subordinate regions and settlements

Figure 3: Hierarchical data visualization as described in al-Muqaddasī's Aḥsan al-taqāsīm

introduce an individual route section from a place to another one and provide a specific distance expressed in classical units (the *Itinerarium Provinciarum* indicates the distance in Roman *miles* or, in the case of Britain and Gaul, with an additional equivalent in *leagues*; the *Itinerarium Maritimum* provides measurements in *stadia*; al-Muqaddasī specifies the distances in *stage*, *farsakh*, *day*, or *mile*). In some cases, additional details are provided, such as indications of orientation, conditions of travel, etc. Three parts, structurally formed as an individual entity, are essential: start/subject, end/object, distance/predicate. Conceptually, the route sections are defined as a part of an itinerary or a route network: for modeling purposes, we propose this distinction, which is not meant to be a definition of literary or cultural value, but is functional to a specific extraction and modeling workflow according to the type of dataset. In this definition, we indicate as itinerary a linear route in the characteristic shape of a travelogue, where places are listed in a progressive way, often connected through the respective distances; we indicate as route network the comprehensive description of the travel system in a spatial extent, e.g. a road network or the narration of several connecting routes having a central place in common. By definition, an itinerary is linear, whereas a route network is ramified throughout multiple centers, which are connected by multiple lines.

3 TAGGING DATA

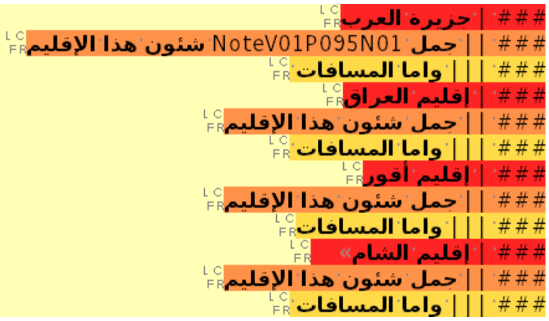
Tagged data in a text allows one to make it usable in computational analysis. Annotating spatial narrative in premodern sources needs an appropriate vocabulary or standard, which is, at the moment, still missing. Widely used standards like TEI/EpiDoc offer a basic scheme for named entities, but do not provide any semantic information about geographical relations and classification [28]. Ontology-based attempts, in the case of premodern geographies, are still seldom explored [34]; however, the adaptation of current ontologies, designed for modern cartographic frameworks, for the semantic annotation of spatial entities in premodern sources, implies the risk of not showing what is outside the adopted standard [19]. Therefore, we use a simple tagging system, named `OpenITI mARkdown` [5] which is built on regular expressions[1] to provide easily customizable tagging options of recurrent and regular linguistic patterns. It has been developed to tag a variety of logical, structural, and analytical patterns. Particularly, we extend it to tag complex semantic patterns such as geographical units, i.e., place relations, especially hierarchies and connections, which provide the backbone to any type of geographical narrative. Moreover, it defines the vocabulary of the source through a bottom-up approach, where systematic patterns are semantically tagged and classified with minimum intervention on the text. More description is given in 3.1.

3.1 OpenITI mARkdown

`OpenITI mARkdown` consists of a language-independent tagging system through which data collection, information extraction and conversion of raw text into machine-actionable formats are easily achievable[2]. As [30] states, it has been developed to overcome the

[1] The underlying regular expression implemented for each pattern in this scheme is given in https://alraqmiyyat.github.io/mARkdown

[2] It is implemented in **EditPad Pro** (https://www.editpadpro.com) and can be downloaded and used via https://github.com/OpenITI/mARkdown_scheme.



```
### | جزيرة العرب
### || جمل NoteV01P095N01 شئون هذا الإقليم
### ||| واما المسافات
### | إقليم العراق
### || جمل شئون هذا الإقليم
### ||| واما المسافات
### | إقليم أقور
### || جمل شئون هذا الإقليم
### ||| واما المسافات
### | إقليم الشام«
### || جمل شئون هذا الإقليم
### ||| واما المسافات
```

Figure 4: Logical units tagging in al-Muqaddasī's Aḥsan al-taqāsīm

issues of XML tagging applied to right-to-left scripts, but can be applied to any language and bi-directional texts and sources to avoid excessive complexity of simple editing tasks and providing a lightweight and easy-to-use scheme for working with voluminous textual collections. Fig. 4 illustrates an example of tagging logical units of a text highlighted in various colors according to different level of headings. The highlighting scheme is customized using **EditPad Pro**[3]. The red color corresponds to the header of the chapter (tagged with "### | "), the orange color represents the sections (tagged with "### || "), and the yellow color highlights sub-subsection (tagged with "### ||| "). This structure can go further to show any arbitrary structure of a text.

We use `OpenITI mARkdown` and extend it by proposing analytical patterns and semantic tagging of geographical narratives, hierarchical and route sections data. More details is discussed in the following sections.

