

Spatiotemporal Mapping of Wikipedia Concepts

Adrian Popescu
Institut Télécom/TELECOM Bretagne
Technopôle Brest-Iroise
29238 Plouzané
+330229001435

adrian.popescu@telecom-bretagne.eu

Gregory Grefenstette
Exalead
10 Place de la Madeleine
75008 Paris
+330155352766

gregory.grefenstette@exalead.com

ABSTRACT

Space and time are important dimensions in the representation of a large number of concepts. However there exists no available resource that provides spatiotemporal mappings of generic concepts. Here we present a link-analysis based method for extracting the main locations and periods associated to all Wikipedia concepts. Relevant locations are selected from a set of geotagged articles, while relevant periods are discovered using a list of people with associated life periods. We analyze article versions over multiple languages and consider the strength of a spatial/temporal reference to be proportional to the number of languages in which it appears. To illustrate the utility of the spatiotemporal mapping of Wikipedia concepts, we present an analysis of cultural interactions and a temporal analysis of two domains. The Wikipedia mapping can also be used to perform rich spatiotemporal document indexing by extracting implicit spatial and temporal references from texts.

Categories and Subject Descriptors

H.m [MISCELLANEOUS]

General Terms

Algorithms, Experimentation, Human Factors.

Keywords

Wikipedia, spatial-temporal, concept, multilinguism, cultural, interaction.

1. INTRODUCTION

“Where” and “when” are important implicit aspects of a wide variety of concepts. When we read a story, we place naturally characters in time and space that provide us with further context to understand. Assuming that spatial and temporal facets of concepts are potentially useful not only in human understanding but also in computing applications, we introduce a technique for automatically associating time and space to all concepts found in Wikipedia, providing what we believe to be the largest scale spatiotemporal mapping of concepts yet attempted.

Wikipedia is a well known community generated encyclopedia, available in many languages with variable levels of granularity. Derived works such as DBPedia provide structured access to some of the information found in Wikipedia [3]. Though spatial information is often associated with geographically related

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

JCDL '10, June 21–25, 2010, Gold Coast, Queensland, Australia.
Copyright 2010 ACM 978-1-4503-0085-8/10/06...\$10.00.

articles, and birth and death dates are included in article concerning people, temporal and spatial information concerning other concepts (for example, Romanticism, Scholasticism) is rarely explicit and certainly not readily available in a structured form such as DBPedia. Here we present a technique for associating both geographical locations and temporal intervals for every article in Wikipedia, including those without any explicit geotagging or time periods.

In section 2, we show how to create exhaustive lists of important locations from Wikipedia, and use these locations to identify the place associated with each concept appearing as a Wikipedia article by following outlinks of the article. In Section 3, we also show how to associate people with each concept and then how to use these persons' lifespans to associate characteristic time periods to each concept.

We then illustrate, in sections 4 and 5, how our spatiotemporal mapping of Wikipedia can be used in two sociological and historical applicative scenarios:

- Analysis of cultural interactions between different regions of the world – see how strong are the cultural relations (in a broad sense) between pairs of regions via a quantification of cultural interactions. This analysis is carried out in different cultural domains, such as literature, science, philosophy etc.
- Highlighting important regions for particular domains in different periods.

2. SPATIAL MAPPING

We extract locations related to a concept by examining outgoing links in the concept's Wikipedia article, using seven versions of the article from English, German, French, Spanish, Italian, Dutch and Portuguese.

2.1 Location list construction

First we construct a multilingual reference list of geographical locations. We use the list of geotagged English Wikipedia articles available from DBPedia [3] as a starting point. This list contains only titles and latitude/longitude pairs for geotagged articles, which are insufficient for our purpose. Before proceeding with the list construction, we eliminate extraterrestrial entities (found on Moon or on other planets) which are included in the DBPedia dataset, using patterns found in the categories (i.e. “on the Moon”). After this filtering, we find 326,667 locations remaining (down from 339,000 in the initial DBPedia list). This earth-bound geotagged dataset is completed with multilingual name variants and encompassing entities. Name variants in languages other than English are necessary in order to perform the link analysis in German, French, Italian, Spanish, Dutch and Portuguese. When possible, they are extracted using Wikipedia's translation links.

Encompassing entities concern spatial inclusion (*partOf*) which allows our system to determine that *Empire State Building* is situated in *New York City*, which is a part of the *United States*. When *Empire State Building* is mentioned in an article, we wish to infer that *New York City* and *United States* are encompassing entities. To build this inclusion hierarchy, we first extracted a list of countries from the Wikipedia article *List_of_countries*. Then, to find “important cities”, we extract geotagged articles items which have translations in at least 20 Wikipedia languages and are categorized under one of the following English categories: *cities*, *capitals*, *towns* or *settlements*. We thus found 14,000 important cities. To find the encompassing countries of each city, we matched the categories and the infobox of their geotagged article to our list of countries. Some geotagged articles are not associated to countries this way¹ or are associated to several countries². In a second pass, for each unlabeled or multiply labeled city we searched for the closest toponym, within 50 km of its geocoordinates, annotated from with a single country, and attribute this country to the target city.

Table 1. Snapshot of the geotagged articles dataset used for mapping Wikipedia concepts.

	Latin Quarter, Paris	Aachen
Name	Quartier Latin	Aix-la-Chapelle,
Variants (FR et ES)	(quartier parisien) ; Barrio Latino de Paris ...	Aquisgran, Aquisgrana
Latitude	48.851417	50.775278
Longitude	2.343167	6.082778
Categories	Districts of Paris, 5th arrondissement of Paris ...	Matter of France, Belgium–Germany border crossings ...
Encompassing City	Paris	Aachen
Country	France	Germany

For the remaining geotagged articles, we attempted to choose an encompassing city by first selecting “important” cities within a radius of 20 km of the coordinates and then searching these cities’ names in the article’s categories and infobox. To avoid errors, we exclude categories containing terms such as *St. Louis County* because they reference a region broader than the city itself. If several city names are found in the categories, the one closest to the entity’s coordinates is selected. Not all articles have associated encompassing cities and countries either because such entities do not exist (i.e. a continent is larger than a country and a city; a state is larger than a city) or because they are situated outside countries (i.e. underwater entities) or cities (i.e. mountains, forests).

