International Journal of Geo-Information

MDPI

Article

Spatiotemporal Information Extraction from a Historic Expedition Gazetteer

Mafkereseb Kassahun Bekele [1,*], Rolf A. de By [2] and Gaurav Singh [3]

[1] Department of Computer Science, University of Cape Town, Cape Town 7700, South Africa
[2] Department of Geo-Information Processing (GIP), Faculty of Geo-information Science and Earth Observation (ITC), University of Twente, 7514 AE Enschede, The Netherlands; r.a.deby@utwente.nl
[3] Regional Integrity Management Systems, ROSEN Europe B.V., 7575 EJ Oldenzaal, The Netherlands; gsingh@rosen-group.com
* Correspondence: mafkereseb@hotmail.com; Tel.: +27-72-925-0688

Academic Editor: Wolfgang Kainz
Received: 17 October 2016; Accepted: 24 November 2016; Published: 29 November 2016

Abstract: Historic expeditions are events that are flavored by exploratory, scientific, military or geographic characteristics. Such events are often documented in literature, journey notes or personal diaries. A typical historic expedition involves multiple site visits and their descriptions contain spatiotemporal and attributive contexts. Expeditions involve movements in space that can be represented by triplet features (location, time and description). However, such features are implicit and innate parts of textual documents. Extracting the geospatial information from these documents requires understanding the contextualized entities in the text. To this end, we developed a semi-automated framework that has multiple Information Retrieval and Natural Language Processing components to extract the spatiotemporal information from a two-volume historic expedition gazetteer. Our framework has three basic components, namely, the Text Preprocessor, the Gazetteer Processing Machine and the JAPE (Java Annotation Pattern Engine) Transducer. We used the Brazilian Ornithological Gazetteer as an experimental dataset and extracted the spatial and temporal entities from entries that refer to three expeditioners' site visits (which took place between 1910 and 1926) and mapped the trajectory of each expedition using the extracted information. Finally, one of the mapped trajectories was manually compared with a historical reference map of that expedition to assess the reliability of our framework.

Keywords: Geographic Information Retrieval; Temporal Information Retrieval; Natural Language Processing; temporal inference

1. Introduction

Historic expeditions are journeys made in the past with exploratory, scientific, military or geographic intentions [1]. The spatiotemporal and thematic properties of such historic expeditions are likely to be represented, often in printed documents, which are contextual in nature. In general, the contexts that exist in historic expedition documents are spatial, temporal and descriptive. Element extraction from the textual documents will provide alternatives to represent and visualize those historic events in a spatiotemporal environment. Historic expeditions are past strings of events that are likely documented in unstructured text formats and have possibly left their traces in history. Reading such documents is not adequate for visualizing the events with a full spatiotemporal perspective or for conducting further studies; extracting the spatiotemporal and descriptive contents from the documents is required to get the associated contexts.

Expedition gazetteers (here seen as documents that provide a narrative of places and events related to expeditions) often have three basic characteristics: spatial, temporal and descriptive. Historic

expeditions were carried out for different purposes. However, the gazetteers of such expeditions have common characteristics: they all have spatial, temporal and descriptive phrases in their respective texts. Hence, the main objective of this article is to present a spatiotemporal information extraction framework that consumes those gazetteers and extracts the spatial and temporal entities from the texts.

The extraction of spatiotemporal entities from an expedition gazetteer is challenging because it may contain *endonyms*, names given to places by local people, or *exonyms*, names given to places by outsiders, or they may have phrases that express spatiotemporal relationships. Moreover, a gazetteer text may display spatial and temporal vagueness. A spatial entity might be characterized with a vague phrase such as "*a few miles from place X*", place names or coordinates may be missing, which leads to ambiguous information extraction results; the scope of this article does not cover both spatial relationships and spatial vagueness. In addition, temporal vagueness in gazetteer texts may cause ambiguity when extracting such items. For instance, a time-marker such as "*January 1922*" is vague because start and end dates of the events described are not explicit, and such vagueness leads to inconclusive duration of the event. The recognition and extraction of a crisp—explicitly mentioned—temporal and spatial entity is relatively easy; nevertheless, a successful extraction of spatiotemporal information needs to resolve this spatial and temporal vagueness. In our framework, we only address the temporal vagueness in the expedition gazetteer texts. The framework has a temporal inference and reasoning tool to determine, where possible, missing temporal boundaries.

Approximately 80% of all the world's information is stored as unstructured textual documents, and 85% of this has spatiotemporal traces [1]. Consequently, a high demand exists for methods to structure and extract such contents. For instance, the *Brazilian Ornithological Gazetteer* [2,3]—the corpus that we use for this research project—identifies approximately 6000 Brazilian sites where birds were observed or collected (this paper focuses particularly on three expeditions in the years 1910–1926). Reading the text does not fully satisfy the need to visualize the undertaken historic expedition from a spatiotemporal perspective, because it is full of entities such as people's names, place names, institute names, and spatial and temporal markers described by natural language (see Figure 1). A deepened spatiotemporal understanding can help to make actual timing or location of events explicit, or otherwise, can possibly help to restrict time/place options.

We thus need Natural Language Processing (NLP) and Information Retrieval (IR) methods to extract these spatial/temporal/spatiotemporal entities from the text to visualize the events in a spatiotemporal environment, and allow pinpointing [1]. The general aim of this article is to present and discuss a semi-automated spatiotemporal information extraction framework with multiple NLP and IR techniques that can

- extract spatiotemporal information from historic expedition gazetteer texts;
- help understand the temporal relationships between vague timeframes; and
- infer relative timeframes.

Our approach is not restricted to the *Brazilian Ornithological Gazetteer*; it is supposed to work for any expedition gazetteer that comprises spatial, temporal and descriptive phrases in its text.

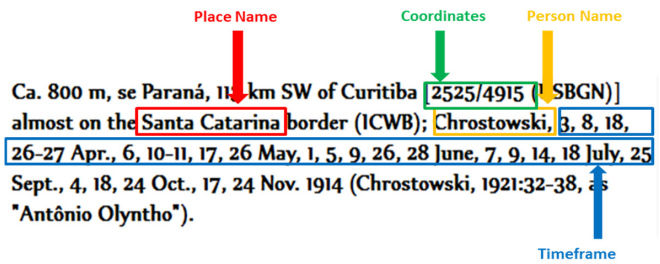

Figure 1. Typical expedition gazetteer entry that describes one location and its history of visits (source: [2,3]). Note: these sources are shared under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 license.

2. Related Work

2.1. Geospatial Information Extraction from the Web and Text Documents

Standard web search engines treat geospatial terms just like descriptive terms used as key words to search for specific documents, information or services. This may lead to failure in finding relevant search results. However, association of spatial and textual indexing has been proposed in [4] as a solution. The study in [5] uses addresses and postal codes, telephone numbers, geographic feature names, and hyperlinks as sources of geospatial context to discover geospatial contents in web pages. Even though many academic studies in the geographic search technology area have focused primarily on techniques to extract geographic knowledge from the web, Chen et al. [6] study the problem of efficient query processing in scalable geographic search engines and proposes several query processing algorithms that compute the score of documents that contain query terms. The study investigates how to maximize the query throughput for a given problem size and amount of hardware.

As a result of the conventional Internet acquiring a geospatial dimension, web documents are becoming geo-tagged objects. Considering both spatial proximity and text relevancy in such objects, Wu et al. [7] propose a new indexing framework and query that are achieved by the fusion of geo-location and text.

Much research has been carried out to extract geospatial information from different text contents, such as web queries, micro-text messages, metadata and Wikipedia entries [8–12], and some research [13] has been conducted on retrieving temporal information from text documents. Our effort focuses on introducing geospatial or temporal information extraction methods from relatively structured text contents. However, a significant challenge remained in bridging the semantic gap between structured geospatial data as held in a GIS and hard-to-analyze spatial information, expressed in natural language [14]. The study in [14] uses a natural language information processing platform called GATE (General Architecture for Text Engineering) to extract geographical named entities and associated spatial relations in natural language, based on syntactical rules from a large-scale annotated corpus. The study in [15] introduced self-annotation as a new supervised learning approach for developing and implementing a system that extracts fine-grained relations between entities (such as geospatial relations). The main benefit of self-annotation is that it does not need manual labeling. Studies have been conducted to extract both spatial and temporal information from documents. For instance, Strötgen et al. [16] present an approach that combines the Temporal Information Retrieval (TIR) and Geographic Information Retrieval (GIR) domains in the context of document exploration and information extraction tasks. In addition, our framework focuses on extracting spatiotemporal (geospatial and temporal) information from historic expedition gazetteers.

GIR is the interaction of GIS and IR. Brisaboa et al. [17] present two types of information retrieval approaches that fall to such domain, specifically, a textual technique and a spatial technique, targeting linguistic and spatial aspects of documents, respectively. On the other hand, Boguraev and Ando [13] describe a temporal analysis framework to discover the temporal dimension of a corpus.

Travel guides and travel diaries were used in [18] to correctly recognize geographic information and construct actual trajectory datasets that can be visualized on a map. In this research project, the extraction of relative and absolute geographic information has been achieved. The main advantage of the method used in [18] is that only the linguistic, semantic and contextual information contained in the provided documents are used. The study in [19] came up with a system that adds a spatio-linguistic reasoner to interpret the spatial language mentioned in image captions. The system helps to determine the location of images based on the spatial information contained in their captions.

2.2. Geo-Parsing

Geo-parsing is a method that identifies and annotates geospatial entities in text documents [20]. A geo-parsing web service was developed by [21] to extract geospatial information from travel narratives using Yahoo! Placemaker as a geo-tagging tool. The service has two main steps: entity

extraction and disambiguation. However, the issue of relative positioning of spatial objects was not addressed. The service can extract geospatial entities and visualize them, but the spatiotemporal relationships between entities were not under study in this approach. Such narratives often contain vague temporal entities, which require a temporal inference tool to resolve. Unlike this approach, our framework includes a temporal inference tool to resolve temporal vagueness in the expedition gazetteer texts. In addition, the framework focuses on extracting spatiotemporal information from historic expedition gazetteers. To this effect, the framework depends on the linguistic and contextual information contained in the provided gazetteer.

2.3. Temporal Reasoning

A reasoning activity in a dynamic domain needs to include a temporal perspective [22]. The *time semi-interval* is a temporal primitive that is the *start* or *end* point of an event (an event is a location visit in our case). In [23], *time semi-intervals* and their relationships are used as the basic units of temporal knowledge. Temporal reasoning between *time semi-intervals* requires a reasoning capability to compute the missing temporal member of the primitive, either the *start* or *end* of an event. The Brazilian Ornithological Gazetteer has location descriptions with non-crisp temporal marks, for instance, "*Aug. 1922,*" a vague temporal entity because the exact start and end date of the event represented by this mark are not explicit. In such cases, a temporal inference method is required to infer the relative temporal boundaries of a given location visit. To do so, our framework has a tool that infers a relative timeframe for the vaguely defined location visits relative to other crisply defined location visits.