3.2 Administrative Divisions

The pattern that we introduce here considers the coverage are as the `WORLD` and divides it into the highest level of divisions, called `PROVINCE`. Each `PROVINCE` can be divided into multiple subordinate regions where other subordinate divisions or settlements are placed. Each division can be characterized by a type, which in our case are the premodern type of divisions or settlements. The following scheme represents this explanation:

`WORLD:` **`PROVINCE`** `> TYPE >` **`(REGION)`** `> TYPE >` **`SETTLEMENT`**

As the scheme suggests, each two entities are connected through a `TYPE` and using them all together, we tag relevant information as triples of `SUBJECT > PREDICATE > OBJECT`. This patterns simply create a regulated system to be treated computationally.The following schemes shows how these triples look like:

- `#$#PROV` **`toponym`** `#$#TYPE` **`type_of_region`** `#$#REG1 (`**`toponym`** `#)+`
- `#$#REGX` **`toponym`** `#$#TYPE` **`type_of_region`** `#$#REGX (`**`toponym`** `#)+`
- `#$#REGX` **`toponym`** `#$#TYPE` **`type_of_settlement`** `#$#STTL (`**`toponym`** `#)+`

We preserve the source vocabulary for toponyms and type of places (reflected in *type_of_region/settlements*) to record them as they appear in the source. Fig. 5 depicts piece of a source describing divisions and the corresponding tagged information. Each (purplish) highlighted line represents the inline tagging of the information in the text. As an instance, the first highlighted line carries the list of subordinate regions in the Arabian Peninsula as below:

`#$#PROV` **Jazīraŧ al-ʿarab** `#$#TYPE` **Kūraŧ** `#$#REG1` **al-Ḥiǧāz # al-Yaman # ʿUmān # Hajar**

And the next two annotation lines hold the capital (Qaṣabŧ) and the major cities of al-Ḥiǧāz subregion respectively, shown below:

`#$#REG1` **al-Ḥiǧāz** `#$#TYPE` **Qaṣabŧ** `#$#STTL` **Makka**

`#$#REG1` **al-Ḥiǧāz** `#$#TYPE` **Qaṣabŧ** `#$#STTL` **Yaṯrib # Yanbuʿ # Qurḥ # Khaybar # al-Marwaŧ # al-Ḥawrāʾ # Judda # al-Ṭāʾif # al-Jār # al-Suqyā (Yazīd) # al-ʿAwnid # al-Juḥfa # al-ʿUshayraŧ**

All these individual triples of tagged data form the hierarchical description.

3.3 Route Sections and Itineraries

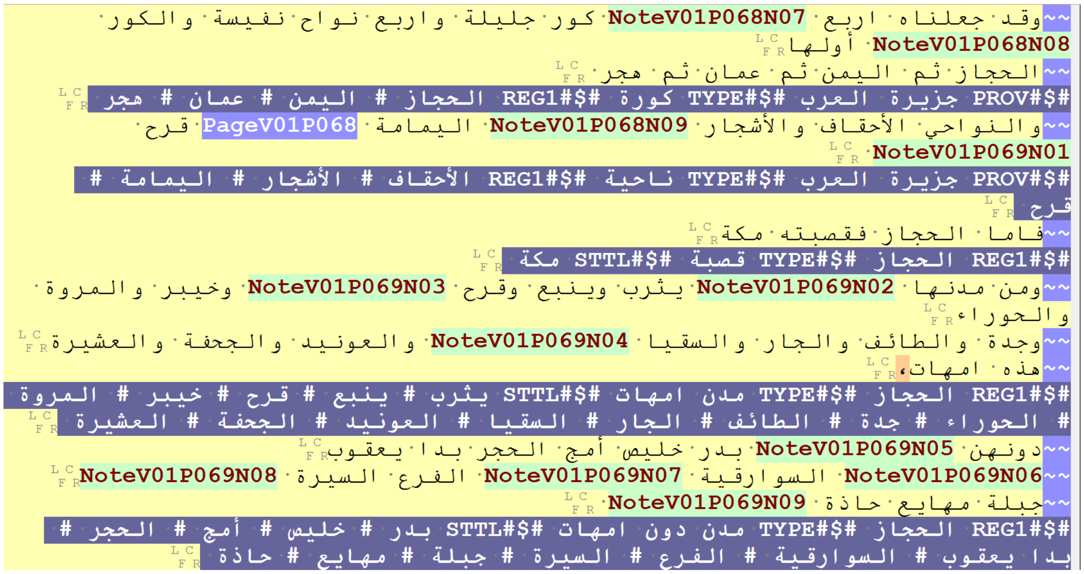
Route sections describe connections between two places, often with a distance. Similar to the hierarchical data tagging scheme, we form triples of data using the source vocabulary following the scheme below:

`#$#FROM` **`toponym`** `#$#TOWA` **`toponym`** `#$#DIST` **`distance_as_recorded`**

All these triples can be put together to form a bigger network or a set of itineraries. Fig. 6 shows how route section descriptions are tagged in the *Antonine Itinerary* using a revised version of `OpenITI mARkdown` implemented for left-to-right scripts. The annotated lines are highlighted in blue by which the individual route sections are represented.