We give an example of the content of the resulting geotagging dataset in table 1. *Quartier Latin* and *Aachen* are characterized by their English name, names in other languages, coordinates, categories and encompassing entities. Cities, such as *Aachen*, are considered self-included. In figure 1, we present a distribution of Wikipedia geotagged articles by country.

¹ For example, http://en.wikipedia.org/wiki/Latin_Quarter,_Paris

² For example, <http://en.wikipedia.org/wiki/Aachen>

The results in figure 1 are well correlated to the distribution of a dataset of over 30 million geotagged images presented in [6]. A majority of geotagged articles describes places in Europe and North America whereas there are few such articles describing Africa and South America. Countries with a detailed representation in the geotagging dataset: *United States* (88365), *France* (38304), and the *United Kingdom* (31237). Also well represented are *Germany*, *Poland* or *Italy*, countries with a good representation of local languages in Wikipedia. Although the distribution is generally intuitive, we found a small number of geotagged articles related to large countries such as *China* (819) or *Russia* (2849 articles). This came as a surprise given the local language versions of these Wikipedia are well developed.

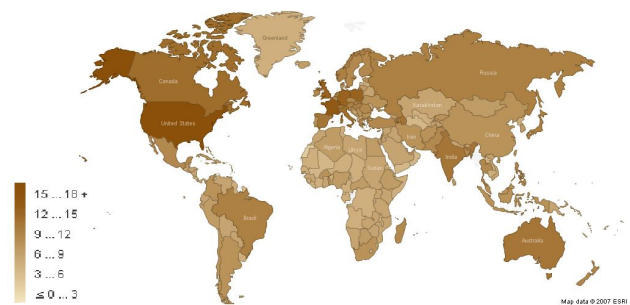


Figure 1. Distribution of geotagged articles in different countries (log2 values are plotted). Map created with <http://manyeyes.alphaworks.ibm.com/manyeyes/>

Spatial inclusion was also extracted at city level and we present top 20 cities by number of associated geotagged articles in figure 2. The high number of geotagged articles describing US locations determines a important presence of American cities among top cities (9 out of 20 in figure 2). Geotagged datasets in languages other than English are poorly described in DBPedia and we built the geotagging dataset using geotagged articles from the English version of Wikipedia. Only 5 cities represented in figure 2 are from countries where English is not the mother tongue. *London* is the city which is described in most detail, with nearly 2000 geotagged articles. A large number of articles are also associated to *New York* and *Paris* (753 and 613). Surprisingly, *Perth* (8th) and *Bristol* (11th) have a larger number of associated articles than larger cities like *Tokyo* or *Moscow*. One explanation for this situation is that *Perth* and *Bristol* are situated in English speaking countries, whereas *Tokyo* and *Moscow* aren't. However, *Perth* and *Bristol* also have more detailed descriptions than *San Francisco*. A second explanation could be that there are active communities that write and geo-tag articles about the former cities.

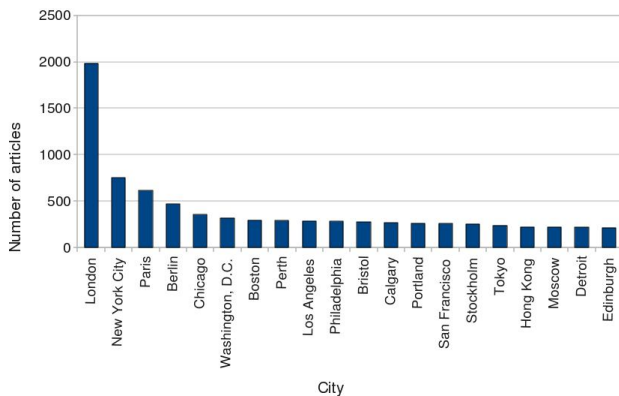


Figure 2. Top 20 cities by number of geotagged articles.

2.2 Location ranking

Not all locations mentioned in Wikipedia articles are equally pertinent for concepts and we rank locations based on their occurrence in the articles. Many Wikipedia concepts have dedicated articles in several languages and we exploit these variants in order to propose a first pertinence score. We consider the pertinence of a location to be proportional to the number of different languages in which it appears - this number constitutes the most important component of location ranking (R1 in table 2). A location which is linked in 7 languages will be considered more important than another location which appears only in four languages. The intuition that supports our ranking procedure is that when a link appears in multiple languages, it is inter-culturally pertinent for the target concept. When ties appear, we also count the total number of times a location is linked (R2) in all languages and the total number of mentions of a location (R3). We illustrate location ranking results for the scholastic philosopher *John Duns Scotus* and the Russian author *Mikhail Bulgakov* in table 2.

Table 2. Top locations, with pertinence scores, for *John Duns Scotus* and *Mikhail Bulgakov*. Only locations appearing in at least two languages are shown.