Historical descriptions have time as a fundamental concern when representing information [24]. If the temporal description of a historical event is vague, then the temporal information to be extracted is subject to uncertainty [24]. Historical events are not always represented with crisp temporal phrases, but with imprecise and subjective ones [24]. Nagypál and Motik [24] state that existing approaches for temporal modeling are based on the assumption that representation of time is crisp. These approaches therefore cannot be applied to all temporal modeling tasks. To overcome the difficulties of vague temporal information representation, Nagypál and Motik [24] present a *fuzzy interval-based temporal model* that is capable of capturing vague temporal information.

Time instants and *time intervals* are mentioned as basic time primitives in [24]. However, a *time instant* becomes a *time interval* if temporal granularity is increased and the interval is one of the usual well-known time intervals (such as day, week, and month). For instance, a month is considered a time instance when it is counted in a given year, but a month itself is a time interval when the days of a given month are considered time instants. Temporal statements are common in historical expedition texts, but they are not always crisp. As a result, we may end up with vague temporal information. This is the main reason for including a temporal inference tool in our framework.

3. Data Source, Tools and Methods

3.1. Data Source

The Ornithological Gazetteer of Brazil, which has more than 6000 descriptions of sites where ornithological expeditions operated throughout Brazil, was compiled by Paynter and Traylor [2,3]. The gazetteer has records of site visits by known expeditioners (here, by "expeditioner" we mean a person who conducts expeditions). Tadeusz Chrostowski (1878–1923, see Figure 2), Maria Emilie Snethlage (1868–1929) and Emil Heinrich Snethlage (1897–1939) are three well-known expeditioners whose names are mentioned many times in the gazetteer. The texts of the gazetteer that mentioned the names of those expeditioners were used to experiment with our framework. For instance, the name "*Chrostowski*" is mentioned in 58 entries.

Figure 2. Tadeusz Chrostowski: 1878–1923 (source: Wikipedia).

3.2. GATE Developer

GATE (General Architecture for Text Engineering) is a text-processing platform used to develop applications that process natural language [25]. The platform consists of processing components that can be used for information extraction systems. GATE has various component types, known as resources, which are reusable, specialized JavaBean types, components that can be manipulated visually in a builder tool [25]. These resources come in three varieties: Language Resources (LRs), Processing Resources (PRs) and Visual Resources (VRs).

3.3. ANNIE

ANNIE (A Nearly-New Information Extraction System) is an information extraction tool distributed with GATE that relies on the basics of text-processing algorithms that focus on sentence chunking, splitting, POS (Part of Speech) tagging and transducing, and the JAPE (Java Annotation Pattern Engine) language that is used to define patterns of items in a textual representation [25].

3.3.1. ANNIE Tokenizer

The ANNIE tokenizer is a tool that chunks a text into a number of typed tokens such as words and numbers [25]. The tokenizer uses a rule that has LHS (Left-Hand Side) and RHS (Right-Hand Side) parts. The LHS is always a regular expression that has to be compared against an input text, whereas the RHS contains the action to be carried out when the LHS expression is matched with the input text [25]. The token types created by the ANNIE tokenizer on input texts are *Word*, *Number*, *Punctuation*, and *SpaceToken*.

3.3.2. ANNIE Sentence Splitter

The ANNIE sentence splitter is a transducer that chunks an input text into a number of sentences. (In the context of this paper, a transducer is a method with input and output phases.) In most cases, a sentence splitter is preceded by a tokenizer because the punctuations in a text are used to split the document into sentences. The sentence splitter uses a gazetteer list of abbreviations to

help it identify a sentence-marking full stop [25]. For instance, consider the sentence "*Mr. Johnson was born in Feb 1989*"; the full stop after "*Mr*" is not a sentence-marking stop. The gazetteer list of abbreviations is application-dependent and subjected to the characteristics of the text-processing machine. After splitting, each sentence is annotated as *Sentence* and each sentence break is annotated as *Sentence Split* [25].

3.3.3. ANNIE POS Tagger

The ANNIE POS tagger follows the tokenizer and the splitter. The tagger produces a POS tag as an annotation class on each *Word* or *Number* token. The annotation class produced by the tagger is used by a pipeline module to extract Named Entities. Each POS tag is considered as a token category by other applications, assuming the applications need a tagged POS that follows the POS tagger in the information extraction pipeline.

After a sentence is tagged by the POS tagger, the output annotation classes along with the POS categories are used in JAPE grammar rules to define the LHS rules of the entity pattern expressions. Here, it is worth noting that annotation classes over the actual sentence and the POS categories have execution orders; the latter is always executed before annotating the entities—spatial, temporal and descriptive in our case—in the actual text. Therefore, the POS categories created by the POS tagger along with the annotation classes created by the ANNIE gazetteer are inputs for the pattern—such as patterns of date *"January 14, 1921"* and coordinates *"9999/9999"*—definition of entities in the expedition gazetteer text.

3.3.4. ANNIE Gazetteer

The ANNIE gazetteer is the part of ANNIE that identifies entity names in the text based on lists. It tags entities in a text—place names, person names and months—using a method that matches the text against lists of items—place names, person names and months. It identifies entity names in the input text by checking their existence in the item list. The lists are plain text files with one entry per line. Each list file represents a set of entity names such as cities, organizations, days of the week and months. Entities of similar categories must be stored with their kinds only. This tagging resource can be tuned to be case-sensitive or insensitive.

The lists of entity names are stored as a ".list" file. An index file is used to access the ".list" files. The "lists.def" file provides the definition of each list file. The definition includes the *file name*, *major type*, *minor type*, *language* and *annotation type* as *columns one to five*, respectively.

3.4. JAPE: Regular Expressions over Annotations

The JAPE (Java Annotation Pattern Engine) allows the recognition of predefined regular expressions of annotation classes over textual documents: a regular expression is set of strings—it does not include graphs. The JAPE transducer always follows the tokenizer, splitter, POS tagger, and/or gazetteer processing module. The tagged POSs of an input text and annotation classes created by the gazetteer processor and the JAPE grammar rules are used by the JAPE transducer to annotate an input text. This set of grammar rules is one of the basic modules in our framework.

3.4.1. JAPE Grammar Rule

A JAPE grammar is a set of pattern-based rules, each of which consists of a set of phases. These rules are stored as a ".jape" file. An index file is used to access the JAPE grammar phases—if multiple phases are defined. Each of the phases consists of a set of pattern/action rules. The rule has a LHS and RHS part. The LHS contains a pattern of entities in a given sentence. The RHS rule contains the action to be taken whenever the pattern on the LHS is matched in a sentence (input texts). In general, the JAPE grammar rules use the following LHS operators:

(or) |
(0 or more occurrences) *

(0 or one occurrence) ?

(1 or more occurrences) +

The following is an example of a JAPE rule that identifies a distance represented by a combination of word, number and punctuation tokens, such as *"ca. 45.1 km"*. Here, *"ca."* means approximately (see Figure 3).

```
1 Phase : distancefinder
2 Input :
3 Options : control = appelt
4 Rule : distance
5 Priority : 50
6 ((
7   ({Token.kind == word, Token.category == MD, Token.string == "Ca"}|
8    {Token.kind == word, Token.category == MD, Token.string == "ca"})
9   ({Token.kind == punctuation, Token.string == "."})?
10  ({SpaceToken}))?
11  (
12  {Token.kind == number, Token.length == "1"}|
13  {Token.kind == number, Token.length == "2"}|
14  {Token.kind == number, Token.length == "3"}|
15  {Token.kind == number, Token.length == "4"}|
16  {Token.kind == number, Token.length == "5"}
17  )
18  (({Token.kind == punctuation, Token.string == "."})?
19  ({Token.kind == number, Token.length == "1"})?
20  ({SpaceToken})
21  ({Token.kind == word, Token.category == NN, Token.orth == lowercase, Token.string == "km"}))
22 )
23 :distance
24 -->
25 :distance.Distance = {rule = "distance"}
```

Figure 3. Example of JAPE grammar rule, the purpose of this rule is to demonstrate how the JAPE rules are defined. The explicit JAPE rules of our semi-automated framework are provided as a separate dataset.

Line 1 defines the phase name. Each of the phases in the JAPE grammar must have a unique name, for instance, here, the phase is named *distancefinder*.

Line 2 defines the input annotations, which the LHS rule uses for pattern-matching, and which must be defined at the start of each grammar. In the absence of an explicit definition of the input annotations, the defaults are *Token*, *SpaceToken* and *Lookup*.

Line 3 defines the option. There are two types of options (control and debug) that can be set at the beginning of each grammar rule:

1. *control* is a rule-matching method. The control options are *Appelt*, *Brill*, *All* or *Once*. For instance, the *Appelt* forces the JAPE grammar rule to trigger a rule with higher priority first.
2. *debug* can be set to either true or false. It notifies a conflict between more than one possible match if it is set to true.

Line 4 defines the name of the rule; in this example, the name is *distance*.

Line 5 defines the priority of the rule. If there are multiple rules in a single phase, the rules with higher priority are triggered and matched prior to the rest.

Line 6–23 is the LHS of the rule. Here, the rule searches for a part of an input text that is a combination of word and number. This LHS pattern rule has three subpatterns:

1. *Subpattern one* matches a combination of word, punctuation and white space that equals *"Ca."* or *"ca."*; note the white space before the closing quotations *(Line 6–10)*.
2. *Subpattern two* matches a string of digits in one of the following formats: *"9"*, *"99"*, *"999"*, *"9999"*, *"99999" (Line 11–17)*.
3. *Subpattern three* matches a combination of punctuation, number, white space and word that resembles *"0.1 km" (Line 18–21)*.

The subpatterns in combination create a pattern rule that matches a distance in a text (e.g., *"ca. 45 km"*). When a part of a text is matched with this pattern, the LHS rule tags the matched part with a temporary label; in this example, the temporary label is *distance*.

Line 23 defines the temporary annotation class.
Line 24 separates the LHS and RHS.
Line 25 is the RHS of the rule renames the temporary label (*Line 23*) into a permanent annotation class. In this example, the temporary label *distance* is renamed into a permanent label (*Distance*). The new label is recognized as an annotation class by other JAPE phases.