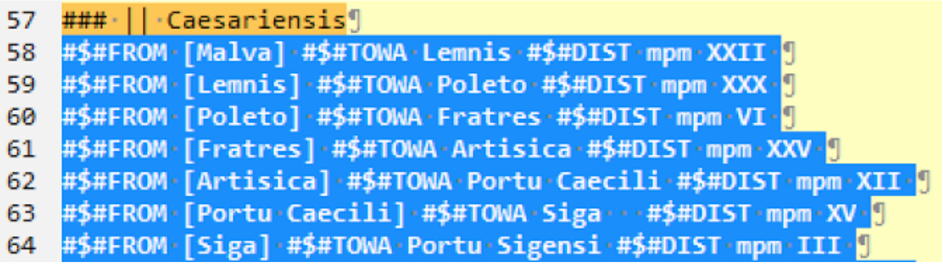
There are cases, as in the *Antonine Itinerary*, where a hierarchy of routes is provided, namely, micro-regional connections are encapsulated in a longer, major route. Such micro-regional connections are provided for two reasons: to provide precise and detailed information about the stations and distances on a certain route, and to give alternative paths. For instance, *It. Ant. 363-4* gives two alternative routes to travel from Durocortorum (Reims) to Divodurum (Metz?) in Gaul: the alternative route is systematically introduced with the words "*alio itinere*", the repetition of the main connecting points, and the distance given in each case: "*Item a Durocortoro Divodorum usque m. p. LXXIIII [...] Alio itinere a Durocortoro Divodorum usque m. p. LXXXVII, sic [...]*" ("The travel from Durocortorum to Divodorum measures 74 miles [...] Another itinerary from Durocortorum to Divodorum measures 87 miles, this way [...]")

[3] `OpenITI mARkdown` is not dependent on any particular editing environment. The current version is implemented in **EditPad Pro** which supports custom highlighting and navigation schemes.



```
~~وقد جعلناه اربع NoteV01P068N07 كور جليلة واربع نواح نفيسة والكور
NoteV01P068N08 أولها
~~الحجاز ثم اليمن ثم عمان ثم هجر
#$#PROV جزيرة العرب #$#TYPE كورة #$#REG1 الحجاز # اليمن # عمان # هجر
~~والنواحي الأحقاف والأشجار NoteV01P068N09 اليمامة PageV01P068 قرح
NoteV01P069N01
#$#PROV جزيرة العرب #$#TYPE ناحية #$#REG1 الأحقاف # الأشجار # اليمامة #
قرح
~~فاما الحجاز فقصبته مكة
#$#REG1 الحجاز #$#TYPE قصبة #$#STTL مكة
~~ومن مدنها NoteV01P069N02 يثرب وينبع وقرح NoteV01P069N03 وخيبر والمروة
والحوراء
~~وجدة والطائف والجار والسقيا NoteV01P069N04 والعونيد والجحفة والعشيرة
~~هذه امهات،
#$#REG1 الحجاز #$#TYPE مدن امهات #$#STTL يثرب # ينبع # قرح # خيبر # المروة
# الحوراء # جدة # الطائف # الجار # السقيا # العونيد # الجحفة # العشيرة
~~دونهن NoteV01P069N05 بدر خليص أمج الحجر بدا يعقوب
~~NoteV01P069N06 السوارقية NoteV01P069N07 الفرع السيرة NoteV01P069N08
~~جبلة مهايع حاذة NoteV01P069N09
#$#REG1 الحجاز #$#TYPE مدن دون امهات #$#STTL بدر # خليص # أمج # الحجر #
بدا يعقوب # السوارقية # الفرع # السيرة # جبلة # مهايع # حاذة
```

Figure 5: Hierarchical data inline tagging in al-Muqaddasī's Aḥsan al-taqāsīm



```
### || Caesariensis
#$#FROM [Malva] #$#TOWA Lemnis #$#DIST mpm XXII
#$#FROM [Lemnis] #$#TOWA Poleto #$#DIST mpm XXX
#$#FROM [Poleto] #$#TOWA Fratres #$#DIST mpm VI
#$#FROM [Fratres] #$#TOWA Artisica #$#DIST mpm XXV
#$#FROM [Artisica] #$#TOWA Portu Caecili #$#DIST mpm XII
#$#FROM [Portu Caecili] #$#TOWA Siga  #$#DIST mpm XV
#$#FROM [Siga] #$#TOWA Portu Sigensi #$#DIST mpm III
```

Figure 6: Route sections inline tagging in Antonine Itinerary

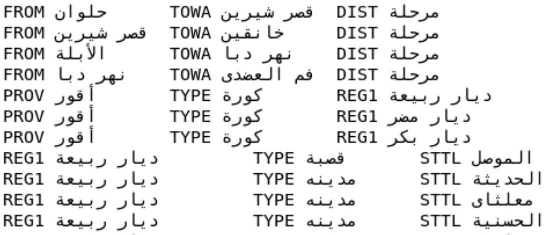
Since this kind of macro-division appears systematically in our source, we introduce a new tag set to notate the presence of a macro-route in order to better circumscribe the area:

```
#$#FRIT toponym #$#TOIT toponym #$#DIIT
distance_as_recorded
```

4 INDIVIDUATING AND MATCHING TOPONYMS

Having extracted the tagged data (Fig. 7), we proceed to identify places as the primitive building blocks of the space described in the source. The first step is to find existing references of premodern places. For the Islamic world Georgette Cornu's *Atlas du monde arabo-islamique à l'époque classique: I^{Xe}–X^{e} siècles* [12] as the reference; for the Latin *Itinerarium Antonini* we use the digital gazetteer of *Pleiades* [3] as the major reference for the ancient and classical world, enriched with the additional data extracted from *Trismegistos Places* [1].