Concept	Top locations (R1, R2, R3)
John Duns Scotus	Scotland (6, 16, 20), Paris (6, 11, 25), Cologne (5, 12, 20), Oxford (5, 9, 16), Cambridge (3, 6, 8), Germany (4, 6, 6), Duns (3, 5, 126), University of Oxford (2, 4, 4)
Mikhail Bulgakov	Moscow (6, 20, 70), Kiev (6, 17, 31), Soviet Union (6, 14, 27), Ukraine (6, 12, 12), Paris (5, 10, 12), Russia (4, 6, 14), Bolshoi Theatre (3, 6, 6), Moscow Art Theatre (2, 4, 6), Smolensk (2, 4, 4), Novodevichy Cemetery (2, 4, 4)

Top locations for Duns Scotus include his home country (*Scotland*), places where he taught (*Cambridge*, *Oxford*, *Paris*), his death place (*Cologne*), his birth place (*Duns*) but also a more precise mention of a related institution (*University of Oxford*). Locations related to Mikhail Bulgakov include: his birth and death cities (*Kiev* and *Moscow*) with corresponding countries (*Ukraine* and *Russia* – later *Soviet Union*), related institutions (*Bolshoi Theatre* and *Moscow Art Theatre*), and his burial site

(*Novodevichy Cemetery*). Top locations in table 2 are those which are intuitively associated to *Duns Scotus* and *Mikhail Bulgakov* and this shows that our ranking procedure succeeds in capturing the most important locations associated to a concept. Related places include entities with variable spatial extent, with a strong presence of cities and countries, which seem to constitute a “basic level of representation” [15] for spatial information. The geotagging dataset allows an expansion of locations using spatial inclusion. For instance, the mention of *Oxford* also implies that *Duns Scotus* is related to *Oxford* and to the *United Kingdom*.

3. TEMPORAL MAPPING

Temporal mapping is performed in a way similar to the spatial mapping, with people replacing locations. People are chosen because they are well represented in Wikipedia [3] and can be situated in time. The hypothesis supporting our approach is that links to people point toward relevant periods for the analyzed article. Where article variants exist, results from multiple languages are aggregated in order to produce a ranked list of related people. A temporal mapping of the concept is then extracted based on the periods when associated people lived. We are primarily interested in a loose temporal mapping of Wikipedia concepts and related periods at a century scale.

3.1 People list extraction

Since one of our goals is to analyze different cultural domains over time and space, we are interested in having domain, time, and country related information, which are often specified as occupations, birth or/death years, and nationalities.

DBPedia [3] provides a list of persons obtained by analyzing infoboxes. As of early 2010, DBPedia’s list contains around 282,000 names. We found that a larger list of people could be extracted using categories rather than infoboxes. In order to do this, we first extracted a list of over 300 occupations from the Wikipedia article “List_of_occupations”, and a list of nationalities and associated countries from the article “Adjectivals and demonyms for countries and nations” With these two lists we revisit each Wikipedia article. If one of the article’s categories matches an occupation, or one of the terms: *people*, *birth*, or *death*, we place the article in a first list of candidate people. This first list of over 600,000 elements contains a number of non-persons: characters from movies, TV series, or books; mythological beings; errors from ambiguous occupation names – *editor* is also used to describe various software editors; groups of people, since “people” is to categorize persons but also ethnic groups, given names, or surnames. For each type of non-person, we create a simple pattern to recognize and eliminate the group, yielding a second, smaller list of 500,896 names, still almost twice the number of persons extracted in DBPedia. We randomly extracted 500 of these names, and found them all to be pertinent person names.

Next we sought to associate temporal references to each person in the list. We extracted each person’s approximate life period (within the right century) either from categories or from the first sentence of the article, in this order of priority. The extraction of temporal information from categories is easy if the relevant century/centuries are directly marked, or if the birth or/and death years are included as a category. Since most notable contributions of a person are related to adult life, we consider a century to be relevant for a person if that person lived at least 20 years during that century. A second century is considered relevant if the person

lived at least five years after its start. A similar procedure is applied to the first sentence, with a large number of lexical patterns used to detect birth/death dates. Using these simple heuristics, from the 500,000 persons in the initial list, we associated temporal information to 423,846 names. Figure 3 depicts a distribution of people by century.

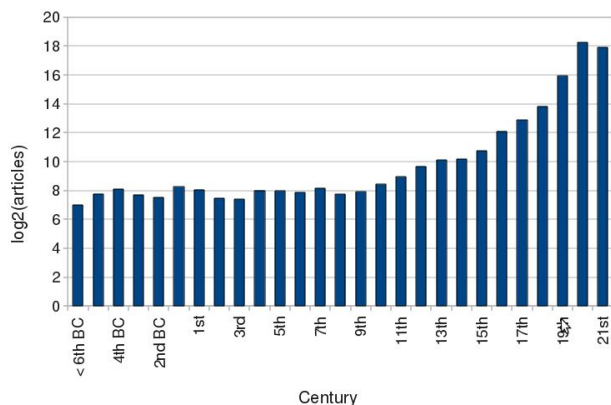


Figure 3. Distribution of Wikipedia person articles per century of lifespan. Log₂ values are shown. The total number of century-people associations is higher than 423,846 because many persons are associated with two centuries

The number of people per century described in Wikipedia articles is highly variable, with the largest values concentrated in the 20th and the 21st century. A simple explanation of the prevalence of articles about contemporary people is that Wikipedia contributors are more knowledgeable with recent information. Writing an article about a historic personality demands sufficient knowledge about the subject as well as a desire to share that information. It is also clear that the number of available written sources is greater for 19th, 20th and 21st century lives than for earlier periods. Understandably, the smallest number articles concerns people living in the 6th century BC or earlier (under 200 articles). The number of biographical articles is roughly 500 per century until the end of the High Middle Ages (11th century). During Antiquity, there is a local maximum for the 4th century BC, corresponding to the flowering of the Hellenistic civilization, for the 1st century BC, which witnessed the rise of the Roman Empire, and during the 1st century AD, during its expansion. The number of articles remains slow until the 11th century, a period which covers most of the “Dark Ages” in Western Europe. After the 11th century, the number of person-specific articles grows steadily until the 14th century. At the end of the Middle Ages that growth is accelerates, possibly related to Gutenberg’s invention of movable type printer and subsequent mass-production of written documents, leaving more traces of people’s lives. Significant growths appear between the 18th and the 19th century, and between the 19th and the 20th century, with four fold multiplications of person articles at each period. The large number of articles concerning living people makes one wonder whether Wikipedia recommendations about autobiographies³ and noteworthiness⁴ are being respected. This

³ <http://en.wikipedia.org/wiki/Wikipedia:Autobiography>

⁴ http://en.wikipedia.org/wiki/Wikipedia:Biographies_of_living_persons

aspect deserves a separate investigation which falls outside the scope of this paper.