3.4.2. LHS Macros

The LHS Macros are methods that allow creating a definition of a regular expression that can be used multiple times in the JAPE rules. The LHS macros are not independent rules that annotate an entity, but they are used as subpatterns of the JAPE grammar rule that matches the parts of a given text. These macros are called inside the rule defined to match a specific entity.

3.4.3. JAPE Transducer

A transducer translates the contents of its input, the LHS rule, to new the content of output, the RHS rule. In our context, it takes an input text, the expedition gazetteer, and returns a text with annotation classes, the annotated expedition gazetteer.

4. Spatiotemporal Information Extraction Framework

The semi-automated framework we presented in this section has multiple components. Most of the components were constructed from the default components of the GATE text-processing application. However, we believe that the framework has two contributions to the GIR and TIR fields of research. The contribution of the framework for the GIR field is showing how spatiotemporal information can be extracted from an expedition gazetteer using pattern- and list-matching techniques. In addition, the contribution of the framework for the TIR is the Temporal Inference algorithm (see Section 4.7)—we consider the temporal inference as the most innovative part of our paper.

All the components of ANNIE are used to build the spatiotemporal information extraction framework. The framework has three basic components, namely text preprocessor, gazetteer processing machine and JAPE transducer (see Figure 4). These components constitute the contextual spatiotemporal information extraction framework. The text preprocessor module is a preliminary annotator that chunks the expedition gazetteer text into tokens and performs POS tagging. On the other hand, the gazetteer processing machine and JAPE transducer are the main modules that recognize and annotate spatial and temporal entities from the expedition gazetteer texts. After we extract the spatiotemporal information from the expedition gazetteer texts, we stored the information in a PostgreSQL database that we developed for this task.

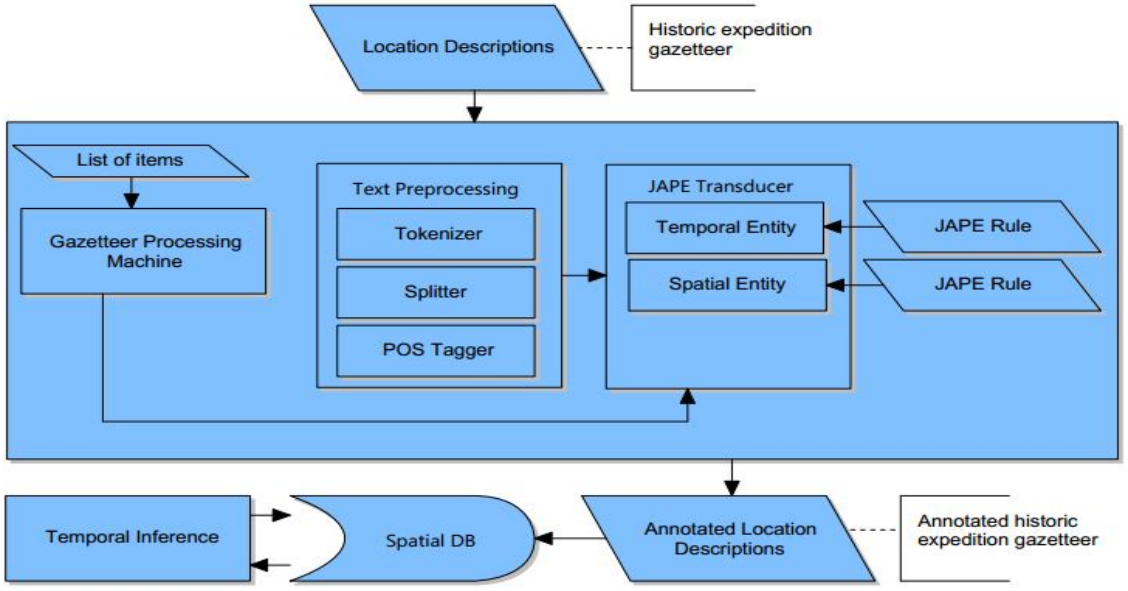

Figure 4. A framework to extract spatiotemporal information from a historic expedition gazetteer.

4.1. Raw Data Extraction (Location Descriptions)

Some entries in our dataset contain descriptions of visits by a single expeditioner (see Figure 5) while others contain descriptions of visits by multiple expeditioners (see Figure 6).

> **Ca. 800 m, se Paraná, 113 km SW of Curitiba [2525/4915 (USBGN)] almost on the Santa Catarina border (ICWB); Chrostowski, 3, 8, 18, 26–27 Apr., 6, 10–11, 17, 26 May, 1, 5, 9, 26, 28 June, 7, 9, 14, 18 July, 25 Sept., 4, 18, 24 Oct., 17, 24 Nov. 1914 (Chrostowski, 1921:32–38, as "Antônio Olyntho").**

Figure 5. Entry with single expeditioner. This entry is extracted from the dataset used for this research (paynter database). The entry ID is 251 (Source: [2,3]).

> **173 m (GRB); extreme sw Paraná on Rio Paraná [3343/5817 (USBGN)] just before confluence with Rio Iguaçu [2536/5436 (USBGN)] at junction of Brazil, Paraguay, and Argentina (MHA, as "Foz do Iguassú", ICWB); Chrostowski, et al., 18–25 Mar. 1923 (Jaczewski, 1925:348, as "Foz do Iguassú, Sztolcman, 1926:113, as "Foz Iguassú"); Snethlage, Oct. 1928 (Snethlage, 1936:90, as "Foz do Iguassú"); Kaempfer, 16 –21, 28, 30 May, 2 June 1930 (Naumburg, 1935:463, as "Foz do Iguassu", 1937:149, as "Fazenda do Iguassú," p. 192, as "Foz do Iguassu"), Partridge, Dec. 1955 (Partridge, 1961, Neotrópica, 7:26, as "Foz do Iguazú"), Zimmer, 1955, Amer. Mus. Novit., no. 1749, p. 17, as "Foz de Iguassú").**

Figure 6. Entry with multiple expeditioners mentioned. This entry is extracted from the dataset used for this research (paynter database). The entry ID is 3130 (Source: [2,3]).

We developed a tool to extract raw data—location descriptions—from the gazetteer. This is a preparatory process for the main spatiotemporal information extraction framework. The tool extracts location descriptions of expeditions that are assumed to be associated with a given expeditioner—in case the name of the expeditioner is provided—and stores the extracted location descriptions as an XML (Extensible Markup Language) document in which one XML element contains a sentence that has the temporal, spatial and attributive phrases of particular locations visited.

The tool parses a location description that is in a form of a paragraph into a number of sub-paragraphs using a semicolon as a separator mark between two subparagraphs. Our gazetteer treats location descriptions as a single paragraph, each of which commonly uses a semicolon to separate the spatial description from the historic description. Within historic descriptions, the semicolon is

often also used to separate location visits by different expeditioners (see Figure 6). There are, however, inconsistent cases where a comma is used instead. Figure 7 shows the XML document with extracted spatial, temporal and attributive phrases from the location descriptions of Figures 5 and 6.

```
<?xml version="1.0" encoding="utf-8" ?>
<Expedition>
    <LocationEntry>
        <Entry>
        2525/4915 Chrostowski, 3, 8, 18, 26-27 Apr., 6, 10-11, 17, 26 May,
        1, 5, 9, 26, 28 June, 7, 9, 14, 18 July, 25 Sept., 4, 18, 24 Oct., 17, 24 Nov. 1914 ANTÔNIO OLINTO
        </Entry>
    </LocationEntry>

    <LocationEntry>
        <Entry>
        3343/5817 Chrostowski, et al., 18-25 Mar. FOZ DO IGUAÇU
        </Entry>

        <Entry>
        3343/5817 Snethlage, Oct. 1928 FOZ DO IGUAÇU
        </Entry>

        <Entry>
        3343/5817 Kaempfer, 16-21, 28, 30 May, 2 June 1930 FOZ DO IGUAÇU
        </Entry>
    </LocationEntry>

</Expedition>
```

Figure 7. Extracted raw data.

4.2. Spatiotemporal Entities

The expedition gazetteer texts we used for the experimentation of the framework have multiple descriptive dimensions. We focused on extracting the spatial and temporal entities. A combination of the spatial (location of the visit), temporal (timeframe of the visit), and attributive (name of the expeditioner) dimensions gives us the triplets of the expedition route.

4.2.1. Triplets with Crisp Timeframe

The temporal dimension of a triplet that is extracted from a location description with explicit *date*, *month* and *year* is always crisp. A location visit description that mentions a single date is considered as a single day event; hence, both the start and end dates are then the same. On the other hand, location visits with a range of dates, such as "*12–28 March, 14 July–December 1817*", are considered as multiple date events. The first has a crisp timeframe, but the second has not. We use the *crisp-triplet* to represent triplets with crisp timeframes.

4.2.2. Triplets with Vague Timeframe

Unlike triplets that mention crisp temporal entities with explicit *date*, *month*, and *year*, those with a vague timeframe have only the *month* and *year* of location visits mentioned explicitly. For instance, consider a location visit description that has a timeframe of "*January 1922.*" The expeditioner who visited this location could have started and ended the visit at any time between 1 and 31 January 1922, or could have stayed at the site for the whole month. Unless we are provided with additional information regarding this particular visit or other site visits by the same expeditioner within the same timeframe (same month and year), there is no way of telling relative timeframes for the event. However, provided we know other site visits (that have crisp timeframes) by the same expeditioner between 1 and 31 January 1922, we can use these to infer a more precise relative start and end date, better than our default assumption of 1 and 31 January. For instance, if the same expeditioner visited another location Y from 15–25 January, the logical timeframe for the visit at location X must either be from 1–15 or from 25–31 January. Note: In this article, "*vague-triplet*" represents triplets with vague timeframes.

4.3. Text Preprocessor

This preliminary annotator produces temporary annotations of certain classes, namely POS, and precedes the JAPE transducer; the annotations created by the text preprocessor are used as input references by the JAPE transducer. The text preprocessor contains the ANNIE tokenizer, ANNIE splitter and ANNIE POS tagger. All this chunking of paragraphs into sentences, sentences into tokens and tokens into POS categories is performed here. We use this module to detect word and number tokens from the expedition gazetteer. For instance, as Figure 1 shows, a typical expedition gazetteer text has spatial elements described by a combination of number and word tokens (*"Santa Catarina, 2525/4915 (USBGN)"*) and temporal elements described similarly (like *"24 November 1914"*). Using the text preprocessor, we tokenize the gazetteer text into numbers and words, and finally these tokens are used by the JAPE transducer to extract the spatiotemporal information from similar expedition texts.