Particularly, we should consider that toponyms might be included in two contexts of hierarchical data and route networks in al-Muqaddasī's book. We define the places in the same region of the hierarchical data as neighbors while the neighborhood concept in route networks has its classic definition. Despite the considerable number of common places mentioned in both hierarchical and route network data in this source, we treat them in distinct processes. The reason is that the neighborhood concept, which is the contextual information to conduct the matching process more accurately, is different in each context. More clearly, utility of connectivity and adjacency of the toponyms is a key to partially contextualize a toponym and limit the area of our search to its related neighborhood. This will prevent mixing two different toponyms with similar names. For example, Aṭrābulus (or Ṭarābulus) is a settlement in north Africa in a region called Barqaŧ (Lybia), but also a settlement in al-Šām (Syria): the absence of such data causes an ambiguous or wrong match. Considering this point, the general idea is first to match toponyms of the sources against entries in the relevant references (Cornu's *Atlas* for Arabic and *Pleiades* for Latin), using exact matches. Then, we proceed the search with string similarity algorithms for matching the toponyms that do not have any exact matches. The current section is structured as follows. First, we prepare the data for applying matching in 4.1. Then, the main matching process is reported in 4.2 and 4.3.



```
FROM حلوان        TOWA قصر شيرين  DIST مرحلة
FROM قصر شيرين    TOWA خانقين      DIST مرحلة
FROM الأبلة        TOWA نهر دبا     DIST مرحلة
FROM نهر دبا       TOWA فم العضدى  DIST مرحلة
PROV أقور          TYPE كورة       REG1 ديار ربيعة
PROV أقور          TYPE كورة       REG1 ديار مضر
PROV أقور          TYPE كورة       REG1 ديار بكر
REG1 ديار ربيعة            TYPE قصبة        STTL الموصل
REG1 ديار ربيعة            TYPE مدينه       STTL الحديثة
REG1 ديار ربيعة            TYPE مدينه       STTL معلثاى
REG1 ديار ربيعة            TYPE مدينه       STTL الحسنية
```

Figure 7: Route sections extracted from al-Muqaddasī's Aḥsan al-taqāsīm after tagging

4.1 Data Preparation

Here, we consider three steps of data preparation. First, we target the language-dependent issues. Second, we explain the alignment of the regions of the premodern source and the reference gazetteer if that applies. Finally, the neighborhood concept will be added to the matching process.

Before starting any matching approach, it is essential to address and regulate language-dependent issues for Arabic and Latin. For Arabic toponyms, we apply a set of generic normalization rules by conflating *y* and *ý*; *h* and *ŧ(ḥ)*; all *alif* variants (A, Â, Ǎ, Ā); w and ŵ; ŷ and '. These rules simply form various forms of a character or vowel into one.

Concerning Latin as an inflected language, toponyms tend to appear in a particular declension according to the context. However, since currently available lemmatizers do not give reliable results for place names, we apply various string similarity matching criteria without doing lemmatization.

Since the regions of the coverage area in al-Muqaddasī, is slightly different from the Cornu's *Atlas* used as reference, we map the regions and provinces from various levels of these two sources. For instance, al-Muqaddasī describes "The Arabian Peninsula" as a region, called Jazīraŧ al-ʿarab, while Cornu's *Atlas* divides the peninsula into two regions, Jazīraŧ al-arab and al-Yaman. Another example would be the part of western provinces in north Africa and southern Europe. Cornu's *Atlas* divides this area into three regions: al-Andalus (Spain), Barqaŧ (Lybia), and Siqiliyyaŧ (Sicily) while al-Muqaddasī calls this whole area as "The Region of al-Maġrib". On the other hand, we can map the subordinate regions introduced by al-Muqaddasī to the corresponding regions in Cornu's *Atlas*.

For engaging the neighborhood information from the route network data, we create the network's representative graphs for both premodern and reference gazetteer graphs and extract the neighbors of each toponym up to the 3^{rd} step in the graph. Additionally, we go beyond the connections in the gazetteer for those places that are not a capital and find the other geographically close places to them at a certain distance [4]. This gives all other neighbors of a place which are not achievable through the defined connectivity in the graph. More contextual information for the gazetteer records can also be obtained, particularly for those places that do not have many connections in the network graph or for the matches that the connected places of toponyms in the source text and the gazetteer record are not similar at all. For the hierarchical data, siblings in the corresponding hierarchical tree are taken as neighbors of a place. In al-Muqaddasī's text, this idea works for finding the neighbors of a single micro-region that the toponyms belong to. In Cornu's *Atlas* we use the same geographically close places from the route network data in the abovementioned way since there is only the provincial level regions.