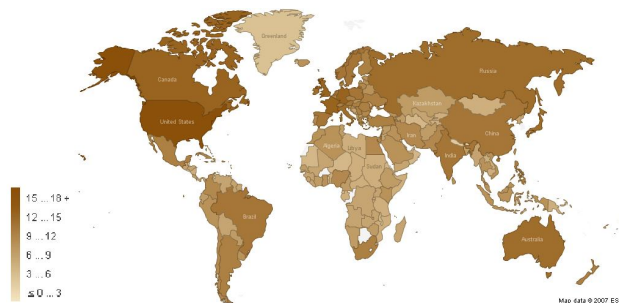


Figure 4. Distribution of Wikipedia biographies by nationality.

Often a person’s nationality is included among their Wikipedia categories and we can therefore derive a map of nationalities represented in the encyclopedia (figure 4). As with geotagged articles, a large majority of articles describe people from Europe or North America. The highest values per country are associated to: *United States* (119,168), *United Kingdom* (60,618), and *Germany* (23,080). Also well represented are *Canada*, *France*, *Italy*, and *Australia*. The predominance of English speaking countries is more pronounced than it was for geotagged articles. Compared to the distribution of geotagged, the number of articles about *Russia* (8903) and *China* (4783) is larger but still tiny considering these countries’ size, population and contribution to cultural history.

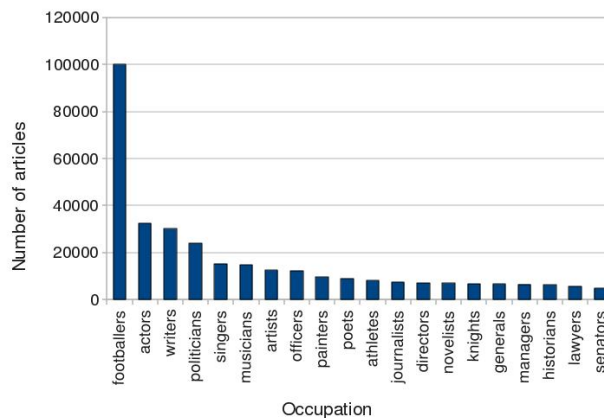


Figure 5. Distribution of Wikipedia biographies by occupation.

From our list of extracted people, we present the top 20 occupations represented in Wikipedia in figure 5. Since a person can have several occupations, the total number of occupation-person pairs is higher than 423,846. Occupations follow a long-tail distribution. Top occupations of Wikipedia people include a large number of sportsmen and people from the artistic world. *Footballers* (soccer and American football players) are by far the most represented category (around 100,000 articles), followed by *actors* and *writers* (around 30,000 articles each) and *politicians* (over 20,000 articles). Three of the top 20 occupations are related to war (*officers*, *generals*, and *knights*). Interestingly, except for *historians*, no scientific profession appears among top 20

categories (*scientists* comes 23rd). This finding might be explained by the high fragmentation of scientific professions.

In [22], the author underlines that, whereas popular culture is very well covered in Wikipedia, the same cannot be said about “high” culture. The focus of [22] was on museums but we can draw a similar conclusion about the Wikipedia representation of people. Many active communities gravitate around subjects such as *football, movies, music* or *politics* –subjects that often appear in everyday life of many people– and less focus is put domains such as science or philosophy.

3.2 People ranking

As we can rank a Wikipedia in term of most frequent geographical references, we can also use the same procedure to rank the people most closely associated with an article. We rank a person’s pertinence to a concept by considering: the number of different languages the person is cited in (R1), the total number of outgoing links (R2) over all language versions and the total number of mentions (R3).

Concept	Top 5 related people (R1, R2, R3)
Scholasticism	Thomas Aquinas (7, 26, 38); Peter Abelard (7, 18, 18); Albertus Magnus (7, 15, 17); Duns Scotus (7, 15, 15); Anselm of Canterbury (7, 15, 15)
Romanticism	William Blake (7, 16, 16); George Gordon Byron (6, 16, 17); William Wordsworth (5, 15, 15); Victor Hugo (5, 14, 17); Eugène Delacroix (5, 13, 18)

Table 3. Top 5 people, with pertinence scores, for scholasticism and romanticism.

Table 3 shows top personalities found for the articles *scholasticism* and *romanticism*. The top four philosophers in table 3 are considered some of the main figures of *scholasticism*. *Anselm of Canterbury* is one of the three founders of the movement in its medieval form while *Thomas Aquinas*, ranked first for scholasticism, being probably the best known philosopher of the Middle Ages as he was widely studied since the 19th century. Romanticism is a comprehensive cultural movement and this is reflected in table 3 with the presence of *Blake, Byron, Wordsworth* (poets), *Hugo* (novelist, poet, dramatist) and *Delacroix* (painter). As with locations, our ranking procedure generally succeeds in extracting very representative elements.

From these references to people, we derive important periods for Wikipedia concepts, in the following way. The century/centuries associated with a personality are weighted using the R1 scores, with a century’s total score obtained by aggregating individual R1 scores of persons linked in the article. In our two examples, the best ranked centuries for scholasticism are the 13th, the 12th and 11th centuries, the period when this approach was at its height in occidental philosophy. Also important are the 4th century BC (*Aristotle* is a major source of inspiration for scholastics) and the 3rd and 4th centuries AD (when precursors of scholasticism, such as *Ambrose* and *Augustine* lived). The 18th and 19th centuries are prevalent for romanticism, which corresponds to the description

of the current as being characteristic to the second half of the 18th century and the first half of 19th century⁵.

In this way we associate with every Wikipedia concept, both spatial and temporal references. In the next sections we present two application scenarios that exploit the newly created connection.