4.4. Gazetteer Processing Machine (List Matching)

Named entities such as person names, place names and organization names are common in expedition gazetteer texts and are easily confused. Defining a pattern to extract these entities from the text with the JAPE transducer can be ambiguous, because some items may have identical patterns. For instance, both place names and person names are written with initial capitals; the JAPE transducer cannot be explicit enough to tell which is what. The best way to avoid the ambiguity is to use the gazetteer processing machine (list-matching technique) to recognize the named entities, such as place and person names (see Table 1).

Table 1. Entities annotated by the Gazetteer Processing Machine.

No	Example	Annotation Class
1	Paraná	State
2	City of Manacapuru	City
3	USBGN	Organization
4	Chrostowski	Person
5	Feb.	Month

The list-matching process needs input reference datasets—place name, temporal (list of month), organization and person name datasets. We prepared the place name dataset using GeoNames (http://www.geonames.org) consisting of Brazilian place names. Since the experimental dataset mentioned the place names in their Portuguese form, we copied the reference place names from GeoNames written in Portuguese to resolve problems when matching the entities through the gazetteer processing machine. The temporal reference dataset consists of a list of months. The organization and person name datasets consist of a list of organizations and person names that were extracted from the expedition gazetteer, respectively (these datasets were prepared manually from smart pattern searches). The list-matching process checks every token of the expedition gazetteer text on whether it has a match in the reference datasets. If that is the case, that token will be annotated with the matching annotation class. The annotation classes created by this component of the framework along with the tokens from the text preprocessor are used as inputs by the JAPE transducer.

The list-matching process in our framework is fully dependent on the reference dataset. If the framework is to be used for more general information extraction applications, larger datasets—newly created gazetteers—need to be included to update the reference datasets continuously. We acknowledge this as a limitation of the framework when used for other applications.

4.5. JAPE Transducer (Pattern Matching)

Assuming the spatiotemporal entities are mentioned in the expedition gazetteer texts, the patterns of such entities are defined by JAPE grammar rule. Once the patterns of the spatiotemporal items

are defined, the JAPE transducer matches the predefined patterns of entities against the expedition gazetteer text contents. The defined patterns are explicit representations of the possible entities in a text, for instance dates, coordinates, or abbreviations. Such entities can be annotated with their respective classes if the patterns are well-defined. The completeness of the JAPE rules—defining rules for every pattern of the spatiotemporal items in the expedition gazetteer—will affect the performance of the framework in extracting the spatiotemporal entities. To complete our JAPE rules, we defined rules for all spatiotemporal item patterns that we identified (see Table 2). Hence, our framework—the JAPE transducer specifically—has an infinitesimal chance of leaving the spatiotemporal parts of the expedition text unidentified because JAPE rules are defined for most if not all of the spatiotemporal text items. JAPE can be used in combination with the text-processing resource (ANNIE components) to handle the spatiotemporal information extraction task.

Table 2. Entities annotated by the JAPE Transducer.

No	Entity Type	Pattern	Annotation Class
1	Coordinate	9999/9999, ca. 9999/9999, 9999N/9999, ca. 9999/9999? or Place Name 9999/9999	Coordinate
2	Unknown Coordinate	Not located or location?	CoordinateUnknown
3	Date	99–99 Month	DateMonth
4	Date	99 Month–99 Month 9999	DateMonthDuration
5	Date	99 Month–Month 9999	DateMonthMonthDuration
6	Date	99–99 Month 9999	DateMonthYear
7	Date	99 Month 9999–99 Month 9999	DateMonthYearDuration
8	Date	Month 9999	MonthYear
9	Date	Month–Month 9999	MonthDuration
10	Date	99, 99–99 Month, 99, 99–99 Month 9999	DateMonthListYear
11	Date	Month (?) 9999 (?)	DateVague

The main tasks of the JAPE transducer are to annotate the spatial and temporal entities from the gazetteer text. It has two transducers, namely the spatial entity transducer and temporal entity transducer. The spatial entity transducer uses an explicitly defined JAPE rule that is capable of matching coordinate (latitude/longitude) patterns. The typical patterns of a coordinate in the expedition gazetteer are listed in Table 2. Similarly, the temporal entity transducer uses a single-phase JAPE rule to annotate nine different patterns of temporal entities in the expedition gazetteer. This transducer uses the *Month* annotation class created by the gazetteer processing machine and the token categories created by the text preprocessor (ANNIE tokenizer) as inputs to define the LHS parts of the JAPE grammar rule. In the gazetteer texts, the possible patterns for any temporal entity are those listed in Table 2. Figure 8 shows the annotated version of the location descriptions (depicted in Figure 7) extracted from our dataset. The figure shows the annotation classes created on the input text using the gazetteer processing machine and the JAPE transducer.

As Figure 1 shows, expedition gazetteer texts are, most of the time, rich with detailed contents of the spatial, temporal and attributive entities. For instance, place names, dates, months and years are mentioned explicitly often, except in some cases where the spatial and temporal entities are vague and ambiguous to extract. For instance, when vague temporal entities, such as "*January 1921*", are encountered, the JAPE transducer assigns 1 and 31 January as the start and end dates of the visit, and once all the spatiotemporal and other attributive entities are extracted from the gazetteer text, the temporal inference tool will infer the possible relative temporal boundaries considering other visits undertaken by the same expeditioner within the same month and year. The scope of this paper does not address the spatial ambiguity; however, some temporal vagueness is resolved using the tool we developed for a number of temporal inference tasks (see Section 4.7).

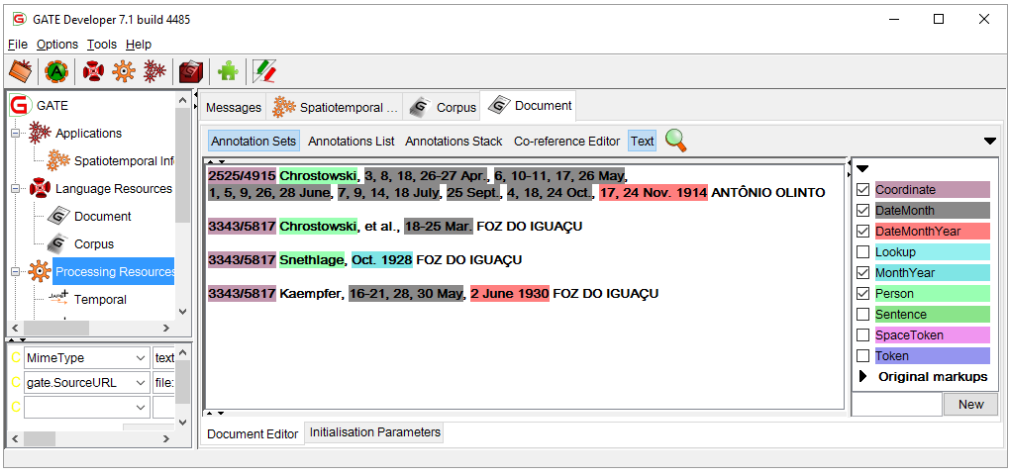

Figure 8. The spatiotemporal entities annotation pipeline.

4.6. Spatial Database

A database was designed and implemented in PostgreSQL. Figure 9 shows part of its data model. The designed database stores the elements of extracted triplets (location, expeditioner and timeframe). JDBC (Java Database Connectivity) was used as a bridge between the spatiotemporal information extraction framework and the database. It enables the automation of extracted triplet storage. It is possible that a single location description mentions visits by multiple expeditioners. This requires a data model that captures the triplets in a separate relation and allows creating a trajectory on demand.

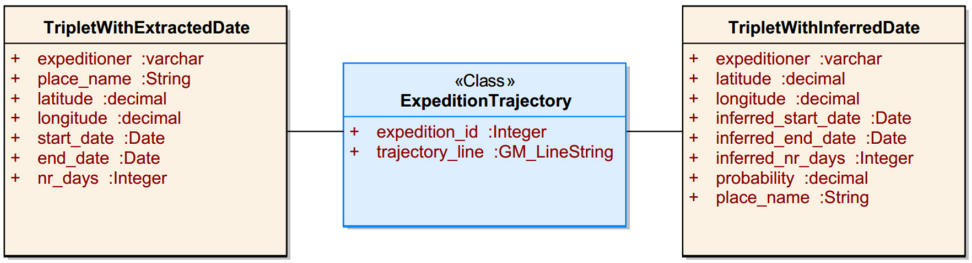

Figure 9. Data model for the extracted expeditions.

4.7. Temporal Inference

In the context of this article, temporal inference is defined as a process of interpolating a relative temporal boundary. The result is a set of temporal scenarios for the extracted vague triplets. The temporal inference process, as depicted in Figure 10, entertains two-way communications with the spatial database to fetch crisp reference triplets and store the inferred ones. This process interpolates alternative timeframes and determines the probability of a given location visit to occur within the inferred timeframes. For instance, assume an expeditioner visited three sites (X, Y and Z) within a month. Assume he visited site X and Y with crisp temporal boundaries of 5–21 January 1921 and 25–31 January 1921, respectively. Additionally, he visited site Z with a vague temporal boundary (January 1921). The third visit must have been started and ended between 1 and 5 January 1921, or 21 and 25 January 1921. However, in the case of our framework, the default start and end dates assigned by the JAPE transducer for the timeframe January 1921 are 1921/01/01 and 1921/01/31, respectively, but after running the temporal inference algorithm (see Figure 11), the start and end dates will be two scenarios, A: (1921/01/01–1921/01/05) and B: (1921/01/21–1921/01/25). However, these inferred triplets can be refined by considering their distance to the crisp visits. The more realistic inferences would be those close to one of the crisp triplets. Moreover, in cases where the inferred triplets are equally distant from the crisp triplets, one can associate a probabilistic value to the inferred triplets.

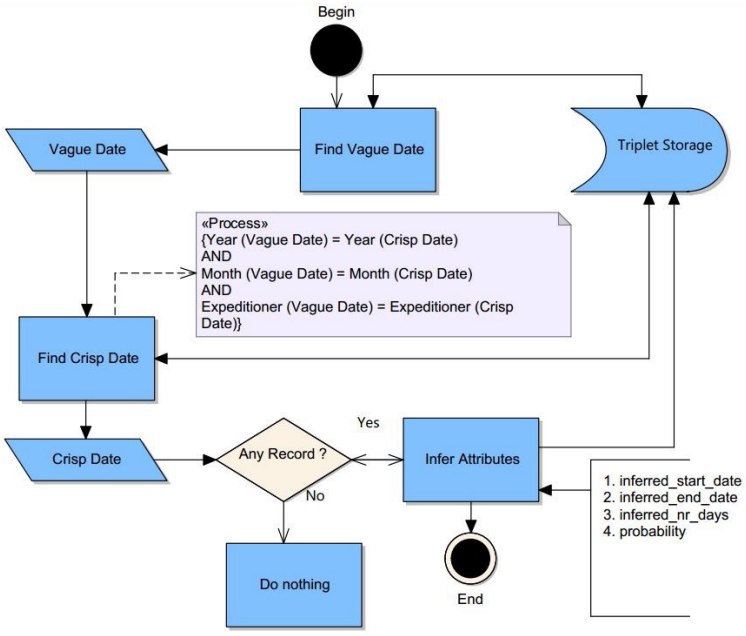

Figure 10. The temporal inference process.