4.2 Matching Process

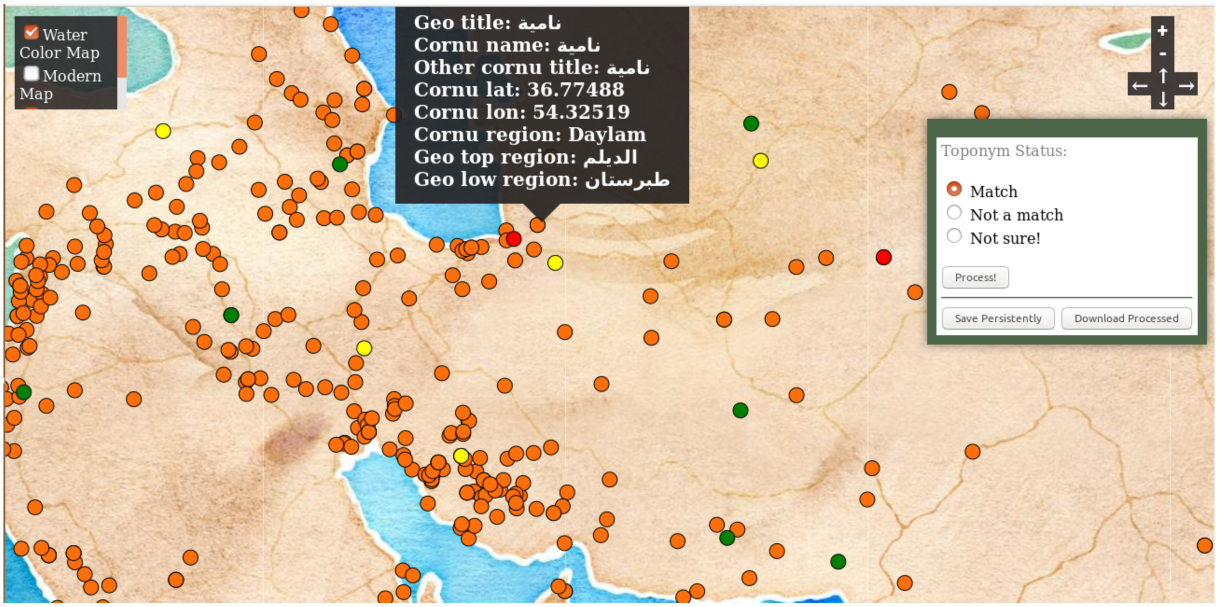
Before starting string matching process on our data set, we apply a number of string matching metrics on a small sample data set (see 4.3. Then, we start the matching process by searching for the exact matches. The first set of results requires a validation step to check all the toponyms matching more than one entry in the reference gazetteer, or multiple toponyms that match a single gazetteer entry. An example could be the places that share the same name in a single region. The key piece of information that contributes to clarify the results is the subordinate regions that are already engaged in the process, as well as neighbors that include the toponyms mentioned in the same micro-region as the toponym in question. Despite absence of explicit connectivity mentioned between the toponym and its neighbors, we can verify neighborhood of the chosen toponym on gazetteers, such as al-Ṯurayyā [4]. We also have developed an online tool particularly for this purpose. This tool gets an input set of matching data (toponyms and positions) and visualize them on a map (Fig. 8). Based on the visualized matches and all the related information on the map, the user can confirm/reject the proposed match or assigned undefined status for uncertain ones that are destined to further inquiries. The status of the matches are color-coded: orange circles are the matches that are not processed/evaluated yet; green color shows those matches that are verified as a correct match by the user; red represent the wrong matches; yellow highlights those which need more investigation. The status of each matches together with all the other information can be saved in a file and used for further analyses.

This step ignores some exact matches, since not all the places common to both sources are categorized in the same region. The toponym in question, e.g. placed in the region "R_1" in the source "S_1", might be in the neighboring region "R_2" in the the source "S_2". Therefore, excluding the enforcement of exact regional categorization from the process yields a list of identical toponyms that are not located in the same region in both sources. This refinement tweaks the matching process and guarantees that we do not lose matches due to arbitrary changes in the administrative divisions (related to historical period or specificity of the source); we solve this problem regarding the granularity level of divisions of the sources. However, the above-mentioned problem of identical names in a micro-region will still need to be addressed.

Remaining toponyms after verifying the exact match results are either totally distinct place names or the same place, but variously spelled. Hence, we continue the matching process by using string similarity metrics. Before starting this experiment, we specify the various cases of similar strings for identical toponyms in our sources:

- various spellings of the same name:
  - Ṭazar or Ṭazarŧ - Bahnasā of Bahnasaŧ
  - Ptandari or Tanadaris - Callipoli or Kallipolis

[4] We define this distance by experimenting an initial value of 10000 meters and increasing it where it is necessary until we find at least one capital in this area.



☑ Water Color Map
□ Modern Map

Geo title: نامية
Cornu name: نامية
Other cornu title: نامية
Cornu lat: 36.77488
Cornu lon: 54.32519
Cornu region: Daylam
Geo top region: الديلم
Geo low region: طبرستان

Toponym Status:
● Match
○ Not a match
○ Not sure!

Figure 8: The tool developed for manual verification of the matches

  - Callicome or Kallikome - Iovavum or Iuvavum
  - Beroia or Beroa - Acinquum or Aquincum
- various names for the same place:
  - Maḥallaṣurad or Maḥallasurad
  - Baġdād or Madinaŧ al-Salām
  - Vax or Villa Repentina - Bathnai or Markopolis
- toponyms made up of different parts, only partly appearing in a given context:
  - Raḥba Mālik Ibn Ṭawq
  - Colonia Agrippina or Colonia Claudia Ara Agrippinensium
  - Ratiaria legio XIIII Gemina or Ratiaria or Ratiaria Legio XIIII