4. CULTURAL INTERACTIONS

The term culture is used here in a broad sense - it includes popular and “high” culture. People described in Wikipedia can be seen as exponents of the cultures⁶ to which they belong. Given the large number of people from different regions of the world described there, with their origin (nationality) and their associated locations, we can study cultural interactions at a global scale. Knowing the strength of cultural relations between regions of the world is certainly useful for having a global view of cultural exchanges but can also help us better understand individual cultures and discover their sources. Differing from the direct study of migrations is that here we are interested in the influence a person has in a region rather than her direct link with that region.

We compute incoming and outgoing migratory flows for each country in the world and then aggregate results in order to see how strong the relations between the two countries are. To find relations between pairs of countries, we cross nationality and location information available. Statistics are calculated using only the most strongly related locations associated to a person, which are usually most relevant ones. Outgoing relations (to which countries people from a particular country are related to?) are computed by selecting all people categorized with the country’s nationality and by analyzing their associated locations as extracted during spatial mapping of Wikipedia.

More specifically, we retain the country field of each location and weight it with R1, the number of different languages the location appears in. We then sum results for all people from the analyzed country to obtain a list of most related countries.

Since we are interested in interactions between countries, home locations are discarded from the obtained list. Incoming relations (where do strangers related to a country come from?) are computed inversely. We first select all concepts related to locations in a particular country and then sum the nationalities of people related to the respective country. Relations between any pair of countries can be obtained this way and aggregated in a global map of cultural interactions. A excerpt of this cultural interaction map is presented in figure 6, where top related countries are presented for 10 target countries. The size of tags displayed in figure 6 is computed using the same scale for all countries and one can notice important differences between the 10 countries – Algeria and Lebanon are much less present in Wikipedia than the USA, France or Italy. Incoming and outgoing relations show that European countries and the USA represent a vast majority of sources of cultural influence

⁵ <http://en.wikipedia.org/wiki/Romanticism>

⁶ The acceptance of culture here includes both popular and “high” culture.

Country	Incoming	Outgoing
Algeria	France, Spain, USA, UK, Italy	France, Italy, USA, Spain, Tunisia
Australia	UK, USA, France, New Zealand, Germany	UK, USA, France, Germany, Canada
Brazil	Portugal, USA, UK, Italy, France	USA, France, Italy, Spain, UK
France	UK, Germany, USA, Italy, Spain	USA, Italy, UK, Germany, Canada
Italy	USA, UK, France, Germany, Spain	USA, France, UK, Germany, Spain
Japan	USA, UK, Germany, France, Italy	USA, UK, China, France, Germany
Lebanon	USA, UK, Palestinian territories, Syria, France	USA, France, Syria, UK, Canada
Mexico	USA, Spain, UK, France, Argentina	USA, Spain, France, Italy, UK
Poland	Germany, USA, UK, Russia, France	USA, Germany, France, Russia, UK
USA	UK, Ireland, Germany, Italy, France	UK, Germany, France, Canada, Italy

Figure 6. Top 5 sources (incoming) and destinations (outgoing) of cultural interactions for 10 countries. Statistics computed from the study of locations associated to persons present in Wikipedia. The size of displayed countries is proportional to the log₂ of the country's score.

The only exceptions appear for Lebanon, which is strongly related to the *Palestinian Territories* and to *Syria* in reality and this proximity is translated into the results presented here. The analysis of incoming relations for *Japan* came as a surprise because *China* is historically related to this country but does not appear among the top 5 related countries. This situation is probably an effect Wikipedia's focus on recent information, correlated to the fact that the relations between two countries were difficult after World War II and the weak representation of China in Wikipedia.

The prominence of European countries and of the *USA* might be explained by their better representation in Wikipedia which is clearly shown in sections 2 and 3. However, their prominence is primarily an effect of historical contexts. Most incoming relations correspond to one's intuition analysis of incoming relations and their analysis helps us understand which the roots of particular cultures are. For instance, *Algeria*, *Brazil*, and *Australia* were parts of *France*, *Portugal*, and the *United Kingdom* and the latter countries constitute the main source of cultural immigration for the former ones. American culture is shaped mainly by its interaction with the *UK*, *Ireland*, *Germany*, *Italy* and *France*. The first four countries also represent a source of massive migration towards the *United States*, while the relation to *France* is probably due mainly to the influence of French culture on the American one. Top sources for American culture correspond to countries that are historically linked to the *United States* but which no longer provide a large number of immigrants. The main sources of immigration over the last 10 years are *Mexico*, *China*, and the *Philippines*⁷, neither of which appear in table 6. One explanation

is that recent immigration is mainly economically motivated. We could also conclude that cultural interactions are slower to appear compared to economic interactions. Globally, the greatest influence on the countries represented in table 6 is that of the *USA*, a country which had a leading role in the world during the 20th century. The prominence of American culture is obvious for popular culture. Its main vectors are products such as Hollywood movies, mainstream music, and popular literature. As for American "high" culture, [8] notes that it grew at an astonishing rate in the first part of the 20th century. This growth is reflected in its influence on other cultures and a domain-centered view of this dominance is presented in the next section.

Outgoing relations show how a certain culture radiates in other regions. In most cases, top countries are also countries where important diasporas from the target country exist. Again, most top countries are the *United States* and a small set of European countries (*UK*, *France*, *Germany*, *Italy*, *Spain*). Such countries are also home to major cultures but also constitute major emigration destinations and host important communities of people that influence their home countries. Relations between countries are generally asymmetrical. *Algeria's* incoming and outgoing relations to *France* are much stronger than links to other countries and this indicates that the first culture is strongly related to the second. However, the inverse seems not to be true because *Algeria* does not appear among *France's* top related countries. This illustrates a common relation between dominated and dominant cultures. The latter are usually Western countries, which have a history of spatial expansion and have and propensity for dissemination outside their own borders.