Algorithm: Temporal Inference

Data: CT (triplets with crisp dates) = Set $\{T1, \ldots, Tn\}$ such that $T1, \ldots, Tn = (e, l, sd, ed)$;
where e = *expeditioner*, l = *location*, sd = *start_date* and ed = *end_date*
Data: VT (triplet with vague date) = (e, l, sd, ed);
where e = *expeditioner*, l = *location*, sd = *start_date* and ed = *end_date*
Result: IT (triplet with inferred date) = Set $\{T1, \ldots, Tn\}$ such that
$T1, \ldots, Tn = (e, l, isd, ied)$;
where e = *expeditioner*, l = *location*, isd = *inferred_start_date and*
ied = *inferred_end_date*

1 **begin**
2 *expeditioner* ⟵ *VT.expeditioner;*
3 *location* ⟵ *VT.location;*
4 *vague_triplet* ⟵ *VT;*
5 *crisp_triplets* ⟵ *CT;*
6 *inferred_triplet* ⟵ *NULL;*
7 **for** $c \in$ *crisp _triplets* **do**
8 **if** *c.sd.month* = *vague_triplet.sd.month* **and**
c.sd.year = *vague_triplet.sd.year* **then**
9 $inferred\_triplet_1.isd \longleftarrow vague\_triplet.sd;$
10 $inferred\_triplet_1.ied \longleftarrow c.sd;$
11 $inferred\_triplet_2.isd \longleftarrow c.ed;$
12 $inferred\_triplet_2.ied \longleftarrow vague\_triplet.ed;$
13 $probability(visit, inferred\_time) = \frac{1}{number\ of\ inferred\ temporal\ boundaries}$
14 **end**
// Add all the inferred start date and end date along with the
expeditioner name and location to the IT set.
15 **end**
16 **end**

Figure 11. The temporal inference algorithm.

Assuming there are chronologically close crisp triplets for a given vague triplet, the temporal inference tool interpolates relative temporal boundaries. This process has three basic steps (see Figure 11), and is discussed below. Note that the last day of the specific month analyzed must be taken into consideration while conducting the inference process. For instance, the 31st was taken as the last day of the month for the illustration below. If a reference crisp triplet does not exist for a given

vague triplet, the inference process may not be successful and the default vague temporal boundary remains as only option.

Data: The extracted vague and crisp triplets (of the same expeditioner) upon which the relative temporal boundaries for the vague triplets are inferred.
Process: The inference process discussed here is applicable only for the vague triplets which timeframes are captured with the *MonthYear* annotation class (see Table 2) by the JAPE transducer.
Result: The result of this algorithm is a set of triplets with inferred temporal boundaries. The triplets with the inferred temporal boundaries are stored in the database.

Step 1: Finds a parsed and stored vague triplet.
Step 2: Finds crisp triplets; the crisp dates are constrained to be about the same expeditioner, same month and same year as the vague triplet in ***Step 1***.
Step 3: Infers relative temporal boundaries and determines their probability of occurrence for those vague triplets in ***Step 1*** relative to those crisp triplets in ***Step 2***.
Line 8–14 (see Figure 11): Let the vague triplet be VT and the crisp triplet be CT. If the month and year of the VT and CT are similar, for every given VT, a minimum of one or maximum of two temporal boundaries are inferred. If the given CT starts at the first day of the month or ends at the last day of the month, only one temporal boundary is inferred. If the given CT starts and ends between the first and last days of the month, two temporal boundaries are inferred. Given the VT and CT, the following holds (see Figure 12).

1. *If start_date of _CT > 1* and *end_date of _CT < 31*
inferred_start_date of VT_1 *= start_date of VT*
inferred_end_date of VT_1 *= start_date of CT*
inferred_start_date of VT_2 *= end_date of CT*
inferred_end_date of VT_2 *= end_date of VT*
2. *If start_date of _CT > 1* and *end_date of _CT = 31*
inferred_start_date of VT_1 *= start_date of VT*
inferred_end_date of VT_1 *= start_date of CT*
3. *If start_date of _CT = 1* and *end_date of _CT < 31*
inferred_start_date of VT_1 *= end_date of CT*
inferred_end_date of VT_1 *= end_date of VT*

Figure 12. Constraints of the temporal inference algorithm.

4.8. Expedition Route Production

After all the triplets of a given expeditioner are extracted and stored, a process follows to produce a trajectory that depicts expedition routes. The extracted triplets of a given expeditioner are grouped into a number of expeditions. The grouping depends on the detection of temporal gaps between location visits. Our framework has three methods to handle the expedition trajectory production. The first method finds boundary triplets between two expeditions of a given expeditioner based on a predefined temporal gap. Given such boundaries, the second and third methods produce the expedition trajectory. Here, assigning the temporal gap is subject to a specific use of the framework. For instance, the temporal gap could be 60 days, assuming that the expeditioners back in the 1900s would have to stock up and prepare before heading out for a next expedition.

5. Results and Discussion

We discussed that our experimental dataset, the Brazilian Ornithological Gazetteer, consists of described named places that featured in the historic expeditions of many expeditioners. Tadeusz Chrostowski (1878–1923), Emil Heinrich Snethlage (1897–1939) and Maria Emilie Snethlage (1868–1929) were among these. We used our framework to extract the spatiotemporal information from the expedition gazetteer texts for these expeditioners.

5.1. Expeditioner: Tadeusz Chrostowski

Tadeusz Chrostowski (1878–1923) is one of the expeditioners mentioned in the Brazilian Ornithological Gazetteer. According to Wikipedia, he conducted three expeditions in Brazil during the period 1910–1923. His first expedition took place in the year 1910 along the River Iguaçu after which he returned to Poland in 1911; his second expedition ran from 1913 to 1915, and then he returned to Poland in 1915, due to the news of the outbreak of World War I. In [26], it is mentioned that Chrostowski conducted his third expedition from 1921 to 1923. However, after extracting the spatiotemporal information from the 58 entries where his name has been mentioned, we were able to produce six expedition routes with a *temporal gap* of two months between two consecutive expeditions (see Table 3). As the Table shows, Expeditions II, III and IV and Expeditions V and VI are close to each other as measured in time. Based on this closeness, we suggest the following aggregations to arrive at three expedition routes only.

Case 1: Looking at the expeditions in Table 3, it can be observed that Expedition I is far from the other expeditions as measured in time. The gap between end date of the first expedition and start date of the second is more than two years: from 26 August 1911 to 22 January 1914; it is not likely that a single expedition went on so long. This gives us a reason to keep Expedition I as it was derived.

Case 2: The temporal gap between Expeditions II and III is about two months (13 July 1914 to 25 September 1914) and the temporal gap between Expeditions III and IV is also about two months (2 December 1914 to 10 February 1915). Moreover, the number of location visits in Expeditions III and IV is fewer than in Expedition II. A typical expedition is expected to involve a considerable number of location visits. Hence, it might be less convincing that Expedition IV, which has only one location visit, is actually an expedition. It can be argued that expedition teams in those days needed time to stock up again before continuing fieldwork. As a result of the stocking-up days, one expedition route might be produced as two in our framework. This gives us a concrete reason to aggregate Expeditions II, III and IV. Considering the number of location visits and the temporal gap among these expeditions, they could be conveniently produced as one if the temporal gap were set to two or three months. The closeness in time among these three expeditions leads us to an expedition route aggregation. The aggregated expedition route, therefore, merges Expeditions II, III and IV and it covers the period from 22 January 1914 to 10 February 1915.

Case 3: Expeditions V and VI are chronologically apart from each other by four months; the temporal gap is from 31 August 1921 to 1 January 1922. Expedition V has only one triplet while Expedition VI has 38. It may not be feasible to keep Expedition V while the number of triplets is one; instead the triplets of both routes can be aggregated to produce one. The aggregation gives us an expedition that covers the time period from 1 August 1921 to 5 May 1923. However, expeditions with just one triplet could have appeared as such due to temporal mislabeling. In such a case, further consistency checking required.

Table 3. Expeditions of Chrostowski as produced by our framework.

Expedition	Start Date	End Date	No. Triplets
I	26 May 1910	26 August 1911	64
II	22 January 1914	3 July 1914	25
III	25 September 1914	2 December 1914	7
IV	10 February 1915	10 October 1915	1
V	1 August 1921	31 August 1921	1
VI	1 January 1922	5 May 1923	38

The Third Expedition of Chrostowski: 1921–1923

Straube and Urben-Filho mention that Chrostowski's third expedition was carried out from 1921–1923 [26]. We used the expedition gazetteer texts that mentioned his name to experiment on our spatiotemporal information extraction framework. Accordingly, we managed to extract the spatial and temporal elements from the given text and created spatial features represented by the point data type

using the extracted spatiotemporal information, and then we mapped the trajectory of the expedition route by connecting the extracted points chronologically. According to the extracted information, this expedition consists of 39 (see Table 4) site visits, out of which only 35 were identified as distinct site visits, meaning the remaining four visits were made to identical locations at different times (see the purple, blue and green colored records in Table 4), or these visits might have been mislabeled.

Table 4. Extracted spatiotemporal information (Chrostowski's third expedition, 1921–1923).