The concept of using regions in the matching process can also be implemented in the route network data. Having multiple places with the same name in our data, the generated route network graph presents a model where distinct nodes bearing the same name are unified. In order to differentiate those toponyms, extra information is required. If the route network is described at regional level (or any contextual data available in the source), preserving regions as a property of each toponym will help to differentiate identical names in a graph and assign an individual node for each. However, this might raise another problem; in the inter-regional route descriptions, the route sections which connects two neighboring regions include toponyms belonging to two distinct regions. That means a route section starts in a regions and ends in another one. The automatic assignment of the regions to which this route section belongs will assign the source and destination the same region while one of them practically belongs to another region. For instance, al-Muqaddasī describes a route section in the province of al-Šām by which two neighbor provinces—al-Šām and Aqūr—are connected through two neighbor settlements from each province:

"..., then (from al-Ruṣāfaŧ) to al-Raqqaŧ half a stage."
Accordingly, the toponyms al-Ruṣāfaŧ and al-Raqqaŧ, will automatically belong to al-Šām while al-Raqqaŧ is a settlement in the province of Aqūr regarding the al-Muqaddasī’s description of divisions. Moreover, al-Raqqaŧ will be duplicated with a new province as a property, when al-Muqaddasī describes a route section starting from/ending at this place in Aqūr. Consequently, an individual place might be included in multiple regions while they are the same. A workaround would be specifying the identical places in multiple regions in order to consider this issue in further analyses of the graph.

Having the graph of route network and the tree of hierarchical data by which each toponym is assigned a region, we apply a number of the widely-known [2, 11] string metrics—for both Arabic and Latin data sets—from the list that [11] have applied for matching toponyms in Roman alphabet:

- Jaro Distance [17, 18]
- Jaro-Winkler [36]
- Monge Elkan [26]
- Jaccard Index [15]
- Dice [33]
- Overlap Coefficient [14]

4.3 Experimental Results

Applying the aforementioned information to our data structure, we build toponymic entities so that each entry carries a name, with a region and a subregion if available. As mentioned before, we start our experiment by running some string similarity metrics on a sample data set of 100 toponyms from al-Muqaddasī’s source by applying various thresholds. Each toponym is being checked against all 2532 entries in Cornu’s *Atlas*. The results (Table 1) show two numbers for each threshold of a metric. The left value is the

Table 1: First experiment of matching toponyms with string similarity metrics on the Arabic dataset

	Threshold[5]			
Method	1	2	3	4
Levenstein	69+1343	65+909	52+21	42+6
Jaro	58+905	55+99	45+17	40+0
Jaro W.	73+1544	59+73	51+36	42+6
Jaccard	59+173	52+81	45+26	41+12
Monge E.	64+2150	59+308	52+149	43+7
Overlap Co.	71+11	57+70	46+23	40+2
Dice	38+21	35+17	31+14	27+12
Affine	55+76	42+6	39+1	34+3
Needleman W.	57+40	43+52	40+55	38+19
Smith W.	58+147	44+90	41+89	39+89
Hamming	63+418	59+308	49+152	41+9
Nr. of topos	100			

Table 2: Matching results for Latin toponyms

Method	Matches	Percentage
Exact Match	1707	78,08
Jaro Distance	190	39,6
Jaro Winkler	25	8,6
Jaccard	34	12,8
Monge Elkan	41	17,8
Overlap Coefficient	3	1,5
Dice	38	2,04
Not available or Ambiguous	148	6,7
Total number of toponyms	2186	

number of correct matches and the right one is the number of wrong matches. Hence, the sum of these two numbers is the proposed matches of each method.

Based on this results, we then apply a selected list of metrics on our main datasets choosing the most efficient threshold for each. String length of the toponyms for some metrics like Levenstein is not considered for applying to the main datasets. As an instance, for a wide range of string lengths from 3 to 18, the threshold of 2 will match all the toponyms of length 3 to almost any toponym of length 3 with only one different character. This yields a huge list of trash results. Therefore, string length should be considered for choosing the threshold. We skip those metrics in this experiment and leave it for future work.

The results show different trends in the Arabic (Table 3, Table 4) and Latin dataset (Table 2). As we see here, a considerable amount of toponyms are still not correctly matched to any gazetteer entries even after applying all these metrics. It should be mentioned that these unmatched toponyms might not be included in the reference gazetteer.

Clearly, more Latin toponyms were available in the authority reference: we used the *Pleiades* [3] dataset, enriched with additional data retrieved from *Trismegistos Places* [1], where the province of each place of the Itinerary is also indicated (e.g. "Raetia" province, assigned to the place "Abusina" found in the text. This type of information is not provided by *Pleiades*). On the other hand, the lack of an exhaustive reference for Arabic besides Cornu's *Atlas* has generated much less exact results.