⁷http://en.wikipedia.org/wiki/Immigration_to_the_United_States#Origin

5. DOMAIN ANALYSIS

Domain analysis can help one visualize important locations quickly and can be used to support understanding that domain's structure. We extract people-related articles using a list of occupations and can link occupations to different cultural domains. Associated to temporal information, lists of domain-related occupations constitute the basis of a spatiotemporal domain analysis. Here the question that is answered is: which are the important locations for a cultural domain in a given range of time? We consider that a cultural domain is represented by the people that are active in that domain and we perform a statistical analysis using spatial and temporal information related to these people in order to aggregate knowledge about the domain.

To illustrate our approach, we have chosen philosophy and literature, two major cultural domains which are reasonably well represented in Wikipedia. We first retain pertinent occupations for each domain (philosophers, logicians, ontologists etc. for philosophy; writers, novelists, dramatists, poets etc. for literature) and search for people categorized under at least one of these categories. Temporal information is available for around 3000 people related to philosophy and for over 34,000 people representative for literature. Literature can be considered a border domain between popular and high culture whereas philosophy is a typical "high" culture domain. The ratio between these two domain representations is yet another proof that Wikipedia's content is more related to popular than to "high" culture. Given each domain's Wikipedia representation, but also its inherent distribution over time, we decided to use different time scales for the analysis. Five century slots are used for philosophy and one century slots for literature. Also, since spatial inclusion information is available at city and country level, we use both levels to find representative locations. The reader should notice that results are provided in terms of current country limits because it would be very difficult to mine a country's spatial extent over time automatically. Results in figures 7 and 8 are obtained by aggregating results for individuals who are representative of each domain.

Top locations related to philosophy follow one's intuition about the domain's history. Mediterranean countries dominate the periods up to the 5th century, with *Greece* coming first until the 1st century BC and *Italy* coming first for the subsequent period. *Greece* and *Italy* were home to the two major European Antic cultures, *Classical Greece* and the *Roman Empire*. *Egypt* is an even older culture but its flourishing period is less well documented. After the 4th century BC, *Egypt* and *Turkey* were in the Greco-Roman cultural sphere. *China* is the only non-Mediterranean country that appears in figure 7 and it developed a philosophical tradition independent of (and older than) the Greco-Roman sphere of influence. Top cities are all related to Mediterranean countries and we notice a growth of importance of *Rome* (3rd during the first period, 1st during the second) compared to *Athens* (1st during the first period, 2nd afterwards). The shift of importance from Greece to the Roman Empire follows more general historical movements. Four of top 5 cities from the first period are Greek cities or colonies whereas only *Athens* remains important in the second period and even this city is incorporated to the *Roman Empire*. The period between the 6th and the 10th

century comes after the decline of the Roman Empire. It is then that the rise of Arab caliphates happened. The *High Middle Ages* saw a decline of civilization in Europe and this decline is reflected in the results from figure 7. Compared to Antiquity, the number of philosophy related articles is significantly smaller during the *High Middle Ages* (reflected by the size of tags). Top countries (*Iraq*, *Spain*) and cities (*Baghdad*, *Kufa*, *Tehran*) are under Arab rule. The results of our statistical analysis confirm the key role of Arab civilization in the transmission (and augmentation) of philosophical knowledge from Antiquity to Modern Age. From the 11th to the 15th century, Europe becomes again the prevailing domain related region. The period marks a regain of interest for philosophy. Antic philosophy, particularly *Aristotle*, is rediscovered and is studied by Western philosophers. Important countries include *Italy*, *France* and *Spain*. This is also the period when the first universities are founded and those in *Paris* and *Oxford* play an important role in the development of philosophy (reflected by the apparition of the two cities among the top 5 for the period). Italian cities flourished at the end of the Middle Ages and were dominant centers of Renaissance culture. Three of them are found among major philosophical centers of the analyzed period.

Between, the 11th and the 15th we notice the apparition of *Germany* and of the *United Kingdom* among top countries. These two countries become central to the development of philosophy starting with the 16th century. The *United States* also play a prominent role starting with the 19th century and they are ranked 2nd for the most recent period analyzed here. Despite the slow decline of *France*, *Paris* remains the most important philosophical center. This is not the case for Italian cities, which disappear completely from the top locations. They are replaced by *London*, *Berlin* and *Vienna*. Even though the *USA* are ranked 2nd, no American city appears among top 5 results during the last period. This shows that changes in philosophy are slow and that it takes a lot of time for a place to lose or gain notoriety. Literature (figure 8) is examined starting with the 15th century. All top 5 countries and cities are situated in Europe or North America. Results for the 15th and 16th centuries (roughly the height of Renaissance), indicate that Italy plays a leading role in literature. This is particularly true for the 15th century, when all top 5 cities are from Italy, with particularly important roles for *Florence* and *Rome*. The *United Kingdom* has a leading role in literature during the 1600s and 1700s, to be subsequently replaced by the *United States*. We notice a decline of *Italy* starting with the 17th century and of *France* starting with the 19th century, correlated with an importance growth of the *United States* and of *Canada*. With the *UK*, *Germany* is the country which has the most constant presence among top countries. If *Paris* dominates the last centuries in philosophy, the same can be said about *London* and literature. Related to a larger context, *Vienna* appears to play an important role during the 18th and the 19th centuries, periods of development of the *Austrian Empire*. A similar observation is true for *New York City*, which rises to prominence during the 20th century, a period of fast development of the *United States*.

Period	Countries	Cities
<= 1st BC	Greece, Italy, Turkey, Egypt, China	Athens, Alexandria, Rome, Syracuse, Sparta
1st - 5th	Italy, Greece, Egypt, Turkey, China	Rome, Athens, Alexandria, Carthage, Cappadocia
6th - 10th	Iraq, Spain, China, Italy, Egypt	Baghdad, Rome, Kufa, Tehran, Cairo
11th - 15th	Italy, France, Spain, Germany, UK	Paris, Florence, Rome, Oxford, Venice
>= 16th	Germany, USA, UK, France, Italy	Paris, London, Berlin, Oxford, Vienna

Figure 7. Top 5 cities and countries in philosophy for different periods. Starting with the 1st century, five century slots were used.