No	Place Name	Lat	Lon	Start Date	End Date	No Dates
1	FAZENDA FERREIRA	26.01	51.36	1922-03-12	1922-03-28	17
2	UBA, SALTO	24.3	51.28	1922-11-18		1
3	SALVADOR	12.59	38.31	1921-08-01	1921-08-31	31
4	CORONEL QUEIROZ	25.22	52.1	1923-05-05	1923-07-04	60
5	CARA PINTADA	24.88	51.26	1922-05-15	1922-06-04	20
6	CONCORDIA,RIO	25.43	51.17	1922-03-01	1922-03-12	12
7	COBRE, SALTO DO	23.53	51.53	1922-12-11	1922-12-19	9
8	SAO DOMINGOS	25.43	51.17	1922-02-15	1922-02-28	14
9	APUCARANA	24.47	51.1	1922-08-01	1922-08-31	31
10	PARY, CORREDEIRA DO	23.38	52.19	1923-01-04	1923-01-06	3
11	PINHEIRINHO	25.25	53.55	1923-03-28	1923-04-23	26
12	MUTUM, ILHA DO	23.15	53.43	1923-01-14	1923-01-15	2
13	FAZENDA WISNIEWSKY	26.03	50.38	1922-02-01	1922-02-28	28
14	MANGUINHOS	22.47	41.56	1922-01-01	1922-01-31	31
15	GUARAPUAVA	25.23	51.27	1922-04-28	1922-05-14	17
16	BOM, RIO	23.56	51.44	1922-12-20	1922-12-22	3
17	AFONSO PENA	25.32	49.06	1923-01-25		1
18	MALLET	25.55	50.5	1922-01-10	1922-02-02	23
19	FENIX	23.54	51.57	1922-12-23	1923-01-02	10
20	FERRO, CORREDEIRA DO	23.12	52.54	1923-01-07	1923-01-13	7
21	BANANEIRAS,SALTO DAS	23.4	52.13	1923-01-03		1
22	FOZODO IUACU	25.33	54.35	1923-03-18	1923-03-25	8
23	FAZENDA ZAWADSKI	25.43	51.17	1922-02-15	1922-02-28	14
24	FAZENDA FIRMIANO	26	50.32	1922-03-01	1922-03-12	12
25	PINDAHURA, CACHOEIRA DE	24.08	51.31	1922-11-28	1922-12-06	9
26	UBZINHO, RIO	24.35	51.2	1922-10-12	1922-11-20	39
27	AREIA, RIO DA	26.01	51.36	1922-03-29	1922-04-12	14
28	ARIRANHA, CACHOEIRA	24.22	51.27	1922-11-23	1922-11-26	4
29	VERMELHO	24.61	51.26	1922-06-06	1922-07-05	30
30	BANHADOS	25.3	51	1922-04-13	1922-04-17	5
31	PORTO XAVIER DA SILVA	23.25	53.47	1923-01-15	1923-01-17	3
32	CANDIDO DE ABREU	24.35	51.2	1922-08-02	1922-10-11	70
33	SETE QUEDAS, SALTO DAS	24.02	54.16	1923-01-23	1923-02-26	34
34	CONCORDIA RIO	24.42	51.24	1922-03-01	1922-03-12	12
35	TERESA CRISTINA	24.48	51.07	1922-07-08	1922-07-31	24
36	RIO DE JANEIRO	22.54	43.14	1922-01-01	1922-01-31	31
37	CLARO,RIO	25.55	50.74	1922-02-03	1922-02-14	12
38	PORTO MENDES	24.3	54.2	1923-02-27	1923-03-16	20
39	URA, SALTO	24.3	51.28	1922-11-14		1

To assess the reliability of the extracted spatiotemporal information, we developed a simple algorithm that calculates the distance between two chronologically consecutive site visits and compares the result with an average distance and the *temporal gap* we predefined; here it is worth noting that setting the average distance and the *temporal gap* was subjected to the assumption made by us that people back in the 1920s could travel 30 km in a day. Then we conducted the assessment setting the average travel distance to 30 km a day, meaning if the distance between two site visits is more than 30 km and the *temporal gap* is less than a day, then the route is considered unreliable. After running the algorithm on the 35 distinct site visits, the routes through records 33 (light blue–colored), 17 (orange-colored) and 38 (light blue–colored) of Table 4 came up as unreliable. At this point, a manual intervention was necessary to investigate the unreliability. Hereafter, let records 33 (light blue–colored) be point A, 17 (orange-colored) point B and 38 (light blue–colored) point C for the sake of simplicity.

The distance measured as the crow flies from point A to B is about 535 km whereas the distance from point B to C is another 531 km. The expeditioner visited these points as follows: point A (1923/01/23–1923/02/26), point B (1923/01/25) and point C (1923/02/27–1923/03/16) (see Figure 13). Assuming the distance that could be covered is 30 km a day, the expeditioner must have traveled for 18 days from point A to B and another 18 days from point B to C. However, the story we see in the extracted information is that the expeditioner traveled through these points in a single day, which is very unlikely to have happened back in 1923, assuming that expeditioners back in those days traveled on horseback. Such inconsistency may be related to mislabeling or extracting wrong information. Considering these, the followings are possible scenarios.

Scenario one: The extracted information might actually be of the right expeditioner, but the description could be of another expedition, for instance instead of 1923/01/25, the visit date might have been 1921/01/25.

Scenario two: The extracted information might be of the right time and expeditioner, but the problem could be the extracted location. In this case, if point B were near points A and C, we could have believed that the route via these points is reliable.

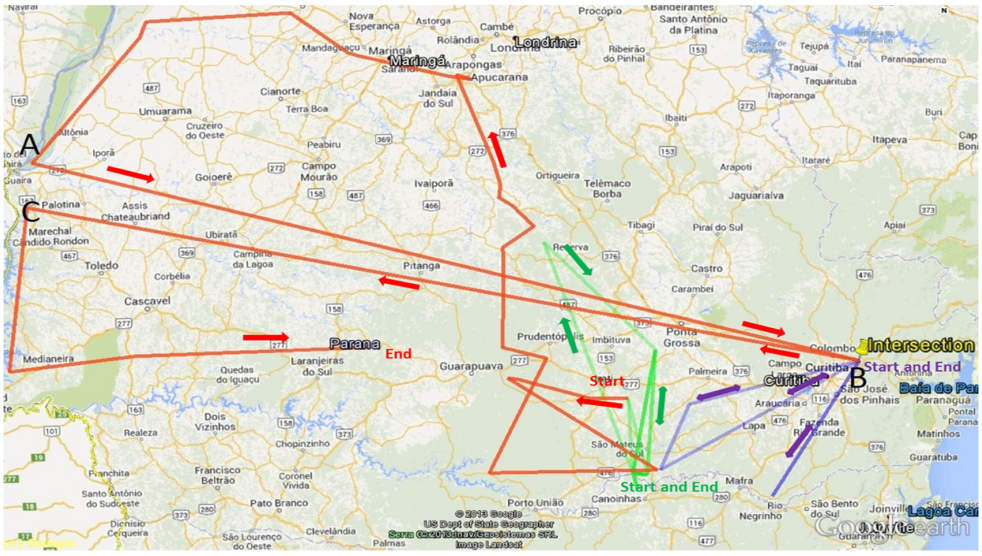

Figure 13. Expeditions of Chrostowski as produced by our framework: Expedition I (Green line), Expedition II (Blue line) and Expedition III (Red line). See the external dataset to refer the details of the expeditions (refer to the KML files in "analyzed" and "Chrostowski" folders of the Supplementary file).

If the unreliability of the route under assessment has the cause of *scenario one*, the right solution seems to check if the other expedition route has an attributive, temporal and spatial intersection at the outlier point (point B) of the assessed expedition. Figure 13 shows the three expedition routes of Chrostowski; the same figure shows the spatial intersection between Expedition III (red line) and Expedition II (blue line). Looking at the intersection points, we can infer that these two points could be extracted from an identical location description and they share an identical location. Hence, the reason that point B of the assessed expedition route is an outlier must be either due to extracting the triplets from the wrong location description or having a wrongly written description. To support this claim, we have to look at the description from which the triplets are extracted. The paragraph below is the same visit description from which the information is extracted. According to this description, the outlier point is extracted correctly; "*25 January 1923*" is, of course, there. In the same description there is a phrase that reads "*although it was not mentioned by Chrostowski*"; here we have to be suspicious about the credibility of "*25 January 1923*". Therefore, the author of this description might have made an attributive misreporting. Assuming point B was completely an outlier and may belong to another

expedition (Chrostowski's Expedition II in this case), we excluded it from the assessed expedition route, and we modified the expedition route by connecting points A and C. Figure 14 shows this modified expedition route.

"Ca. 900 m, on S side of Rio Iguassu [Rio Iguaçu, 2536/5436 (USBGN)], ca. 12 km SE of Curitiba [2525/4915 (USBGN)], Chrostowski, 22, 31 January, 11, 14, 19–20, 22 February, 15 March 1914, 10 February 1915[?], 25 January 1923 (Chrostowski, 1921:31–34, as "Affonso Penna"; 1922, Ann. Zool. Mus. Polonici Hist. Nat., 1:400, as "Affonso Penna"; Sztolcman, 1926:119); description places this near São José dos Pinhais [2531/4913 (USBGN)], although it was not mentioned by Chrostowski."

Figure 14 shows two expedition routes; the red line shows the route connected by straight lines passing through each point, and the blue line shows the same route connected by the road network (for convenience sake, we used the present-day way-finding tool of Google Earth) passing through each point. The expedition route depicted by the blue line is considered as a reasonable representation of the third expedition conducted by Chrostowski during the period 1921–1923. We compared this route map visually—manual intervention was necessary at this point—with a reference map of the same expedition that was prepared manually by one of Chrostowski's friends, Jaczewski, in 1925. The objective of this visual comparison is to confirm whether the framework is reliable and the spatiotemporal information is extracted correctly. The centers of this comparison are geometrical and spatial situation similarities between the route we produced and the reference route. Figure 15 shows the reference route; we colored the original reference route as blue to enhance its visibility and make the visual inspection easy, and Figure 14 shows the expedition route we extracted from the expedition gazetteer text. The figures show that the expedition routes in both cases are geometrically and spatially similar. The resemblance of these routes gave us a compelling reason to believe that our framework is reliable and can be used to extract spatiotemporal information from similar expedition gazetteer texts. Note: while reviewing the biography of the expeditioner (Chrostowski), we discovered a surprising fact. According to [27], Chrostowski died on 4 April 1923. However, the spatiotemporal information we extracted from his expedition gazetteer text shows that his last visit was conducted from 5 May 1923 to 4 July 1923 (see Table 4, record 4), which contradicts the fact that he died prior to this very visit. The only possible explanation to this contradiction is misreporting the site visit, which might have happened while the gazetteer was written.