Table 3: Matching results for Arabic toponyms—hierarchical data

Method	Matches	Percentage
Exact Match	693	52,2
Jaro Distance	122	19,2
Jaro Winkler	17	3,3
Jaccard	11	2,2
Monge Elkan	2	0,4
Overlap Coefficient	1	0,2
Dice	0	0,0
Not available or Ambiguous	514	38,7
Total number of toponyms	1327	

Table 4: Matching results for Arabic toponyms—route network

Method	Matches	Percentage
Exact Match	789	65,2
Jaro Distance	141	33,4
Jaro Winkler	1	0,35
Jaccard	18	6,4
Monge Elkan	6	2,3
Overlap Coefficient	1	0,39
Dice	8	3,1
Not available or Ambiguous	260	21,4
Total number of toponyms	1223	

However, the results from the *Itinerarium Antonini* (Table 2) also bear a substantial percentage of ambiguity and unresolvable toponyms. The reason for this phenomenon lies in the multiple occurrences of the same name, which, in fact refer to different places in the text: very common names in the Roman Empire occur frequently in different areas (Mediolanum, Alexandria, etc.) or are only vaguely indicated by the author himself (Novas, Castra, In Medio, ad Fines, Portus, Limes, etc.). The disambiguation of these cases can only be performed through the extraction of the neighbors, according to how they appear in the annotated source for each occurrence of the same toponym. The combined reference to the region of appearance of the ambiguous toponym and of its neighbors results in disambiguation on geographical basis: the same regional criterion was applied for disambiguation in our Arabic source (see 4.2).

5 CONCLUSION AND FUTURE WORK

We have described the main characteristics of geographical narrative in premodern sources, used and extended a lightweight and easy-to-use tagging system for annotating and extracting semantically meaningful linguistic patterns expressing spatial information. Having extracted the data, we have proposed an implementation of a process to identify places by using contextual information and experimenting exact and approximate matching algorithms. We have discussed all the issues we faced when testing two very different use cases, and introduced a set of heuristic solutions to solve problems related to the specificities of natural languages. The offered approach is practically applicable to any other source that deals with complex patterns of geographical descriptions and data sets with similar issues. The detailed experiment can be used to envisage the problems or issues for works on new sources as well.

The tagging system is developed to cover the major geographical descriptive patterns that appear frequently and systematically for specific uses. To include specifications of directions and other granular descriptions, which appear in the description of routes, this tagging system needs to be expanded and improved. Resolving ambiguities—e.g. when a route shifts from one province to the next—requires improvements on the computational side.

In the process of identifying places, our first attempt of matching toponyms provides a list of canonical place names in both hierarchical and network data. Having these toponyms, we can take advantage of combining the two data sets, for cases where such information is available. The idea is to exploit canonical and individual toponyms and their closeness and connections using state-of-the-art approaches, such as machine learning, in order to identify the non-matching toponyms.

REFERENCES

[1] 2005. Places. http://www.trismegistos.org/geo/index.php. (2005). Accessed: 2017-09-20.

[2] 2013. A Comparison of String Similarity Measures for Toponym Matching. In *Proceedings of The First ACM SIGSPATIAL International Workshop on Computational Models of Place (COMP '13)*. ACM, New York, NY, USA, Article 54, 8 pages. https://doi.org/10.1145/2534848.2534850

[3] 2013. pleiades.stoa.org. https://pleiades.stoa.org. (2013). Accessed: 2017-09-25.

[4] 2016. al-Ṯurayyā Project. https://althurayya.github.io. (2016). Accessed: 2017-09-20.

[5] 2017. OpenITI mARkdown. https://alraqmiyyat.github.io/mARkdown/. (2017). Accessed: 2017-09-23.

[6] al-Muḥammad ibn Aḥmad Muqaddasī. 1906. *Kitāb Aḥsan al-taqāsīm fī maʿrifat al-aqālīm*. Bibliotheca Geographorum Arabicorum, Vol. 1-3. Brill, Leiden.

[7] al-Muḥammad ibn Aḥmad Muqaddasī. 1994. *The best divisions for knowledge of the regions: a translation of Aḥsan al-taqāsīm fī maʿrifat al-aqālīm*. Reading, UK : Centre for Muslim Contribution to Civilization : Garnet Pub.

[8] K. Brodersen. 1995. *Terra Cognita: Studien zur römischen Raumerfassung*. Georg Olms Verlag AG. https://books.google.de/books?id=3cV-AAAAMAAJ

[9] Mauro Calzolari. 1996. *Introduzione allo studio della rete stradale dell'Italia romana: l'itinerarium Antonini*. Roma: [Accademia Nazionale dei Lincei].

[10] G. Cheng, F. Wang, H. Lv, and Y. Zhang. 2011. A new matching algorithm for Chinese place names. In *2011 19th International Conference on Geoinformatics*. 1–4. https://doi.org/10.1109/GeoInformatics.2011.5980801

[11] William W. Cohen, Pradeep Ravikumar, and Stephen E. Fienberg. 2003. A comparison of string distance metrics for name-matching tasks. 73–78.

[12] Georgette Cornu. 1985. *Atlas du monde arabo-islamique à l'époque classique: I^{Xe}–X^{e} siècles*. Brill, Leiden.

[13] Otto Cuntz and Gerhard Wirth. 2012. *vol. 1: Itineraria Antonini Augusti et Burdigalense, Accedit tabula geographica*. De Gruyter, Berlin/Boston.

[14] Henry F. Inman and Edwin L.Bradley Jr. 1989. The overlapping coefficient as a measure of agreement between probability distributions and point estimation of the overlap of two normal densities. *Communications in Statistics - Theory and Methods* 18, 10 (1989), 3851–3874. https://doi.org/10.1080/03610928908830127 arXiv:http://dx.doi.org/10.1080/03610928908830127

[15] P. Jaccard. [n. d.]. Distribution de la flore alpine dans le Bassin des Dranses et dans quelques regions voisines. 37 ([n. d.]), 241–272.