Period	Countries	Cities
15th	Italy, France, UK, Germany, Spain	Florence, Rome, Ferrara, Venice, Bologna
16th	Italy, UK, France, Spain, Germany	London, Paris, Rome, Venice, Florence
17th	UK, France, Germany, Italy, Spain	London, Paris, Oxford, Cambridge, Amsterdam
18th	UK, France, Germany, Italy, USA	London, Paris, Vienna, Berlin, Oxford
19th	UK, USA, France, Germany, Italy	London, Paris, Berlin, Vienna, Oxford
20th	USA, UK, France, Germany, Canada	London, Paris, New York City, Berlin, Oxford
21st	USA, UK, Canada, Germany, France	London, New York City, Paris, Oxford, Chicago

Figure 8. Top 5 cities and countries in literature from the 15th century to nowadays. One century time slots were used.

Interestingly, at a country level, results for philosophy between the 11th and the 15th centuries are well correlated with results for literature during the 15th century. For the subsequent period, the correlation between the two domains is smaller since *Germany* is ranked after the *United Kingdom* and the *United States*. The leading role of the USA in philosophy and literature is more nuanced than the one revealed by aggregating results over all domains of activity. In philosophy, a “high” culture domain, the dominance of the United States is subject to debate whereas in literature, a domain in between “high” and popular culture, the *United States* dominates the 20th and the 21st century.

The reader should recall presented results represent general tendencies and were obtained using the English version of Wikipedia as a base. A bias in favor of English speaking countries may appear because they may be better represented in Wikipedia. However, if such a bias exists, it is minimized by the multilingual

filtering of results. Our purpose is not to judge the value of philosophical or literary creations but rather to show places with intense activity related to these domains through a quantitative analysis. Rather, we show that it is possible to produce meaningful domain-related results by using the spatiotemporal mapping of Wikipedia concepts. Also, results presented here are generic but the same methodology can be applied to smaller scale regions and to more specific domains if enough data are available in Wikipedia.

6. RELATED WORK

Our work stands at the intersection of domains such as information extraction from semi-structured sources, spatial annotation or temporal annotation. The fast development of Wikipedia and its accessibility generated burgeoning research in

areas like information extraction, entity ranking, semantic similarity or information retrieval [12].

An early tentative of creating and maintaining location and person gazetteers based on Wikipedia is presented in [19]. DBPedia [3] is one of the most coherent Wikipedia-related research efforts. Structured parts of the articles (infoboxes, categories) are analyzed in order to extract valuable information about concepts and to enable complex queries over the encyclopedia's content. Geotagged articles are extracted in several languages but the extraction is efficient only for English. For other languages, "geo-patterns" are not well processed and the number of extracted entities is small. As we mentioned the list of English geotagged articles needs to be cleaned in order to remove extraterrestrial entities. Also, spatial inclusion relations are not explicitly defined and we showed how to specify them at a city and country level. A list of people, with associated properties, is extracted in DBPedia using the infoboxes and it currently contains over 282,000 items. We presented a method that extends the coverage of the list (while maintaining a similar precision) by exploiting categorical information. While DBPedia extraction is generic, we focus on spatial and temporal information and provide more detailed and structured information for these two domains.

The categorical structure of Wikipedia is analyzed in [14] and the authors present a method for extracting taxonomy from Wikipedia. Their initial observation is that many categories do not follow Wikipedia guidelines and are noisy. However, when restraining the extraction to temporal and spatial information from categories, the amount of noise is not significant and categories can be reliably used.

Also related to our work are named entity extraction and ranking. Named entity extraction is used for disambiguation in open text domains [4] and [7]. Ambiguity is important for open text domains and resolving it should be part of future work on text annotation but is not a problem during spatiotemporal mapping because links are already disambiguated. Entity ranking using entity containment graphs and a Web search engine is explored by [23]. Their research problem is different from ours because they rank entities whereas we rank relations between entities.

One of the most appealing usages of Wikipedia is the extraction of semantic relatedness between concepts. WikiRelate! [18] exploits categories, Explicit Semantic Analysis (ESA) [9] is based on the computation of similarities between concept-related vectors and the method described in [13] uses links. Relatedness between two concepts is computed inside monolingual versions of Wikipedia. All three methods are applicable to any Wikipedia concept but do not type extracted relations. In contrast, we focus on providing deeper insight into spatial and temporal dimensions of concepts. Cross-lingual extensions of ESA are introduced in [17] and [10]. Reported results in information retrieval [17] or word sense disambiguation [10] tasks are not very encouraging. Whereas many aspects of Wikipedia are well studied, the full strength of its multilingual alignment is still to be unleashed [12] and we presented an application to relatedness ranking.

Methods exist for automatically delimiting borders of particular entity types: neighborhoods [16] or vernacular names [20]. One common conclusion of [16] and [20] is that borders are often imprecise. Sometimes it is preferable to replace delineations by granular spatial representations of concepts. We show how to extract concept footprints with Wikipedia and reuse them for spatial indexing. Spatial indexing of documents [11], [21] relies

on the detection and disambiguation of place names in written documents. Existing approaches demand for a location to be explicitly referenced in a text. We hypothesize that location information is also implicit to concepts other than locations and that mentions of these concepts implicate the mention of associated locations. Whereas [11] is interested in direct mentions of geographical information, we are more interested in the spatial implications the use of a concept has.

The identification of temporal references in texts is a well studied problem [1], [2], [5] but, similarly to location extraction, only explicit references are extracted. While such an approach is useful and accurate when mining precise temporal information, it has limitations when considering coarse-grained periods related to a concept. We identify a class of concepts (people) which are well situated in time and often cited in texts that have an important temporal dimension. Explicit and implicit references detection are in fact complementary in both spatial and temporal domains and combining could improve indexing precision and coverage.