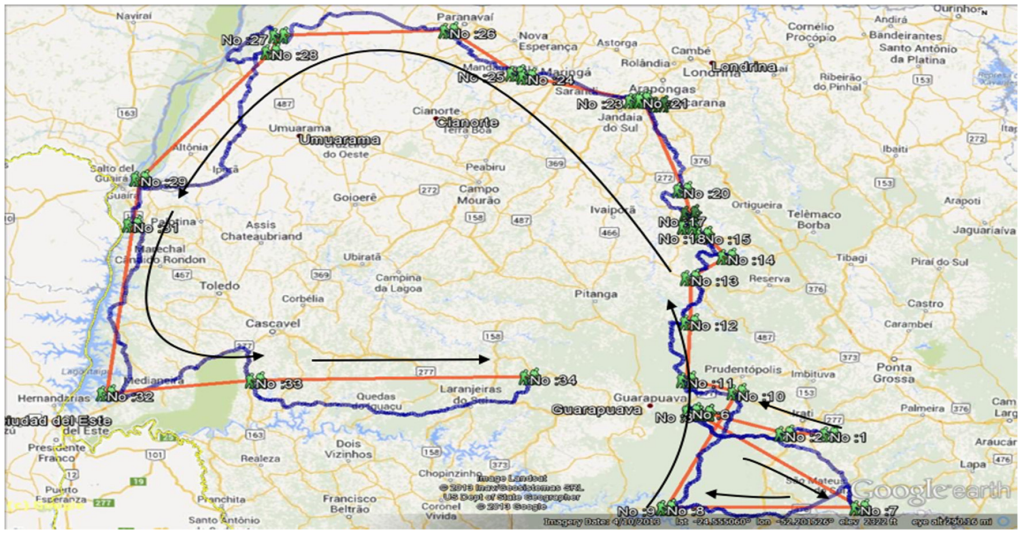

Figure 14. Straight line (red line) and road network (blue line) extracted expedition routes. The red route is created by connecting the point through a straight line, and the blue route is created by connecting the same points through the road network we managed to obtain from Google Earth. See the external dataset to refer the details of the expeditions (refer to the KML files in "analyzed" and "Chrostowski" folders of the Supplementary file).

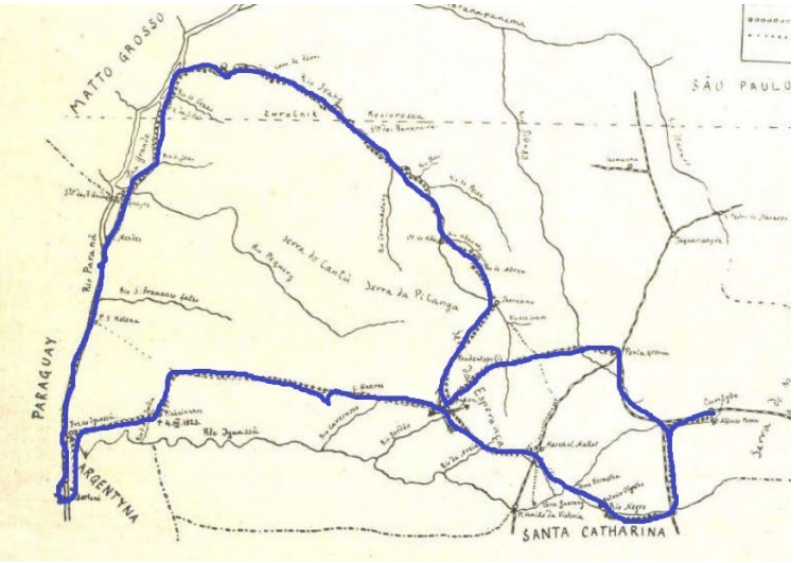

Figure 15. Reference expedition route (Jaczewski, 1925 [26]). See the external dataset to refer the details of the expedition (refer to the KML files in "analyzed" and "Chrostowski" folders of the Supplementary file).

5.2. Expeditioner: Emil Heinrich Snethlage and Maria Emilie Snethlage

Maria Emilie Snethlage (1868–1929) was an ornithologist who undertook many expeditions from 1905 until her death. Emil Heinrich Snethlage (1897–1939) was a zoologist and ethnologist. He was a nephew of Maria who was influenced by his aunt's work, which inspired him to travel with her. According to [28], Emil conducted his first expedition in the years 1923–1926 in association with his aunt. He started his expedition in Maranhão state, northeastern Brazil. However, the two Snethlages did not travel the whole period together; from March 1924 to 1926 Emil traveled alone [28]. This brings an interesting assumption that Emil and Maria must have had expedition routes that intersected someplace in the years 1923/1924; since Emil traveled alone after March 1924 to 1926, the likely place to have met his aunt must have been one of the places he visited in the years 1923/1924. Based on the extracted information from their respective location descriptions, Maria visited the Rio de Janeiro, Bahia, Minas Gerais and Espírito Santo states in the years 1923–1926 while Emil traveled to places in the Maranhão, Ceara and Piaui states.

One of the location visit descriptions (see Figure 16) mentioned that Maria and Emil visited a place around the "*coast of northwestern Maranhão*" in the years 1923/1924. The same description gives a clue about the possible location that might be an intersection point between the expedition routes of these expeditioners. Another location description (see Figure 17) mentioned that Emil visited the "*northwest of Maranhão*" in the years 1923/1924. The attributive intersection between these two descriptions gives us a preliminary proof that the "*northwest of Maranhão*" is the place these expeditioners could have traveled together. Having this assumption, we produced the expedition routes of these expeditioners.

"Sea level, on coast of northwestern Maranhão, on left side of mouth of Rio Turiaçu [0136/4519 (USBGN)] (ICWB); E. Snethlage, latter half of 1923, H. Snethlage, 3, 5-6, 8-10, 12-13, 15-19, 23-27, 29-31 Oct., 3,5-9,12-17,19-23,26-30 Nov., 4-7, 10-15, 17, 19-20, 23, 27, 29 Dec. 1923, 5 Jan. 1924 (Snethlage, 1927d:58, as "Tury-assú"; 1936:87, 88, 91, as "Tury-assú", Hellmayr, 1929a:238ff, as "Tury-assú"; Novaes, 1947:246, as "Turiassú"; Ruschi, 1951b:2, as, "Turry-assu," pp. 5ff, as "Tury-assu"); collector?, 17 Feb. 1945 (Ruschi, 1949b:7); SIlva and Oren, sometime between 1980-1988 (Silva Oren, 1990:311); Pinto, 1944a:296, as "Turiassu.""

Figure 16. E.H. Snethlage's visit description.

"northwestern Maranhão, SW of Turiaçu [0141/4521 (USBGN)], 40 km inland, H. Snethlage, 7-10, 13-17, 19-22 Nov., 6 Dec. 1923,4 Jan. 1924 (Hellmayr, 1929a:map, 239, 318, 319, 333, 343, 355, 363, 370, 380, 416, 419, 450; Snethlage, 1936:85, as "Alto da Alegria, Turiassu"; Ruschi, 1951b:5, as "Tury-assú, Alto da Alegria"); correct spelling unknown; apparently not the Alto Alegre farther S at 0307/4550 (USBGN)."

Figure 17. M.E. Snethlage's visit description.

The produced expedition routes are visually analyzed to support our assumption on the attributive, temporal and spatial intersection of the expeditions. Figure 18 shows the expedition routes of Emil and Maria, as well as the possible intersection point of these expedition routes. The expedition route of Maria runs from the southeast of Brazil (*Rio de Janeiro)* to the northeast of Brazil *(Maranhão)* and intersects with the expedition route of Emil and gets back to the southeast of Brazil *(Espírito Santo)*. Here, we can think of two scenarios about the two expedition routes intersecting at the starting location visit of Emil's expedition:

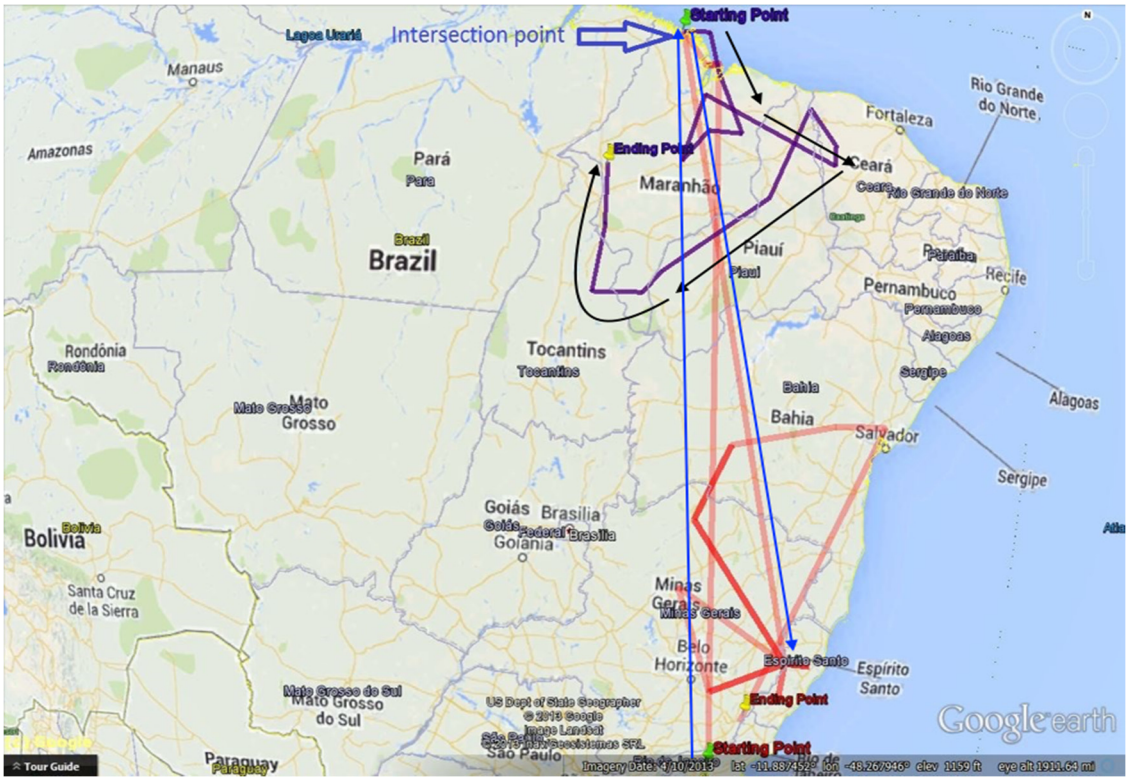

Figure 18. Expeditions of E.H. Snethlage (purple line); and M.E. Snethlage (red line), 1923–1926. See the external dataset to refer the details of the expeditions (refer to the KML files in "analyzed" and "Snethlage" folders of the Supplementary file).

Scenario one: These two expeditioners visited the intersection point at different times, and Figure 18 only shows an intersection in space without temporal proof that the expeditioners met at that intersection point.

Scenario two: The expeditioners actually met at the intersecting point. Therefore, to assess validity of the two scenarios, we need to check if their visits share a similar timeframe at the intersecting location. Given the fact that these expeditioners met in 1923/1924 and Emil traveled alone after March 1924 [28], we need at least two triplets at the intersecting point, one from each expedition route, with either identical or close temporal marks.