[16] P. Janni. 1984. *La mappa e il periplo: cartografia antica e spazio odologico*. Bretschneider. https://books.google.de/books?id=sXUSAQAAIAAJ

[17] Matthew A. Jaro. 1989. Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. *J. Amer. Statist. Assoc.* 84, 406 (1989), 414–420. https://doi.org/10.1080/01621459.1989.10478785 arXiv:http://www.tandfonline.com/doi/pdf/10.1080/01621459.1989.1047878 5

[18] Matthew A. Jaro. 1995. Probabilistic linkage of large public health data files. *Statistics in Medicine* 14, 5-7 (1995), 491–498. https://doi.org/10.1002/sim.4780140510

[19] Marinos Kavouras, Margarita Kokla, and Eleni Tomai. 2005. Comparing categories among geographic ontologies. *Computers & Geosciences* 31, 2 (2005), 145 – 154. https://doi.org/10.1016/j.cageo.2004.07.010 Geospatial Research in Europe: AGILE 2003.

[20] Deniz Kılınç. 2016. An accurate toponym-matching measure based on approximate string matching. *Journal of Information Science* 42, 2 (2016), 138–149. https://doi.org/10.1177/0165551515590097 arXiv:https://doi.org/10.1177/0165551515590097

[21] Jochen L Leidner. 2008. *Toponym Resolution in Text : Annotation, Evaluation and Applications of Spatial Grounding of Place Names*. Universal Press, Boca Raton, FL, USA.

[22] Lin Li, Xiaoyu Xing, Hui Xia, and Xiaoying Huang. 2016. Entropy-Weighted Instance Matching Between Different Sourcing Points of Interest. *Entropy* 18, 2 (2016).

[23] Bruno Martins. 2011. *A Supervised Machine Learning Approach for Duplicate Detection over Gazetteer Records*. Springer Berlin Heidelberg, Berlin, Heidelberg, 34–51. https://doi.org/10.1007/978-3-642-20630-6_3

[24] Grant McKenzie, Krzysztof Janowicz, and Benjamin Adams. 2014. A weighted multi-attribute method for matching user-generated Points of Interest. *Cartography and Geographic Information Science* 41, 2 (2014), 125–137. https://doi.org/10.1080/15230406.2014.880327 arXiv:http://dx.doi.org/10.1080/15230406.2014.880327

[25] Ludovic Moncla, Mauro Gaio, Javier Nogueras-Iso, and Sébastien Mustière. 2016. Reconstruction of itineraries from annotated text with an informed spanning tree algorithm. *International Journal of Geographical Information Science* 30, 6 (2016), 1137–1160. https://doi.org/10.1080/13658816.2015.1108422 arXiv:http://dx.doi.org/10.1080/13658816.2015.1108422

[26] Alvaro Monge and Charles Elkan. 1996. The field matching problem: Algorithms and applications. In *In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. 267–270.

[27] M. Monmonier. 2014. *How to Lie with Maps*. University of Chicago Press. https://books.google.de/books?id=7pHeBQAAQBAJ

[28] CHIARA PALLADINO. 2016. NEW APPROACHES TO ANCIENT SPATIAL MODELS: DIGITAL HUMANITIES AND CLASSICAL GEOGRAPHY. *Bulletin of the Institute of Classical Studies* 59, 2 (2016), 56–70. https://doi.org/10.1111/j.2041-5370.2016.12038.x

[29] Nicholas Reed. 1978. Pattern and Purpose in the Antonine Itinerary. *The American Journal of Philology* 99, 2 (1978), 228–254. http://www.jstor.org/stable/293648

[30] Maxim Romanov. 2017. Algorithmic Analysis of Medieval Arabic Biographical Collections. *Speculum* 92, S1 (2017), S226–S246. https://doi.org/10.1086/693970 arXiv:https://doi.org/10.1086/693970

[31] João Santos, Ivo Anastácio, and Bruno Martins. 2014. Using machine learning methods for disambiguating place references in textual documents. *GeoJournal* (2014), 1–18. http://dx.doi.org/10.1007/s10708-014-9553-y

[32] Rui Santos, Patricia Murrieta-Flores, and Bruno Martins. 2017. Learning to combine multiple string similarity metrics for effective toponym matching. *International Journal of Digital Earth* 0, 0 (2017), 1–26. https://doi.org/10.1080/17538947.2017.1371253 arXiv:http://dx.doi.org/10.1080/17538947.2017.1371253

[33] T. Sørensen. 1948. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. *Biol. Skr.* 5 (1948), 1–34.

[34] TeX Users Group 2016. *Ontologies and the Cultural Heritage. The case of GO!* TeX Users Group.

[35] M. Thiering. 2015. *Spatial Semiotics and Spatial Mental Models: Figure-Ground Asymmetries in Language*. De Gruyter. https://books.google.de/books?id=E4jnBQAAQBAJ

[36] William E. Winkler. 1999. *The state of record linkage and current research problems*. Technical Report. Statistical Research Division, U.S. Bureau of the Census.

[37] Johan Åhlfeldt, Merrick Lex Berman, and Marc Wick. 2016. *Historical Gazetteer System Integration : CHGIS, Regnum francorum Online, and GeoNames*. Indiana University Press, 110–125.