7. CONCLUSION

We discussed methods for spatial and temporal mapping of Wikipedia concepts which are based on the encyclopedia's linking structure. We also presented methods for extracting rich spatial and temporal annotation datasets and presented an analysis of Wikipedia content reported to space, time and people's occupations. The relatedness between a concept and a location/period is computed using multilingual versions of articles. Our hypothesis was that the relatedness of a location/person is proportional to the article variants in which the location/person is used. Compared to existing spatial/temporal indexing schemes, we propose an extension of the indexing to any recognizable concept in order to provide a richer context for documents. To illustrate the utility of the spatiotemporal mapping, we analyze cultural interactions and the evolution of two cultural domains over time.

There are a lot of potential developments of the work presented here. First, we are interested in qualifying spatial and temporal relations. This is partially done for semi-structured information [3] but not free text. It would be interesting to apply the same ranking schema to links other than locations and people in order to extract semantic relatedness between concepts. This would constitute a multilingual counterpart of the method described in [13]. A third work direction is the analysis of Wikipedia versions in order to see how different cultures represent themselves and how they are related to other cultures. Of particular interest for such an analysis are *ism articles, which define abstract concepts with which we operate frequently. We find that questions such as "what does anarchism/scholasticism/behaviorism mean for American/Spanish/Italian people?" can help us better understand people with different background.

The resources described in this paper are created from freely accessible content and are accessible at the following site: http://comupedia.org/wiki_mapping.

8. ACKNOWLEDGMENTS

This research is partially funded by Georama and e-Diasporas, two French research projects funded by ANR (ANR-08-CORD-009, respectively ANR-08-CORD-0061).

9. REFERENCES

- [1] Alonso, O., Gertz, M., Baeza-Yates, R. On the value of temporal information in information retrieval. *SIGIR Forum*, 41(2), 2007.
- [2] Alonso, O., Gertz, M., Baeza-Yates, R. Clustering and exploring search results using timeline constructions. *Proc. of CIKM 2009*. Hong Kong, PRC.
- [3] Auer, S., Bizer, C., Lehmann, J., Kobilarov, G., Cyganiak, R. and Ives, Z. 2007. DBpedia: A Nucleus for a Web of Open Data. *Proc. of ISWC 2007* (Busan, Korea, November 2007).
- [4] Bunesco, R., Pasca, M. Using Encyclopedic Knowledge for Named Entity Disambiguation. *Proc. of EACL 2006*.
- [5] Catizone, R., Dalli, A., Wilks, Y. Evaluating Automatically Generated Timelines. *Proc. of LREC 2006*, Genova, Italy.
- [6] Crandall, D., Backstrom, L., Huttenlocher, D., Kleinberg, J. Mapping the World's Photos. In *Proc. of WWW 2009* (Madrid, Spain).
- [7] Cucerzan, S. Large-Scale Named Entity Disambiguation Based on Wikipedia Data. *Proc. of EMNLP-CoNLL 2007*, Prague, Czech Republic.
- [8] Franklin, W., Steiner, M. (eds.), *Mapping American Culture* (Iowa City: University of Iowa Press, 1992).
- [9] Gabrilovich, E., Markovitch, S. Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. *Proc. of IJCAI 2007*, Hyderabad, India.
- [10] Hassan, S., Mihalcea, R. Cross-lingual Semantic Relatedness Using Encyclopedic Knowledge, *Proc. of EMNLP 2009*.
- [11] Jones, C. B., Abdelmoty, A. I., Finch, D., Fu, G., Vaid, S. The SPIRIT Spatial Search Engine: Architecture, Ontologies and Spatial Indexing. *LNCS vol. 3234*, 2004.
- [12] Medelyan, O., Milne, D., Legg, C., Witten, I. Mining Meaning from Wikipedia. *Intl. Journal of Human-Computer Studies*, 67(9), Sept. 2009.
- [13] Milne, D., Witten, I. H. An effective, Low-Cost Measure of Semantic Relatedness Obtained from Wikipedia Links. *Proc. of WIKIAI 2008*, Chicago, USA.
- [14] Ponzetto, S. P., Strube, M. Deriving a large scale taxonomy from Wikipedia. *Proc. of AAI 2007*, Vancouver, Canada.
- [15] Rosch, E.H., Mervis, C.B., Gray, W.D., Johnson, D.M. and Boyes-Braem, P. (1976) Basic objects in natural categories. *Cognitive Psychology* 8: 382-439.
- [16] Schockaert, S., De Cock, M. Neighborhood Restriction in Geographical IR. In *Proc. of SIGIR 2007* (Amsterdam, The Netherlands, July 2007).
- [17] Sorg, P., Cimiano, P. Cross-lingual Information Retrieval with Explicit Semantic Analysis. *Working Notes of CLEF 2008*.
- [18] Strube, M., Ponzetto, S. P. WikiRelate! Computing Semantic Relatedness Using Wikipedia. *Proc. of AAI 2006*, Boston, USA.
- [19] Toral, A. and Munoz, R. 2006. A proposal to automatically build and maintain gazetteers for named entity recognition by using wikipedia. In *Proc. of Workshop on NEW TEXT Wikis and blogs and other dynamic text sources* (Trento, Italy, April 2006).
- [20] Twaroch, F., Jones, C., Abdelmoty, A. Acquisition of a vernacular gazetteer from Web sources. In *Proc. of LocWeb Workshop – WWW 2008*. (Beijing, China, 2008)
- [21] Vaid, S., Jones, C. B., Joho, H., Sanderson, M. Spatio-textual Indexing for Geographical Search on the Web. *LNCS*, vol 3633, 2005.
- [22] Wyatt, L. Wikimedia & Museums – why we need each other and what we can do about it. *Proc. of Wikimania 2009*, Buenos Aires, Argentina.
- [23] Zaragoza, H., Rode, H., Mika, P., Atserias, J., Ciaramita, M., Attardi, G. Ranking very many typed entities on Wikipedia. *Proc. of CIKM 2007*, Lisbon, Portugal.