The visit dates shown in Figures 16 and 17 are chronologically close, which is a sound reason to believe that Maria and Emil actually met in the "*northwest of Maranhão*". Since the intersection between the expedition routes is in space and time, our story about the two expeditioners meeting in the "*northwest of Maranhão*" in the years 1923/1924 is true and supported by a fact.

6. Conclusions

In this article, we presented a semi-automated framework to extract spatiotemporal information from a historic expedition gazetteer. The approach was implemented and experimented on a sample dataset acquired from the Brazilian Ornithological Gazetteer. Mainly, we used the pattern-based JAPE rules and the gazetteer list-matching processes for the annotation and information extraction tasks. If patterns of entities in a text (for instance, patterns of dates and coordinates) or the list of reference items (such as months, place names and person names) to match with entities in the text are not explicitly prepared, the information extraction process may not be successful. However, defining patterns and listing reference items for all entities that we could find in the expedition gazetteer texts is not ideal. Our framework is totally dependent on the pattern-based JAPE rules and gazetteer list-matching. As a result, some items may not be annotated and extracted if the patterns are not defined and the list of items for reference is not prepared in advance. We also included a temporal inference tool in the framework to suggest relative temporal boundaries for vague triplets extracted from the expedition gazetteer texts. However, this tool is dependent on the availability of crisp triplets that can be used as references to compute the relative temporal boundaries. The framework was tested on three datasets from the Brazilian Ornithological Gazetteer. Expedition route maps were produced using the spatiotemporal information extracted from these datasets. Manual intervention was necessary during the route map production processes and reliability assessment tasks. One of the produced expedition route maps was manually compared with a previously produced map—this reference map was drawn in 1925 by Jaczewski [26]—to assess the reliability of the framework. To improve the information extraction quality of our framework, we suggest incorporating a semantic spatial annotation module, a spatial inference module, spatiotemporal consistency checking module, and spatial ontology to the framework.

Spatial ontology represents spatial knowledge as a hierarchy of concepts within a defined spatial context, using a shared vocabulary and place names to denote the types, properties and spatial interrelationships of those concepts. The semantic spatial annotation approach will use a spatial knowledge base and graph database rather than a database or items dataset, because the basic input for the semantic annotation is a representation of spatial associations among items, and databases are not capable of representing this knowledge. Semantic spatial annotations can be integrated with the spatial inference module to interpret spatial relationships such as "*40 km south of Maranhão*". Parts of the expedition gazetteer text (we have seen the entities in this article) that depict location are not definitive all the time; there are cases where these entities are vague, semantic and relative, for instance the spatial phrase "*on the coast of northwestern Maranhão*" is vague and annotating this location requires knowing the context of the phrase. Hence, the semantic spatial annotation approach will include spatial machine learning, spatial ontology, a spatial graph database and spatial indexing as basic features. We suggest designing spatial machine learning, a spatial inference tool and a spatial graph database and fusing them with the pattern-based and gazetteer-based (our approach) framework, which will improve the information extraction quality and the spatiotemporal disambiguation tasks.

If semantic spatial annotation, along with the spatial ontology, is included in the framework, spatiotemporal disambiguation, temporal inference and information extraction tasks will be easy, and the framework could be used to extract spatiotemporal information from any expedition gazetteer texts without changing the framework much. Therefore, we recommend further research to focus on semantic spatial annotation and spatial ontologies.

Supplementary Materials: The following are available online at www.mdpi.com/2220-9964/5/12/221/s1.

Acknowledgments: The first author would like to thank the University of Twente for provision of the data during his master's thesis work, and Debebe Ayele Dessalegn for preparing the extracted information table.

Author Contributions: All of the contents of this article are the contributions of the authors. Some portions of this article have been presented in the Mafkereseb Kassahun Bekele's thesis (Spatial tracing of historic expeditions: from text to trajectory) and Gaurav Singh's PhD thesis (From location description to map: Understanding VGI from the past) [29]. Rolf de By conceived the historic gazetteer project at ITC, developed the baseline data for its execution, suggested the research work that this paper presents, and supervised its execution.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Bekele, M.K. Spatial tracing of historic expeditions: From text to trajectory. Master's Thesis, University of Twente, Enschede, The Netherlands, 2014.
2. Paynter, R.A.; Traylor, M.A. *V.1—Ornithological Gazetteer of Brazil*; Harvard University: Cambridge, MA, USA, 1991.
3. Paynter, R.A.; Traylor, M.A. *V.2—Ornithological Gazetteer of Brazil*; Harvard University: Cambridge, MA, USA, 1991.
4. Subodh, V.; Christopher, B.J.; Hideo, J.; Mark, S. Spatio-textual indexing for geographical search on the web. In *Advances in Spatial and Temporal Databases*; Springer: Heidelberg, Germany, 2005.
5. McCurley, K.S. Geospatial mapping and navigation of the web. In Proceedings of the 10th International Conference on World Wide Web, Hong Kong, China, 1–5 May 2001.
6. Chen, Y.Y.; Suel, T.; Markowetz, A. Efficient query processing in geographic web search engines. In Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data, Chicago, IL, USA, 27–29 June 2006.
7. Wu, D.; Cong, G.; Jensen, C.S. A framework for efficient spatial web object retrieval. *Int. J. Very Large Data Bases* **2012**, *21*, 797–822. [CrossRef]
8. Freire, N.; Jos, Y.; Borbinha, Z.; Calado, V.; Martins, B. A metadata geoparsing system for place name recognition and resolution in metadata records. In Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries, Ottawa, ON, Canada, 13–17 June 2011.
9. Gelernter, J.; Balaji, S. An algorithm for local geoparsing of microtext. *Geoinformatica* **2013**, *17*, 1–33. [CrossRef]
10. Guillén, R. Geoparsing web queries. In *Advances in Multilingual and Multimodal Information Retrieval*; Peters, C., Jijkoun, V., Mandl, T., Müller, H., Oard, D., Peñas, A., Petras, V., Santos, D., Eds.; Springer: Berlin, Germany, 2008; Volume 5152, pp. 781–785.
11. Witmer, J.; Kalita, J. Extracting geospatial entities from Wikipedia. In Proceedings of the IEEE International Conference of Semantic Computing, Berkeley, CA, USA, 14–16 September 2009.
12. Zubizarreta, Á.; Fuente, P.; Cantera, J.; Arias, M.; Cabrero, J.; García, G.; Llamas, C.; Vegas, J. Extracting geographic context from the web: Georeferencing in MyMoSe. In *Advances in Information Retrieval*; Boughanem, M., Berrut, C., Mothe, J., Soule-Dupuy, C., Eds.; Springer: Berlin, Germany, 2009; Volume 5478, pp. 554–561.
13. Boguraev, B.; Ando, R.K. TimeML-compliant text analysis for temporal reasoning. In Proceedings of the 19th International Joint Conference on Artificial intelligence, Madrid, Spain, 30 July–5 August 2005; pp. 997–1003.
14. Zhang, C.; Zhang, X.; Jiang, W.; Shen, Q.; Zhang, S. Rule-based extraction of spatial relations in natural language text. In Proceedings of the International Conference on Computational Intelligence and Software Engineering, Wuhan, China, 11–13 December 2009.
15. Blessing, A.; Schütze, H. Self-annotation for fine-grained geospatial relation extraction. In Proceedings of the 23rd International Conference on Computational Linguistics, Beijing, China, 23–27 August 2010.
16. Strötgen, J.; Gertz, M.; Popov, P. Extraction and exploration of spatio-temporal information in documents. In Proceedings of the 6th Workshop on Geographic Information Retrieval, Toronto, ON, Canada, 30 October 2010.
17. Brisaboa, N.; Luaces, M.; Places, Á.; Seco, D. Exploiting geographic references of documents in a geographical information retrieval system using an ontology-based index. *Geoinformatica* **2010**, *14*, 307–331. [CrossRef]
18. Drymonas, E.; Pfoser, D. Geospatial route extraction from texts. In Proceedings of the 1st ACM SIGSPATIAL International Workshop on Data Mining for Geoinformatics, San Jose, CA, USA, 3–5 November 2010.

19. Hall, M.M.; Smart, P.D.; Jones, C.B. Interpreting spatial language in image captions. *Cognit. Process.* **2011**, *12*, 67–94. [CrossRef] [PubMed]
20. Horák, J.; Belaj, P.; Ivan, I.; Nemec, P.; Ardielli, J.; Růžička, J. Geoparsing of Czech RSS news and evaluation of its spatial distribution. In *Semantic Methods for Knowledge Management and Communication*; Katarzyniak, R., Chiu, T.-F., Hong, C.F., Nguyen, N., Eds.; Springer: Berlin, Germany, 2011; Volume 381, pp. 353–367.
21. Abascal-Mena, R.; López-Ornelas, E. Geo-information extraction and processing from travel narratives. In Proceedings of the 14th International Conference on Electronic Publishing, Helsinki, Finland, 16–18 June 2010.
22. Godo, L.; Vila, L. Possibilistic temporal reasoning based on fuzzy temporal constraints. In Proceedings of the 14th International Joint Conference on Artificial Intelligence, Montreal, QC, Canada, 20–25 August 1995.
23. Freksa, C. Temporal reasoning based on semi-intervals. *Artif. Intell.* **1992**, *54*, 199–227. [CrossRef]
24. Nagypál, G.; Motik, B. A fuzzy model for representing uncertain, subjective, and vague temporal knowledge in ontologies. In *On the Move to Meaningful Internet Systems 2003: Coopis, Doa, and Odbase*; Springer: Berlin, Germany, 2003; pp. 906–923.
25. Cunningham, H.; Maynard, D.; Bontcheva, K.; Tablan, V.; Ursu, C.; Dimitrov, M.; Dowman, M.; Aswani, N.; Roberts, I.; Li, Y. *Developing Language Processing Components with Gate Version 5: (A User Guide)*; University of Sheffield: South Yorkshire, UK, 2009.
26. Dicionário Geográfico das Expedições Zoológicas Polonesas ao Paraná. Available online: http://www.ao.com.br/download/polonesa.pdf (accessed on 25 March 2016).
27. Straube, F.; Urben-Filho, A. Tadeusz Chrostowski (1878–1923): Biografia e perfil do patrono da ornitologia paranaense. *Bol. Inst. Hist. Geogr. Paraná* **2002**, *52*, 35–52.
28. Life, Expeditions, Collections and Unpublished Field Notes of Dr. Emil Heinrich Snethlage. Available online: http://www.snethlage.info/ (accessed on 20 September 2015).
29. Singh, G. From Location Description to Map: Understanding VGI from the Past. Ph.D. Thesis, University of Twente, Enschede, The Netherlands, 2014.